

# **Novel Algorithm Development for 'Next- Generation' Sequencing Data Analysis**

Agne Antanaviciute

Submitted in accordance with the requirements for the degree of  
Doctor of Philosophy

University of Leeds

School of Medicine

Leeds Institute of Biomedical and Clinical Sciences

12/2017



The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly given within the thesis where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement

©2017 The University of Leeds and Agne Antanaviciute

The right of Agne Antanaviciute to be identified as Author of this work has been asserted by her in accordance with the Copyright, Designs and Patents Act 1988.

## ***Acknowledgements***

I would like to thank all the people who have contributed to this work. First and foremost, my supervisors Dr Ian Carr, Professor David Bonthron and Dr Christopher Watson, who have provided guidance, support and motivation. I could not have asked for a better supervisory team.

I would also like to thank my collaborators Dr Belinda Baquero and Professor Adrian Whitehouse for opening new, interesting research avenues. A special thanks to Dr Belinda Baquero for all the hard wet lab work without which at least half of this thesis would not exist.

Thanks to everyone at the NGS Facility – Carolina Lascelles, Catherine Daley, Sally Harrison, Ummey Hany and Laura Crinnion – for the generation of NGS data used in this work and creating a supportive and stimulating work environment.

Finally, I am extremely grateful to Sir Jules Thorn Charitable Trust, who funded this work.

## ***Abstract***

In recent years, the decreasing cost of 'Next generation' sequencing has spawned numerous applications for interrogating whole genomes and transcriptomes in research, diagnostic and forensic settings. While the innovations in sequencing have been explosive, the development of scalable and robust bioinformatics software and algorithms for the analysis of new types of data generated by these technologies have struggled to keep up. As a result, large volumes of NGS data available in public repositories are severely underutilised, despite providing a rich resource for data mining applications. Indeed, the bottleneck in genome and transcriptome sequencing experiments has shifted from data generation to bioinformatics analysis and interpretation.

This thesis focuses on development of novel bioinformatics software to bridge the gap between data availability and interpretation. The work is split between two core topics – computational prioritisation/identification of disease gene variants and identification of RNA N6 -adenosine Methylation from sequencing data.

The first chapter briefly discusses the emergence and establishment of NGS technology as a core tool in biology and its current applications and perspectives.

Chapter 2 introduces the problem of variant prioritisation in the context of Mendelian disease, where tens of thousands of potential candidates are generated by a typical sequencing experiment. Novel software developed for candidate gene prioritisation is described that utilises data mining of tissue-specific gene expression profiles (Chapter 3). The second part of chapter investigates an alternative approach to candidate variant prioritisation by leveraging functional and phenotypic descriptions of genes and diseases from multiple biomedical domain ontologies (Chapter 4).

Chapter 5 discusses N6 AdenosineMethylation, a recently re-discovered post-transcriptional modification of RNA. The core of the chapter describes novel software developed for transcriptome-wide detection of this epitranscriptomic mark from sequencing data. Chapter 6 presents a case study application of the software, reporting the previously uncharacterised RNA methylome of Kaposi's Sarcoma Herpes Virus. The chapter further discusses a putative novel N6-methyl-adenosine -RNA binding protein and its possible roles in the progression of viral infection.

# Contents

<b>1. Introduction .....</b>	<b>1</b>
1.1 Nucleic acid sequencing.....	1
1.2 ‘Next-Generation’ Sequencing.....	6
1.3 “Next-Generation” Sequencing Applications .....	10
1.4 Bioinformatics Challenges in ‘Next-Generation’ Sequencing Era.....	12
1.5 Overview of this work.....	14
<b>2. Candidate Disease Gene and Variant Prioritisation .....</b>	<b>16</b>
2.1 Disease Gene/Variant Identification.....	16
2.1.1 Genetic Linkage and Association Studies .....	16
2.1.2 Whole Exome and Whole Genome Sequencing Studies .....	18
2.1.3 Putative disease variant prioritisation.....	22
2.2 Computational Candidate Disease Gene Prioritisation .....	27
<b>3. GeneTiER: gene Tissue Expression Ranker .....</b>	<b>41</b>
3.1 Motivation .....	41
3.2 Methods.....	47
3.2.1 Software Implementation .....	47
3.2.2 User Input Processing.....	48
3.2.3 GeneTiER Database .....	51
3.2.4 The gene prioritization algorithm.....	54

3.3 Results and Discussion.....	55
3.3.1 Performance Assessment .....	55
3.3.2 Case Study Genes .....	65
3.3.3 GeneTiER Application .....	74
3.3.4 Discussion .....	76
<b>4 OVA: Ontology Variant Analysis Tool .....</b>	<b>79</b>
4.1.1 Motivation .....	79
4.1.2 Representing Biomedical Knowledge in Machine-Readable Ways .....	79
4.1.3 Biomedical Domain Ontologies .....	83
4.2 Methods.....	85
4.2.1 Overview.....	85
4.2.2 Software Implementation .....	87
4.3 Results and Discussion.....	103
4.3.1 Performance Assessment .....	103
4.3.2 OVA Application.....	121
4.3.3 Discussion .....	132
<b>5. N6-Methyl Adenosine Sequencing .....</b>	<b>136</b>
5.1 Introduction.....	136
5.1.2 N6-Methyl Adenosine Molecular Biology.....	137
5.1.3 m <sup>6</sup> A detection.....	155

5.1.4 Computational methods for m <sup>6</sup> A-seq data analysis.....	159
<b>Box 1. Hidden Markov models.....</b>	<b>163</b>
<b>Box 2. Expectation Maximisation algorithm.....</b>	<b>166</b>
<b>Box 3. Viterbi Algorithm .....</b>	<b>167</b>
5.1.5 Computational prediction of m <sup>6</sup> A sites .....	172
<b>5.1.6 Summary .....</b>	<b>175</b>
5.2. m6aViewer application methods.....	176
5.2.1 Overview.....	176
5.2.2 Sequence Read Processing .....	176
5.2.3 m <sup>6</sup> A peak-calling .....	190
5.2.4 m6A peak deconvolution.....	195
5.2.5 Technical False Positive m <sup>6</sup> A peak identification .....	209
5.3 Results and Discussion.....	221
5.3.1 m6aViewer implementation.....	221
5.3.2 Evaluation of m6aViewer’s peak calling performance .....	230
5.3.3 Evaluation of m6aViewer’s False Positive Filter Performance .....	232
5.3.4 Comparison with m <sup>6</sup> A prediction algorithm SRAMP .....	240
5.3.5 Integration with m6aViewer software .....	242
5.3.6 Distribution of identified peaks to nearest m6A ‘RRACH’ sequence motif.....	242
5.3.7 Discussion .....	246

<b>6. Characterisation of Kaposi’s sarcoma Herpesvirus-8 m<sup>6</sup>A methylome and identification of a putative novel m<sup>6</sup>A ‘reader’ protein .....</b>	<b>248</b>
6.1 Motivation .....	248
6.2 Kaposi’s Sarcoma Herpesvirus-8 .....	248
6.3 Methods .....	250
6.3.1 Sequence Data Generation .....	250
6.3.2 Publicly available sequencing data .....	251
6.3.3 Sequence data analysis .....	252
6.4 Results and Discussion .....	255
6.4.1 KSHV Methylome .....	255
6.4.2 Characterisation of a new putative m <sup>6</sup> A ‘reader’ .....	265
6.4.3 Identification of transcriptome-wide SND1 binding sites .....	266
6.4.4 RNA gene expression, lifetime profiling and alternative splicing analysis in SND1-depleted BCBL-1 cells .....	280
6.4.5 Discussion .....	291
<b>7. Conclusion .....</b>	<b>297</b>
<b>8. Publications .....</b>	<b>299</b>
<b>9. Bibliography .....</b>	<b>300</b>

***List of Tables***

Table 1.....30

Table 2.....59

Table 3.....66

Table 4.....66

Table 5.....89

Table 6.....105

Table 7.....108

Table 8.....120

Table 9.....121

Table 10.....178

Table 11.....220

Table 12.....265

Table 13.....282

## ***List of Figures***

Figure 1.....	3
Figure 2.....	5
Figure 3.....	9
Figure 4.....	20
Figure 5.....	45
Figure 6.....	48
Figure 7.....	50
Figure 8.....	53
Figure 9.....	59
Figure 10.....	60
Figure 11.....	62
Figure 12.....	63
Figure 13.....	64
Figure 14.....	65
Figure 15.....	67
Figure 16.....	70
Figure 17.....	71
Figure 18.....	72
Figure 19.....	73
Figure 20.....	75
Figure 21.....	82
Figure 22.....	86
Figure 23.....	103

Figure 24.....	106
Figure 25.....	112
Figure 26.....	113
Figure 27.....	114
Figure 28.....	115
Figure 29.....	117
Figure 30.....	119
Figure 31.....	123
Figure 32.....	138
Figure 33.....	142
Figure 34.....	143
Figure 35.....	150
Figure 36.....	158
Figure 37.....	164
Figure 38.....	169
Figure 39.....	178
Figure 40.....	182
Figure 41.....	184
Figure 42.....	187
Figure 43.....	188
Figure 44.....	191
Figure 45.....	196
Figure 46.....	197
Figure 47.....	199
Figure 48.....	201
Figure 49.....	202
Figure 50.....	203

Figure 51.....	206
Figure 52.....	208
Figure 53.....	212
Figure 54.....	215
Figure 55.....	216
Figure 56.....	217
Figure 57.....	219
Figure 58.....	222
Figure 59.....	231
Figure 60.....	233
Figure 61.....	234
Figure 62.....	235
Figure 63.....	237
Figure 64.....	238
Figure 65.....	241
Figure 66.....	244
Figure 67.....	245
Figure 68.....	256
Figure 69.....	259
Figure 70.....	261
Figure 71.....	262
Figure 72.....	262
Figure 73.....	264
Figure 74.....	267
Figure 75.....	268
Figure 76.....	270
Figure 77.....	271

Figure 78.....	273
Figure 79.....	274
Figure 80.....	277
Figure 81.....	278
Figure 82.....	279
Figure 83.....	281
Figure 84.....	282
Figure 85.....	284
Figure 86.....	286
Figure 87.....	287
Figure 88.....	289
Figure 89.....	290
Figure 90.....	291
Figure 91.....	292
Figure 92.....	294

## *Abbreviations*

<b>AUC</b>	Area Under The Curve
<b>ATP</b>	Adenosine Triphosphate
<b>BAM</b>	Binary Alignment Map
<b>BIC</b>	Bayesian Information Criteria
<b>BLAST</b>	Basic Local Alignment Search Tool
<b>CESSM</b>	Collaborative evaluation of semantic similarity measures
<b>CDS</b>	Coding sequence
<b>ChIP-Seq</b>	Chromatin Immunoprecipitation Sequencing
<b>CNV</b>	Copy Number Variant/Variation
<b>ddNTP</b>	Dideoxynucleoside triphosphate
<b>dNTP</b>	Deoxynucleoside triphosphate
<b>DO</b>	Disease Ontology
<b>EBV</b>	Epstein–Barr virus
<b>eCLIP</b>	Enhanced cross-linking and immunoprecipitation
<b>EER</b>	Enhanced entity-relationship
<b>EM</b>	Expectation Maximisation
<b>ENCODE</b>	Encyclopedia of DNA Elements
<b>FDR</b>	False Discovery Rate
<b>fRIP-Seq</b>	Formaldehyde RNA immunoprecipitation sequencing
<b>FTO</b>	Fat mass and obesity-associated protein
<b>GEO</b>	Gene Expression Omnibus
<b>GO</b>	Gene Ontology
<b>HIV</b>	Human Immunodeficiency Virus
<b>HMM</b>	Hidden Markov Model
<b>HNRNPC</b>	Heterogeneous nuclear ribonucleoprotein C
<b>hnRNPs</b>	Heterogeneous nuclear ribonucleoproteins
<b>HPO</b>	Human Phenotype Ontology
<b>IC</b>	Information Content
<b>IP</b>	Immunoprecipitation
<b>KEGG</b>	Kyoto Encyclopedia of genes and genomes
<b>KSHV</b>	Kaposi's Sarcoma Herpes Virus
<b>lncRNA</b>	Long non-coding RNA
<b>m<sup>6</sup>A</b>	N6-Methyl Adenosine
<b>MeDIP-Seq</b>	Methyl DNA Immunoprecipitation sequencing
<b>MeSH</b>	Medical Subject Headings
<b>MIAME</b>	Minimum Information About a Microarray Experiment
<b>MICA</b>	Most informative common ancestor

<b>MIRA-Seq</b>	Methylated-CpG island recovery assay sequencing
<b>miRNA</b>	Micro RNA
<b>mRNA</b>	Messenger RNA
<b>MTD</b>	Mixture transition distribution model
<b>NCBI</b>	National Center for Biotechnology Information
<b>NGS</b>	Next generation sequencing
<b>OBO</b>	Open Biological and Biomedical Ontology
<b>OMIM</b>	Online Mendelian Inheritance in Man
<b>PAR-CLIP</b>	Photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation
<b>PARS</b>	Parallel analysis of RNA structure
<b>PCR</b>	Polymerase chain reaction
<b>PO</b>	Pathway Ontology
<b>RAM</b>	Random Access Memory
<b>RF</b>	Random Forest
<b>RIP-Seq</b>	RNA immunoprecipitation sequencing
<b>RISC</b>	RNA-induced silencing complex
<b>RNA-Seq</b>	RNA sequencing
<b>ROC</b>	Receiver operating curve
<b>rRNA</b>	Ribosomal RNA
<b>RTA</b>	Replication and transcription activator
<b>SAM</b>	Sequence Alignment Map
<b>SCARLET</b>	Site-specific cleavage and radioactive-labeling followed by ligation-assisted extraction and thin-layer chromatography
<b>SNP</b>	Single nucleotide polymorphism
<b>snRNA</b>	Small nuclear RNA
<b>SNV</b>	Single nucleotide variant
<b>SRA</b>	Sequence Read Archive
<b>SVM</b>	Support Vector Machine
<b>TLC</b>	Thin layer chromatography
<b>tRNA</b>	Transfer RNA
<b>TSS</b>	transcription start site
<b>TTS</b>	Transcription termination site
<b>UTR</b>	Untranslated region
<b>VCF</b>	Variant call format
<b>WES</b>	Whole exome sequencing
<b>WGCNA</b>	Weighted gene co-expression network analysis
<b>WGS</b>	Whole genome sequencing
<b>YTH</b>	YT521-B homology





# **1. Introduction**

## **1.1 Nucleic acid sequencing.**

At the heart of molecular biology lies the central dogma – DNA encodes RNA; RNA encodes proteins. Detailed in a landmark 1970 paper (Crick 1970) by Francis Crick, this philosophy of information flow from nucleic acids to protein cemented the role of biological polymers as the foundation of all life. The central dogma was the culmination of much of the work earlier in the 20<sup>th</sup> century – perhaps the most noteworthy of which was the discovery of the double-helix structure of DNA (Watson and Crick 1953b, 1953a), which unravelled the molecular basis of heritability.

The importance of sequence in biological polymers, however, was first recognised somewhat earlier by Frederick Sanger in his studies of bovine insulin protein, which established that proteins have a defined amino acid composition (Sanger and Tuppy 1951; Sanger 1949). The link between protein and DNA sequence was recognised soon after (Gamow 1954), though it is interesting to note that the ‘coding problem’ was solved more than a decade prior to the determination of the first DNA sequence.

Indeed, while the structure of DNA provided insight into the importance of nucleotide sequences, it was some years before the exact order of nucleotides could be determined – early methods could provide insights into nucleotide composition of DNA, but not its sequence (Holley et al. 1961). Moving forward from his work on protein sequences, Frederick Sanger began development of methods for nucleic acid sequence determination – but was beaten to the publication of the first nucleic acid sequence, that of alanine tRNA (Holley et al. 1965). This first generation of sequencing methods relied on ribonuclease treatments to produce partially-digested RNA fragments which could be radioactively labelled and separated using two dimensional ionophoresis (Sanger et al. 1965) and was used to determine a number of ribosomal and transfer RNA sequences (Adams et al.

1969; Brownlee and Sanger 1967; Cory et al. 1968). The chemical cleavage-based Maxam-Gilbert method (Maxam and Gilbert 1977) dominated the early sequencing efforts, where radioactively end-labelled DNA fragments would be cleaved in four separate, base-specific reactions. The DNA sequence could then be determined by size-separating the cleaved fragments produced by each reaction.

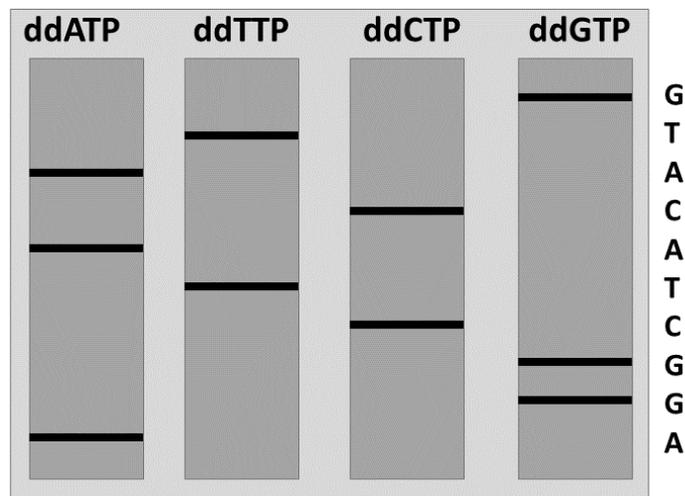
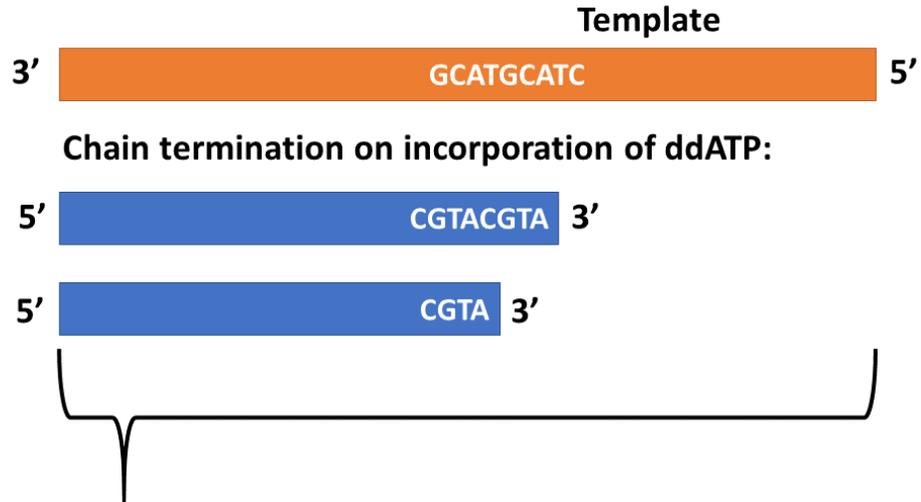
However, it was the later sequencing methods developed by the Sanger lab that would become the mainstay of biological research, which relied on the premature termination of DNA elongation by DNA polymerases, rather than chemical cleavage. The first such sequencing method pioneered by Sanger and Coulson was termed “plus-minus” sequencing and was used to determine the sequence of bacteriophage phiX174 (Sanger and Coulson 1975; Sanger et al. 1978), the first complete genome to be sequenced. The technique uses an initial reaction that generates all possible radioactively labelled DNA products of increasing length, which can be used in eight subsequent reactions where synthesis is terminated in a sequence-specific manner by limiting the supply of nucleoside triphosphates - four ‘plus’ reactions are supplied with only one of the four nucleotides, while the four ‘minus’ reactions use three of the four. These fragments are then resolved on a polyacrylamide gel in the order of increasing chain length. Thus, DNA strands differing by only a single nucleotide could be resolved as discrete bands across 8 lanes on the resulting autoradiograph of the polyacrylamide gel.

While not without issues – for example, difficulty in resolving homopolymer runs – the chain termination approach formed the basis of modern DNA sequencing methods. A few years later, Sanger *et al* (1977) further refined the chain-termination approach by using dideoxy-nucleotides, which had been shown some years earlier to inhibit DNA polymerase activity if incorporated in place of deoxy-nucleotides during chain synthesis (Atkinson et al. 1969). Thus, a polymerase reaction incubated with a mixture of four dNTPs and one ddNTP would yield varying length chains all terminating at a particular base, which can then be resolved by gel electrophoresis (**Figure 1**).

# DNA Polymerase

+ four dNTPs

+ ddATP



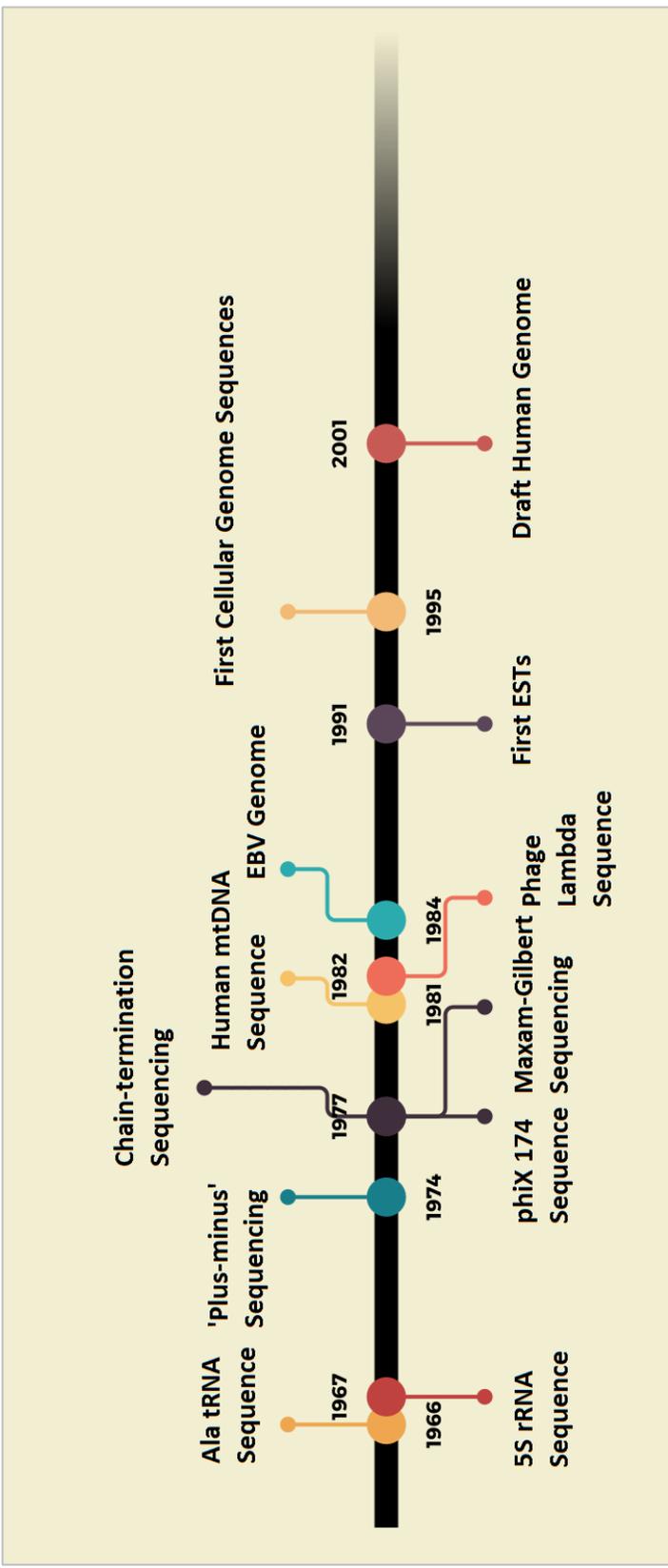
**Size-separation of fragments**

**Figure 1.** Sanger sequencing method by ddNTP chain-termination. In the presence of four nucleotides, DNA polymerase extends the primer based on the template DNA strand. An incorporation of a dideoxynucleotide, however, results in chain termination. Four parallel reactions, each including a different dideoxynucleotide, will generate fragments of varying length, which can then be resolved by electrophoresis and the base sequence of the template read.

The refined technique was widely adopted and subsequently used to sequence human mitochondrial DNA in 1981 (Anderson et al. 1981), the lambda phage genome in 1982 (Sanger et al. 1982) and EBV genome in 1984 (Baer et al. 1984). The next few years saw a number of improvements made to Sanger sequencing that allowed ever-increasing automation of the approach, including elimination of radioactive labelling in favour of fluorescent dyes (Ansorge et al. 1986).

In parallel with the advances in the sequencing reactions, the first commercial DNA sequencing instruments, which appeared in the mid-1980s, also underwent development. In 1986, Applied Biosystems (ABI) developed an automated Sanger sequencing machine using patented fluorescent dye-labelled ddNTPs which allowed sequencing in one reaction rather than four (Smith et al. 1986). The report also demonstrated that sequence data could be read directly by a computer by recording the sequence of colours as DNA fragments passed a detector at the end of the gel. Automation thus enabled the sequencing of more complex genomes, and the following decade saw the first cellular genome sequences published, starting with the bacterium *Haemophilus influenza* in 1995 (Fleischmann et al. 1995) (**Figure 2**).

The ABI 370A DNA sequencer launched in 1986 could produce approximately 1000bp of sequence per day. By 1995, the new ABI PRISM 377 instrument had optimised the method, but it wasn't until the following year that a breakthrough which saw the replacement of slab gels with capillary electrophoresis dramatically improved throughput. By 1998, the ABI PRISM 3700 96 capillary system could produce approximately 900 kbs of sequence data per day – a substantial leap forward, yet a single instrument would still have required more than 45 years to sequence the 3 billion bases of the human genome at 5X coverage.



**Figure 2.** The timeline summarises the early milestones in nucleic acid sequencing. The first nucleic acid sequence was determined in 1966, that of Alanine tRNA. The first widely used sequencing methods, plus-minus, chain-termination and Maxam-Gilbert, emerged in the 1970s, which paved the way for the sequencing of whole organisms, starting with viral and bacterial genomes and culminating in the first draft of the human genome on 2001.

Nevertheless, by this time the efforts to sequence the human genome were well under way. The Human Genome project was conceived as a collaborative, publicly funded endeavour and, by the late 1990s, significant strides had been made towards its completion. The advent of truly automated capillary sequencing, however, attracted commercial attention and Celera Genomics was founded with the aim of completing the sequence of the human genome cheaper and faster than the publically funded efforts. By 1999, the first human chromosome sequence was published by The Human Genome Consortium (Dunham et al. 1999), now in direct competition with Celera Genomics. Despite the controversy, the completion of the project was announced in 2000 – officially a tie between the publicly funded efforts and Celera Genomics (Lander et al. 2001; Venter et al. 2001). This began a new “-omics” era in biology, with a proven capability to study whole genomes and transcriptomes.

### ***1.2 'Next-Generation' Sequencing***

When the first human genome sequence was nearing completion, the National Human Genome Research Institute of the US National Institutes of Health outlined plans for the future of genomics research – among which was the then near fictional goal of sequencing individual human genomes at the cost of less than \$1000 per genome (Bennett et al. 2005; Schloss 2008).

Yet in the 15 years since the completion of the human genome draft sequence, the technological progress has been explosive. With commercial interest, massively parallel methods of sequencing emerged that were finally able to challenge the traditional Sanger approach.

The first of the ‘Next Generation’ sequencing (NGS) instruments were produced by 454 Life Sciences (Margulies et al. 2005) and used the pyrosequencing approach conceived some 20 years earlier (Nyrén and Lundin 1985). Initially, sheared DNA molecules are captured on a bead array and amplified in an emulsion droplet. As with the Sanger method, sequencing is carried out via primed DNA polymerase synthesis – the array is exposed to each dNTP in turn and the

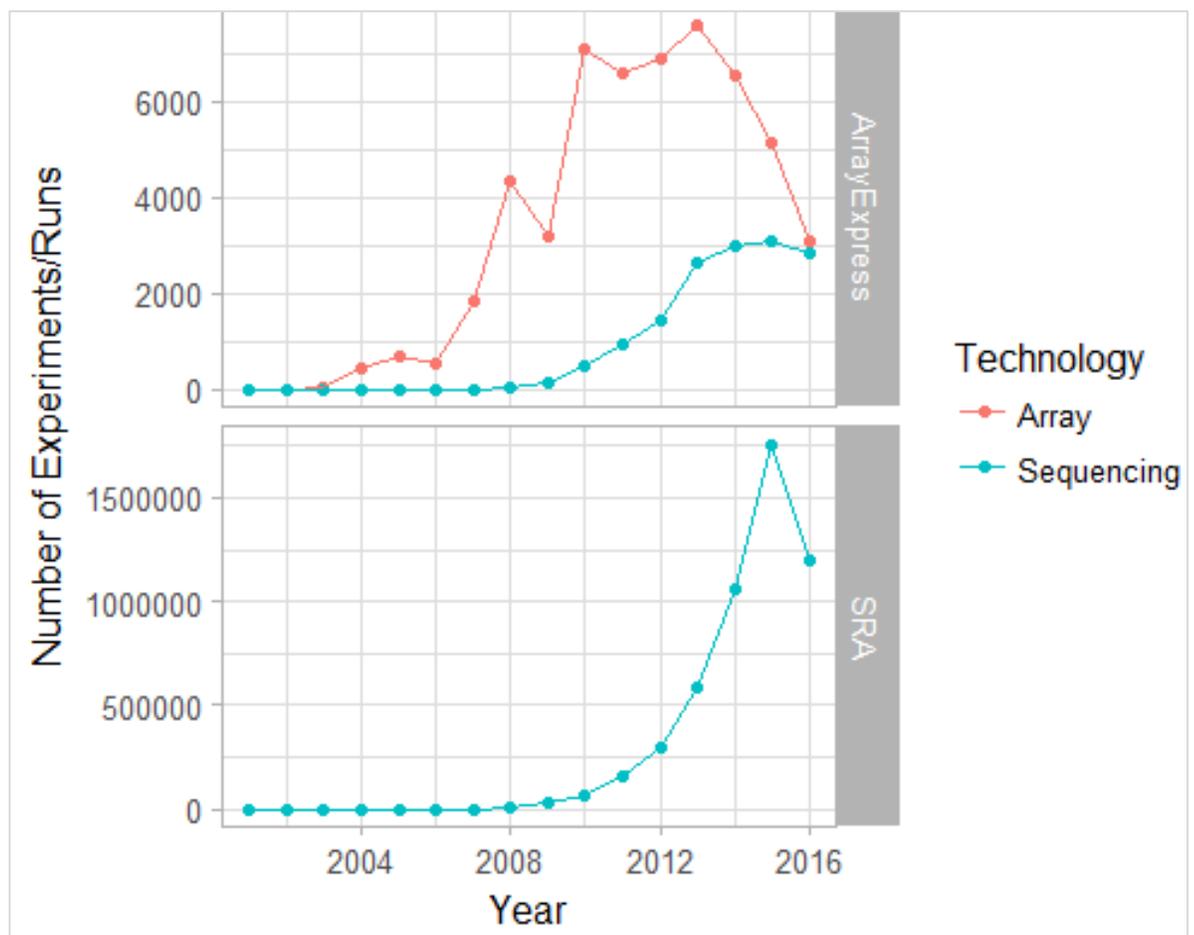
amount of incorporation is monitored by the amount of pyrophosphate produced in a two enzyme process: ATP sulfurylase converts pyrophosphate into ATP, which acts as the substrate for luciferase, and produces light proportional to the amount of pyrophosphate incorporated which can be monitored in real time.

Following the commercial success of 454 instruments, a number of competitors quickly emerged. One of the most important developments was in the Solexa (which was later acquired by Illumina) instruments, that differed from 454 technology in several key areas. In contrast to the bead emulsion PCR method used in 454 instruments, the company patented an approach dubbed 'bridge amplification', wherein adapter-ligated DNA molecules are passed over a flow cell surface of complimentary oligonucleotides. Subsequent PCR cycles would generate neighbouring clusters of clonal populations of the original DNA strands, where each replicating DNA molecule would arch over in order to prime the next round of amplification off the flowcell surface bound oligonucleotides (Bentley et al. 2008). Additionally, rather than measuring pyrophosphate incorporation, the instrument still relied on a traditional Sanger approach, with improved, reversible chain-termination chemistry (Turcatti et al. 2008): after the incorporation and identification of each fluorescently-labelled nucleotide, the 3' terminator which would normally inhibit further polymerisation is cleaved off and thus a new 'cycle' can be performed.

A number of competitors in the early sequencing market emerged (and some also disappeared), including ligation-based chemistry of SOLiD systems (McKernan et al. 2009; Shendure et al. 2005), DNA 'nanoballs' technology (Drmanac et al. 2010) by Complete Genomics, and Ion Torrent, wherein nucleotide incorporation is measured by the difference in pH caused by release of protons during polymerisation, rather than conventional light detection (Rothberg et al. 2011). However, none have yet matched the success of the widely adopted Illumina instruments (Greenleaf and Sidow 2014), which have brought rapid improvements to quality, through-put and cost of sequencing (Check Hayden 2014).

The rapid development of ever-improving DNA sequencing technologies has even outpaced the 'computer revolution', with the sequencing cost per nucleotide halving every five months between 2004 and 2010 (Stein 2010). Recent years have now seen the emergence of 'third generation' instruments, which are capable of single molecule sequencing (thus negating the requirement for the amplification of DNA and avoiding amplification-related biases), as well as producing much longer reads than the established Illumina technologies (Schadt et al. 2010).

The developments in DNA sequencing instrumentation have drastically decreased not just the cost, but also the ease of sequencing, and as such, the numbers of sequencing experiments performed have increased exponentially. The first major sequencing project following The Human Genome Project – the 1000 Genomes Project (Project Consortium et al. 2012) – generated twice as much sequencing data in the first 6 months than had been deposited in the entire GenBank database (Benson et al. 2005) in the preceding 30 years (Stein 2010) **(Figure 3)**.



**Figure 3.** Annually deposited ‘Next Generation’ sequencing data in public data repositories. Top panel shows the number of annually deposited datasets in ArrayExpress database (Kolesnikov et al. 2015). The numbers of experiments using array technologies have been steadily declining, while sequencing experiments have become more popular. Similarly, the annually submitted sequencing runs within SRA sequence archive (Leinonen et al. 2011a) have been increasing (bottom panel).[Date accessed: 09/02/2017]

### ***1.3 “Next-Generation” Sequencing Applications***

The increasing levels of automation and ease of library preparation together with the rapidly decreasing sequencing costs have also driven the application of NGS beyond simple determination of DNA sequences.

In human genetics, DNA sequence variant characterisation has proved hugely successful in disease gene identification. In the past, as the whole genome of an individual was simply too big, costly and time-consuming to sequence, large families were required in order to first map the disease-causing locus. However, NGS-based approaches, such as Whole-Exome sequencing (WES) (Rabbani et al. 2014) of parent-child trios, have led to many successes in identifying causative genes for *de novo*, dominant and recessive diseases (Vissers et al. 2010; Ng et al. 2010b; Zhang et al. 2013). Similarly, NGS has enabled DNA structural/copy number variation detection via break-point mapping and/or coverage-based methods, with successful applications in both human genetics (Coe et al. 2014) and cancer studies (Wang et al. 2015a). Targeted sequencing has also become a routinely adopted clinical diagnostics tool (Trujillano et al. 2017).

The field of epigenetics has also benefitted extensively from the advent of NGS technologies. Various types of genome-wide DNA methylation analyses have gained popularity in recent years, including the sequencing of bisulfite treated DNA or immunoprecipitation-based techniques, such as MeDIP-Seq (Wilson et al. 2012) and MIRA-Seq (Jung et al. 2015). Novel insights into genome regulation have also been gained via other NGS applications, such as genome-wide mapping of DNase I hypersensitive sites (Crawford et al. 2006) or transcription factor binding sites via ChIP-seq (Ren et al. 2000). The ENCODE project (ENCODE Project Consortium et al. 2012) was launched shortly after the completion of The Human Genome Project with the aims to characterise functional genome elements largely via the use of NGS technologies such as transcription factor ChIP-Seq, histone mark enrichment ChIP-Seq and DNase-Seq, has proven – in terms of amount of data generated - a landmark success. Indeed, the original aims of the project have since

expanded, and now encompass cataloguing of RNA gene expression and RNA-protein interactions.

In the studies of RNA gene expression, NGS has come to replace microarray technologies as a transcriptome profiling method of choice, due to its higher sensitivity and specificity. Transcriptome-wide RNA sequencing has enabled not only accurate quantification, but also characterisation of alternative splicing events and allele-specific expression (Edsgård et al. 2016; Ding et al. 2017). Many areas of molecular biology were re-discovered, including the study of small RNA species (e.g. microRNAs) by small RNA-sequencing methodologies (Persson et al. 2017; Campbell et al. 2015); or circular RNAs via the sequencing of RNase-treated samples (Salzman et al. 2012).

More recent additions to the NGS toolbox for the study of RNA include ribosome/polysome profiling (Ingolia et al. 2009) for the identification of actively translated RNA species; nascent RNA sequencing for characterising transcriptional events (Carrillo Oesterreich et al. 2010); antibody pull-down based methods for the identification of RNA-protein interaction sites, such as eCLIP, PAR-CLIP, RIP-Seq and fRIP-Seq (Van Nostrand et al. 2016; Spitzer et al. 2014; Wessels et al. 2016; G Hendrickson et al. 2016); characterisation of RNA modifications such as pseudouridylation, adenosine methylation or deamination via mutation-detection, reverse-termination or antibody-based methods (Carlile et al. 2015; Dominissini et al. 2013; Chepelev 2012); or even transcriptome-wide RNA secondary structure analysis via methods such as PARS (Wan et al. 2013).

The rise in the variety of NGS applications have naturally resulted in unprecedented amounts of data generated. Sequence data in public repositories has grown exponentially from a small handful of assembled genomes less than two decades ago: NCBI now hosts 103,417 complete prokaryotic and 4,586 eukaryotic genome sequences, while SRA boasts 3,179,619,260,782,116 total sequence bases (Leinonen et al. 2011a) [Accessed 29/07/2017]. Thus, it is clear that the

bottleneck in 'omics' research has shifted from data generation to data analysis, bringing about a new set of challenges and opportunities.

#### ***1.4 Bioinformatics Challenges in 'Next-Generation' Sequencing Era***

In any type of NGS experiment - clinical, forensic or research - once samples are sequenced, bioinformaticians must tackle the task of storing, analysing and interpreting the data. Yet, the field of bioinformatics has struggled to keep up with the rapidly advancing NGS technologies. While there are many challenges posed in terms of data storage, quality control, assembly, alignment or annotation, this is mostly felt in the area of analysis and interpretation. Translating huge NGS datasets into interpretable biological hypotheses or actionable clinical diagnoses remains perhaps the biggest challenge in bioinformatics.

As with any new research avenue, algorithms in these areas will take time to develop and mature. To an extent, this has already happened in several key areas, such as short read sequence alignment. Some of the early algorithms of this type, such as Needleman–Wunsch (Needleman and Wunsch 1970), Smith-Waterman (Waterman 1984), BLAST (Altschul et al. 1997) and CLUSTAL (Chenna et al. 2003), are either still utilised today or have served as a basis for more scalable improvements, which balance accuracy, speed and memory requirements. Indeed, there have been few notable improvements on the performance of the gold standard short read sequence alignment software (Engström et al. 2013; Shang et al. 2014), such as BWA (Li and Durbin 2009) and Bowtie (Langmead and Salzberg 2012) (or STAR (Dobin et al. 2013) and Tophat (Trapnell et al. 2009) for split-read alignment), and with only seemingly marginal gains still to be exploited, the problem can be considered largely 'solved' for current datasets

Thus, while a community consensus has been reached in many core areas of current applications in NGS data analysis, much still remains an open problem. Tools to aid the biological interpretation of large datasets are still lacking, and as such, NGS data is largely under-utilised. It is not only the typical research projects that suffer from this failing. Large sequencing projects, such as the 1000 genomes

project (Project Consortium et al. 2011) or ENCODE (ENCODE Project Consortium et al. 2012), focused very much on the data generation aspects and less so on the interpretation of the resulting data. Thus, it will take the bioinformatics community years to fully digest the outcome of these large-scale efforts, which have yielded enormous databases but relatively few biological insights.

This presents a ripe opportunity for biological data mining, which can greatly aid hypothesis generation and data interpretation. How does one extract biological meaning from tables of gene expression values? How can a disease-causing or a susceptibility variant be pinpointed from the tens of thousands of potential candidates identified in an NGS experiment? There are many such open problems in bioinformatics.

However, the lack of available software is not the only bottleneck in NGS data interpretation. It is important for bioinformatics to interface closely with biology specialists, as while much of the NGS data analysis can be automated, data interpretation requires significant human input. Yet there is currently a large accessibility issue— software and algorithms are largely created by and for those with computer science skills and may inhibit their use by bench scientists with little computing training. Projects such as BioPython (Cock et al. 2009), BioPerl (Stajich et al. 2002) and Bioconductor (Gentleman et al. 2004) between them boast many state-of-the-art bioinformatics algorithms, yet require programming knowledge to use. Similarly, the majority of bioinformatics software is limited to Unix-type systems, have complicated installation procedures and typically require the user to be proficient at command line and bash scripting. Thus, in order to overcome the chasm dividing the NGS data generation from data interpretation, not only are novel bioinformatics software and algorithms are needed, but their implementation must also focus on improving accessibility and usability.

## ***1.5 Overview of this work***

In light of the bioinformatics challenges outlined above, this thesis focuses on the development of novel bioinformatics software/algorithms to aid NGS data analysis and interpretation in the biomedical domain. As the current availability of open access NGS data allows for unprecedented integration of data mining approaches, much of this work describes re-purposing public data sets to extract the most out of this under-utilised resource. Furthermore, in contrast to much of the bioinformatics software, this work aims to implement the algorithms described herein as user-friendly, cross-platform software accessible without specialist programming knowledge or the use of command line.

The first part of the thesis presents novel algorithms for candidate disease gene and variant prioritisation and interpretation. Two different approaches are described and implemented as database-driven online web applications:

- GeneTiER: an unbiased, data driven approach explores the possibility of prioritising disease genes based on RNA expression profiles
- OVA: a knowledge-based approach that focuses on data mining biomedical domain ontology resources, which have recently become the standard for describing biological knowledge in machine-readable way

The second part of this work focuses on the development of m6aViewer – a novel application to detect, visualise and interpret transcriptome-wide N6-methyl adenosine (m<sup>6</sup>A) RNA modifications from sequencing data. The software focuses on probabilistic modelling to achieve higher resolution than other currently available approach and uses data mining and machine learning techniques to improve the specificity of detected residues.

Finally, the last part of this work describes the application of m6aViewer to characterise transcriptome-wide cellular and viral m<sup>6</sup>A landscapes during the course of Kaposi's Sarcoma Herpes Virus (KSHV) infection. KSHV transcriptome is

revealed to be heavily m<sup>6</sup>A methylated, with m<sup>6</sup>A present in key viral genes that control progression from latent to lytic stages in the viral life cycle. The work also describes a putative novel m<sup>6</sup>A reader protein, SND1, and its potential role as an effector protein of an m<sup>6</sup>A –mediated regulation of KSHV latent-lytic cycle progression.

## ***2. Candidate Disease Gene and Variant Prioritisation***

### ***2.1 Disease Gene/Variant Identification***

There are more than 6000 rare inherited diseases currently catalogued by Online Mendelian Inheritance In Man (OMIM) (Amberger et al. 2015). While individually they affect less than 0.1% of the population, collectively, rare disease represents a significant public health issue, affecting up to 8% of the population (Forman et al. 2012). Rare diseases comprise a very heterogeneous set of conditions, varying in rareness, onset, prognosis, penetrance and organ systems affected. This presents barriers for accurate diagnosis and treatment, as well as limiting commercial interest in research investment and treatment development. Furthermore, due to the rarity of individual conditions, the numbers of eligible participants for research studies are often extremely limited; such patients may be geographically dispersed; or even be misdiagnosed due to the lack of knowledge by the treating physician about the disease. As a result, more than a quarter (1640 out of 6218) of all rare conditions indexed by OMIM have an unknown molecular basis [accessed 11/11/2015].

#### ***2.1.1 Genetic Linkage and Association Studies***

There are several experimental methods commonly used to elucidate the inherited basis of human disease. The molecular aetiology of the majority of the conditions catalogued by OMIM has largely been determined through linkage mapping, historically a difficult and labour-intensive process.

Linkage analysis relies on the co-segregation of sequences – that is, polymorphisms at the same chromosomal loci tend to be inherited together. Thus, disease-causing variants can be mapped by identifying known polymorphisms that co-segregate with the disease phenotype. By linking the disease to a number of polymorphic alleles, a disease haplotype can be identified and the limits of the disease locus can be narrowed down to the chromosomal regions where affected family members share the same haplotype.

The early linkage maps were composed of a series of microsatellite markers. One of the first linkage maps in humans consisted of some 400 common DNA markers (Donis-Keller et al. 1987), but had grown to more than 5000 in 1996 (Dib et al. 1996), and with it, the number of known disease genes.

Initially, disease gene mapping was hampered by the poor level of annotation of the human genome, which meant it was both difficult to identify genes in a disease locus and to rapidly screen the sequences for possible deleterious variants. However, there were notable early successes, such as Huntingdon's disease (Gusella et al.) and cystic fibrosis (Kerem et al. 1989), as well as identification of Mendelian subtypes of more common diseases, such as diabetes (Bell and Polonsky 2001) and hypertension (Lifton 2004).

In the cases of complex disease, association studies that evolved alongside linkage mapping were less successful. Although many studies claiming associations between genetic variants and affected individuals were reported, the statistical significance would often be weak or the results could not be replicated (Moskvina and O'Donovan 2007; Ioannidis et al. 2006; Lohmueller et al. 2003). Furthermore, even in successful studies, the effect sizes tended to be small, with most identified alleles increasing disease risk by a factor of less than 1.5. A few notable exceptions have been the linking of APOE4 to Alzheimer's disease (Strittmatter and Roses 1996) and CFH to age-related macular degeneration (Klein et al. 2005).

In the last 20 years, both linkage and association studies have been greatly aided by the sequencing and annotation of the human genome, and then by the advances in NGS technologies. Where previously the sequencing of even a small genomic locus was a time and labour-intensive task, modern high-throughput approaches generate large volumes of sequence data, which after comparison to the human genome reference sequence, allow rapid identification of variants on a genome-wide scale in a short time frame. Consequently, purely NGS-based approaches have largely superseded traditional linkage studies.

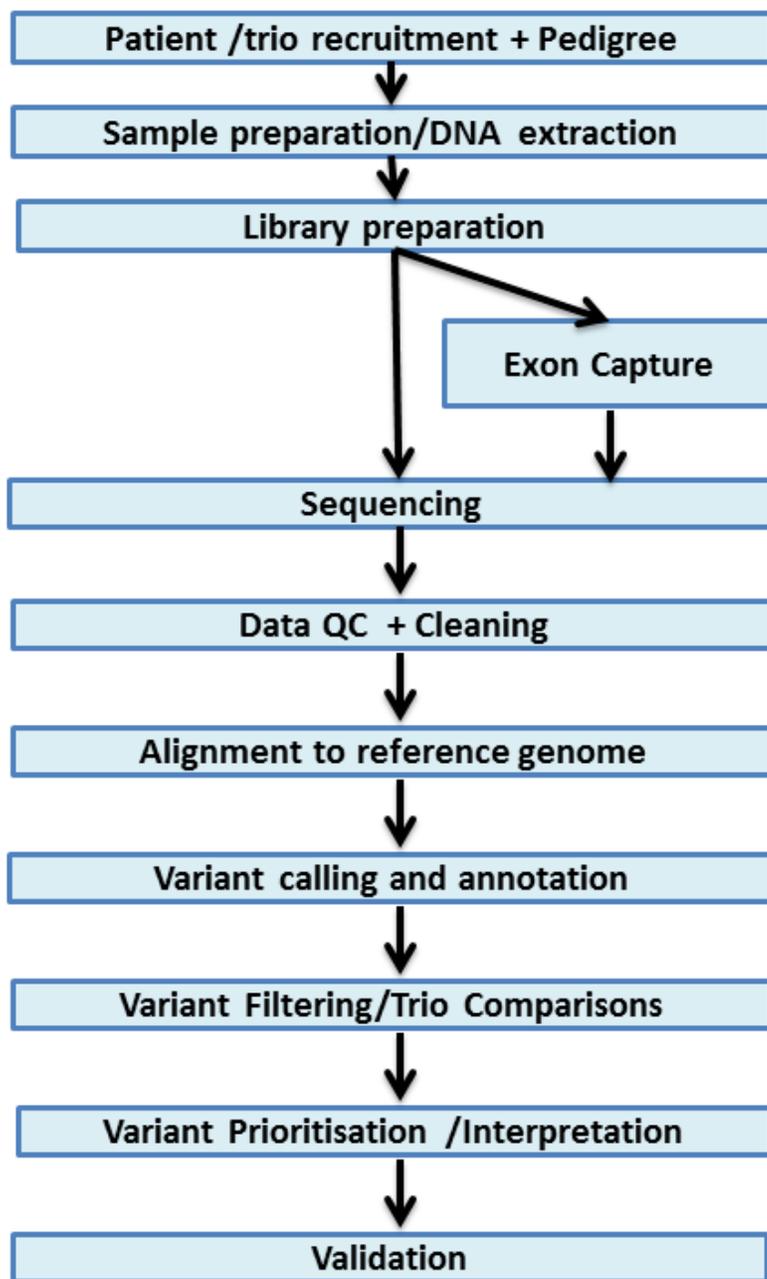
### ***2.1.2 Whole Exome and Whole Genome Sequencing Studies***

Whole Genome Sequencing (WGS) is a relatively unbiased method for disease gene identification and typically involves sequencing the human genome at 30-fold or greater coverage. Whole Exome Sequencing (WES), on the other hand, is the selective sequencing of only the coding parts of the human genome and has been more widely applied due to the dramatically lower cost per sample. WGS and WES share much of the same workflow, which is summarised in **Figure 4**. Downstream of sequencing data generation, the short sequence reads are aligned to the reference genome and analysed to detect and genotype single nucleotide substitutions, insertions and deletions. These can be compared to the unaffected samples and/or allele frequencies in the general population in order to identify candidate disease variants.

Thus far, NGS-based approaches have been hugely successful. In a relatively short time frame, WES has been used to identify hundreds of disease-gene associations in both dominant and recessive cases, including Miller syndrome (Ng et al. 2010b), Kabuki syndrome (Ng et al. 2010a) and Joubert syndrome (Srour et al. 2012).

WES has proved so successful largely due to the tendency of highly penetrant phenotypes to arise from deleterious changes/loss of function in proteins, with only a small proportion of Mendelian disease variants found in non-coding/regulatory regions of the genome (Botstein and Risch 2003). This may be due at least in part to an observational bias, as coding sequences are often the major focus of diagnostic and research projects in human genetics. As WGS is becoming more cost-effective, however, more disease-causing mutations in regulatory regions can be expected to be uncovered. Putative disease variants in non-coding regions have been found for diseases such as autism (Turner et al. 2016); preaxial polydactyly (Furniss et al. 2008); Hirshsprung disease (Emison et al. 2005) and Pierre Robin sequence (Gordon et al. 2014), amongst others. Where in the past novel disease gene discovery would take many years, given a suitable pedigree, NGS-based approaches can identify a deleterious variant in a matter of months.

However, some disease-causing mutations are easier to identify than others. As autosomal recessive diseases are defined by the presence of two deleterious alleles (compared to only one in dominant disease), the link to a causative variant is more easily made. The use of autozygous mapping strategies with either consanguineous families or patients from genetically isolated populations has further aided identification of deleterious recessive variants. For instance, in affected individuals from consanguineous families, disease causing variants are typically inherited from the same ancestral allele via both parents, and can be characterised by an extended run of homozygous polymorphic variants at the disease locus. The higher incidence rate of rare diseases in such populations, combined with reduced search space by isolation of runs of homozygosity, has led to the discovery of many novel autosomal recessive disease genes. Consequently, autosomal recessive diseases have been over-represented in the early WES studies.



**Figure 4.** Overview of typical workflow in WGS and WES studies in disease gene identification. Samples are typically obtained from patients recruited together with unaffected family members. Sequencing libraries are prepared from extracted DNA (WGS), or in the case of WES, post exon capture. The libraries are sequenced, raw intensity values are converted to base calls and stored in Fastq format. Fastq reads are quality checked and trimmed to remove sequencing adapters and poor-quality bases before aligning to the reference human genome. Variants are called by comparing genomic reference positions to the sequenced

data and allelic ratios can be obtained from the proportion of total aligned reads supporting each allele. Typically, detected variants are compared to the unaffected samples and further filtered based on expected inheritance mode, type of variant or allele frequencies in general population. The remaining variants are typically prioritised based on available biological information and the relevance to disease can be validated by functional studies.

---

When data on the patient and both parents is available, *de novo* dominant mutations are typically straight-forward to identify, as comparatively few novel variants exist between the parents and affected children.

One of the most exciting applications of NGS to disease gene discovery, however, has been in identifying mosaic mutations. Mosaic mutations are often identified as an unexpected allele read depth ratio in the sequencing data, for example a *de novo* mutation in PIK3CA gene, attributed as a cause of megalencephaly–capillary malformation, was present in only 11% of reads mapping to the variant site in an affected patient’s leukocytes (Rivière et al. 2012).

While NGS has contributed to many successes in disease gene discovery, the failure rate is harder to estimate as negative findings are rarely published. It is clear that even in familial cases with well-defined pedigrees, NGS approaches can generate thousands of candidate variants with the expected segregation characteristics, with hundreds or more of these resulting in non-synonymous substitutions, most of which will have no effect on the phenotype of interest. Similarly, linkage or association studies may reveal the genomic location of the disease causing gene; however, these intervals are often several mega bases in length and can contain a large number of candidate genes and associated sequence variants. Consequently, it is a non-trivial task, both in terms of time and resources, to deduce the disease-causing mutation(s) via experimental verification of putative disease-causing variant.

### ***2.1.3 Putative disease variant prioritisation***

A more common and economical approach of identifying deleterious variants in a large set of benign polymorphisms relies on examination of information pertaining to each candidate, a wealth of which is available to researchers in the form of biological databases or literature. This includes data from sources such as previous studies, functional annotations, protein-protein interactions, biological pathways or model organism phenotypes, all of which can be taken into consideration in order to choose the candidates most likely to be disease causative.

For cases where similar disease phenotypes have been previously characterised, prioritising candidates based on information pertaining to gene molecular functions is perhaps the most intuitive approach. Mutations in genes acting in the same or similar biological processes are also likely to cause similar phenotypes. A growing number of databases catalogue biological function and pathway information obtained either via automated text-mining approaches or manual literature curation. However, this approach is inherently biased towards well-studied genes and processes, and may miss better candidates solely because our understanding of biological processes is far from complete.

Protein and gene interaction databases could stand in as a less biased alternative to manually curated, but smaller, curated pathway databases such as KEGG (Kanehisa and Goto 2000) and Reactome (Fabregat et al. 2016) are still largely based on scientific literature. Data from a number of high-throughput protein-protein interaction screens are publicly available in databases such as STRING (Jensen et al. 2009); however, such data is mostly obtained via *in vitro* studies such as yeast-two-hybrid screens and may not always be representative of real world protein interactions *in vivo*. For example, while two proteins may exhibit strong interactions *in vitro*, *in vivo* they may be trafficked to separate cellular (or extra-cellular) compartments and participate in largely unrelated pathways. Similarly, important protein interaction may only occur after protein modification, such as phosphorylation, which may not occur in yeast. Furthermore, proteins are

often multi-functional and may have different interacting partners across different tissues.

In general, when prior knowledge about the molecular phenotype of the disease is available, 'guilt-by-association' principle can be employed to identify other candidate genes which, when perturbed, would contribute to the same phenotype. Knowledge-based approaches, such as database/literature search, often fail to identify poorly characterised genes which participate in disease pathways. However, when the molecular mechanism of a disease is known, these genes are often the prime disease gene candidates.

Another common approach to disease gene prioritisation is to consider the effects of mutations/knock-outs of a putative candidate gene in a model organism system, as phenotypic effects of mutations in animal homologs are often similar to those seen in humans. This approach suffers from similar shortcomings, however, in that it is biased towards better studied genes. In the future, this may be at least partially rectified by large scale projects, such as Mouse Phenotyping Consortium (Meehan et al. 2017; Koscielny et al. 2014), which aim to characterise and catalogue the effects of individual phenotypic effects of null mutations in all mouse proteins.

Still, often mutations which have severe phenotypic effects in humans have very different or even no observable effect in model organism counterparts, and *vice versa*. For example, Amish infantile epilepsy syndrome is an autosomal recessive disorder characterised by recurrent seizures, developmental regression and hypo- or hyper-pigmented skin macules, whereas in mice the null mutation of the disease gene *ST3GAL5* manifests in hypoglycaemia and increased insulin sensitivity (Boccuto et al. 2014; Blake et al. 2017). Similarly, null mutations of *TGIF1*, the gene responsible for Holoprosencephaly-4, have no observable phenotypic effect in mice (Petryk et al. 2015; Blake et al. 2017).

Animal model systems are also further complicated by the spectrum of potential mutations – that is, mutations in different parts of a gene may have very different

phenotypic effects and it is infeasible to induce every possible mutation in model organisms to facilitate an unbiased search. As a result, data from knock-out and mutation studies in model organisms may be even more limited for discerning the genetic causes of a gain-of-function disease, rather than a loss-of-function one. However, even in loss-of-function diseases, the essentiality of a gene cannot always be adequately assessed in the lab, for instance in immune-related genes or in late-onset diseases.

Gene expression information can also aid in prioritising candidate disease genes based on a number of different approaches. Gene co-expression analysis is an unbiased way of highlighting genes which are likely to act together and therefore, have closely related functions, participate in the same pathways or directly interact to form protein complexes.

Similarly, where gene expression data is available in both patient and normal samples, differential expression analysis may highlight perturbed pathways and genes. A differentially expressed gene/pathway can point to a mutation in regulatory regions, while RNA sequencing data in particular can also highlight aberrant splicing events, which may arise as a consequence of a splice-site mutation.

However, gene expression data is not frequently used in such a manner, as it can be expensive to generate, is inherently noisy, and often, tissue-specific samples that would be required are impossible to obtain. Furthermore, variant calling and genotyping from RNA-Seq data is difficult due to varying transcript abundances, allele-specific expression or post-transcriptional editing events. Matched WGS or WES data is nevertheless required for variant identification; thus, increased sequencing costs often prohibit matched RNA-DNA sequencing experiments.

In high-throughput sequencing approaches, alternative methods of prioritising disease genes without relying on prior knowledge often compare gene sequence properties between candidates. Intrinsic features such as the length of a gene, or

the number of paralogs it has, are known to be predictive of disease genes. Longer genes, for example, have been shown to be statistically more likely to be disease causative genes, even when normalised for sequence length (Lopez-Bigas and Ouzounis 2004). Genes which have few closely related paralogs may be more likely to be disease causative, as there are likely to exist fewer functionally similar proteins to compensate for loss-of-function mutations, resulting in a more penetrant phenotype (Lopez-Bigas and Ouzounis 2004).

Similarly, highly conserved genes or residues across multiple species can be indicative of their importance, and such genes harbouring putative loss-of-function mutations present a case for a strong candidate gene. On the other hand, genes which are extremely well conserved even across distant species are unlikely to tolerate mutations and thus genetic defects in these genomic regions are more likely to result in perinatal or *in utero* termination, as opposed to presenting as a childhood or later onset disease (Huang et al. 2004). Still, sequence variants altering conserved positions in protein sequences are considered *a priori* more likely to be damaging, as these positions are more likely to have important molecular functions. Indeed, rare human alleles have been shown to be strongly negatively correlated with purifying selection pressure (Kryukov et al. 2007).

Indubitably, sequence is one of the most valuable indicators of the pathogenicity of a variant. While most sequence variation is considered neutral (Dudley et al. 2012), nucleotide substitutions, insertions and deletions can have an effect on gene expression and protein function. Some of these changes can be well tolerated, some beneficial, while others can manifest a deleterious phenotype. In general, non-synonymous variants which change the amino acid sequence of a protein are considered more likely to be deleterious than synonymous substitutions, as this type of variation can have a dramatic impact on the function of a protein. In the OMIM database, which catalogues genetic causes of inherited disease, more than half of all reported deleterious human variation is comprised of non-synonymous substitutions (Amberger et al. 2015) [Accessed 11/11/2015]. This suggests that these types of mutations are more likely to have a disease-causing

effect; however, these variants are also easier to identify and this may also contribute to this observed imbalance.

Naturally, not every amino acid change has an equal effect – the type of substitution and where it occurs is very important. For example, a change from one hydrophobic amino acid to another may not result in as damaging a perturbation to protein function as a change to a hydrophilic one. Similarly, a change from a small amino acid to a larger one may cause steric hindrance and perturb protein folding or protein ligand binding, while a more subtle change may have no effect. Location is just as important – an amino acid change within an active site of an enzyme may be more deleterious to protein function than one occurring in less structured or non-catalytic regions. Similarly, it has been demonstrated that disease-causing mutations are more likely to affect internal protein residues, with solvent accessibility of the mutation site being a key predictor of a deleterious mutation (Chen and Zhou 2005). However, as a typical human genome is thought to contain approximately 24,000 to 40,000 heterozygous non-synonymous SNPs (Auton et al. 2015), it is clear that the majority of non-synonymous variation is neutral.

Similarly, insertions and deletions also occur regularly within the genome and may have differing effects on the fitness of a protein. Small, in-frame insertions or deletions are more likely to be tolerated and are thus more common than indels resulting in a translational frame-shift. Frame-shift mutations towards at the end of a protein may have a less damaging effect than those earlier in the sequence, as the latter are more likely to affect key downstream protein domains.

Nevertheless, variants deleterious to protein function do not necessarily produce a disease phenotype; some cases may be dosage-insensitive, and thus a damaging heterozygous allele may be compensated by the non-mutant copy; similarly, homozygous loss-of-function variation may be compensated by other proteins with paralogous functions. In fact, it is estimated that the average human genome harbours approximately a hundred genuine loss-of-function mutations (MacArthur et al. 2012).

While synonymous mutations are generally considered silent, they may have subtle, but critical effects that are less easily understood. Mueller *et al* (2015) argue that as much as 45% of synonymous mutations may impact pre-mRNA splicing (Mueller et al. 2015). While the integrity of the open reading frame sequence in an mRNA is very important, synonymous changes in transcript sequence can affect a wide range of crucial regulatory functions. Thus, mutations considered silent can still impact a transcript's ability to bind proteins or other regulatory RNAs in a sequence- or secondary structure specific manner. This may then affect RNA splicing, export and localisation, the stability of the mRNA, translation efficiency, or even the final sequence of an encoded protein via aberrant RNA editing/modifications. Indeed, functional synonymous mutations in regulatory regions have been identified in diseases like autism, schizophrenia (Takata et al. 2016) and melanoma (Gartner et al. 2013).

Variants in non-coding regions may also be damaging. A mutation in a promoter or enhancer region may have a severe impact on gene expression and may result in dysregulation of otherwise tightly-controlled proteins.

In summary, candidate disease genes/variants can be prioritised based on multitude of different criteria. Examining biomedical literature and databases, as well as sequence properties, has often proved successful. This approach, however, can be painstakingly slow and prone to human errors and biases, as each candidate is not independently classified; nor is this likelihood quantified when the choice is formed solely by the impressions of the researcher.

## ***2.2 Computational Candidate Disease Gene Prioritisation***

In light of these challenges, various computational gene prioritisation approaches have been proposed in recent years that aim to automate this task to various degrees. Candidate gene prioritisation remains an active area of research despite a considerable number of algorithms and applications that have been developed (Adie et al. 2006; Chen et al. 2011b; Britto et al. 2012; Nitsch et al. 2011; Zhang et al. 2013; Seelow et al. 2008; Aerts et al. 2006; Smedley and

Robinson 2015; Bornigen et al. 2012; Chen et al. 2009; Jiang 2015). **Table 1** summarises candidate disease gene and variant prioritisation tools published prior to this work and the diverse data sources that they utilise. Thus far, while new methods and improvements continue to be introduced, there has been no universally applicable or precise approach.

Classically, gene prioritization tools have been geared towards scrutinizing regional gene sets obtained from linkage studies. However, in recent years, next generation sequencing has become a *de facto* standard method for disease gene discovery in Mendelian diseases. Consequently, a few applications have recently emerged that expand and/or adapt currently used gene prioritization approaches to be more applicable for the evaluation of variants found in exome datasets. For example, the Exomiser tool (Robinson *et al.*, 2014) supplements variant pathogenicity scoring with an algorithm for comparing human diseases with mouse phenotypes, while ExomeWalker (Smedley *et al.*, 2014) incorporates interactome data from STRING (Jensen *et al.*, 2009). PriVar (Zhang *et al.*, 2013) combines variant pathogenicity scores from multiple sources together with pedigree information to rank variants.

Variant prioritization is not a novel concept – established algorithms like SIFT (Kumar *et al.*, 2009) and POLYPHEN (Adzhubei *et al.*, 2013) assess the likelihood of pathogenicity using information such as positional conservation across homologs or the effects the change is likely to have on the protein. However, this approach is not without drawbacks – often variants predicted to be deleterious will produce no visible changes in phenotype due to, for example, the redundancy in the genome, as demonstrated by cases of synthetic lethality (Lord and Ashworth 2017). This approach is also limited to phylogenetically conserved regions; consequently, SIFT and PolyPhen can only score 60-81% of the proteome (Adzhubei *et al.* 2010).

Variant filtering approaches tailored to individual situations provide an alternative to pathogenicity scoring for reducing the size of candidate variant lists. Tools like

AgileExomeFilter (Watson *et al.*, 2014) allow filtering of variants on a variety of criteria, such as inheritance mode, regions of autozygosity, sequencing quality or variant types thought to mostly be benign, for example synonymous substitutions or small in-frame insertions or deletions. However, as human exomes typically contain in excess of 30,000 variants, often neither approach proves adequate for pinpointing the correct mutation.

**Table 1.** Summary of a number of candidate gene and variant prioritisation tools published prior to this work. The right hand side columns indicate the types of information used for prioritisation by each tool.

Tool	Publication	Description	Biological Functions	Gene Expression	Gene Regulation	Text-mining	Protein interactions	Pathways	Sequence	Phenotype	Conservation	Other
aGeneApart	Van Vooren <i>et al.</i> , <b>Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations</b> , <i>Nucleic Acids Res.</i> (2007)	Text-mining of MEDLINE/abstracts to annotate/associate genes to biomedical concepts	No	No	No	Yes	No	No	No	No	No	No
Biomine	Eronen <i>et al.</i> , <b>Biomine: predicting links between biological entities using network models of heterogeneous databases</b> , <i>BMC Bioinformatics</i> (2012)	Integration of data from heterogeneous data sources	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Bitola	Hristovski <i>et al.</i> , <b>Using literature-based discovery to identify disease candidate genes</b> , <i>Int. J. Med. Inform.</i> (2005)	Text-mining of MEDLINE/abstracts to annotate/associate genes to biomedical concepts	No	No	No	Yes	No	No	No	No	No	No
Candid	Hutz <i>et al.</i> , <b>CANDID: a flexible method for prioritizing candidate genes for complex human traits</b> , <i>Genet. Epidemiol.</i> (2008)	Integration of data from heterogeneous data sources	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes
CGI	Ma <i>et al.</i> , <b>CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data</b> , <i>Bioinformatics</i> (2007)	Network-based prioritisation by combination of gene expression and protein interaction data	No	Yes	No	No	Yes	No	No	No	No	No
DIR	Chen <i>et al.</i> , <b>In silico gene prioritization by integrating multiple data sources</b> , <i>PLoS One.</i> (2011)	Integration of data from heterogeneous data sources	No	Yes	No	No	Yes	Yes	No	No	No	No

Tool	Publication	Description	Biological Functions	Gene Expression	Gene Regulation	Text-mining	Protein interactions	Pathways	Sequence	Phenotype	Conservation	Other
DomainRBF	Zhang et al, <b>DomainRBF: a Bayesian regression approach to the prioritization of candidate domains for complex diseases.</b> BMC Systems Biology (2011)	Prioritises genes based on domain information	No	No	No	No	No	No	Yes	No	No	No
Endeavour	Aerts <i>et al.</i> , <b>Integrating Computational Biology and Forward Genetics in Drosophila</b> , <i>PLoS Genetics</i> (2009) ;Aerts et al., <b>Gene prioritization through genomic data fusion</b> , Nature Biotechnology (2006)	Integration of data from heterogeneous data sources	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes
ExomeWalker	Smedley et al, <b>Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases.</b> Bioinformatics (2014)	Integration of pathogenicity scoring with protein interaction network	No	No	No	No	Yes	No	Yes	No	Yes	Yes
Exomiser	Smedley et al, <b>Next-generation diagnostics and disease-gene discovery with the Exomiser</b> , Nature Protocols (2015)	Integration of pathogenicity scoring with protein interaction network and phenotype information	No	No	No	No	Yes	No	Yes	Yes	Yes	Yes
G2D	Perez-Iratxeta <i>et al.</i> , <b>Update of the G2D tool for prioritization of gene candidates to inherited diseases</b> , <i>Nucleic Acids Res.</i> (2007) ;Perez-Iratxeta et al., <b>G2D: a tool for mining genes associated with disease</b> , BMC Genet. (2005)	Integration of data from heterogeneous data sources	Yes	No	No	Yes	Yes	No	Yes	No	No	No
GeneDistiller	Seelow <i>et al.</i> , <b>GeneDistiller--distilling candidate genes from linkage intervals</b> , <i>PLoS</i>	Integration of data from heterogeneous data sources	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	No

Tool	Publication	Description	Biological Functions	Gene Expression	Gene Regulation	Text-mining	Protein interactions	Pathways	Sequence	Phenotype	Conservation	Other
	<i>ONE (2008)</i>											
GeneFriends	van Dam <i>et al.</i> , <b>GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases</b> , <i>BMC Genomics (2012)</i>	Prioritisation via gene co-expression networks	No	Yes	No	No	No	No	No	No	No	No
GeneProspector	Yu <i>et al.</i> , <b>Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases</b> , <i>BMC Bioinformatics (2008)</i>	Prioritisation via highly curated literature database	No	No	No	Yes	No	No	No	No	No	No
GeneRank	Morrison <i>et al.</i> , <b>GeneRank: Using search engine technology for the analysis of microarray experiments</b> , <i>BMC Bioinformatics (2005)</i>	Functional annotation and gene co-expression network prioritisation	Yes	Yes	No	No	No	No	No	No	No	No
GeneRanker	Gonzales <i>et al.</i> , <b>GeneRanker: An Online System for Predicting Gene-Disease Associations for Translational Research</b> , <i>Summit on Translat Bioinforma. (2008)</i>	Integration of data from heterogeneous data sources	Yes	No	No	Yes	Yes	No	No	Yes	No	No
GeneSeeker	van Driel <i>et al.</i> , <b>GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases</b> , <i>Nucleic Acids Res. (2005)</i>	Integration of data from heterogeneous data sources	No	Yes	No	Yes	No	No	No	Yes	Yes	No

Tool	Publication	Description	Biological Functions	Gene Expression	Gene Regulation	Text-mining	Protein interactions	Pathways	Sequence	Phenotype	Conservation	Other
GeneWanderer	Kohler <i>et al.</i> , <b>Walking the interactome for prioritization of candidate disease genes</b> , <i>Am. J. Hum. Genet.</i> (2008)	Network-based prioritisation by combination of gene expression, protein interaction and other data types	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No
Génie	Fontaine <i>et al.</i> , <b>Génie: literature-based gene prioritization at multi genomic scale</b> , <i>Nucleic Acids Res.</i> (2011)	Text mining and sequence homology	No	No	No	Yes	No	No	No	No	Yes	No
GenTrepid	George <i>et al.</i> , <b>Analysis of protein sequence and interaction data for candidate disease gene prediction</b> , <i>Nucleic Acids Research</i> (2006)	Integration of data from heterogeneous data sources	No	No	No	Yes	Yes	Yes	Yes	No	No	No
GLAD4U	Jourquin <i>et al.</i> , <b>GLAD4U: deriving and prioritizing gene lists from PubMed literature</b> , <i>BMC Genomics</i> (2012)	Prioritisation based on resources at NCBI	No	No	No	Yes	No	No	No	No	No	No
GPSy	Britto <i>et al.</i> , <b>GPSy: a cross-species gene prioritization system for conserved biological processes--application in male gamete development</b> , <i>Nucleic Acids Research</i> (2012)	Integration of data from heterogeneous data sources	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
GUILD	Guney <i>et al.</i> , <b>Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization</b> , <i>PLoS One</i> (2012)	Text mining gene-disease associations	No	No	No	No	Yes	No	No	No	No	No

Tool	Publication	Description	Biological Functions	Gene Expression	Gene Regulation	Text-mining	Protein interactions	Pathways	Sequence	Phenotype	Conservation	Other
MedSim	Schlicker <i>et al.</i> , <b>Improving disease gene prioritization using the semantic similarity of Gene Ontology terms</b> , <i>Bioinformatics</i> (2010)	Integration of data from heterogeneous data sources	Yes	No	No	No	Yes	No	No	Yes	Yes	No
MetaRanker	Pers <i>et al.</i> , <b>MetaRanker 2.0: a web server for prioritization of genetic variation data</b> , <i>Nucleic Acids Research</i> (2013) ;Pers <i>et al.</i> , <b>Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes</b> , <i>Genet Epidemiol.</i> (2011)	Integration of data from heterogeneous data sources	No	No	No	Yes	Yes	No	No	No	No	Yes
MimMiner	Van Driel <i>et al.</i> , <b>A text-mining analysis of the human phenome</b> , <i>Eur. J. Hum. Genet.</i> (2006)	Integration of data from heterogeneous data sources	Yes	No	No	Yes	Yes	No	Yes	Yes	No	No
PGMapper	Xiong <i>et al.</i> , <b>PGMapper: a web based tool linking phenotype to genes</b> , <i>Bioinformatics</i> (2008)	Text mining OMIM and Pubmed	No	No	No	Yes	No	No	No	No	No	No
Phenodigm	Smedley <i>et al.</i> , <b>PhenoDigm: analyzing curated annotations to associate animal models with human diseases</b> , <i>Database</i> (2013)	Prioritisation based on model organism phenotypes	No	No	No	No	No	No	No	Yes	No	No
Phenolyzer	Yang <i>et al.</i> , <b>Phenolyzer: phenotype-based prioritization of candidate genes for human diseases</b> . <i>Nature Methods</i> (2015)	Prioritisation based on disease phenotypes	No	No	No	No	No	No	No	Yes	No	No

Tool	Publication	Description	Biological Functions	Gene Expression	Gene Regulation	Text-mining	Protein interactions	Pathways	Sequence	Phenotype	Conservation	Other
PhenoPred	Radivojac <i>et al.</i> , <b>An integrated approach to inferring gene-disease associations in humans</b> , <i>Proteins</i> (2008)	Integration of data from heterogeneous data sources	Yes	No	No	No	Yes	No	Yes	Yes	No	No
Pinta	Nitsch <i>et al.</i> , <b>PINTA: a web server for network-based gene prioritization from expression data</b> , <i>Nucleic Acids Res.</i> (2011)	Integration of data from heterogeneous data sources; requires gene expression data set	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No
PolyPhen 2.0	Adzhubei <i>et al.</i> , <b>A method and server for predicting damaging missense mutations</b> . <i>Nature Methods</i> (2010)	Sequence information, homology and conservation	No	No	No	No	No	No	Yes	No	Yes	No
PolySearch	Cheng <i>et al.</i> , <b>PolySearch: a web based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites</b> , <i>Nucleic Acids Research</i> (2008)	Integration of data from heterogeneous data sources	Yes	No	No	Yes	Yes	Yes	No	No	No	Yes
PosMed	Makita <i>et al.</i> , <b>PosMed: ranking genes and bioresources based on Semantic Web Association Study</b> , <i>Nucleic Acids Research</i> (2013); Yoshida <i>et al.</i> , <b>PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning</b> , <i>Nucleic Acids Research</i> (2009)	Integration of data from heterogeneous data sources	Yes	No	No	Yes	Yes	Yes	No	Yes	No	Yes

Tool	Publication	Description	Biological Functions	Gene Expression	Gene Regulation	Text-mining	Protein interactions	Pathways	Sequence	Phenotype	Conservation	Other
PRINCE	Vanunu <i>et al.</i> , <b>Associating genes and protein complexes with disease via network propagation</b> , <i>PLoS Computational Biology</i> (2010)	Integration of data from heterogeneous data sources; network analysis	No	No	No	No	Yes	No	No	No	No	Yes
ProDiGe	Mordelet <i>et al.</i> , <b>ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples</b> , <i>BMC Bioinformatics</i> (2011)	Integration of data from heterogeneous data sources; network analysis	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	No
ProphNet	Martínez <i>et al.</i> , <b>Network-based gene-disease prioritization using PROPHNET</b> , <i>EMBnet.journal</i> (2012)	Integration of data from heterogeneous data sources; network analysis	Yes	No	No	Yes	Yes	No	No	No	No	No
S2G	Gefen <i>et al.</i> , <b>Syndrome to gene (S2G): in-silico identification of candidate genes for human diseases</b> , <i>Hum Mutat.</i> (2010)	Integration of data from heterogeneous data sources	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No
SIFT	Kumar <i>et al.</i> , <b>Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm</b> , <i>Nature Protocols</i> (2009)	Sequence information, homology and conservation	No	No	No	No	No	No	Yes	No	Yes	No
SNPs3D	Yue <i>et al.</i> , <b>SNPs3D: Candidate gene and SNP selection for association studies</b> , <i>BMC Bioinformatics</i> (2006)	Integration of data from heterogeneous data sources	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	No
TargetMine	Chen <i>et al.</i> , <b>TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery</b> , <i>PLoS One.</i> (2011)	Integration of data from heterogeneous data sources	Yes	No	Yes	No	No	Yes	No	Yes	No	Yes

Tool	Publication	Description	Biological Functions	Gene Expression	Gene Regulation	Text-mining	Protein interactions	Pathways	Sequence	Phenotype	Conservation	Other
ToppGene	Chen <i>et al.</i> , <b>ToppGene Suite for gene list enrichment analysis and candidate gene prioritization</b> , <i>Nucleic Acids Res.</i> (2009); Chen et al., <b>Improved human disease candidate gene prioritization using mouse Phenotype</b> , <i>BMC Bioinformatics</i> (2007)	Integration of data from heterogeneous data sources	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No

Many gene and variant prioritisation approaches rely extensively on data mining, taking advantage of vast amounts of data available in public repositories, databases or from in-house experiments. Along with this wealth of data come a number of inherent limitations stemming from a universal lack of standardisation across and within databases. Reliance on the open access data makes it extremely difficult to perform any standardised quality controls, and thus any such data are vulnerable to accumulation of errors. Often, datasets are incomplete or do not provide enough information about how the data was generated. Discarding such data would limit the build-up of errors further downstream; conversely, it also results in the loss of valuable information. Furthermore, poor quality data is not always easy to detect and therefore filter – often, it is impossible to deduce the quality of the dataset. A poorly described dataset may not be constituted of poor quality data, and *vice versa*.

Large datasets can be assembled by automated or semi-automated text-mining pipelines. However, even recognition of named entities such as genes in free-form text still presents challenges that arise from inaccuracies and ambiguities in text. Consequently, automated data mining often sacrifices accuracy to improve coverage (Cohen and Hersh 2005). For example, while Gene Ontology (GO) – one of the biggest gene functional annotation resources – provides a set of manually curated annotations, most annotations are still electronically inferred and thus prone to errors (Ashburner et al. 2000).

Despite these difficulties, computational gene prioritisation tools boast several successes. Established web tools such as ToppGene (Chen et al. 2009), Endeavour (Aerts et al. 2006) and CANDID (Hutz et al. 2008) employ a modular approach to prioritization, scoring candidates based on a consensus from multiple data sources. Even though it has been demonstrated that consensus methods are more accurate than approaches utilizing fewer data categories (Bornigen et al. 2012), the former have been criticized for both the ‘guilt-by-association’ bias (i.e. genes already known to be important in a disease are scored highest) and failure

to exploit the best performing methods for each component (Chen et al. 2011c). Knowledge-based approaches are often supplemented by the use of high-throughput data. Some algorithms integrate gene expression data into a heterogeneous (Chen et al. 2011c; Nitsch et al. 2011) or homogenous (van Dam et al. 2012) network, where distance between genes can be derived from and/or weighted by differential expression or co-expression values. Alternatively, some methods (Chen et al. 2009; Masotti et al. 2008; Seelow et al. 2008) consider gene co-expression in a non-network context, utilizing common statistical vector correlation measures to rank candidate disease genes based on how well their expression patterns correlate with those of genes known to be directly or indirectly linked to the disease.

Fewer applications attempt to apply tissue-specific expression patterns for gene prioritization tasks. Endeavour (Aerts et al. 2006) incorporates gene expression data from 79 normal human tissues found in the Gene Expression Atlas dataset (Kapushesky et al. 2010), comparing gene expression between candidate and user-supplied seed genes across tissues. A recent update to PhenoDigm (Smedley et al. 2013; Robinson et al. 2014) has incorporated tissue-specific, binary (expressed/not expressed) mouse gene expression data from a small number of mouse tissues and derived phenotype-tissue associations in order to supplement its phenotype-based queries.

Similar to PhenoDigm, Phenolyzer (Yang et al. 2015) also focuses on phenotype information obtained from disease databases and ontologies to score candidate genes. Each gene is scored on the confidence of the gene-disease relationship; thus, the system is more suited for diagnostic use than novel disease gene predictions. Similarly, text-mining applications which focus largely on disease databases, such as MinMiner (van Driel et al. 2006), are also largely unsuitable for novel disease gene discovery.

Thus, it is clear that development of disease gene prioritisation methods that focus on less biased data types and are more universally applicable data would be of great benefit to the scientific community.

### ***3. GeneTiER: gene Tissue Expression Ranker***

#### ***3.1 Motivation***

Gene expression data presents a particularly attractive data mining opportunity for candidate gene prioritisation, mostly due to the ever-increasing publicly available data in repositories such as ArrayExpress (Kolesnikov et al. 2015) or Gene Expression Omnibus (GEO) (Barrett et al. 2013). Recent methods allow relatively inexpensive characterisation of thousands of mRNAs at a time and therefore the number of experiments performed and made publicly available is on the rise. This trend has in part been driven by a growing number of biological journals requiring all published data to be made available via public data repositories. Furthermore, due to the large volume of data being deposited and in the interest of enhancing experiment reproducibility, stricter data submission standards have been established to comply with the MIAME (Minimum Information About a Microarray Experiment) conventions (Brazma et al. 2001). As a result, more of the recently released datasets now contain useful meta data, such as experimental design and methodology.

Utilising gene expression data in order to prioritise genes largely overcomes the biases that affect information-based methods. Approaches which rely on annotations and literature, or any data sources which are derivatives thereof, will almost always score the better understood, more studied genes ahead of novel candidates for which little information is known. On the other hand, gene expression is bound only by the experimental platform limitations. RNA sequencing, for example, is able to quantify expression at a level of high sensitivity and completely independently of any prior knowledge.

While the use of high throughput sequencing methodologies to measure gene expression is on the rise, hybridisation-based technologies, such as microarrays, are still highly prevalent. Microarray-derived expression data is limited to

preselected sequences and does not cover the entire transcriptome; however improvements in array and probe design have allowed for quantification of a vast number of transcripts. A typical microarray design can incorporate tens of thousands of probes which correspond to most known transcripts in the given genome. While microarray data does not allow for detection of expression to the level of specificity of NGS due to the inherent limitation of measurement of light intensity produced by the hybridisation, it still gives an insightful measure of gene expression.

Gene expression information is often utilised to supplement other data types (**Table 1**), with, to the best knowledge of the author, no standalone gene prioritisation tool in existence at the start of this thesis. One of the most common approaches of this type relies on coupling gene expression data with some type of network-based analysis. Typical network-based methods prioritize genes under the assumption that a disease gene will exist in a local network of genes which are highly differentially expressed between affected and unaffected tissues; or, co-expressed with genes in a known disease pathway.

An expression network can be constructed from heterogeneous data types, such as functional or phenotypic associations, often sourced from ontological annotations; regulatory pathways; or, known protein-protein interaction networks, such as STRING (Jensen et al. 2009), Reactome (Fabregat et al. 2016) or KEGG (Kanehisa and Goto 2000). Random walker algorithms, distance-based measures or even search engine technologies such as PageRank algorithm are then used to identify the most likely candidate genes (Smedley et al. 2014; Morrison et al. 2005). Algorithms of this type judge the relatedness of all candidate genes to a set of query nodes of the network – which can be known disease genes, phenotypes or other entities. Edges in the network, representing physical interactions or indirect associations between genes, can also be weighted by gene expression data.

Similarly, some methods use network analysis approaches to consider co-expression rather than differential expression patterns. Co-expression with genes already known to be associated with the query disease (or genes functionally related to the query) can be used to implicate a candidate. Typical analysis, however, requires the user to supply 'seed' genes, thus making the assumption that pre-existing knowledge about the disease is either available or relevant to the particular disease phenotype. This may not be the case; for example, OMIM currently contains several thousand disease entries for which no contributing genes are yet known.

Alternatively, protein-protein interaction data allow for clustering of molecular pathways and differential gene expression or co-expression data in a network of known interactions, thus allowing identification of upstream or downstream candidates which might not be otherwise linked to the query. This approach has proven successful, and a number of algorithms and web tools have been implemented that allow this type of analysis. However, the majority of applications of this type nevertheless suffer from bias towards better characterised genes, as any interaction or functional network, no matter how extensive, does not currently contain complete knowledge for all genes and pathways. Alternatively, less biased approaches, such as implemented by weighted gene correlation network analysis (WGCNA) algorithm (Langfelder and Horvath 2008), can derive the gene network structure entirely from gene expression data using hierarchical gene clustering, which can then be used to identify closely linked modules of co-expressed genes.

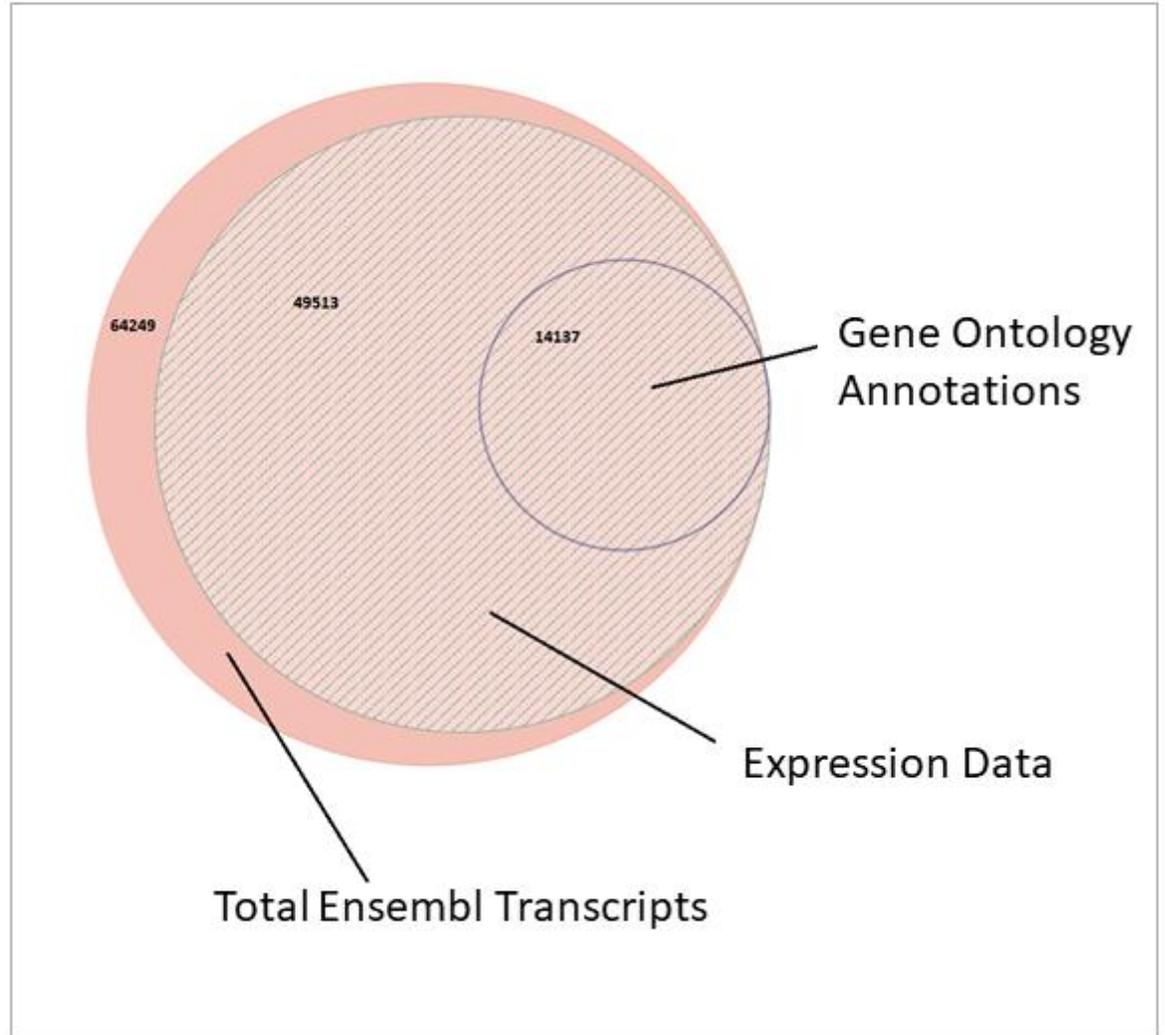
However, there are often difficulties in wider applicability of such approaches. The algorithm implementations can be disease-specific, drawing the underlying data from a small number of specific experiments. In cases of generalised use, it is a typical requirement for the user to supply their own expression datasets for the analysis. MetaRanker 2.0 supports the integration of tissue-specific baseline as well as differential gene expression datasets, however these must be provided by the user (Pers et al. 2013). This is in contradiction to a major aim of computational

candidate prioritisation methods - to reduce the number of experiments, not require the user to perform further studies.

The methods that do not fall into the categories described – i.e. do not require the user to supply an expression dataset and are generalised methods – often do not differentiate between datasets. For example, GeneFriends (van Dam et al. 2012) performs large scale co-expression analysis utilising thousands of microarray experiments across different conditions, however the data for each gene are pooled regardless of the user query. While the application has been benchmarked and proven to have significant classification power, it is of some concern that predictions may suffer from increased false positive rate due to the utilisation of data from vastly heterogeneous conditions, such as normal and cancer tissues, as genes expression patterns in cancer tissues may not accurately represent normal cell biology.

In contrast to network-based methodologies, ‘data fusion’ approaches often use expression data to effectively supplement knowledge-based data sources; expression data can serve as a weighting to confirm or contradict a predetermined link between gene and disease. Similarly to the well-established gene prioritisation software Endeavour (Aerts et al. 2006), POCUS (Turner et al. 2003) - a now retired tool -integrated a basic, non-quantitative level of expression information obtained from UniGene database (Wagner and Agarwala 2013). Therefore, these types of approaches still share all the failings of ‘guilt-by-association’ methodologies.

As the human genome becomes increasingly saturated with annotations, the problem of data completeness will dwindle in importance. At the time of writing, however, this is still very much an issue – Gene Ontology annotations, for example, while representing the most complete set of functional gene annotations currently available, still comprise less than half of the total human genes in the current reference human genome assembly by Ensembl (Cunningham et al. 2014) (**Figure 5**). On the other hand, quantitative high throughput sequencing and microarray data is available for the majority human genome transcripts.



**Figure 5.** Venn diagram (to scale) illustrating the incompleteness of Gene Ontology annotations in comparison to gene expression data availability. hg19 human reference genome release comprises 64,249 Ensembl human gene annotations. Out of these, 49,513 contain some quantitative expression values in the set of 141 human and mouse ortholog normal tissue expression datasets used in this work. In comparison, Gene Ontology annotations downloaded using BioMart (Smedley et al. 2015) tool (download date: 10/03/2014) are only available for 14,137 Ensembl genes. This illustrates the limitations of the information-based approaches, such as those which consider functional information derived from literature or ontologies.

Thus, while a number of methods have previously been employed for the prioritization of candidate disease genes, none are universally applicable. Typical gene prioritisation software tends to rely heavily on prior knowledge about the disease, phenotype and/or genes, making them unsuitable for classifying novel and/or poorly characterized genes. The best performing methods have been shown to rely on a variety of information sources to compensate for inadequacies in knowledge in any single domain (Bornigen et al. 2012); however, gene prioritisation tools are most commonly benchmarked using testing data sets generated from OMIM disease genes. Due to circularity of knowledge in literature and public databases, validation results in these cases might not reflect true performance with respect to novel disease genes. Our knowledge of even well-understood disease mechanisms is rarely entirely complete, and many novel disease cases could benefit from an unbiased prioritisation approach. Therefore, particularly where little prior knowledge about the disease and/or gene is available, prioritisation of putative disease genes remains a challenge.

Consequently, there is value to be found in approaches which distance themselves entirely from the 'guilt-by-association' principle and instead use algorithms that depend solely on large genome-wide datasets generated in a hypothesis free manner, such as high throughput gene expression data.

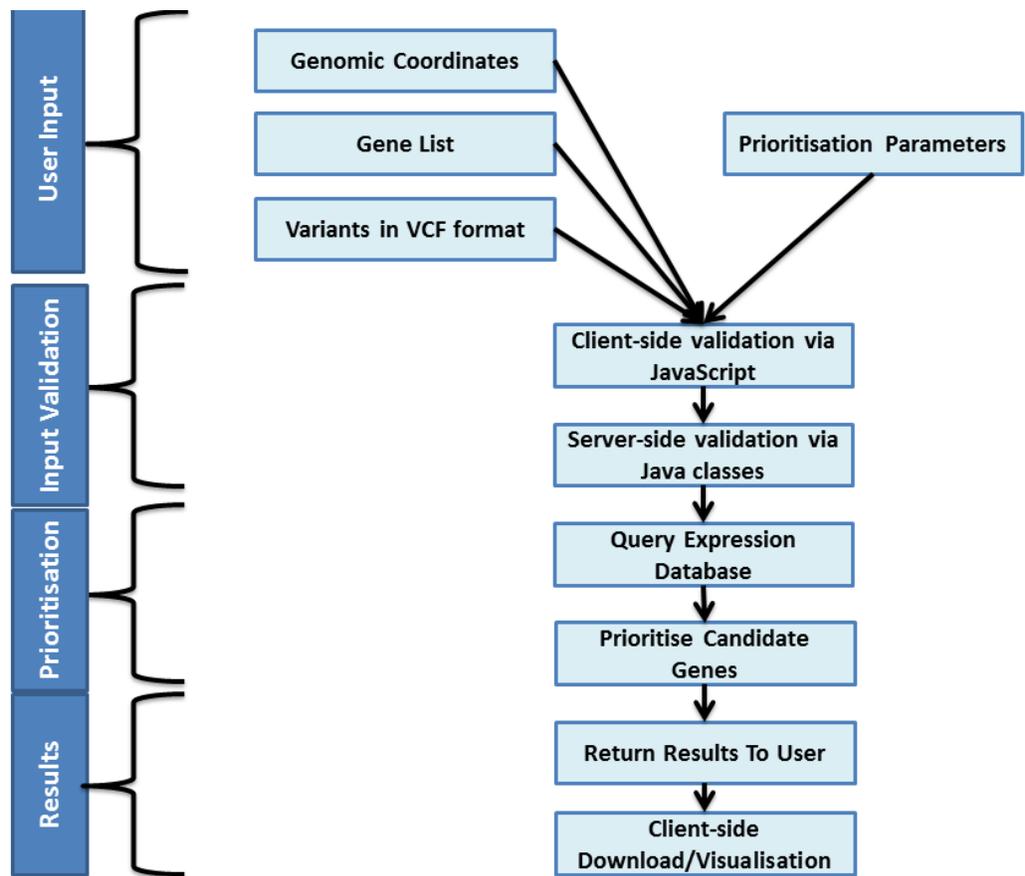
In this work, a novel application for candidate disease gene prioritization is presented that aims to address these shortcomings by taking advantage of both microarray and RNA sequencing data available in the public domain to create an extensive tissue-specific expression database that can support a wide variety of gene prioritization queries. The use of publicly available gene expression data is investigated as the sole means of prioritizing candidate disease genes. The resulting web application, GeneTiER (Gene Tissue Expression Ranker), scores candidate disease genes based on the hypothesis that genes responsible for a tissue(s)-specific phenotype are expected to be more highly expressed in affected

than unaffected tissues. GeneTiER depends on an extensive database that has been built using publicly available microarray and RNA sequencing datasets and is comprised of several million expression values for numerous healthy tissues. This enables the creation of a global, cross-tissue expression profile for each candidate disease gene, permitting expression profile-based prioritization without reliance on or requirement for other prior knowledge about the disease or candidate genes. GeneTiER should thus be suitable for prioritization of candidates for poorly characterized diseases.

## ***3.2 Methods***

### ***3.2.1 Software Implementation***

The GeneTiER methodology was implemented as a web-based application, which is currently hosted on an instance of the Apache Tomcat 8.0 web server running on a CENTOS server and is freely accessible at [dna2.leeds.ac.uk](http://dna2.leeds.ac.uk). The implementation follows the classical design of logical separation of different functions into four software tiers and is summarised in **Figure 6**. The user interface has been implemented using a mixture of HTML, CSS and JavaScript, and accepts and validates user input before passing it to server-side tiers for data processing. Java Server Pages links the client-side interaction to the server-side data processing by further validating user input and generating dynamic web content in response to user queries. The 3<sup>rd</sup> tier is implemented in Java and handles all the 'business' logic: the execution of algorithms and analysis tasks after receiving validated user input and the required data from the MySQL database (4<sup>th</sup> tier). The results are then passed back to the user using Java Server Pages, which generate dynamic content for further client-side interaction.



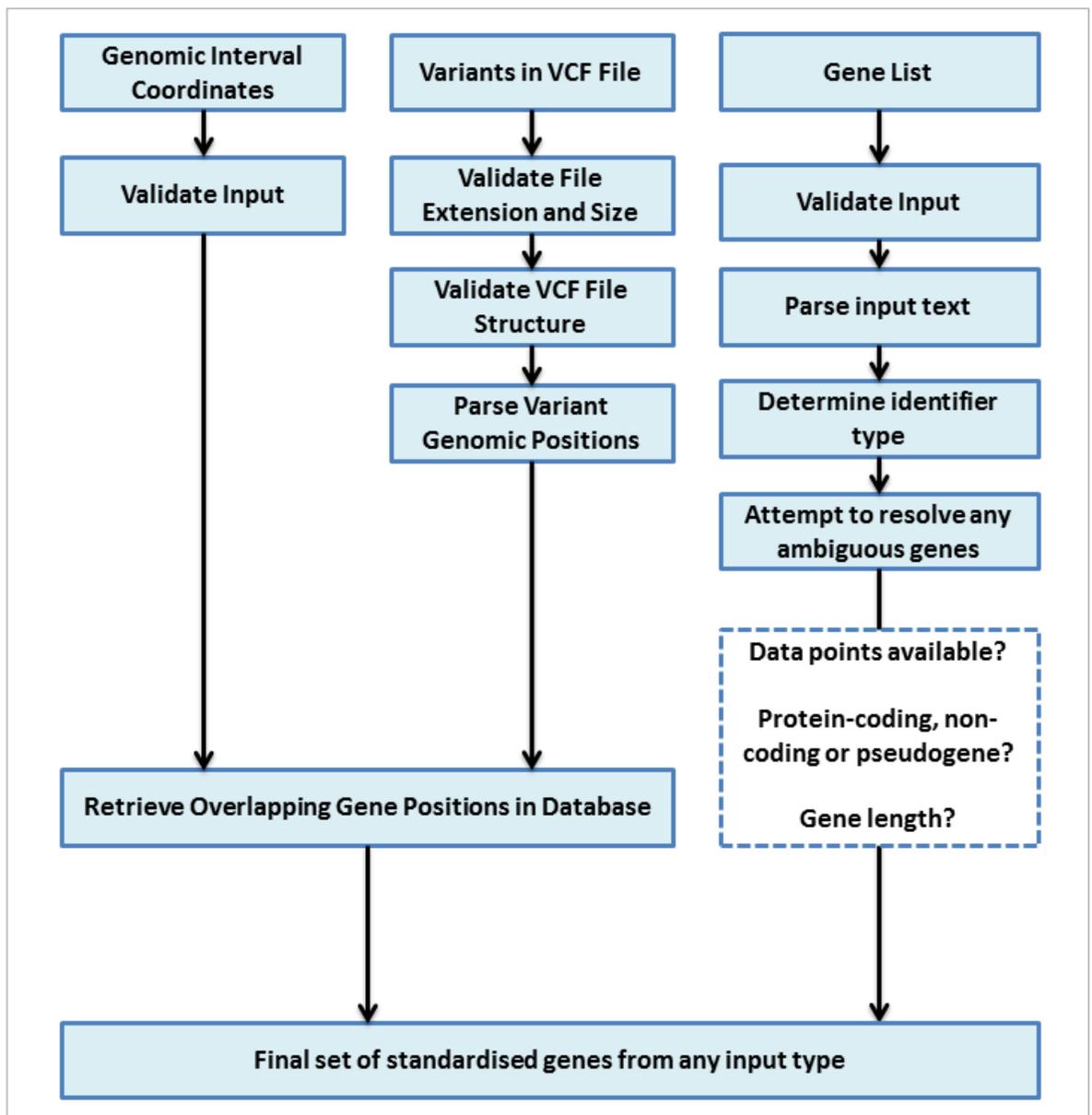
**Figure 6.** Overview of GeneTiER implementation. User input is validated both client- and server-side. The required data based on the query is retrieved for prioritisation from the MySQL database and user input genes are prioritised according to the input parameters. The results are returned as dynamic web pages and can be visualised and downloaded by the user.

### ***3.2.2 User Input Processing***

All user input is subjected to two validation steps. Initially, client-side validation code implemented in JavaScript checks that all the appropriate information is supplied and is in the correct format, while the more stringent server-side validation implemented in Java classes checks all of the uploaded data for the correct format, structure and composition. For instance, client side JavaScript is used to ensure that the supplied gene list is not empty, while the server-side validation would check that an uploaded VCF file is correctly formatted.

The selection of query tissue types is dynamically generated and limited by the web interface to those present in the GeneTiER database, and thus requires little validation. However, the list of candidate genes can be supplied in multiple formats (list of genomic regions, list of gene or transcript identifiers or variants contained in a .VCF file), and therefore more extensive parsing logic is required. Each input type requires a different validation method; however, all inputs are eventually mapped to the same standardised set of internal database genes (**Figure 7**).

Genomic regions may be specified by the user via hg19 human reference genome coordinates. All the genes contained within these regions are retrieved for prioritisation via database look-up. The coordinates are reversed in the case of regions smaller than 0 base pairs. A region must contain at least 2 genes in order to proceed. VCF files are parsed for genomic positions of variants and the corresponding genes retrieved. Gene list input is parsed for common gene names, aliases or commonly used gene database identifiers. These can be any of Ensembl (Flicek, Amode et al. 2014), Entrez (Maglott, Ostell et al. 2011) or Refseq (Pruitt, Brown et al. 2014) accessions and HGNC-approved gene names (Gray, Daugherty et al. 2013), as well as common aliases. Conversions between human genes and their mouse orthologs are automatic and based on Homologene (Sayers et al. 2012). Any ambiguous gene input can either be discarded, resolved manually by the user or resolved automatically.



**Figure 7.** Following client and server-side validation, different input types are all mapped to a standardised set of human genes.

### ***3.2.3 GeneTiER Database***

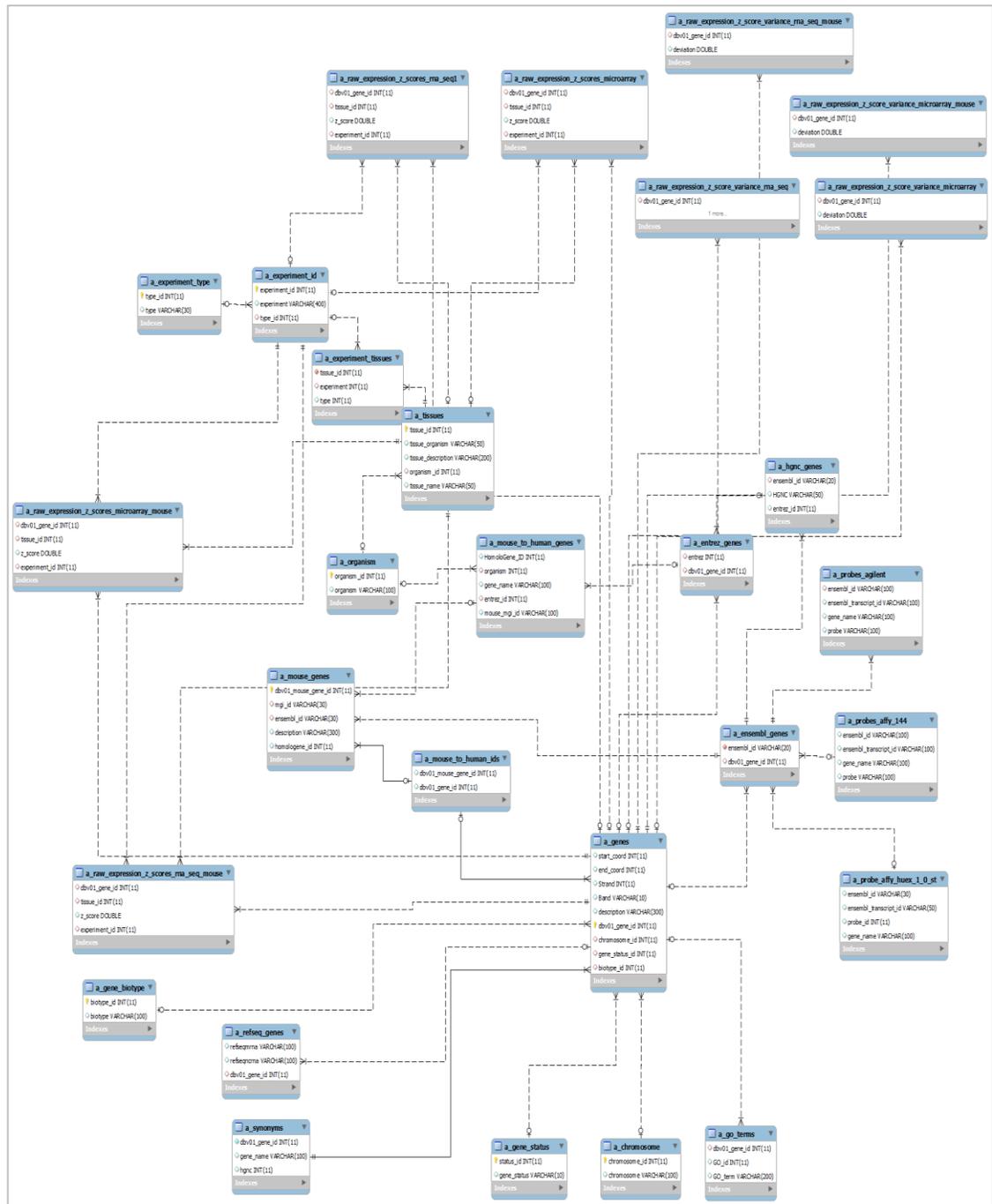
The GeneTiER expression database contains 9,972,862 baseline gene expression values from microarray and RNA-Seq experiments, encompassing 140 different mouse and human tissue types. The database was assembled from public domain sources and includes datasets from Gene Expression Atlas (Kapushesky et al. 2010), RNA-Seq Expression Atlas (Krupp et al. 2012), ArrayExpress (Kolesnikov et al. 2015) and Gene Expression Omnibus (Barrett et al. 2013).

For RNA-Seq datasets, where available, raw read count tables per transcript were downloaded from the respective databases and high abundance RNA species (rRNA and tRNAs) counts were filtered out. Otherwise, raw data in Fastq format was downloaded, quality trimmed using Cutadapt (Martin 2011) software and aligned to human hg19 reference genome using STAR aligner (Dobin et al. 2013). HTSeq-Count software (Anders et al. 2015) was used to obtain raw read counts.

Microarray datasets were downloaded either as within-array normalized intensity tables, or raw intensity CEL, txt or IDAT files for Affymetrix, Agilent or Illumina array designs, respectively. Any raw microarray data was first normalized for within-array comparability using Bioconductor package 'limma' (Smyth 2005).

Microarray probes were mapped to Ensembl gene transcript identifiers using the Biomart resource (Smedley et al. 2015). As per recommended practice, ambiguous data arising from microarray probes which hybridize to more than one distinct gene were discarded (Ramasamy et al. 2008). Similarly, HGNC, Ensembl, and Entrez and RefSeq gene identifiers were obtained from Biomart. UCSC gene names and exon boundary coordinates were downloaded using the UCSC Genome Browser's 'Table Browser' tool (Karolchik et al. 2004) using the hg19 human genome assembly. Mouse-human gene orthologs were downloaded from MGI (Blake et al. 2014) and mapped using HomoloGene (Sayers et al. 2012).

The database was implemented using MySQL Community Server version 5.6.15. **Figure 8** shows the enhanced entity–relationship (EER) diagram of database design (raw data and intermediate tables are not shown for clarity purposes). The database contains three main data components –tissue, gene and gene expression data. These components are linked between all members of the schema, enforced by the use of foreign keys. The database design is such that the majority of tables conform to recommended database schema normalisation practices (3<sup>rd</sup> normal form, (Codd and F. 1982)) in order to facilitate referential integrity and reduce data redundancy. Some tables were exempt from this requirement in order to optimise query speed by removing the need for complex table joins.



**Figure 8.** GeneTiER MySQL database schema. [PLACEHOLDER FOR FOLDOUT FIGURE IN PRINT]

### 3.2.4 The gene prioritization algorithm

Candidate genes are ranked based on several factors derived from gene expression data. These comprise the expression levels in the affected tissues; variance in expression across all tissues; and expression level differences between affected and unaffected tissues. The base score,  $S_g$ , for each candidate disease gene is calculated as follows:

$$S_g = \sum_{t \in T} \begin{cases} \bar{z}_t & \text{if } \bar{z}_t = 0 \\ \bar{z}_t \cdot \left(1 + \ln \frac{\bar{z}_t}{\tilde{z}}\right) & \end{cases}$$

Where  $t$  is an affected tissue in a set of all affected tissues  $T$ ;  $\bar{z}_t$  is the mean of modified z-scores (see below) for tissue  $t$ , and  $\tilde{z}$  is the median modified z-score across all tissues. If gene expression in an affected tissue is greater than its baseline expression the natural logarithm ratio is positive; otherwise the value is negative. The value of  $S_g$  is a fractional modifier, favoring genes which show elevated gene expression in disease-associated tissues, compared to tissues not linked to the disease phenotype, even if the expression value is relatively low. The score can be further adjusted for highly expressed genes which takes into account the level of variance in expression across all tissues in order to reduce the contention of highly ubiquitously expressed housekeeping genes. When included in the analysis, the results from human RNA sequencing, human microarray, mouse RNA sequencing and mouse microarray data are each considered separately, and combined to generate the final ranking score. When the final ranking score is derived from human and mouse data, the relative contribution of mouse tissue datasets relative to the human datasets can be adjusted by the user.

Modified z-scores for all RNA sequencing and microarray datasets were calculated as shown below:

$$Z_{e \in E} = \frac{0.6745 \cdot (e - \bar{E})}{\text{median} (|e - \tilde{E}|)}$$

Where E denotes a set of normalized expression values in an experiment, with individual elements e;  $\bar{E}$  is thus the mean value of E and the denominator is the median absolute deviation, where e is an individual element of E and  $\tilde{E}$  is the median of all elements in E.

The modified z-scores enable the transformation of non-normally distributed gene expression data and measure how each data point differs from the typical observations within the dataset. This transformation serves both to aid the prioritization and to facilitate better comparability between microarray datasets, as it has been suggested that rank-based transformations of microarray data alleviate some of the issues associated with comparing cross-platform, cross-laboratory data (Irizarry et al. 2005).

### ***3.3 Results and Discussion***

#### ***3.3.1 Performance Assessment***

##### ***3.3.1.1 Benchmarking Datasets***

In order to assess how applicable the strategy detailed above can be for disease gene prioritisation in practice, the performance of the GeneTiER application was evaluated using two test sets of gene-disease associations. These were generated using the Human Phenotype Ontology (HPO) annotations (Kohler, Doelken et al. 2014) as a source for disease genes and associated phenotypes. The HPO is a

curated ontology, organizing human disease phenotypes described in OMIM (Amberger et al. 2015), as well as Orphanet (Pavan et al. 2017) databases and medical literature using a structured, controlled vocabulary. This enabled the generation of large testing datasets while bypassing any inaccuracies that can arise from the lack of precision when text-mining unstructured entries in OMIM.

Initially, a sub-group of all HPO phenotypes was selected based on the following criteria: terms with high specificity (defined as the distance of a term from the root of the ontology) and terms which could be unambiguously mapped to tissues through axiomatic links to an anatomical ontology (Golbreich et al. 2006) (Kohler, Doelken et al. 2013) (Hoehndorf, Oellrich et al. 2010) or manual assignment. 2922 distinct known disease genes in total were found to be annotated as associated with these HPO disease phenotypes. To further improve the testing dataset, phenotypes with the associated frequency modifier for the annotated disease denoted as 'very rare' and/or 'occurring in fewer than 2% of all cases reported', were not considered. Ultimately, from the resulting data, 1000 disease-genes associations were selected at random for testing (**Dataset 1**).

Additionally, in order to ascertain how tissue selection affects prioritization with GeneTiER, diseases with a distinct, localized phenotype were categorized based on Disease Ontology (Kibbe et al. 2015) annotations, using definitions which are descendants of the term '*disease of anatomical entity*' (DOID:7). This has led to the selection of another dataset consisting of 500 disease-gene associations with strong links to a single, specific tissue type (**Dataset 2**).

All test genes were prioritized using GeneTiER together with a set of control genes selected by random sampling from the GeneTiER gene database. Gene rankings were collated and the results were processed in R using ROCR package (Sing et al. 2005).

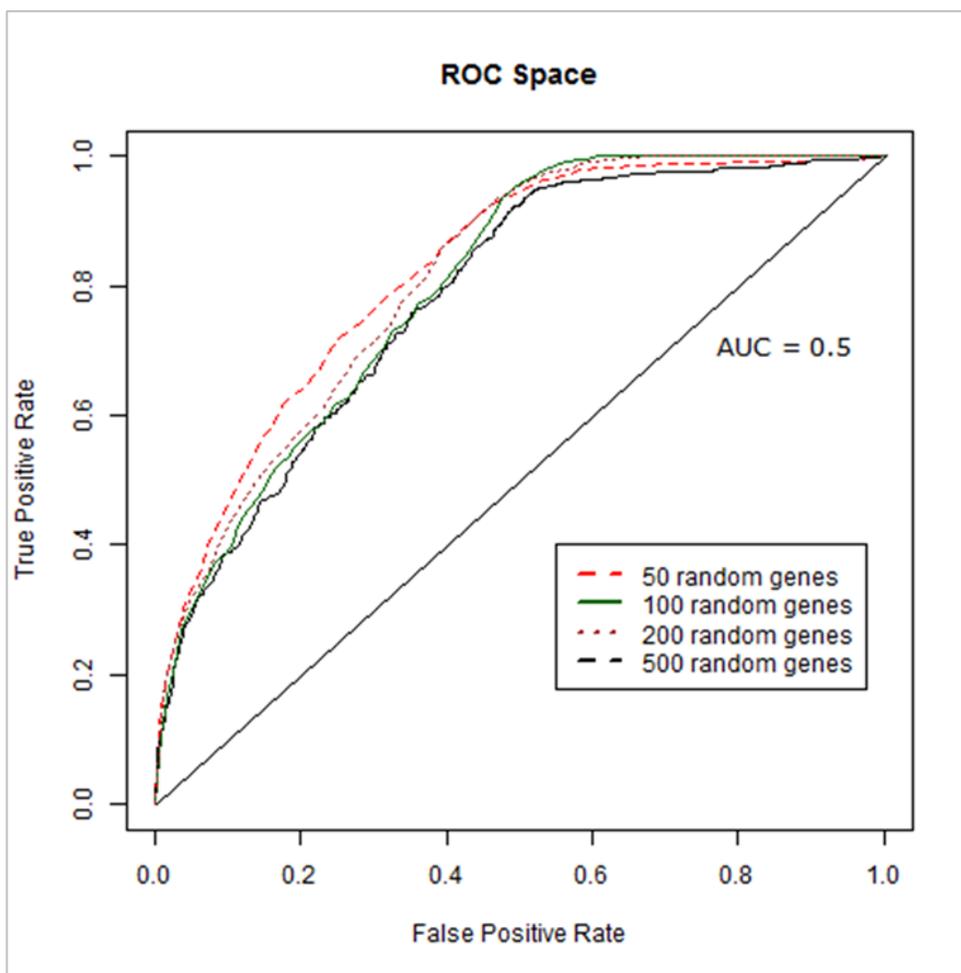
### ***3.3.1.2 Performance Assessment Using Receiver Operating Curve (ROC) Analysis***

The algorithm implemented in GeneTiER works on the assumption that tissue-specific phenotypes often manifest due to disruption of tissue-specific genes, and as such, a disease gene's expression would be higher and/or more localized to affected tissues compared to unaffected tissue. To test the generality of this assumption, expression values were retrieved from the geneTiER database for all genes in the benchmarking data sets and a two-sample Kolmogorov-Smirnov test for non-normally distributed data was performed, using an alternative hypothesis that the cumulative frequency distribution function of modified z-scores from unaffected tissues lies below that of modified z-scores from disease-associated tissues. For RNA-sequencing data this resulted in statistic  $D = 0.1517$ , with respective p-value  $< 2.2e^{-16}$  and for microarray data  $D = 0.1334$ , p-value  $< 2.2e^{-16}$ .

In order to assess how well this observation translates into practicable gene prioritization, Receiver Operating Curve (ROC) analysis was carried out. ROC curves provide a way to visualize and compare classifier performance. Here, the candidate gene prioritization algorithm can be viewed as a non-binary scoring classifier, where disease-linked genes are positive instances and other candidates are negatives. The values -or ranks- from the classifier output can be converted into binary positive and negative scores using cut-off thresholds. Thus, a confusion matrix can be calculated for every integer rank cut-off value from which comparison metrics, such as sensitivity and specificity values are to be derived. ROC graphs allow the visualization of sensitivity and specificity, as well as how the trade-offs between the two are linked for different cut-off values. The line running from the origin (0,0) to the maximum point of 1,1 ( $Y=X$ ), which corresponds to an area under the curve (AUC) of 0.5, thus represents gene prioritisation performance that is no better than random predictions. Points on a ROC curve that occur above this line represent an algorithm with better than random classifier performance, while those below the line have worse than random results, i.e. a bias towards classifying positives as negatives. An algorithm with an AUC of 1 represents perfect classifier performance.

ROC analysis was initially carried out using the benchmarking dataset (see Methods section) comprised of 1000 known associations between disease genes and tissues expected to be affected by each gene's dysfunction. For each disease gene, four sets of random genes were generated by sampling from the GeneTiER database, each comprising 50, 100, 200 and 500 genes in order to simulate gene prioritisation tasks of varying difficulty. The disease genes were prioritized against the genes in the randomly generated gene sets using GeneTiER and the results analyzed using ROC analysis.

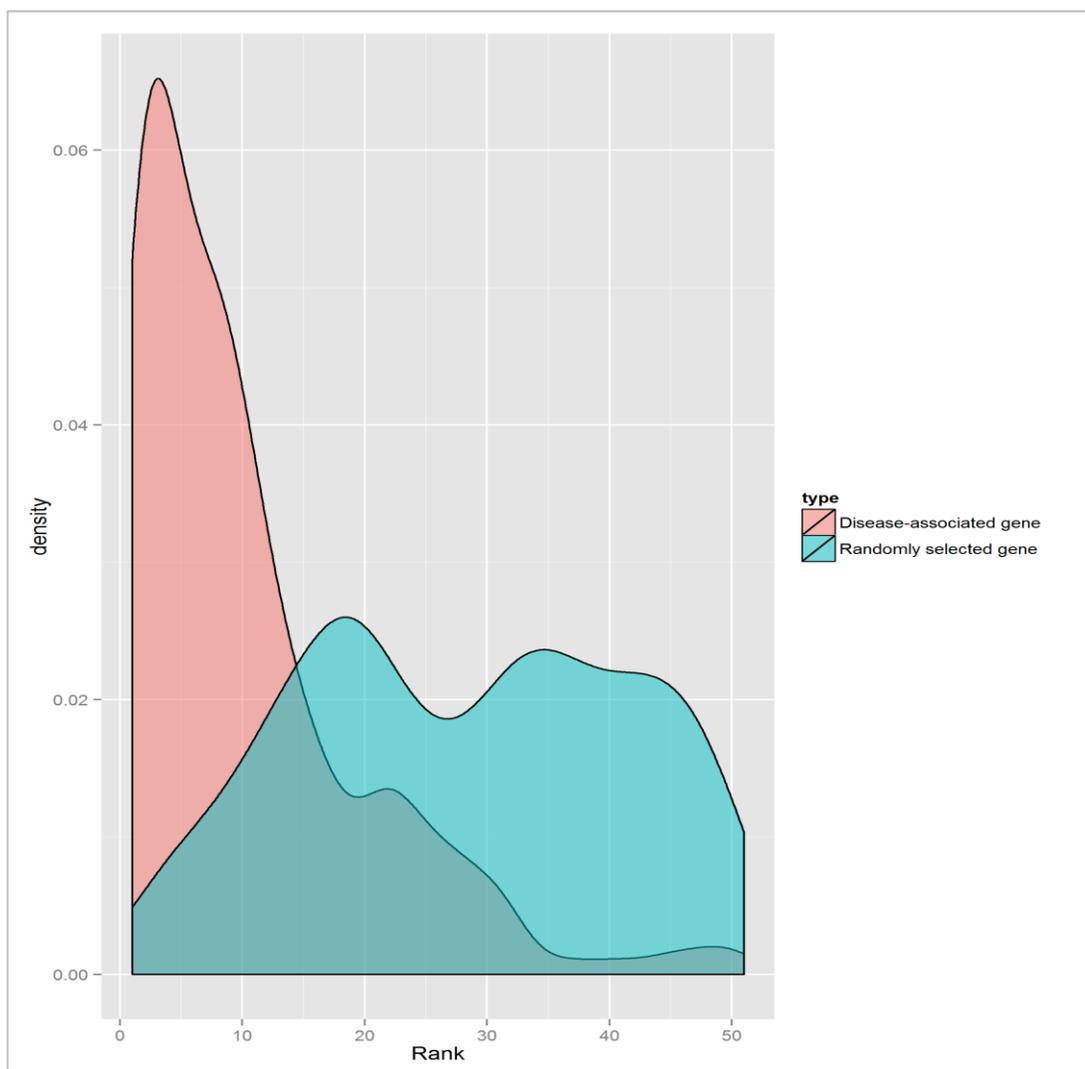
**Figure 9** shows the resultant ROC graph, while **Table 2** shows the corresponding AUC scores. This analysis suggests that the algorithm's performance is inversely related to the number of non-disease genes in the analysis, but does not decline in a linear manner. In fact, the differences in performance when assessed on candidate lists consisting of 100, 200 or 500 candidates are minor and do not suggest that in practice there is a maximum candidate gene list size that will be exceeded in typical gene mapping experiments. Overall, the obtained AUC values are sufficiently high to suggest that disease genes are typically ranked considerably higher than the randomly selected genes in each data set by this algorithm. **Figure 10** visualizes this rank distribution.



**Figure 9.** ROC curve showing classifier performance on different size input generated using disease genes from the benchmarking **Dataset 1** (see **Methods**).

Random Gene Sample Size	Area Under The ROC Curve
<b>50</b>	<b>0.83</b>
<b>100</b>	<b>0.80</b>
<b>200</b>	<b>0.81</b>
<b>500</b>	<b>0.78</b>

**Table 2.** AUC scores for classifier performance when assessed using 1000 known disease genes.



**Figure 10.** Gene rank distributions generated using the benchmarking **Dataset 1** and 50 randomly selected genes.

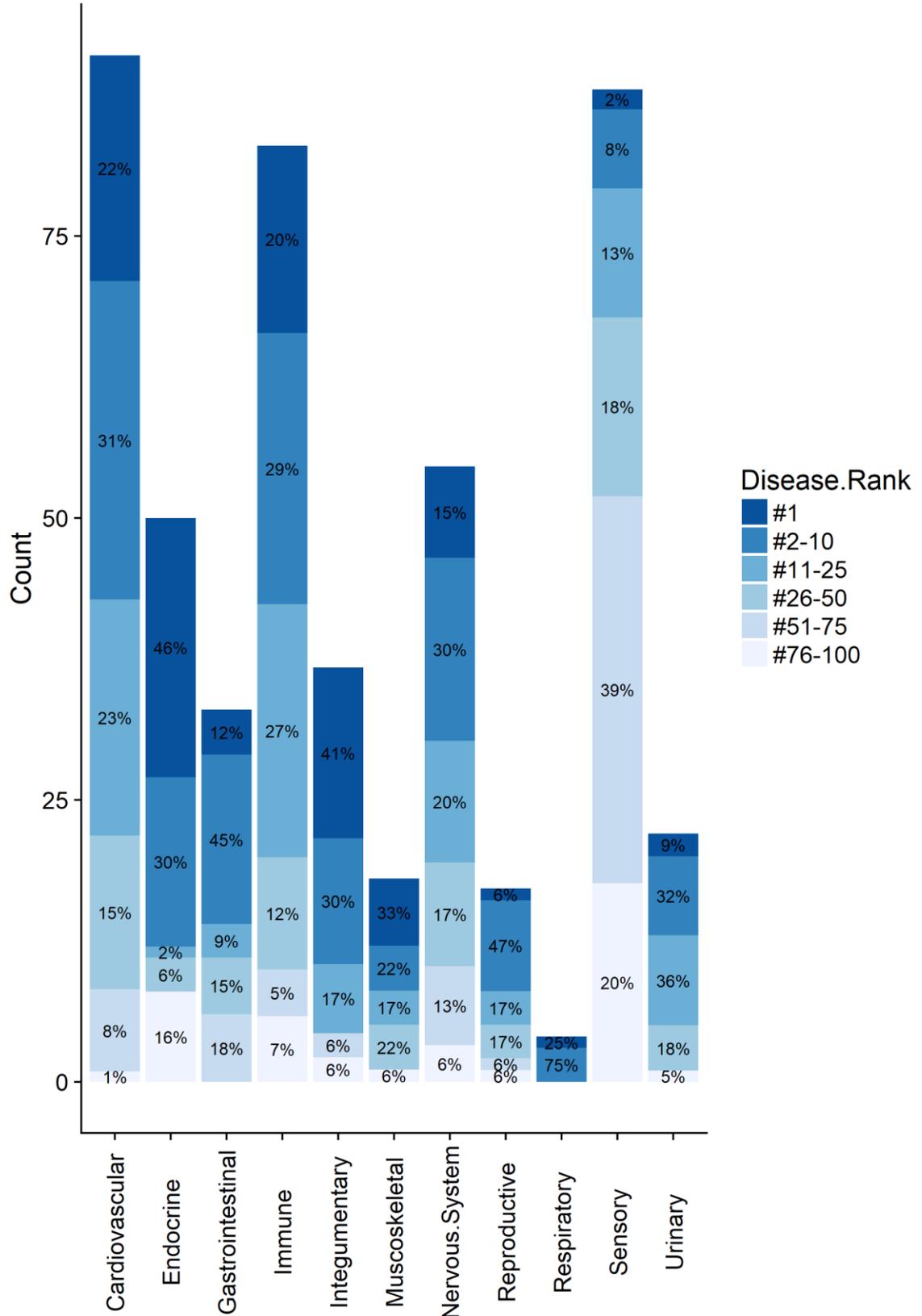
In order to ascertain whether GeneTiER is more appropriate for certain disease types, geneTiER performance was also analysed across a range of tissues. **Figure 11** highlights that GeneTiER can accurately prioritise genes across many tissue categories, recognising endocrine and integumentary system-specific genes particularly well, with 73% and 76% of disease genes respectively ranked in the top 10. However, genes in the sensory category, comprising mostly of eye-related disorders, ranked poorly.

Which data type - RNA-seq or microarray – enabled more accurate prioritisation results was also tested. Initially, all diseases where both RNA-seq and microarray data were available for all the identified affected tissue types were selected from the benchmarking dataset. **Figure 12** shows the ROC curves obtained using only RNA-Seq data or only microarray data when the dataset was prioritised together with 100 random genes. The difference in performance between RNA-seq and microarray data is minimal, with RNA-sequencing data giving better results (ROC 0.80 vs 0.78), but slightly worse than the combined score approach (ROC 0.81). This is in concordance with a recent study by Wang et al, (Wang *et al*, 2014), who found that while more differentially expressed genes identified by RNA-sequencing than microarray studies could be verified by qPCR, the gain was mostly from the improved quantification of low abundance transcripts.

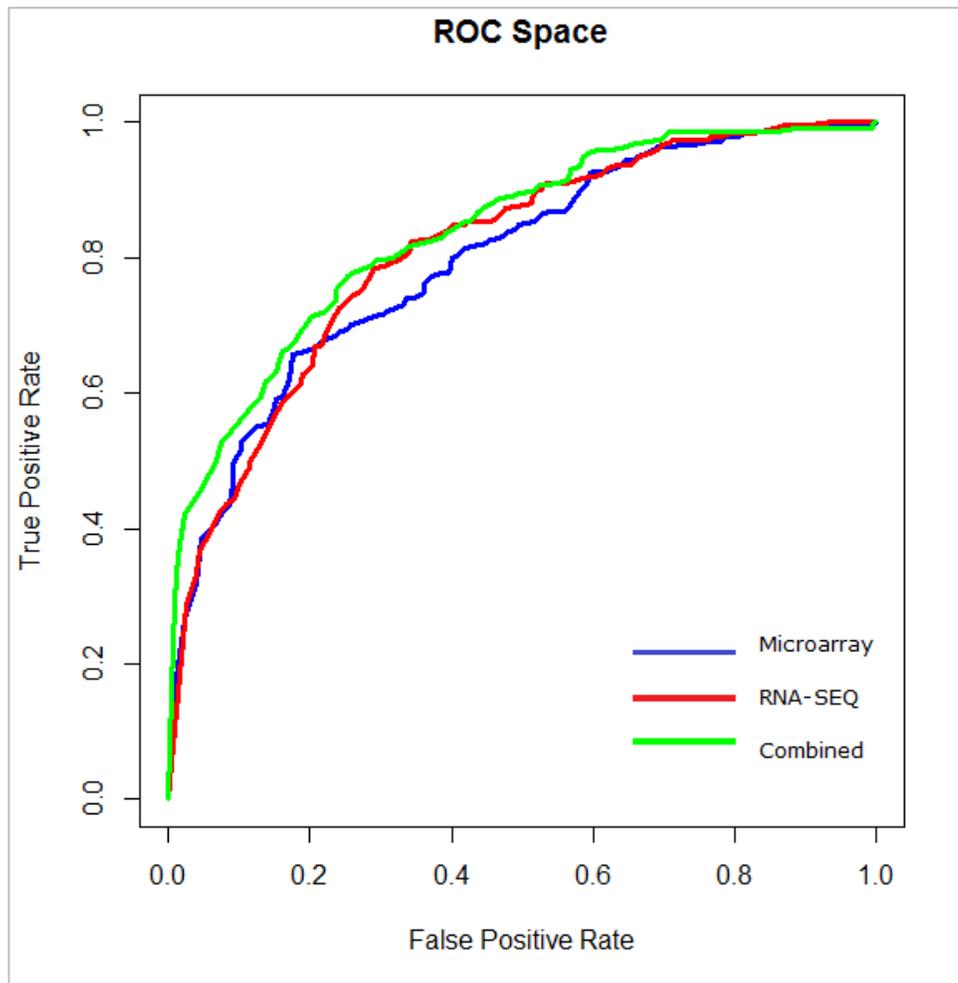
Furthermore, while sequencing data does provide a small improvement over microarray data in prioritisation, this is offset by a more limited public availability of sequencing datasets. At the time of developing GeneTiER, 41,124 microarray datasets were deposited in ArrayExpress database - in contrast to only 5,745 RNA-sequencing experiments (accessed 01/03/2014).

It was further investigated whether GeneTiER is able to prioritize genes equally well across different inheritance modes (**Figure 13**) and different disease onset times (**Figure 14**). While no markedly large differences in performance emerged, it is notable that young adult onset diseases were prioritized with lower accuracy than other types; however, the test dataset contained only six young adult onset diseases, thus this could be due to a random sampling effect arising due to small sample size.

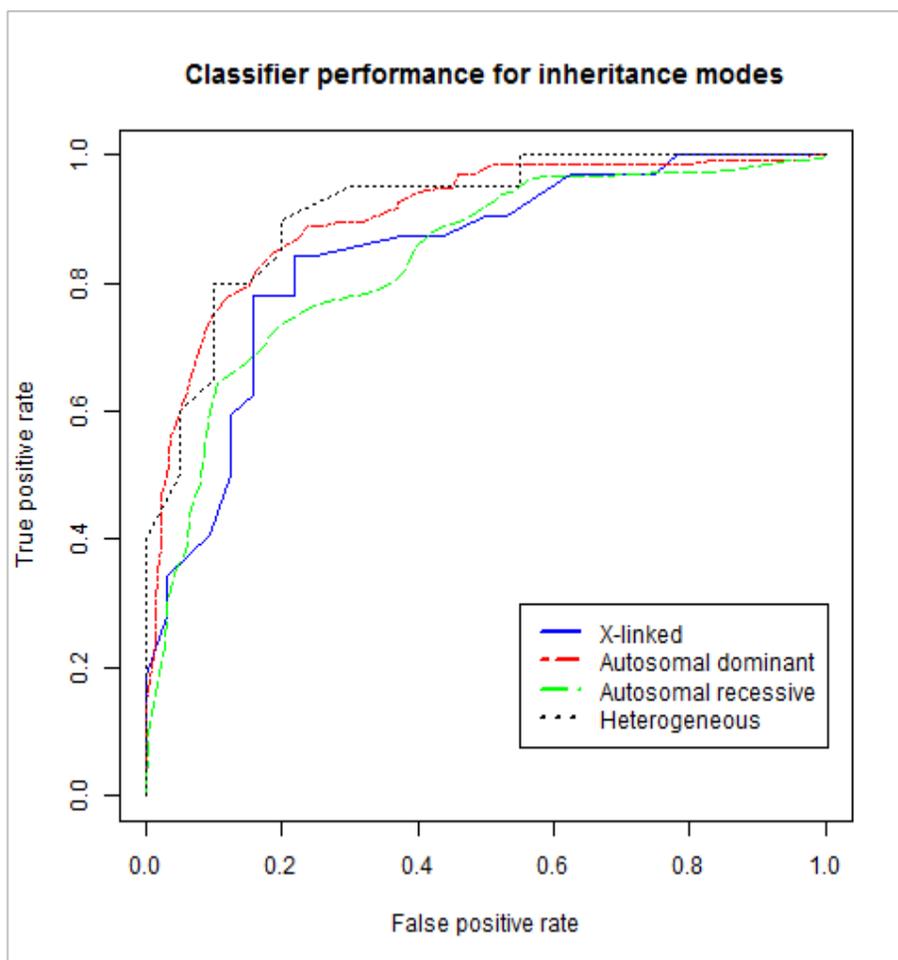
## Prioritisation Results by Anatomical Category



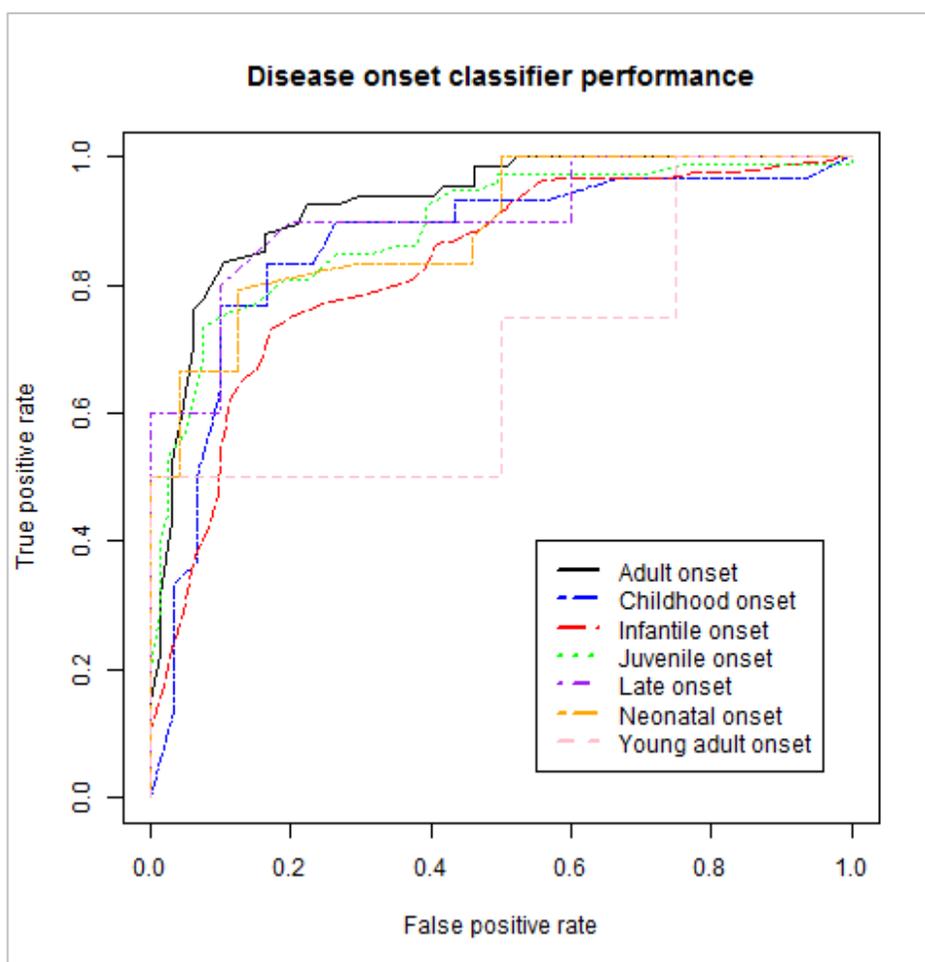
**Figure 11** (previous page) shows GeneTiER performance across a range of tissue types using **Dataset 2**. 500 disease genes known to cause a localised phenotype were prioritised together with 100 randomly selected genes. Tissues are grouped by Disease Ontology terms which are descendants of 'disease of anatomical entity'. The percentage of disease genes in each category is shown by rank distribution – e.g. darkest blue represents the percentage of disease genes in each category ranked first.



**Figure 12.** ROC curve comparison of prioritization using RNA-Seq data only; Microarray data only; or combined data using benchmarking **Dataset 1**.



**Figure 13.** GeneTiER performance visualised as ROC curves for different disease inheritance modes using **Dataset 1**.



**Figure 14.** ROC curves showing GeneTiER prioritisation performance across a range of diseases with different onset age using **Dataset1**.

### **3.3.2 Case Study Genes**

In order to further investigate the circumstances where the methodology presented here either failed or succeeded, a case study of the global expression patterns was conducted using genes implicated in retinitis pigmentosa (OMIM:610282), a degenerative eye disease causing severe vision impairment.

**Figure 15** shows the expression profiles across multiple normal tissues of 5 disease genes known to underlie retinitis pigmentosa (Ali et al. 2017), while **Table 3** and **Table 4** shows the summary of the mean and mean reciprocal ranks

obtained using the GeneTiER methodology. Mean reciprocal ranks is a common metric for evaluating ranking algorithms, which is calculated as:

$$\text{Mean Reciprocal Rank} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{Gene Rank}_i}$$

where  $N$  is the number of ranked lists used for evaluation.

Gene	PR1		ROM1		PRPF6		PRPF31		PRPF3	
Input Size	Mean Rank	Standard Deviation								
50	34.7	11.03	8.16	11.7	8.8	4.2	2.9	1.78	5.4	3.01
100	66.03	8.9	17.07	6.38	22.7	5.53	4.13	2.21	7.1	2.54
200	172.07	5.75	53.87	6.47	28.33	3.2	20.3	3.91	28	4.08
500	288.11	13.33	67.4	10.03	140.65	15.07	39.24	4.45	41.65	6.51

**Table 3.** Mean ranks and standard deviations of 5 case-study genes shown in **Figure 15**. Each gene was ranked 30 times against a set of 50, 100, 200 and 500 randomly generated genes.

Mean Reciprocal Rank					
Input Size	PR1	ROM1	PRPF6	PRPF31	PRPF3
50	0.78	0.09	0.09	0.08	0.08
100	0.66	0.17	0.23	0.04	0.07
200	0.86	0.27	0.14	0.10	0.14
500	0.58	0.13	0.28	0.08	0.08

**Table 4.** Mean reciprocal ranks of 5 case-study genes assessed against a set with 50,100, 200 and 500 randomly generated genes; 30 replicates.



**Figure 15** (previous page). Expression profiles of PRPF3, PRPF31, PRPF6, ROM11 and RP1 genes, associated with retinitis pigmentosa (OMIM: 610282) across a selection of tissues from RNA-Seq data in GeneTiER database.

---

The genes PRPF3 and PRPF31 show distinct, tissue-specific expression in eye tissues with negligible expression in non-ocular tissue; disease genes with similar expression profiles are ranked very highly by GeneTiER. While PRPF6 is also highly expressed in eye tissues, unlike PRPF3 and PRPF31 its expression is not limited to ocular tissues, resulting in a reduced, but still strongly suggestive ranking. ROM1 is expressed in a number of non-ocular tissues as well as in corneal epithelial cells but still ranked highly. This was in spite of its lower expression in corneal epithelial cells than that of the PRPF genes and comparable expression in adipose tissue. Unsurprisingly, in view of its ubiquitously low expression levels, the methodology failed to identify PR1.

The eye is a complex organ with many specialised tissue types. While the geneTiER database contains expression data from corneal epithelial cells, stromal cells and lens from a mouse, these don't encompass all the diverse cell types present in the eye. For example, rod photoreceptor rhodopsin gene, a major cause of retinal dystrophy (Broadgate et al. 2017), is not expressed in any of the eye tissues in the geneTiER database. Conversely, diseases affecting vision can be neurodegenerative in nature, where the causative gene does not have a function in the eye.

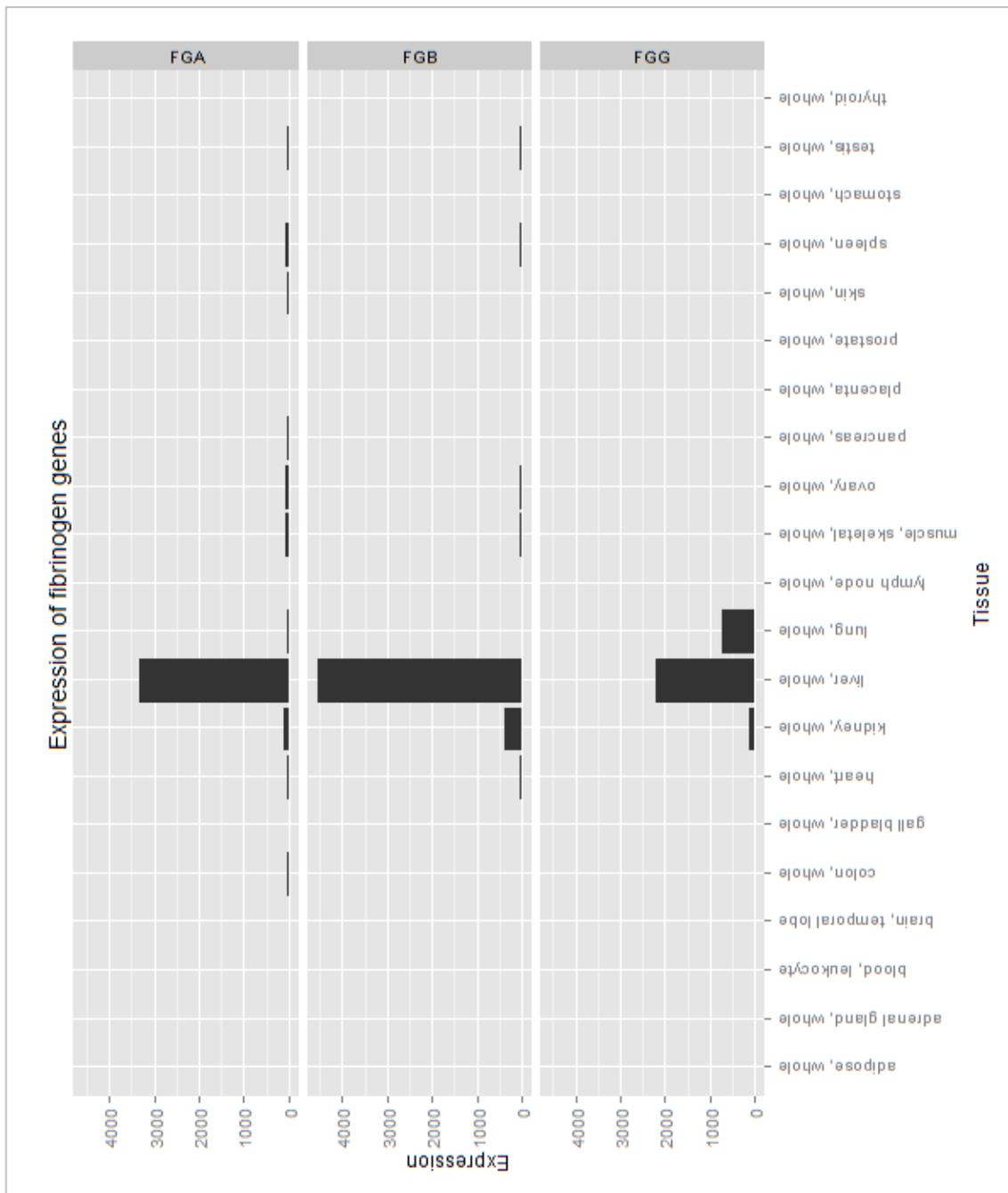
While this work shows that GeneTiER is capable of accurate disease gene prioritization through ROC analysis (with AUC values of up to 0.83), it should be noted that the disease gene is rarely ranked first in the output. This ranking should therefore be used as a guide to the order in which candidate genes should be analyzed further. Even so, it must be noted that not all disease gene expression patterns conform to the assumptions underlying our model. For example, some disease genes show universally high or low gene expression across all tissues

(see RP1 in **Figure 15**). Indeed, ectopic expression of genes can result in a disease phenotype, as is the case in many cancers. However, in order to detect these patterns, the differential expression change must be observed between the normal and affected state. While including differential expression data from normal and affected patients would no doubt improve GeneTiER performance, public availability of such data is mostly limited to a small number of well-studied diseases and therefore would enhance the results for only a small proportion of cases.

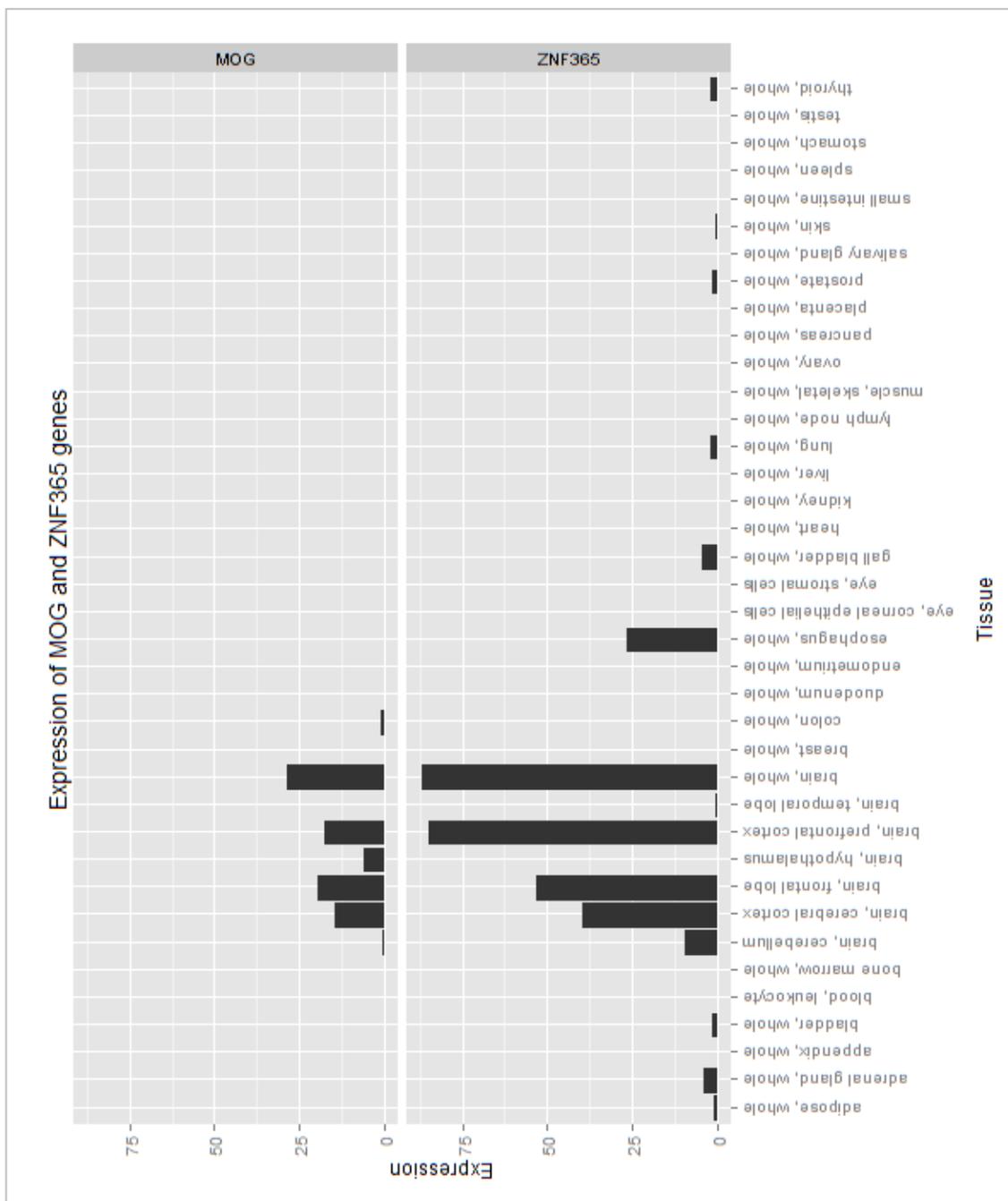
Furthermore, as Oellrich *et al* (2014) note in their analysis, the site of gene expression and the visible phenotype do not always coincide. Consequently, the limitations of this method must be understood and taken into consideration when examining the final gene rankings. This is especially true where the link between tissue and phenotype may not be immediately obvious. For example, congenital dysfibrinogenemia (OMIM:616004) is a blood clotting disorder caused by defective fibrinogen genes FGB, FGG and FGA. Circulating factors affecting blood clotting are synthesized by hepatocytes, and indeed, data collated in GeneTiER database shows that fibrinogen genes are highly and exclusively expressed in the liver (**Figure 16**). However, GeneTiER would not identify these disease genes if the user failed to take this into account and selected blood, rather than liver, as the affected tissue.

Narcolepsy-cataplexy (ORPHANET:2073) is a sleep disorder with multiple causative genes identified. GeneTiER scores a number of these highly, for example MOG and ZNF365, due to localized expression in parts of the brain (**Figure 17**). However, the disease can have an autoimmune component and in some patients the phenotype has been attributed to the loss of neurons in the hypothalamus due to autoimmune attacks. Consequently, this methodology fails to identify histocompatibility genes HLA-DQB1 and HLA-DRB1 as causative genes for the disease and therefore may also find other phenotypes arising from heterogeneous causes challenging (**Figure 18**). Furthermore, tissue samples are often a heterogeneous mix of different cell types and may also contain other

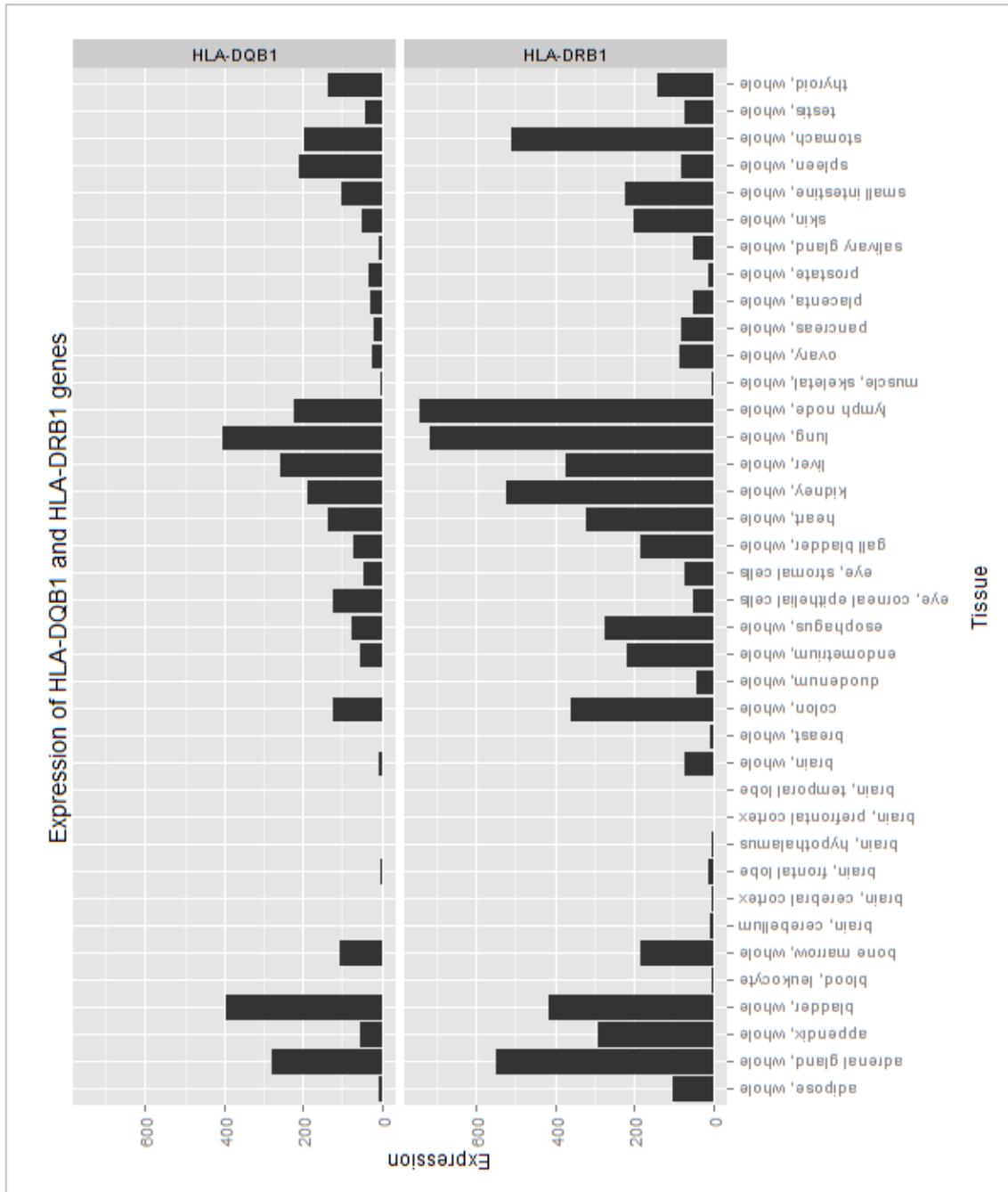
contaminating cell types, in particular immune cells, which can confound the GeneTiER methodology.



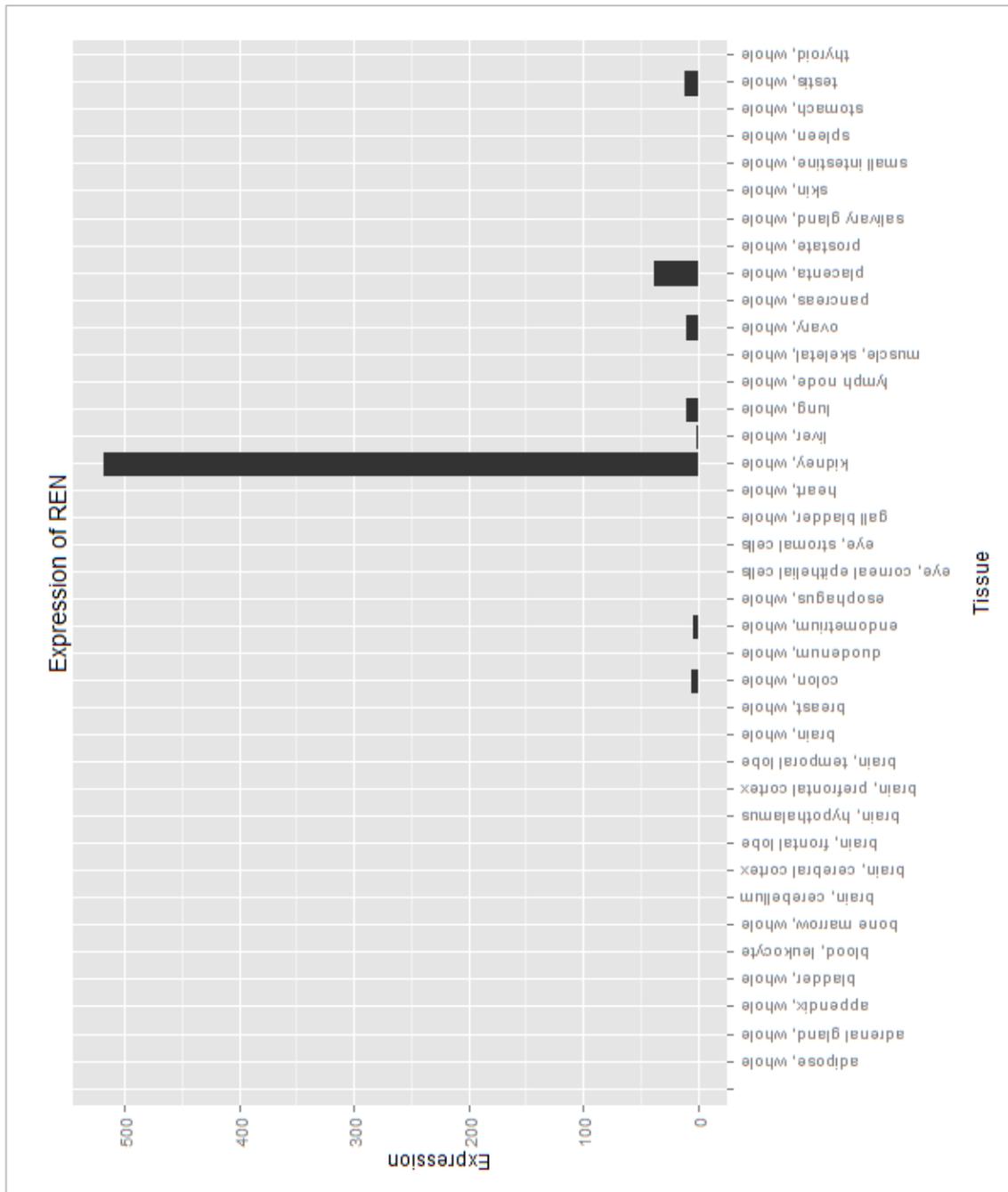
**Figure 16.** Expression of fibrinogen genes across a range of tissues from RNA-Seq data in GeneTiER database.



**Figure 17.** Expression of MOG and ZNF365 genes across a range of tissues from RNA-Seq data in GeneTiER database.



**Figure 18.** Expression of HLA-DQB1 and HLA-DRB1 genes across a range of tissues from RNA-Seq data in GeneTiER database.



**Figure 19.** Expression of REN gene across a range of tissues from RNA-Seq data in GeneTiER database.

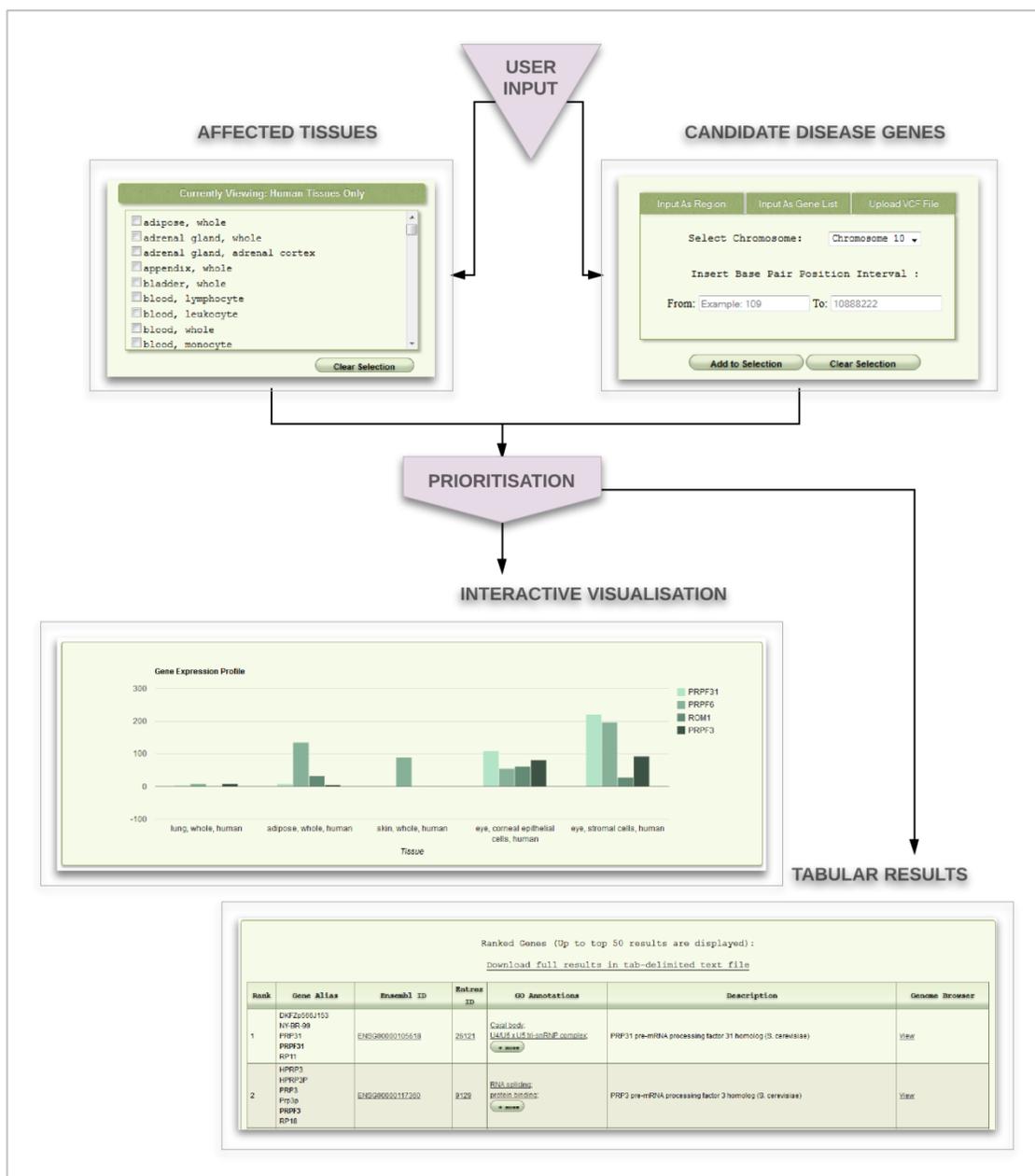
In spite of these challenges, there are numerous cases where the observed phenotype correlates with the site of expression exceedingly well. For example, renal tubular dysgenesis (OMIM:267430) is characterized by a congenital abnormality of the kidneys with low amniotic fluid during pregnancy. The protein associated with the disease, REN, is produced mostly by juxtaglomerular cells of the kidney. The data in GeneTiER database agrees with this, showing elevated expression in the kidney, as well as a secondary major site of expression in the placenta (**Figure 19**).

### ***3.3.3 GeneTiER Application***

GeneTiER algorithm and database access has been implemented and made publicly available via a web interface. This is summarised in **Figure 20**. The workflow is designed to allow the user to select from the database of affected tissues and provide the candidate disease genes in multiple formats. The options include a list of commonly used gene identifiers, region inputs mapped to hg19 (e.g. a linkage region or a region of homozygosity), or a VCF file containing variant coordinates again mapped to hg19. Positional data (variants and regions) are internally parsed and mapped to genes using hg19 annotation metadata.

The prioritisation results are provided in a tabular format, while gene expression profiles across any available tissues can be visualised as an interactive, JavaScript powered chart. The application has been specifically designed to require only minimal user input, and takes care of conversions between a variety of commonly used gene identifiers and between human/mouse orthologs.

GeneTiER does not require the user to have any prior knowledge of the disease, other than the ability to unambiguously identify affected tissues. Organs are made up of many functionally diverse tissue cell types and this can be reflected in the experimental data. Therefore, this work strives to collate data from multiple, distinctive datasets, to enable the user to make tissue cell type-specific queries which are not supported by many of the popular databases.



**Figure 20.** Overview of GeneTiER application. The web-based interface allows the user to supply candidate disease genes to prioritise and to select affected tissues. Top prioritisation results are returned in a tabular form and are available to visualise and compare using an interactive chart. Full results are available for download.

### ***3.3.4 Discussion***

One of the major limitations for data miners attempting to utilise publicly available gene expression data sets stems from the inability to accurately, directly compare the data. In particular for expression data derived from microarray experiments, the raw data is not comparable even within the same experiment until some level of normalisation has been imposed. Furthermore, different array designs make direct comparisons difficult, regardless of the rigidity of the normalisation process. High throughput sequencing data is not exempt from this - sequenced reads must be normalised to library size and/or composition in order to achieve some degree of comparability.

Here, rather than comparing expression values directly, relative expression is compared per experimental condition by calculating modified z-scores. As discussed in the **Methods** section, standard z-scores are used to measure the distance of data points from the mean of the distribution. However, the transformation is only representative if the data are normally distributed – which is not the case for expression data. RNA-Seq data in particular is heavily skewed and can be more optimally represented by a negative binomial distribution. This makes comparisons between RNA sequencing and microarrays particularly difficult, as RNA-Seq is capable of accurately quantifying highly expressed genes where microarrays have an inherent light intensity limit that can be accurately measured.

Modified z-scores take into account the median absolute distance of each data point, and thus the transformation enables quantification of relative gene expression within an experiment. Thus, while direct comparisons between raw expression values could be highly inaccurate, comparing relative expression through modified z-scores is more intuitive and meaningful, as the expression value being compared is relative to the abundance of all other mRNAs measured. Indeed, it has been shown that similar relative rank-based transformations facilitate more accurate comparisons between heterogeneous datasets (Liu et al. 2008; Welsh et al. 2013; Kvam et al. 2012).

Performance assessment using ROC curves presented here indicates that tissue expression-based ranking is capable of meaningful candidate gene prioritization and performs strongly in a substantial proportion of cases tested. However, direct performance comparisons between gene prioritization tools are difficult. Many popular gene prioritization methods that rely on prior knowledge about a disease use either text-mining approaches or Gene Ontology annotations to score candidates based on relevance to query. Thus, to provide a meaningful performance comparison, a cross-validation approach is required – that is, for each test case of a known disease gene, any direct associations to the query disease must be removed from the test. However, to the best of the knowledge of the author, no web gene prioritization application allows for such performance assessment.

Currently, there are still thousands of human genes with no available GO annotations and many more with ‘shallow’ annotations. While this presents a problem for disease gene inference by similarity, the method described here would not be any less applicable. For example, at the time of developing GeneTiER, no Gene Ontology annotations have yet been ascribed to human CDR1 gene, known to contribute to paraneoplastic cerebellar degeneration (OMIM:302650). This gene shows localized expression in brain tissues, in particular in the cerebellum, and as such is scored highly by GeneTiER, whereas approaches reliant on prior knowledge are likely to fail.

On the other hand, it is also worth considering that the data and sample quality collected in the GeneTiER database may have an impact on the accuracy of the prioritization results. Human tissues are often donated through various circumstances. However, these can alter the tissue state or environment, and thus the expression patterns may become perturbed. Tissues donated after removal through various surgical procedures, often due to injury or illness, cannot be considered to be wholly in their native state. Trauma can induce genes to be expressed that would not normally be active in the tissue under normal

circumstances – for example, genes acting in stress response, repair or apoptosis pathways. Furthermore, donated tissue samples are removed from their immediate, normal extracellular environment, the loss of the influence of which will also invariably affect gene expression. Similarly, tissues removed during a *post mortem* may have been deprived of an oxygen supply from several hours to days, perturbing normal gene expression. Thus, this may have a negative impact on gene prioritization accuracy that is difficult to assess.

Similarly, GeneTIER will not be able to identify disease genes that are expressed exclusively in tissues not present in the collected dataset. Likewise, diseases caused by genes that are expressed in response to either an environmental stimulus or within a short development time frame will not perform well if the appropriately stimulated tissue is absent from the database.

Nevertheless, this chapter has highlighted that an unbiased tissue expression prioritisation approach can provide meaningful candidate gene scoring. GeneTIER aims to highlight genes with tissue-specific expression patterns to the user from among other candidate genes in their dataset, and as such will perform best for diseases with distinct, localized phenotypes. A broad selection of tissues allows for scoring of complex phenotypes affecting any combination of tissues. Thus, GeneTIER offers great utility value to the research community and can effectively supplement the *in silico* toolbox of any researcher.

## ***4 OVA: Ontology Variant Analysis Tool***

### ***4.1.1 Motivation***

The candidate gene prioritisation method implemented in GeneTiER is a largely unbiased approach that does not rely on any prior information about candidate genes. This can be particularly useful where knowledge-based approaches fail, as can be in cases of novel phenotypes caused by *de novo* mutations; or mutations present in genes of yet unknown function. However, while this method is capable of prioritising candidate disease genes with better than random accuracy, there are also many cases where it fails and/or is not appropriate to use. Thus, here the author investigates an alternative method for candidate gene prioritisation that could be used as a complement to the GeneTiER application.

As discussed previously, knowledge-based approaches often suffer from biases towards the better characterised genes, as well as often failing to identify disease genes of yet unknown functions. However, as the body of knowledge in biological literature is constantly growing, the issues stemming from gaps in available knowledge are likely to become less of a concern. Thus, as the practical gains that can be obtained from incorporating prior knowledge to candidate gene prioritisation outweigh the potential concerns, here a knowledge-based approach is reconsidered.

### ***4.1.2 Representing Biomedical Knowledge in Machine-Readable Ways***

In order to utilise biomedical domain knowledge for computational candidate gene prioritisation, a machine-readable way of describing biological entities, such as genes, diseases, phenotypes, functions and pathways is required. There are a number of approaches that could be used.

A common method to obtaining data for gene prioritisation is text mining. Text mining could be used to obtain information from free form text databases such as OMIM; PubMed abstracts; or, open access full-text articles. However, text mining

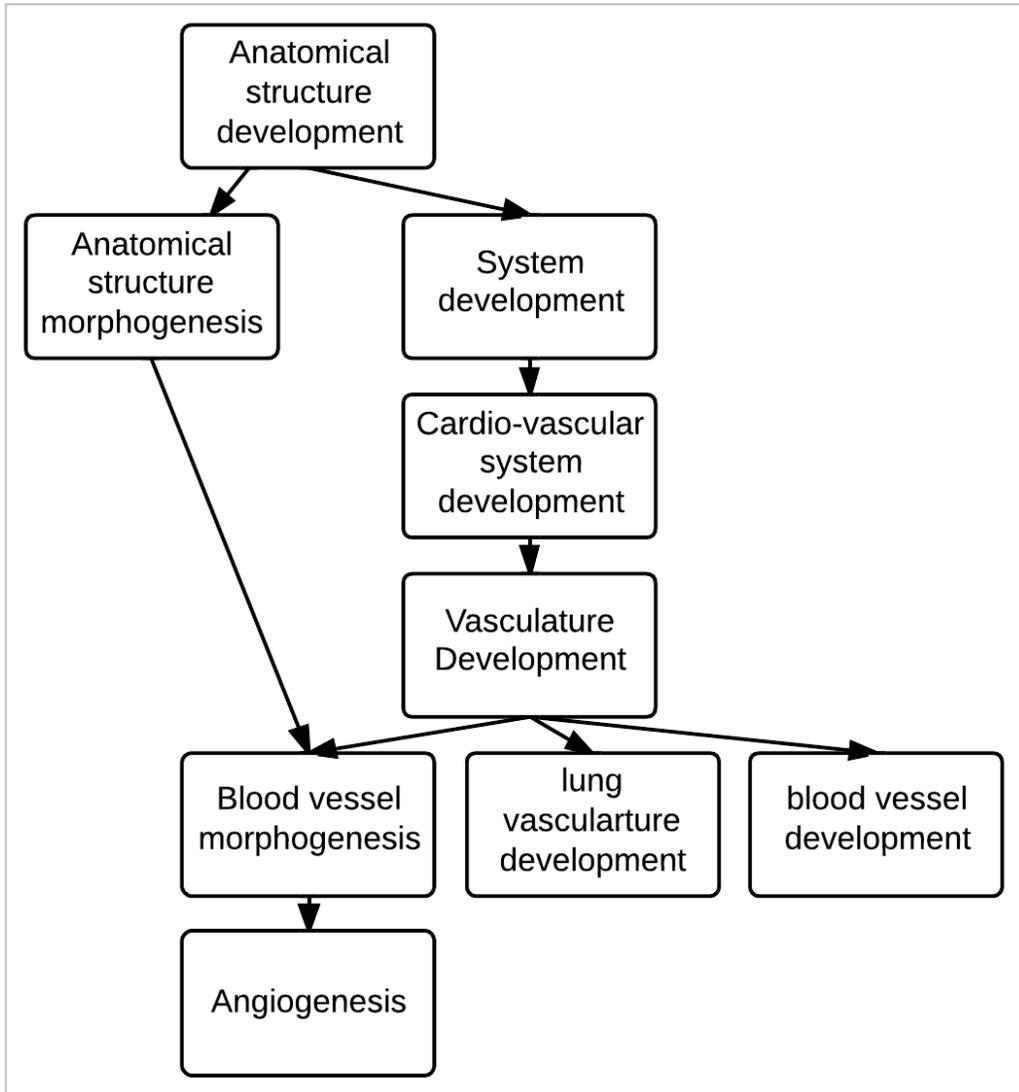
information presents a number of challenges, from recognition of named entities, such as genes or diseases, to interpreting the types of association described in free text. Named entity recognition can be particularly hard in the biomedical domain, as gene and protein names are not fully standardised and are often referred to, particularly in older literature, by legacy names or synonyms. Other terms, such as phenotypes, can be even more difficult to fully characterise. Additionally, text mining can lead to inaccuracies - OMIM entries, for example, can be misleading as some entries may contain information retained for historical overview, which could not be easily distinguished from current scientific consensus by an automated system. Furthermore, it can be difficult to successfully extract associations between biomedical entities when written in natural language, as entity co-occurrence can imply a negative as well as a positive relationship. Even the most advanced natural language processing systems do not achieve 100% precision or recall. In a system which attempts to utilise text-mined data in order to prioritise candidates, this inadequacy would further add to concerns such as incompleteness of data. Furthermore, not all biomedical literature is available for text mining – in fact, only approximately 600,000 full text articles are available for download through Medline. While this number is on the rise and more journals than ever are making full text articles available for free access, currently the majority of the associations to be made by a text mining system would be missed. Alternatively, a text mining tool may choose to limit its scope to Medline abstracts only; however, key information may not always be mentioned in the abstract.

In order to overcome these difficulties, a number of resources exist which attempt to describe biological knowledge in a standardised way. Many databases could function as a controlled vocabulary – for example, GeneTiER already uses Ensembl database (Cunningham et al. 2014) to obtain gene symbols, aliases, synonyms and database identifiers to describe genes. Pathway databases such as KEGG (Kanehisa and Goto 2000) or Reactome (Fabregat et al. 2016) could be used to describe broad molecular functions of these entities. Medical subject header (MeSH) terms, which are used to index articles by their subject at NCBI, is a thesaurus currently containing over 87,000 entries describing biological concepts

and entities such as genes, diseases and chemicals (Dhammi and Kumar 2014), and could also form as part of the controlled vocabulary needed.

However, in recent years, ontologies have become a *de facto* standard for organising knowledge in the biomedical domain in a structured, controlled manner (**Figure 21**). Ontologies serve to organise concepts of a particular domain in a structured, hierarchical way by utilising a small number of relationship types between entities. Collectively, ontology terms represent a controlled vocabulary describing a particular domain, and subsequently, this controlled vocabulary can be used to annotate other entities. Ontology annotations can thus facilitate computational analysis of entities and concepts they relate to within a given domain. Ontological annotations largely circumvent the problems that arise from the use of natural language descriptions, such as ambiguity and subjectivity, and have been invaluable in large scale annotation projects in the biological domains, such as whole genome annotations.

From a computational stand-point, an ontology is a directed, acyclic graph (**Figure 21**) in which vertices generally correspond to controlled vocabulary terms and the edges represent the relationships between terms. Edges can correspond to different relationship types, though often the most common type of edge is an 'is\_a' relationship. As ontology terms are organised in a hierarchical manner, with broad terms nesting towards the root of the ontology, while more specific terms are further away from the root. The directed acyclic graph organisation of concepts allows for the application of several graph-theory algorithms in order to analyse and extract relevant information.



**Figure 21.** Gene Ontology sub-graph example of directed, acyclic graph ontology structure. Broader, less informative terms nest at the top of the hierarchy, while the descendant terms are more specific and informative.

### ***4.1.3 Biomedical Domain Ontologies***

Biological and biomedical ontology development is largely coordinated and organised by the Open Biological and Biomedical Ontologies (OBO) Foundry, a collaborative endeavour that aims to standardise and create controlled vocabularies for use across multiple biological domains (Smith et al. 2007). To date, OBO Foundry contains over a hundred ontologies, 8 of which are considered to be mature.

Three of these, cellular component, biological process and molecular function, comprise Gene Ontology, which attempts to standardise terms used to describe characteristics of genes and proteins across multiple species. Gene Ontology represents perhaps the most extensive collaborative effort to standardise the characteristic descriptions of genes and their products, and as such lends itself to a number of applications. As Gene Ontology has received significantly more development effort than other ontologies that are part of the OBO Foundry project, a number of diverse applications have been developed that utilise this resource, including gene set enrichment analysis (Yu et al. 2012, 2015), interaction network analysis (Maere et al. 2005; Garcia et al. 2007; Vlasblom et al. 2006) and gene prioritisation (Chen et al. 2009).

While Gene Ontology annotations have been exploited by many applications for candidate disease gene prioritisation, to date very little effort has been made to explore the potential of integration of data from multiple ontology types, such as gene, phenotype, disease or pathway ontologies for candidate gene prioritisation. This is perhaps due to the later development of some of these resources. For example, the Mammalian Phenotype Ontology is the oldest phenotype ontology, however, the annotation set available is very limited and a lot of terms are redundant or irrelevant for human data. The Human Phenotype Ontology (HPO) (Köhler et al. 2014) was first proposed in 2008 and is used in a number of different applications (Smedley et al. 2013; Köhler et al. 2009), while the UberPheno (Köhler et al. 2013) – a cross-species ontology integrating phenotype data from human, zebra fish and mouse was not established until 2013. The development of HPO,

and subsequently UberPheno, can facilitate computation analysis of the human phenome in applications where previously text mining of unstructured text-based databases was the only alternative, and as such, was impeded by low precision/recall associated with the application of such techniques to the biological domain.

Currently, to the best knowledge of the author, only a single gene prioritisation tool, PHIVE (Robinson et al. 2014), takes advantage of Uberpheno a recently made available cross-species ontology. PHIVE uses a phenotype similarity measure to score genes from a list of candidates generated by exome sequencing studies relative to the phenotype of the disease and is proposed to be applicable more to clinical diagnostics than gene discovery, as it relies solely on phenotype annotations. Even when taking into account phenotypic annotations that stem from model organism ortholog data, phenotype annotations still present very low coverage of the human genome and as such, it is likely that the majority of novel disease genes would be missed by approaches such as PHIVE.

This work therefore proposes the integration of data and annotations from multiple ontologies, including the Human Phenotype Ontology (Köhler et al. 2014), UberPheno cross-species ontology (Köhler et al. 2013), the three domains of Gene Ontology (Ashburner et al. 2000), the anatomical entity ontology Uberon (Mungall et al. 2012), Disease Ontology (Kibbe et al. 2015) and Pathway Ontology (Petri et al. 2014) to develop a candidate gene prioritisation method that increases the coverage, applicability and precision of standalone phenome-based tools, such as PHIVE (Robinson et al. 2014), Phenolyzer (Yang et al. 2015) or PhenoDigm (Smedley et al. 2013).

An automated method of linking phenotypes to molecular pathways and functions would also be of practical benefit. Currently, the majority of knowledge-based gene prioritisation methods require the user to have extensive domain knowledge, as genes are prioritised with respect to human selected functions and pathways. However, this input is subjective to perceptions of the researcher and

therefore could be inaccurate; furthermore, the user may not be aware of cases where a similar phenotype is caused by perturbation of different biological process, and thus vital associations could be missed, whereas an automated approach is likely to overcome some of these biases.

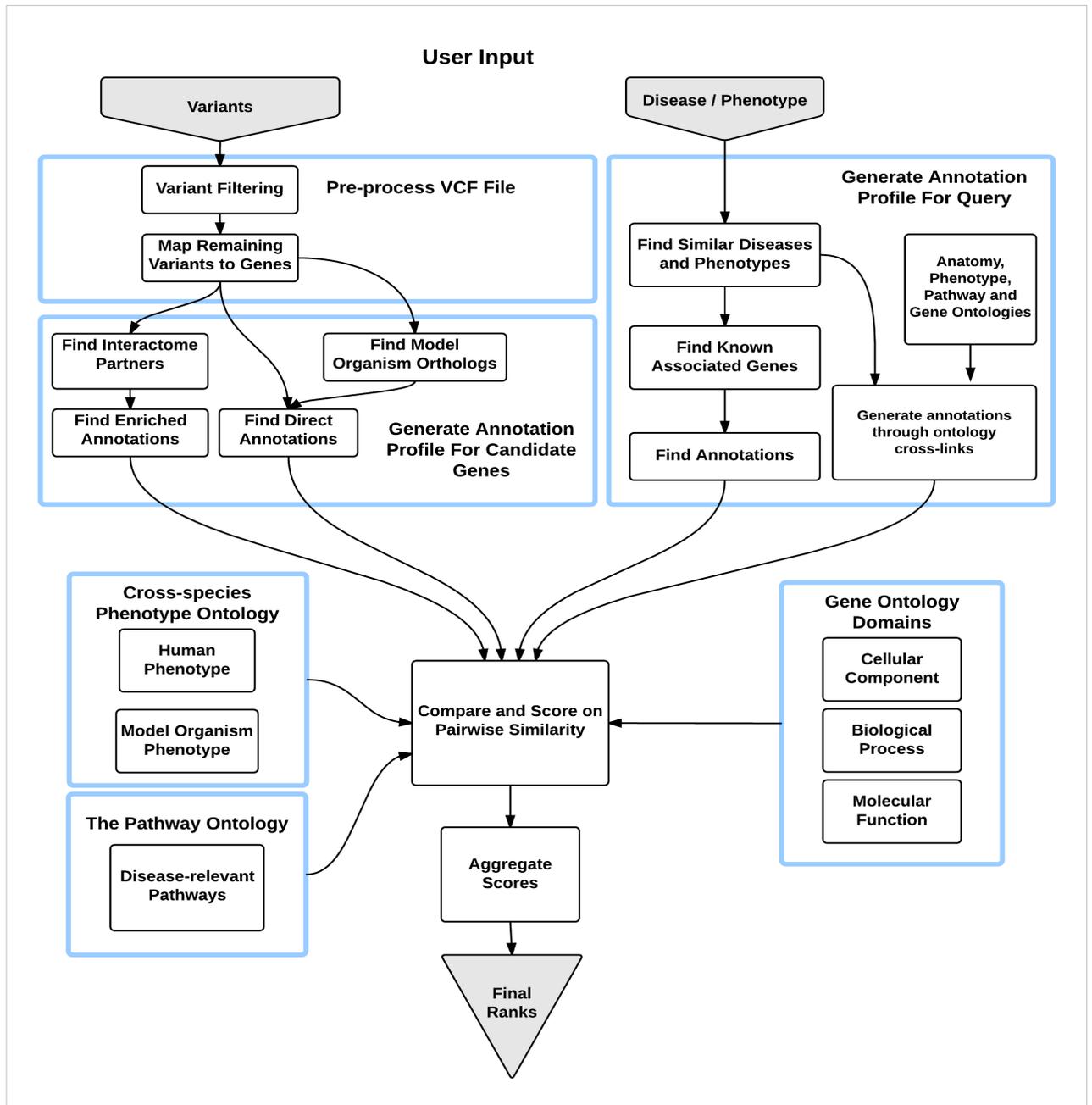
Thus, the remainder of this chapter focuses on a method for close integration of multiple biomedical domain ontologies that can facilitate complex queries and be applied for prioritising candidate disease genes and variants. This approach has been implemented and made available as a web application, Ontology Variant Analysis (OVA) tool.

## ***4.2 Methods***

### ***4.2.1 Overview***

As implemented, OVA workflow consists of three main steps, summarized in **Figure 22**. User may supply candidate genes in the form of a list, a genomic region or a VCF file. In the case of the latter, additional filtering steps are implemented in the software. Uploaded VCF files are passed through custom user variant filters in order to substantially reduce candidate search space by removing likely benign variation. Each remaining variant is mapped to a gene, for which an extensive multi-ontology annotation profile is derived using publicly available annotation data sets consisting of direct annotations and inferred annotations from model organism data and data from the local interactome neighbourhood – i.e. proteins that are known to directly interact with the query. For a given query phenotype or disease, a comparison annotation profile is computed from known phenotype-genotype associations, phenotype similarities and cross-links between multiple ontologies. OVA compares each candidate gene's annotation profile to the phenotype/disease annotation profile and calculates a series of similarity metrics. These scores, together with a number of other related features, are used as an input to a model built using a supervised, decision tree based learning approach, which computes the probability of each candidate gene harbouring the disease-causing variant.

OVA follows the same software multi-tier design patterns as implemented in GeneTiER application, with user input collected and validated client-side and before being passed to the server-side ‘business’ tier for database queries and execution of the analysis algorithms.



**Figure 22.** Overview of OVA workflow for VCF file input.

## ***4.2.2 Software Implementation***

### ***4.2.2.1 User Interface and User Input Processing***

OVA has been made available as a web-based application, currently hosted on a CENTOS server and accessible at **dna2.leeds.ac.uk**. HTML, CSS and JavaScript were used to implement the user interface. JQuery and Java Server Pages were used to facilitate client-server interactions and generate dynamic page content. Gene and variant input validation follows the same procedure as in GeneTiER (**Figure 7**), with gene list, genomic region and variant input types eventually mapped to the same set of reference genes. Disease and phenotype user input is limited to database entries by implementing a database auto-complete free text database search system and as such, limited validation was required. Server-side validation, all algorithms and other data processing tasks were implemented in Java programming language.

### ***4.2.2.2 OVA database***

All required data for candidate gene prioritisation is stored in a MySQL database. Database tables are summarised in **Table 5**. The data in the database is queried through Java classes implementing the JCDB database connector.

Ontology terms were downloaded from respective databases (Smith et al. 2007; Köhler et al. 2014; Mungall et al. 2012; Köhler et al. 2013; Golbreich et al. 2006; Petri et al. 2014; Ashburner et al. 2000; Kibbe et al. 2015). Ontology graph paths were parsed from OBO or OWL formatted data using custom code and stored in MySQL database in a pairwise format, where each row stores a pair of terms, the relationship type between them and distance between. While storing all ancestors of a term in a graph path increases the redundancy in the database, this approach facilitates easier and faster database queries. For each ontology, information content and pairwise term similarity scores were pre-computed using custom code and stored in the database. Ontology cross-link tables were assembled from cross-references made available in the original ontologies and extracted via custom code by:

- 1) reasoning across multiple ontologies;
- 2) key word searches across the ontology graph structure;
- 3) manually curated from the data

and stored in the MySQL database. Ontology annotations were downloaded from their respective web sites and standardised using custom code before insertion into the MySQL database. Disease databases were downloaded from OMIM (Amberger et al. 2015), DECIPHER (Firth et al. 2009) and ORPHANET (Pavan et al. 2017) web sites respectively. Tissue tables were assembled from manually curated Uberon (Mungall et al. 2012) ontology terms. Human and model organism gene information, including identifiers, gene names, coordinates and coding exon sequences were downloaded via UCSC Table Browser tool (Karolchik et al. 2004) or from the Ensembl database via Ensembl Biomart (Smedley et al. 2015). Gene Ontology terms were mapped to GO Slim terms via custom code by traversing the GO and GO Slim graph paths and mapping each GO term to the GO Slim term with highest information content. Interactome data was downloaded from the STRING (Jensen et al. 2009) and mentha databases (Calderone et al. 2013), pre-processed using custom code to standardise the format and stored in the MySQL database. For each human gene, its closest interactome partners were selected and ontology enrichment analysis was performed for each group. Fisher's Exact test was used to select ontology terms significantly enriched in the direct neighbourhood (direct edges only) of each gene, and this was stored as a pre-computed table in MySQL database. A curated transcription factor list was downloaded from UniProtKb database (Magrane and Consortium 2011) by querying reviewed entries for human species and "transcription factor" keywords. Uniprot identifiers were linked to gene identifiers using Ensembl Biomart. Codon tables were downloaded from GenScript web page [<http://www.genscript.com/tools/codon-table> – accessed 01/01/2014].

Semantic similarity table between diseases was computed using custom codes, using a best match average of HPO terms assigned to each disease. Allele frequencies from 1035 individuals in the Born in Bradford cohort were computed

using custom code that processed all VCF files and counted all heterozygous and homozygous alleles, with the results stored in the MySQL database. All database tables were indexed and optimised for query speed due to their large size.

<b>Table Name</b>	<b>Description</b>
<b>Gene Ontology (GO) Tables</b>	
GO Graph Path	Graph path of gene ontology terms, containing term 1, term 2, type of relationship and distance
GO Mouse Annotations	GO mouse ortholog annotations
GO Zebrafish Annotations	Gene ontology annotations for zebrafish
GO Human Annotations	Gene ontology annotations, mapping of gene identifier to term identifier; includes evidence code
GO Rat Annotations	Gene ontology annotations for rat
MF Information Content	Information content of GO Molecular Function ontology terms
BP Information Content	Information content of GO Biological Process ontology terms
CC Information Content	Information content of GO Cellular Component ontology terms
Pairwise Similarity BP	Pre-computed GO pairwise similarity scores between Biological Process terms
Pairwise Similarity CC	Pre-computed GO pairwise similarity scores between Cellular Component terms
Pairwise Similarity MF	Pre-computed GO pairwise similarity scores between Molecular Function terms
GO Slim Graph Path	Graph path of GO slim ontology
<b>Human Phenotype Ontology (HPO) Tables</b>	
Human HPO Annotations	HPO term to gene annotations
HPO Information Content	Information content of HPO terms

HPO Pairwise Similarity	Pre-computed pairwise similarity of HPO terms
HPO Terms	HPO terms and their descriptions
HPO Graph Path	HPO graph path
<b>Uberpheno Ontology Tables</b>	
UberPheno Human Annotations	UberPheno Gene annotations
UberPheno Terms	Uberpheno terms and their full descriptions
UberPheno Graph Path	Uberpheno graph path
UberPheno Information Content	Information content of Uberpheno terms
UberPheno Semantic Similarity	Pre-computed pairwise similarity of Uberpheno terms
UberPheno Mouse Annotations	UberPheno Gene annotations
UberPheno Zebrafish Annotations	UberPheno Gene annotations
<b>Pathway Ontology (PO) Tables</b>	
PO Annotations	PO term to gene annotations
PO Information Content	Information content of PO terms
PO Pairwise Similarity	Pre-computed pairwise similarity of PO terms
PO Terms	PO terms and their descriptions
PO Graph Path	PO graph path
PO Rat Annotations	PO term to rat gene annotations
<b>Uberon Ontology Tables</b>	
Uberon Annotations	Uberon term to gene annotations
Uberon Information Content	Information content of Uberon terms
Uberon Pairwise Similarity	Pre-computed pairwise similarity of Uberon terms
Uberon Terms	Uberon terms and their descriptions
Uberon Graph Path	Uberon graph path
<b>Disease Ontology (DO) Tables</b>	
DO Annotations	DO term to gene annotations
DO Information Content	Information content of DO terms

DO Pairwise Similarity	Pre-computed pairwise similarity of DO terms
DO Terms	DO terms and their descriptions
DO Graph Path	DO graph path
<b>Foundational Model of Anatomy (FMA) Tables</b>	
FMA Annotations	FMA term to gene annotations
FMA Information Content	Information content of FMA terms
FMA Pairwise Similarity	Pre-computed pairwise similarity of FMA terms
FMA Terms	FMA terms and their descriptions
FMA Graph Path	FMA graph path
<b>Gene Tables</b>	
Human Genes	Human genes and common identifiers
Rat Genes	Rat genes and common identifiers
Mouse Genes	Mouse genes and common identifiers
Zebrafish Genes	Zebrafish genes and common identifiers
Orthologs Human to Rat	Mapping between human and rat Orthologs
Orthologs Human to Mouse	Mapping between human and mouse Orthologs
Orthologs Human to Zebrafish	Mapping between human and zebrafish Orthologs
Synonyms	Common synonyms and aliases for human genes
Gene Type	Information about gene type- coding, non-coding, etc.
Exon coordinates	hg19 coordinates of human exon starts and ends
CDS sequences	Sequences of hg19 human coding exons
Codons	Table of human genomic and mitochondrial codons
<b>Cross-link tables</b>	
GO to GO Slim	Mapping of GO ontology terms to less

	descriptive, higher level 'slim' annotation terms
HPO to Uberpheno	Mapping of HPO terms to equivalent Uberpheno terms
PO to DO	Mapping of PO terms to equivalent DO terms
FMA to HPO	Mapping of FMA to HPO equivalent terms
Uberon to FMA	Mapping of FMA to Uberon equivalent terms
UberPheno to GO	Mapping of UberPheno to GO equivalent terms
Uberon to GO	Mapping of Uberon to GO equivalent terms
GO to PO	Mapping of GO to PO equivalent terms
Diseases to HPO	Mapping of diseases to HPO term mapping
HPO to Tissue	Mapping of any HPO terms specific to a tissue
HPO to Uberon	Mapping of HPO terms to equivalent Uberon terms
DO to Disease	Mapping of DO terms to diseases
Uberon to Tissue	Mapping of Uberon terms to a tissue type
<b>Other</b>	
Tissues	List of human tissue types
Diseases	Information about OMIM, ORPHANET and DECIPHER diseases
Disease Annotation	Human genes annotated to OMIM, ORPHANET and DECIPHER diseases
Disease to OMIM Categories	Diseases to disease category
Transcription Factors	Gene table of known transcription factors
Disease semantic similarity	Pre-computed OMIM, ORPHANET and DECIPHER disease to disease semantic similarity based on HPO annotations
Interactome (mentha)	Protein-protein interactions from mentha (mapped to genes)
Interactome (STRING)	Protein interactions from STRING (mapped to

	genes)
Enriched Interactome GO Terms	Pre-computed, enriched GO terms in each gene's local interactome neighbourhood
Born in Bradford	Allele Frequencies for all variants in Born in Bradford dataset

**Table 5.** Summary of OVA MySQL database tables.

#### ***4.2.2.3 Variant Filtering***

OVA combines classic variant filtering techniques together with ontology-based gene prioritisation in a single application. Variant filtering implementation has been designed to be optional, highly flexible and customisable for most datasets. VCF file and user-selected parameters undergo initial client-side validation via custom JavaScript functions before being uploaded to the server for processing. VCF files are read as a data stream from the client and are not stored server-side, thus circumventing some potential data security concerns stemming from long-term user data storage. VCF files are processed using custom Java code with variants passing user-selected screening criteria retained and stored in an intermediate binary, serialised file for later processing. Processed variants which have been stored on the server are automatically deleted after a month.

Variant filtering can be performed on a number of different user-selected criteria. These include variant classes which are often deemed benign variation, such as synonymous single base substitutions, small in-frame insertions or deletions (customisable size) or variation in intronic and/or untranslated regions. Additionally, a chromosomal region filter can be used to exclude variants outside regions of interest, for example, in cases where autozygosity mapping data is available. Genotype filtering integrates support for multi-sample VCF files, which can be used in a number of ways. For instance, in a case of an autosomal recessive phenotype,

only variants which are homozygous or compound heterozygous in the affected but not the unaffected patient samples will be retained.

A number of additional filtering options are also implemented, such as cut-offs based on the variant call quality score or allele frequency from exome data of 1024 mothers in the Born in Bradford project. It is also worth noting that the variant filtering approach implemented in OVA uses transcript, rather than gene, sequences in order to take codon position plurality (Subramanian, 2015) into account.

#### ***4.2.2.4 Semantic Similarity Quantification***

Semantic similarity refers to any metric defined for a set of terms, concepts or entities where the distance between each is quantified based on the meaning/semantic content, rather than from their lexical or syntactical representations.

OVA leverages the rigid, hierarchical structure of the ontology directed acyclic graphs to computationally quantify similarities and differences between biomedical and biological domain entities such as genes, diseases, tissues, pathways and functions. Concepts, or terms, close together on the ontology graph can be broadly considered to have similar meaning, while terms further apart are more semantically different. As an ontology is arranged as a hierarchical graph, terms further away from the root of the ontology are more specific, while terms closer to the root describe broad concepts. Consequently, two specific terms are more similar than two broad terms when separated by the same distance on the ontology graph. Thus, a semantic similarity measure needs to take into account not only the distance between two terms, but also their topographical position.

In order to quantify the semantic similarity between complex entities such as genes that are annotated by multiple ontology terms, one must first quantify the similarity of individual terms used to describe them. Semantic similarity between

two terms,  $a$  and  $b$ , can be described as the amount of information shared by the two terms. Given a hierarchical ontology structure, this can be quantified thus:

$$\mathbf{Sim}(a, b) = \frac{\mathbf{IC}_{MICA(a,b)}}{\mathbf{max}\{\mathbf{IC}_a, \mathbf{IC}_b\}}$$

Where  $\mathbf{IC}_{MICA(a,b)}$  is the information content (IC) of the most informative common ancestor (MICA) of the terms  $a$  and  $b$ . In information theory, the information content of a term  $t$  is often given as:

$$\mathbf{IC}_t = -\ln(\mathbf{P}_t)$$

Where  $\mathbf{P}_t$  is the probability of observing the term in a gold standard corpus (Lord et al. 2003). UniprotKB (Magrane and Consortium, 2011) annotations is a frequently used corpus for estimating information content of Gene Ontology terms and has been shown to facilitate fairly robust semantic similarity measurements (Pesquita, Faria, *et al.*, 2009). However, the annotation corpora that could be used to calculate the information content of terms in other biomedical ontologies are rarely complete or bias-free. Furthermore, there are a number of issues that can arise when estimating information content using a corpus, including bias towards the better characterized concepts, ‘orphan’ terms which cannot be meaningfully scored and the variability of the measure due to the evolution of the corpus (Lord *et al.*, 2003). Thus, in order to quantify similarities between terms within multiple ontologies in an accurate, static and systematic way that enables comparability between different ontologies, a different approach is needed. Consequently, here, a modified version of a previously described topology-based measure (Mazandu and Mulder, 2012) is used as follows. In the absence of a standard corpus, the probability of occurrence of a term can be estimated using the intrinsic ontology

structure - terms further away from the root are expected to be more specific and thus be observed less frequently in a hypothetical corpus. For example, The Human Phenotype Ontology term '*Abnormality of the eye*' (HP:0000478) is less informative than its descendant terms such as '*Glaucoma*' (HP: 0000501).

The level of the term in the ontology graph does not always correlate with its specificity. Fortunately, a number of topological characteristics in the ontology graph can help correct where this is not the case. The number of direct descendants of each term can be interpreted thus: if a term has a large number of children, its children are more specific than those of a term that has fewer children, as it encompasses more branches in the sub-domain. Furthermore, parents and their positions within the ontology graph should be taken into account. A term that descends from highly specific parent terms can be reasoned to be more informative than a term descending from less specific parent terms. The original approach (Mazandu and Mulder, 2012) considers the specificity of all direct ancestors of a term, using a product formula to calculate probabilities of occurrence, which, for 'deeper' ontologies in particular, can result in the inflation of specificity of terms with more than one direct ancestor; an issue that propagates down the ontology tree. To address this issue, here only the most informative direct ancestor of a term is considered in order to model a lower rate of information content gain while traversing down the ontology tree.

An ontology is never cyclic – thus, while a term may have multiple parents, it is impossible for a term to have a parent that is also its descendant. However, multiple direct ancestors of a given term may have child-parent relationships of their own. This property of the ontology graph allows calculating the information content of each term recursively, starting from the root of the ontology:

$$P_t = \begin{cases} 1 & \text{if root} \\ \frac{P_a}{C_a} \cdot 0.2 & \text{if } t \text{ 'is\_a' } a \\ \frac{P_a}{C_a} \cdot 0.4 & \text{if } t \text{ 'part\_of' } a \end{cases}$$

Where  $P_a$  is the probability of occurrence of a direct ancestor of term  $t$  and  $C_a$  represents the number of children direct parents of term  $t$  have. Here, only “is\_a” and “part\_of” relationship types are considered in all ontologies, which comprise the majority of all edges across ontologies used here. Each edge type is given a weight, with more weight (0.2 vs 0.4 multiplier) assigned to “is\_a” type edges. The addition of a weight factor to different edge types also allows for terms which are sole children of their ancestor to be assigned higher information content than that of their ancestor, which otherwise under this scheme would be equal. While it could be argued that a term with only a single descendant has not differentiated, and thus no new information is gained by the descendant, here it is important to consider that not all ontologies used are complete and may not fully describe the domains they attempt to characterise. Consequently, many such branches may exist that evidently gain in specificity, and thus require a measure that captures this.

Using this approach, the similarity of a term to itself is 1, as the most informative common ancestor of a term and itself is itself. Similarly, because the information content of the root is  $-\ln(1) = 0$ , any two terms for which the only common ancestor is root will have a similarity of 0. This definition defines a normalized range of semantic similarity for two terms.

As the estimated probability of occurrence becomes very small at the end of the ontology hierarchy, in particular for ‘deep’ ontologies like HPO, this can become impossible to calculate directly due to floating point precision. Thus, here the

logarithm product rule ( $\log(x \cdot y) = \log(x) + \log(y)$ ) was employed to enable multiplication of small fractions. The estimated probability of occurrence, information content and semantic similarity was thus computed for the three domains of the Gene Ontology, The Human Phenotype Ontology (Köhler et al. 2014), Uberpheno (Köhler et al. 2013), Uberon (Mungall et al. 2012), Disease Ontology (Kibbe et al. 2015), Foundational Model of Anatomy (Golbreich et al. 2006) and The Pathway Ontology (Petri et al. 2014), with information content of each term, and information content of terms and pairwise similarities between all terms within each ontology stored in a pre-calculated MySQL database table.

As each gene (or disease) can be described using ontology annotations, semantic similarity between genes (or diseases) can be computed by comparing their respective annotation sets. A gene/disease is rarely annotated with just a single term – thus, a measure to combine individual pairwise term similarities into a single score is needed. While three approaches are frequently used in the literature (Pesquita *et al.*, 2008) – the average, maximum or best match average of pairwise similarities. Here, the best match average approach, which averages only the highest scoring match for each pair, is used, as this provides the highest score resolution (Pesquita, Pessoa, *et al.*, 2009).

Pairwise semantic similarity approaches for gene semantic similarity can suffer from a bias arising from ‘shallow’ annotations. While a pair of terms deep within the ontology separated by the same distance will have higher semantic similarity than those closer to the root, the semantic similarity between a term and itself is always 1. Thus, two highly functionally divergent genes could contain a high-level annotation such as ‘protein binding’ and the resulting match would lead to an increase in the final pairwise score which may bias the results. In order to address this issue, information content of a term is taken into consideration for perfect matches. Thus, for gene pairwise similarities, if the best match similarity for a pair of terms is 1, then this is modified by factor  $1-M$ , which is the percentile where information content for that particular term falls within the distribution of information content scores for that ontology. Thus, high level, low information content term like

'protein binding' will not result in inflated similarity values, whereas informative terms with high information content will be affected very little:

$$Sim(a, b) = M \cdot \frac{IC_{MICA(a,b)}}{\max\{IC_a, IC_b\}}$$

#### 4.2.2.5 Candidate Disease Gene Scoring

Initially, an annotation profile of functions, processes, cellular and anatomical components, pathways and model organism phenotypes that may be relevant to the query human disease or phenotype is derived by querying the annotation and ontology tables in the assembled OVA database. This is accomplished in two ways. Firstly, using phenotype semantic similarity, all genes are selected which have been previously linked to diseases presenting similar phenotypes to the query. Secondly, given a query disease, further annotations for query are derived where possible by reasoning across ontologies, starting from phenotype terms. For example, UberPheno phenotype term '*abnormal(ly) disrupted determination of left/right symmetry*' (ZP:0000333) can be directly linked to GO Biological Process term '*determination of left/right symmetry*' (GO:0007368); following the GO ontology graph, similar/more informative terms can be inferred as related, for example '*TGF-beta receptor signalling pathway involved in determination of left/right asymmetry*' (GO:0035463). Similarly, Uberon anatomical ontology term '*heart*' (UBERON:0000948) can be linked to multiple Gene Ontology terms such as '*heart morphogenesis*' (GO:0009653), phenotype terms such as '*Cardiomegaly*' (HP:0001640) and pathway ontology terms such as '*cardiovascular system disease pathway*' (PW:0000020). Furthermore, model organism data is also leveraged this way. For example, '*abnormal snout morphology*' (MP:0000443) in UberPheno describing a mouse phenotype can be mapped to its human counterpart '*abnormality of the nose*' (HP:0000366).

For each candidate gene, ontology annotations are then compared to the disease or phenotype derived annotation profile. Using the semantic similarity measure described above, similarity scores are computed between each candidate gene and the query annotation profile of the disease for each ontology domain. For each candidate gene, where possible, model organism ortholog annotations are also queried. Finally, in order to at least partially overcome the challenge presented by incompleteness of gene annotations, for each candidate disease gene, the neighbourhood genes within the human interactome are also considered. Interacting groups of proteins are more likely to participate in the same or similar processes, and thus, if a protein lacking in quality functional annotations is known to interact with a group of proteins for which informative annotations are available, these can be extended to apply to the poorly annotated gene. Here, an interactome neighbourhood is defined as a set of genes sharing direct interactions with the gene in question. These are derived from *mentha* (Calderone *et al.*, 2013), a collection of curated physical protein-protein interactions from several primary databases. For cases where a particular gene is not covered by *mentha* database, the STRING interactome is used, which is comprised of higher coverage but lower confidence gene interaction data. A gene set enrichment approach is then used to select only annotations which are over-represented in the interactome neighbourhood in order to reduce noise and extract common functions. A Fisher's exact test is used to test for term enrichment within the interactome neighbourhood against whole interactome background. Bonferroni correction (Armstrong 2014) is applied to account for multiple testing. Terms with corrected p-values  $< 0.01$  in the interactome neighbourhood are retained for comparison. These results are pre-computed for each gene and stored in the OVA database in order to enable faster gene prioritisation.

To aggregate this information and arrive at a ranked candidate gene/variant list, multiple approaches were considered and implemented: the average of similarity scores across all domains; a weighted average; and a supervised learning approach.

The average of similarity scores simply works out the average across all domains. The weighted average approach gives more weight to features which can be considered more informative – for example, while information about cellular localization of a protein is important, a mouse model with a highly similar phenotype to the query disease is a much stronger predictor of a good candidate disease gene. Additionally, this approach can dynamically adjust the similarity score weights based on information available about the candidate gene. A low score for a gene poorly annotated in a particular domain is not always comparable to a low score derived from multiple informative annotations. Consequently, OVA implementation allows the user to customise domain weights.

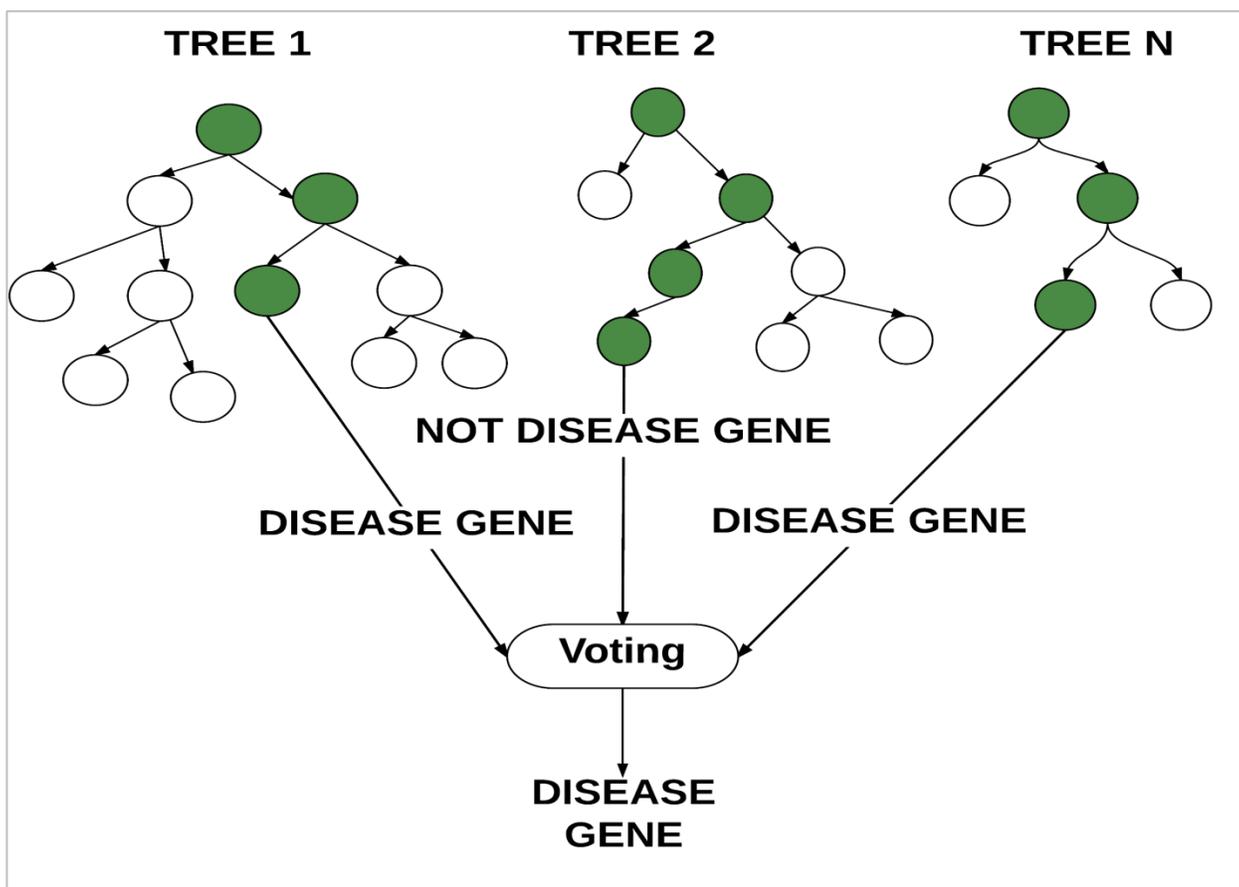
Lastly, in order to find an optimum scoring function for the multiple derived metrics, a supervised learning approach is considered. Supervised learning approaches attempt to classify data based on a learned set of features from a labelled training data set. In order to produce such a data set, all OMIM diseases were selected that have at least one known causative disease gene and split into two sets – two thirds were selected for training, while one third was retained for testing. To serve as negative (i.e. not associated with the disease) training examples, a matched set of random genes not documented by OMIM was selected. For each disease example, both the positive and negative training genes were scored based on semantic similarity profile in each category for disease. For each example, a number of other features were obtained, including the number and informativeness of annotations that support each score; proximity in the interactome to known disease genes; and disease category based on OMIM classifications. Each instance was thus labelled as either ‘disease gene’ or ‘non-disease gene’. Using this labelled data set, a Random Forest (RF) model was trained using Java package WEKA (Hall et al. 2009). RF is an ensemble learning algorithm which constructs and combines information gained from multiple decision trees and is robust to over fitting.

Given a training data set consisting of  $n$  instances of genes, let  $V_i = \{f_1 \dots f_m\}$  be a feature vector describing the  $i^{\text{th}}$  gene and  $L_i$  its class label (“disease gene” or “non-

disease gene”). Briefly, in a RF model, each decision tree is constructed using a bootstrapped sample of instances from the training data  $n$  and a randomly selected subset of features  $f$ , consisting of  $x < m$  features. In a decision tree, a set of rules describing  $L$  are learned from the training data by recursively splitting the feature space at each node until all leaf nodes contain instances from only one  $L$  class. Given an unknown instance, each tree in an RF model ‘votes’ for the likeliest class label, with the percentage of individual trees voting for a given class representing the posterior probability that the instance belongs to that particular class based on the training data (Breiman, 2001).

The optimum performance/accuracy trade off was reached with a model of 600 decision trees, each considering 6 random features. Once the optimum training parameters were obtained, the training dataset was pruned by removing misclassified instances using 10-fold cross-validation of the original model, as these likely represent outliers. The final model was then rebuilt using the optimised training parameters and optimised training data set.

The model was saved as a binary, serialised Java object file. Given a disease or phenotype profile, any given gene can then be classified using this model. While the classification is binary, this is based on the confidence score cut-off, where the confidence score is effectively a proportion of random trees in the model which have ‘voted’ in favour of a particular label. Thus, this score is used directly to provide a confidence-based ranking of genes (**Figure 23**).



**Figure 23.** Schematic representation of Random Forest algorithm. Each tree makes a decision based on information learnt from training data. The confidence score can then be derived from the proportion of trees ‘voting’ in favour of an outcome.

## 4.3 Results and Discussion

### 4.3.1 Performance Assessment

#### 4.3.1.1 Gene Semantic Similarity Measure Assessment

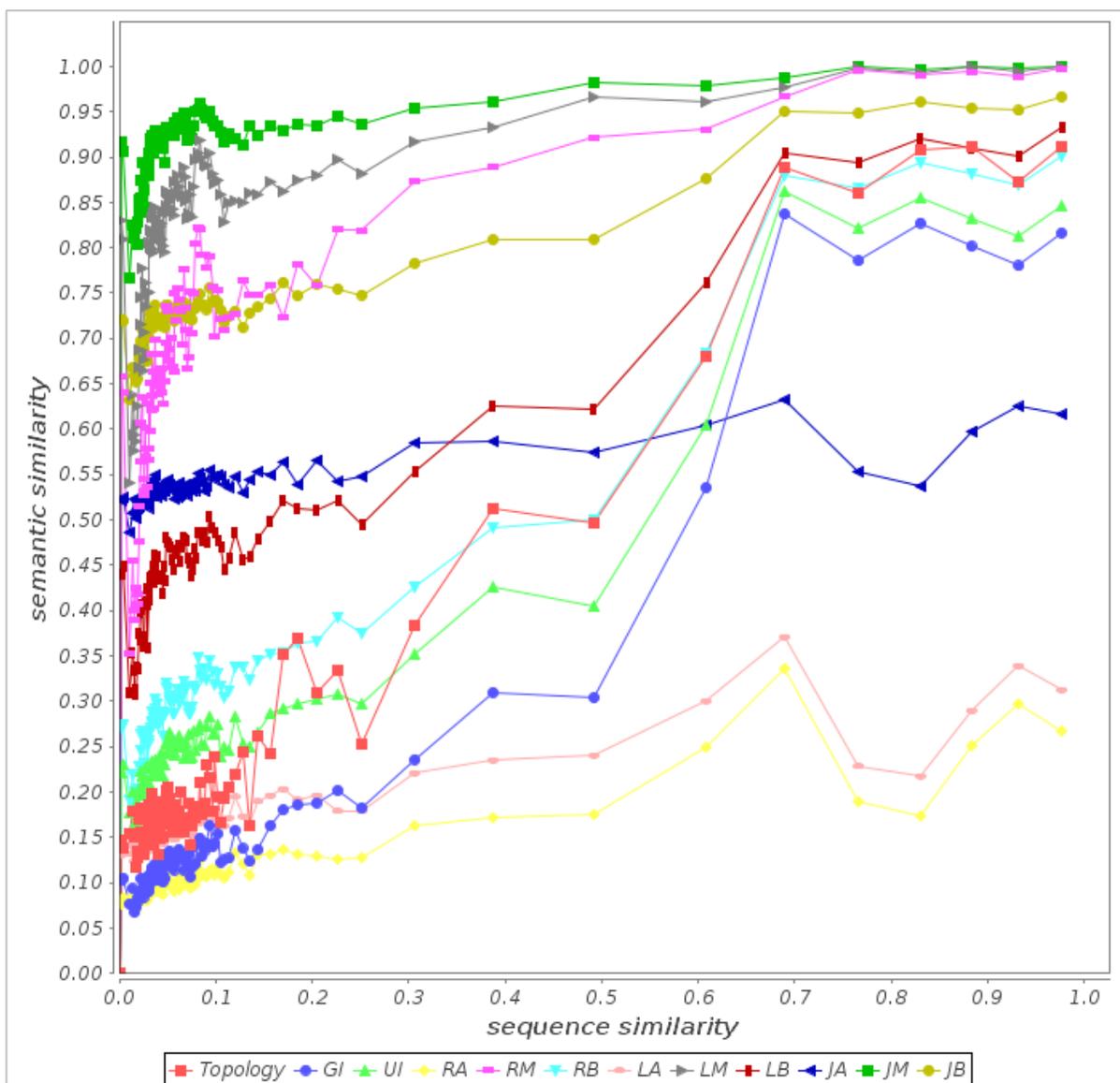
In order to ascertain whether the semantic similarity measure proposed here enables the generation of meaningful comparisons between entities such as diseases and putative causal genes, the measure was first compared to several alternative semantic similarity measures proposed in literature. Gene Ontology-based semantic similarity was computed for 12,430 protein pairs in the CESSM

dataset, a proposed standard for comparing semantic similarity measures (Pesquita et al. 2009). For each protein pair, the semantic similarity score was compared to sequence similarity. This is calculated by CESSM using relative reciprocal BLAST scores (Pesquita et al. 2009, 2008) and has been suggested to be a good general indicator of functional similarity (Joshi and Xu 2007). Furthermore, the semantic similarity measure used here was compared to 11 other frequently used measures: simGIC (Pesquita et al. 2007), simUI (Gentleman, 2005), and the average (Lord et al. 2003), maximum (Sevilla et al. 2006) and best-match average (Couto et al. 2007) combinations of the term similarities by Resnik (1995), Lin (1988) and Jiang & Conrath (1997). Correlations between semantic similarity scores and protein sequence similarity for each measure are shown in **Figure 24**.

Approaches which consider only the maximum of all pairwise ontology term similarities tend to systematically overestimate protein functional similarity; whereas the average-based approaches tend to severely underestimate functional similarity even for very similar proteins. A good semantic similarity measure, then, is one that has high resolution in such a comparison. The Jiang & Conrath average method shows the worst resolution out of all measures considered, with the semantic similarity scores produced by this method covering only 14% of the spectrum of protein sequence similarity scores (**Table 6**). The semantic similarity measure proposed here has the resolution of 91% across the range of sequence similarities. While this is slightly lower than that achieved by Lin's best match average method (93%), it is evident (**Figure 24**) that the semantic similarity measure proposed here correlates better with sequence similarity across the entire range, whereas Lin's method does not score protein pairs uniformly, with very few proteins pairs given a score of 0, and then rising sharply and scoring protein pairs with less than 10% sequence similarity as being 30-50% functionally similar, whereas the method presented here scores the similarity of these proteins far more conservatively (15-20% functional similarity for the same pairings), thus effectively achieving higher resolution across the dynamic range of data.

Measure	Resolution
Topology	0.912130203
simGIC	0.837303181
simUI	0.862813741
Resnik average	0.336733651
Resnik maximum	0.645218076
Resnik best match average	0.900413686
Lin average	0.370578176
Lin maximum	0.458924175
Lin best match average	0.932665166
Jiang & Conrath average	0.145241479
Jiang & Conrath maximum	0.232779576
Jiang & Conrath best match average	0.334555302

**Table 6.** Resolution of semantic similarity methods shown in **Figure 24**, computed using CESSM tool.



**Figure 24.** Sequence similarity vs semantic similarity for a number of different semantic similarity methods considered. **Topology** is the measure presented here, whereas other methods are denoted as follows: simGIC (**GI**), simUI (**IU**), Resnik average (**RA**), Resnik maximum (**RM**), Resnik best match average (**RB**), Lin average (**LA**), Lin maximum (**LM**), Lin best match average (**LB**), Jiang & Conrath average (**JA**), Jiang & Conrath maximum (**JM**) and Jiang & Conrath best match average (**JB**).

#### ***4.3.1.2 Testing Datasets for the Assessment of Prioritisation Accuracy***

Multiple datasets were employed to benchmark the capabilities of OVA. Initially, all OMIM and Orphanet disease entries with a known molecular basis were selected to create a dataset comprised of 1340 disease/gene combinations for which two or more known causative genes have been attributed and 2964 disease/gene pairs with only one known causative gene. As described in section 4.2.2.5, two thirds of these were used to train the random forest model, while the remaining disease/gene pairs form testing **Datasets 3** and **4**, as described below.

**Dataset 3** consists of all OMIM or Orphanet disease entries not used for training with at least two known causative genes attributed. This dataset comprises of 442 disease-gene sets and aims to simulate use cases where a novel disease gene causes a disease with a previously described genetic basis.

**Dataset 4** consists of all OMIM or Orphanet disease entries not used for training where only one known causative gene is known. This dataset is comprised of 978 disease-gene sets and aims to simulate use cases where a novel disease gene causes a disease with no previously known genetic basis.

All VCF files used for testing OVA were generated by simulating the presence of a single known deleterious variant by inserting it into VCF files obtained from WES from healthy individuals.

**Dataset 5** consists of 150 VCF files, each containing a known deleterious variant from ClinVar (Landrum et al. 2014) database. Deleterious variants were selected on the following criteria:

1. Annotated as “Pathogenic”
2. An insertion, deletion or single nucleotide substitution
3. Annotated to an OMIM disease with at least one other known disease gene

**Dataset 6** consists of 20 VCF files, each containing a deleterious variant that has been published since the beginning of January 2015. These variants have been curated from peer-reviewed articles and are summarized in **Table 7**.

Gene	Publication
<b>AFF4</b>	Germline gain-of-function mutations in AFF4 cause a developmental syndrome functionally linking the super elongation complex and cohesin, Krantz ID et al, 2015
<b>CACNA1B</b>	CACNA1B mutation is linked to unique myoclonus-dystonia syndrome, Tijssen MA et al, 2015
<b>CEP120</b>	A founder CEP120 mutation in Jeune asphyxiating thoracic dystrophy expands the role of centriolar proteins in skeletal ciliopathies. Hum. Molec. Genet. 24: 1410-1419, 2015
<b>CHCHD10</b>	Mutation in the novel nuclear-encoded mitochondrial protein CHCHD10 in a family with autosomal dominant mitochondrial myopathy Neurogenetics, Ajroud-Driss et al, 2015
<b>COL17A1</b>	Mutations in Collagen, Type XVII, Alpha 1 (COL17A1) Cause Epithelial Recurrent Erosion Dystrophy (ERED), I. Golovleva et al, 2015
<b>COQ4</b>	COQ4 Mutations Cause a Broad Spectrum of Mitochondrial Disorders Associated with CoQ10 Deficiency, Calvo et al, 2015, American Journal Of Human Genetics
<b>DCDC2</b>	DCDC2 mutations cause a renal-hepatic ciliopathy by disrupting Wnt signaling. Am. J. Hum. Genet. 96: 81-92, 2015 Schueler et al
<b>DDX58</b>	Mutations in DDX58, which encodes RIG-I, cause atypical Singleton-Merten syndrome. Am. J. Hum. Genet. 96: 266-274, 2015
<b>DTNA</b>	Identification of two novel mutations in FAM136A and DTNA genes in autosomal-dominant familial Meniere's disease, Lopez-Escamez JA et al, 2015
<b>ETV6</b>	Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. Nature Genet. 47: 180-185, 2015 Zhang et

	al
<b>KCNA2</b>	De novo loss- or gain-of-function mutations in KCNA2 cause epileptic encephalopathy, Lemke JR et al, 2015
<b>KCNC1</b>	A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. Nature Genet. 47: 39-46, 2015 Muona et al
<b>NALCN</b>	De novo mutations in NALCN cause a syndrome characterized by congenital contractures of the limbs and face, hypotonia, and developmental delay, Chong et al, 2015, American Journal of Human Genetics
<b>PTRH2</b>	Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families, Alazami et al, 2015, Cell Reports
<b>SEMA3D</b>	Disruption of the SEMA3D Gene in a Patient with Congenital Heart Defects, Le Caignec C. et al, 2015
<b>SLC9A1</b>	Mutation of SLC9A1, encoding the major Na <sup>+</sup> /H <sup>+</sup> exchanger, causes ataxia-deafness Lichtenstein-Knorr syndrome. Hum. Molec. Genet. 24: 463-470, 2015
<b>SNRPB</b>	Mutations in SNRPB, Encoding Components of the Core Splicing Machinery, Cause Cerebro-Costo-Mandibular Syndrome, Cormier-Daire V et al, 2015
<b>USP8</b>	Mutations in the deubiquitinase gene USP8 cause Cushing's disease, Komada M et al, 2015
<b>WWOX</b>	WWOX-related encephalopathies: delineation of the phenotypical spectrum and emerging genotype-phenotype correlation J. Med. Genet. 52: 61-70, 2015 Mignot et al
<b>PNKP</b>	Mutations in PNKP cause recessive ataxia with oculomotor apraxia type 4, Bras et al, 2015, American Journal of Human Genetics

**Table 7.** Summary of novel disease gene variants comprising OVA benchmarking Dataset 6.

#### ***4.3.1.2 Candidate disease gene prioritisation accuracy***

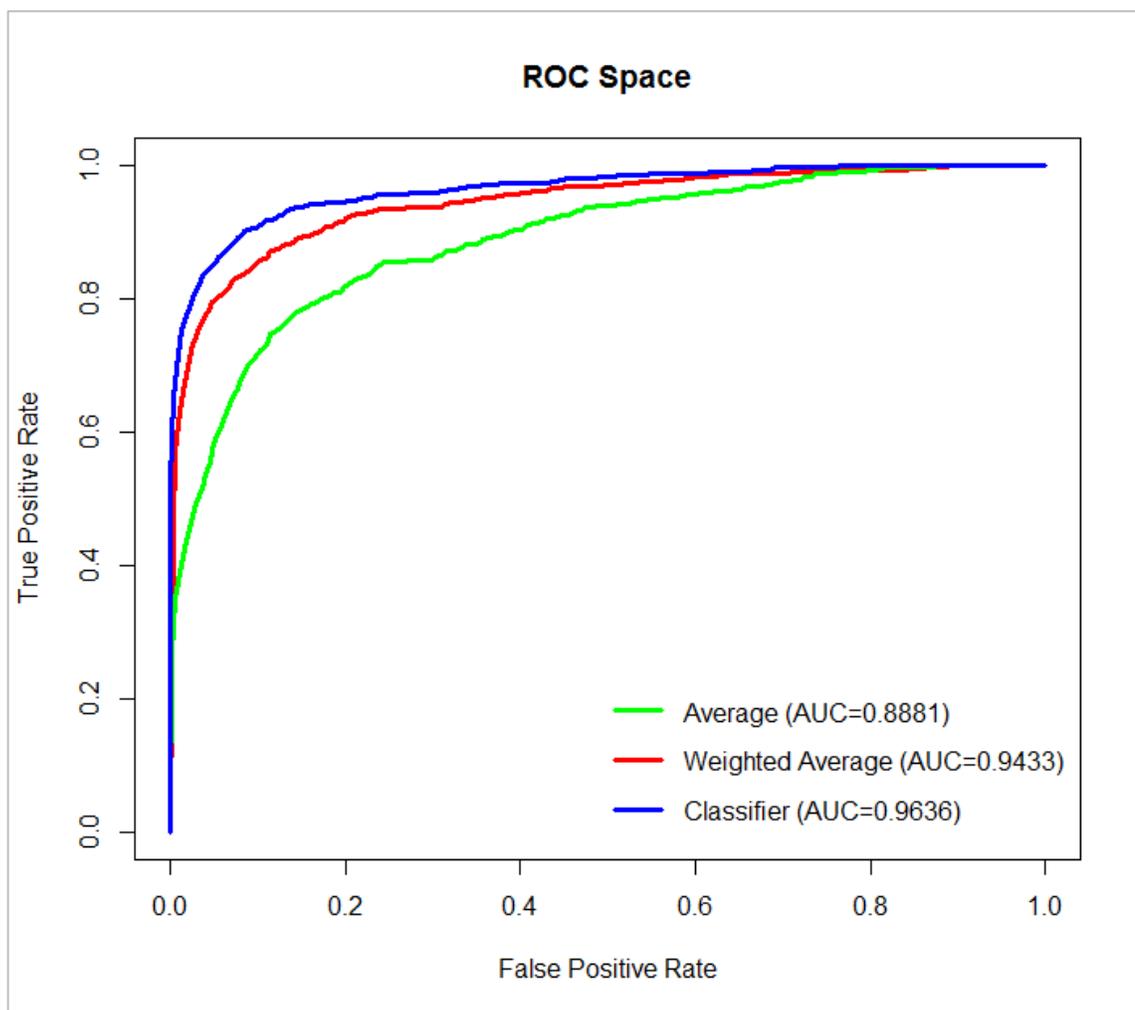
OVA's prioritisation accuracy was assessed using multiple datasets, assembled as described in the previous section. A leave-one-out cross validation type of approach was used, where for each known disease gene in a dataset, all associations between that gene and disease/human phenotype were temporality removed from OVA database, simulating a previously unknown disease gene. While this is not a perfect approach, as knowledge tends to be somewhat circular (i.e. functional annotations often result from further work validating a gene-phenotype association), there are no alternative methods for performance assessment that encompass such a large range of disease-gene associations.

**Dataset 3** and **Dataset 4** were used to assess OVA performance for simulated disease cases both with and without a previous known molecular basis. Each test gene was ranked with respect to disease together with 200 randomly selected genes. The selection of random genes was limited to the pool of all human genes which have at least minimal Gene Ontology annotations (2 or more terms) in order to avoid any potential bias that could lead to overestimation of performance, as known disease genes are rarely entirely unannotated. Three methods for obtaining the final scores were assessed, as discussed in the **Methods** section – average, a weighted average approach and a Random Forest classifier model approach.

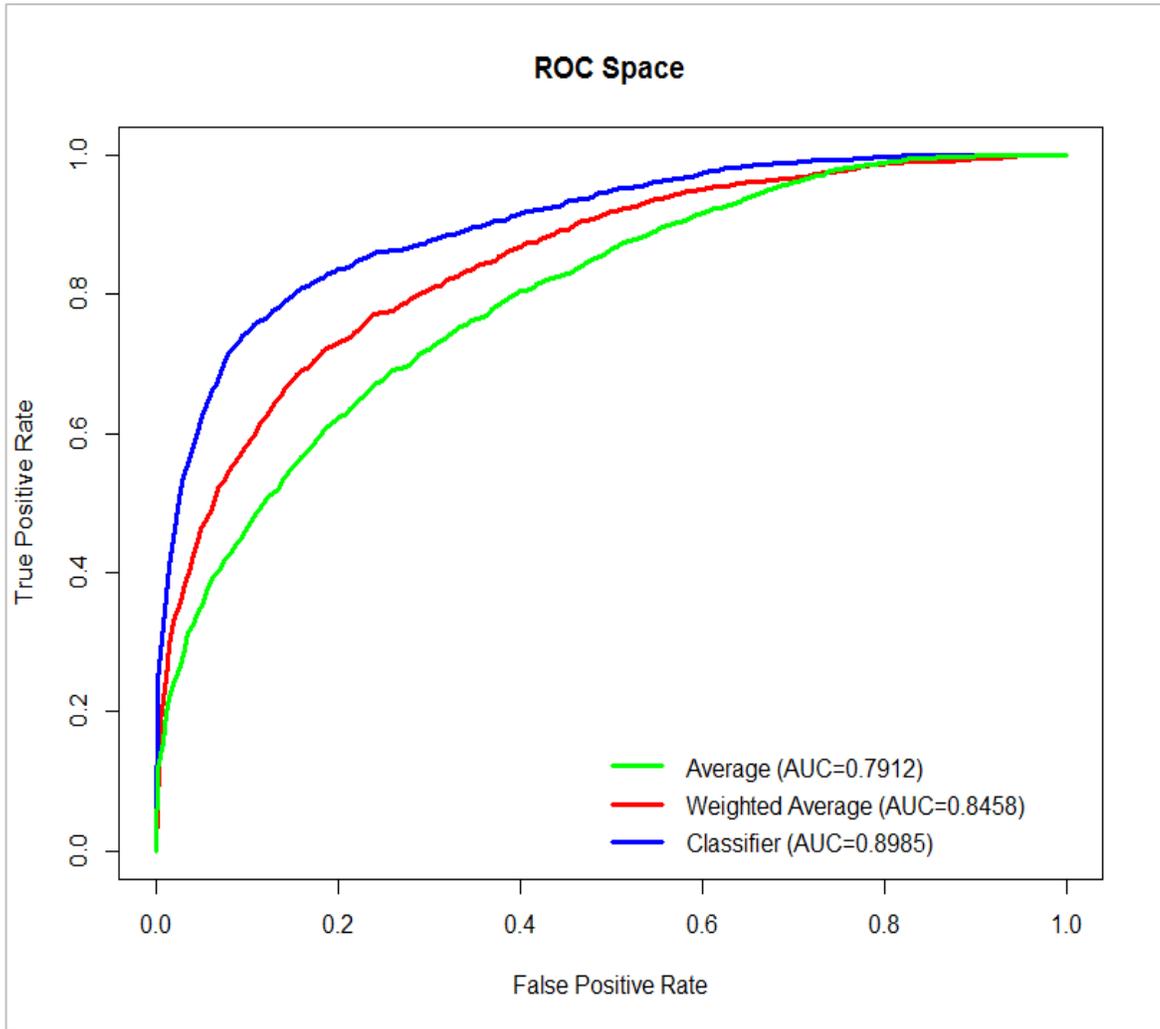
In each case, the datasets were prioritized using OVA and the ranking results were collated and analysed in R using the package 'ROCR' (Sing et al. 2005). **Figure 25** shows the ROC curves obtained using **Dataset 3**, while **Figure 26** shows the ROC curves obtained using **Dataset 4**. There is a notable difference in performance between the three methods that is consistent across the two datasets. The average score method performed the worst, while the Random Forest classifier approach was able to prioritize the test cases with the best accuracy. It is worth re-iterating that the test genes in either dataset did not form part of the training data for the model, thus the increase in classification accuracy is unlikely to be due to biases from improper controls.

Additionally, as expected, there is a notable difference in prioritization accuracy between **Dataset 3** and **Dataset 4**. Performance of OVA is greatly enhanced (AUC up to 0.9636) where knowledge about previously identified molecular causes of the disease is available. However, extending the search to diseases which cause similar phenotypes allows the prioritization of cases where little is known about the molecular causes of a disease that is still robust (AUC up to 0.8985).

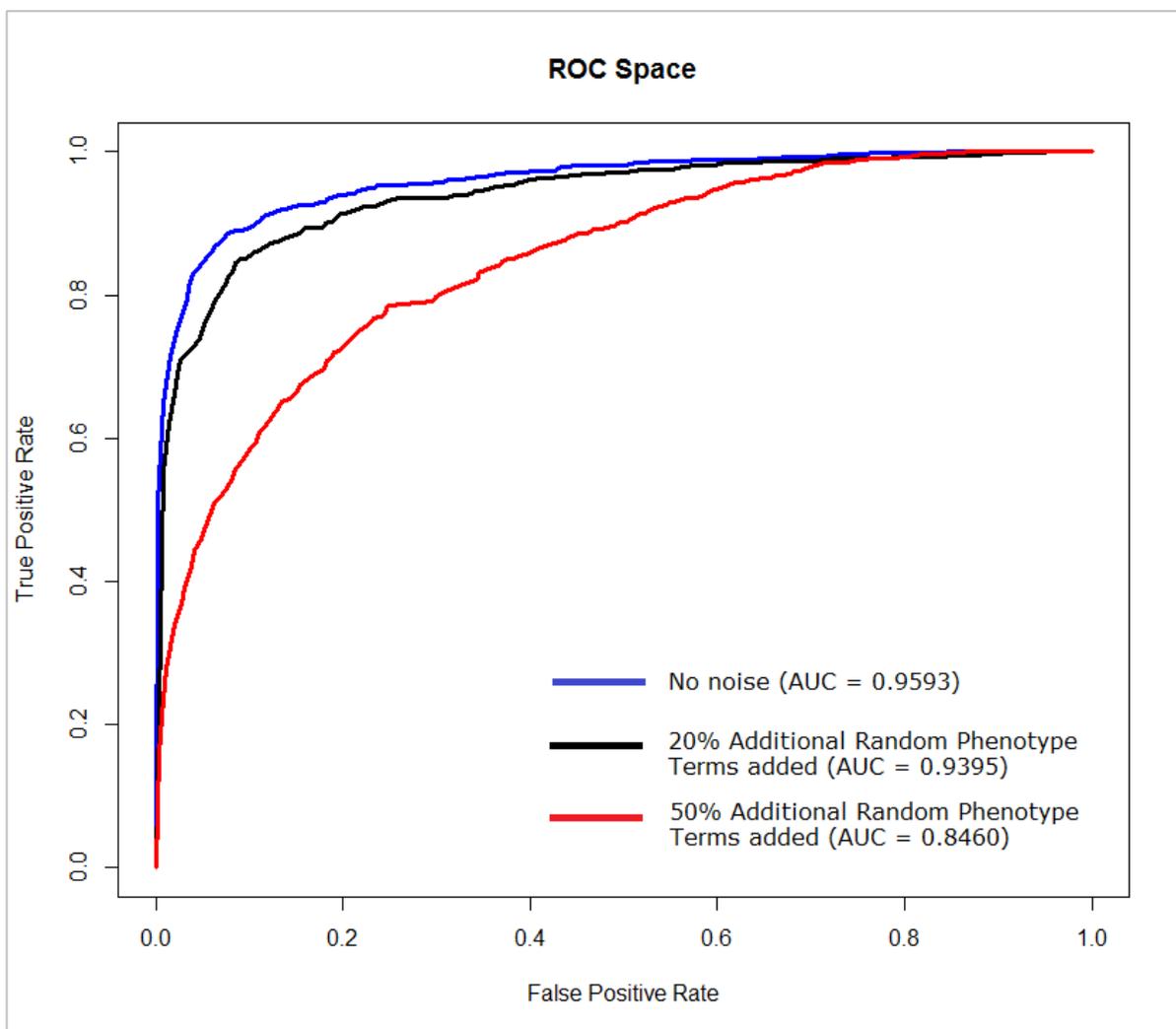
The key parameter in OVA is a phenotype selection step. In the previous step, the OVA algorithm was provided perfectly accurate disease descriptions for each test instance, which might not accurately reflect real world use of the application. In order to ascertain how sensitive the algorithm is to various amounts of input noise, **Dataset 3** was used to simulate cases where the phenotype is inaccurately or inadequately described. To simulate such cases, each input phenotype term describing the query disease in **Dataset 3** was supplemented with additional, randomly selected phenotype terms to simulate inaccurate descriptions; or, some phenotype terms were removed in order to simulate an incomplete phenotype description. As expected, introducing any type of noise to the phenotype input negatively impacts gene prioritization accuracy (**Figure 27** and **28**). Unsurprisingly, higher levels of additional noise negatively impact the method's performance more. Introducing additional irrelevant query phenotypes, however, had less of an impact on the accuracy of the results than excluding relevant terms.



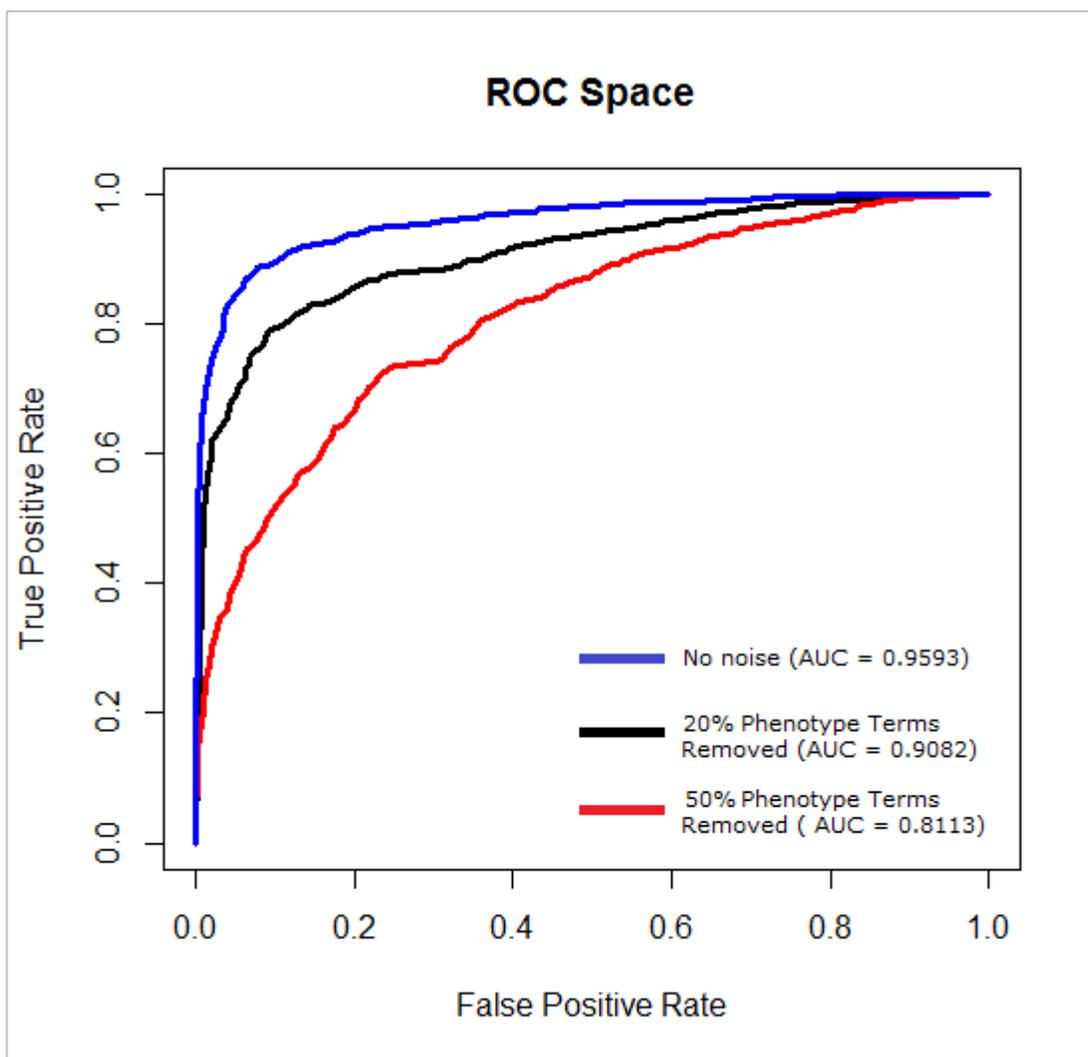
**Figure 25.** ROC curves obtained from prioritizing disease genes in **Dataset 3** using leave-one-out cross validation. Three different approaches of score aggregation are compared – average, weighted average and Random Forest classifier. Area under ROC curve of 0.5 ( $x=y$ ) indicates no better than random performance; area under ROC curve of 1 indicates 100% accuracy.



**Figure 26.** ROC curves obtained from prioritizing disease genes in **Dataset 4** using leave-one-out cross validation. Three different approaches of score aggregation are compared – average, weighted average and Random Forest classifier. Area under ROC curve of 0.5 ( $x=y$ ) indicates no better than random performance; area under ROC curve of 1 indicates 100% accuracy.



**Figure 27.** Dataset 3 was used to assess how well OVA tolerates input noise. For each test disease gene, additional phenotype search terms were randomly generated and added to the pool of accurate phenotype descriptions to simulate inaccurate user input. The original phenotype descriptions for each disease were supplemented with 20% (**black**) and 50% (**red**) noisy input (rounding up, minimum 1 extra phenotype added). The figure compares the resulting ROC curves.



**Figure 28.** Dataset 3 was used to assess how well OVA tolerates input noise. For each test disease gene, some phenotype terms were randomly removed from the disease description to simulate inadequate user input. 20% (**black**) and 50% (**red**) of total phenotype terms used to describe a disease were removed from the input (rounding up, minimum 1 phenotype removed). The figure compares the resulting ROC curves.

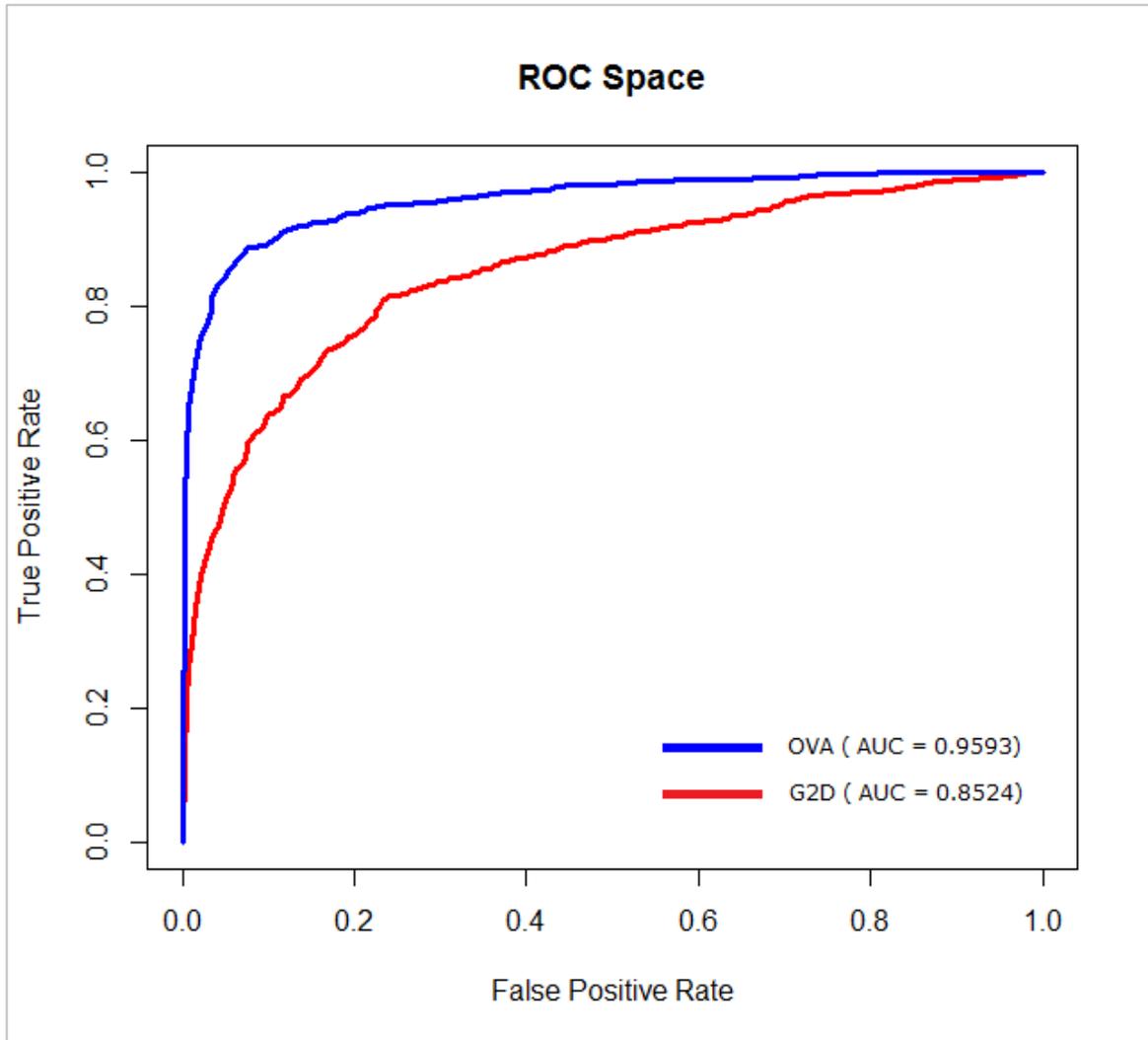
#### ***4.3.1.3 Comparison with other gene prioritisation tools***

It can be useful to compare the performance of a tool with others currently available in literature. Such comparisons, however, are often hard to make. The majority of available tools are implemented as web applications and do not expose an API for large scale queries that would be required to make such a comparison. Additionally, even in cases where large scale queries can be automated, it can be impossible to use a leave-one-out cross-validation approach without direct access to underlying databases.

Here, in order to assess how well the algorithm implemented in OVA compares with other candidate gene prioritization tools, a comparison was made with Genes-2-Diseases (G2D) tool (Perez-Iratxeta *et al.*, 2007). G2D can prioritize candidates using data from GO annotations, sequence similarity, MeSH terms and STRING protein-protein interactions. Additionally, G2D requires known disease genes as input 'seeds', rather than phenotypes; this type of approach, while not directly comparable to OVA, enables a comparison of both tools using a cross-validation, as the test disease gene can be withheld from input 'seed' gene list. Furthermore, the G2D tool is a good choice for comparison due to parallels in data sources also used by OVA, allowing the comparison to be put in context of the capabilities of the methodology rather than the underlying data types.

The G2D web application uses a simple submission form, allowing the application to be automated via HTTP post requests. Custom code was written to facilitate a large number of such requests and to retrieve and parse the results. As G2D requires a genomic region as input, using **Dataset 3**, for each test case, a 100MB genomic region containing the known disease gene was provided as genomic region input, while other genes associated with the disease were provided as 'seeds'. OVA was used to prioritize the same queries after the removal of the test gene/disease associations from the OVA database in each test instance. While OVA significantly outperformed G2D (AUC 0.9593 vs 0.8524, **Figure 29**), some of

the differences in accuracy could be accounted for by somewhat outdated data used by G2D (as of this writing, last updated in 2010).



**Figure 29.** Performance comparison between OVA and G2D, another gene prioritisation tool using **Dataset 3**. Area under ROC curve of 0.5 ( $x=y$ ) indicates no better than random performance; area under ROC curve of 1 indicates 100% accuracy.

#### ***4.3.1.4 Comparison with other variant prioritisation tools***

Gene prioritization approaches can be much more successful if the search space could be effectively reduced, for example by identifying likely candidate disease regions through techniques such as autozygosity mapping, or removal of all common and benign variation from a patient VCF file. Thus, in order to enable a more streamlined approach, an extensive variant filtering step for VCF files was implemented in OVA application.

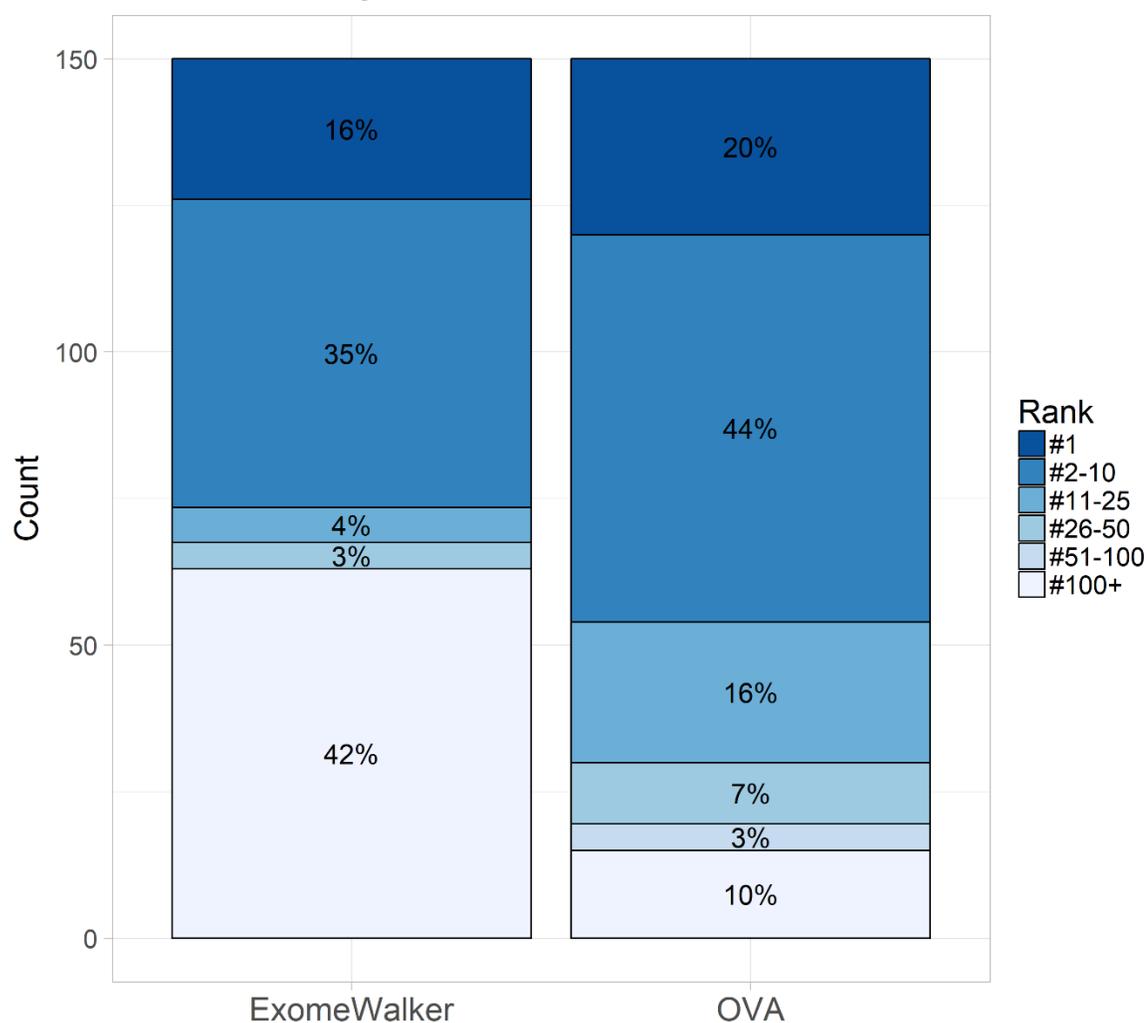
This includes support for multi-sample VCF files with multiple affected/unaffected patients that may be available to researchers from parent-child trios or familial studies. For the purposes of this assessment, however, single patient VCF files in **Dataset 5** were used, comprising 150 pathogenic variants found in ClinVar database (Landrum et al. 2014) that were inserted into VCF files from healthy (unaffected by severe pediatric disease) human WES data, which were obtained from SRA (Leinonen et al. 2011b).

As before, a leave-one-out cross-validation type of approach was employed, except each test VCF file was passed through the OVA variant filter. Variants were filtered based on inheritance mode (e.g. homozygous or compound heterozygous for autosomal recessive), synonymous substitutions, intronic variants and small in-frame deletions and insertions were removed in each case. Splice-site variants were retained. The remaining variants were then prioritized using OVA's RF classifier mode.

In order to compare the accuracy of OVA to other candidate variant prioritization methods, these results were compared to those obtained from prioritisation of **Dataset 5** with ExomeWalker (Smedley *et al.*, 2014), which uses a current state-of-the-art algorithm for network-based gene prioritization coupled with a variant scoring approach. As ExomeWalker is provided as a stand-alone Java application, large scale comparisons between these tools are feasible.

**Figure 30 and Table 8** show the disease gene rank distributions obtained by both tools. Out of 150 VCF files, in 20% of the cases OVA ranked the true disease gene first, with ExomeWalker performing similarly at 16%. A total of 64% of instances were ranked in the top 10 by OVA, compared to 51% by ExomeWalker. While ExomeWalker scored 51% of all cases very accurately (Top 10), 42% ranked very poorly (outside of Top 100), whereas only 10% test cases were ranked outside the top 100 by OVA.

Performance comparison between ExomeWalker and OVA using 150 exomes



**Figure 30.** Performance comparison between OVA and ExomeWalker using 150 exomes. ExomeWalker identified the correct disease variant in 16% (24 out of 150)

cases, whereas OVA prioritized the correct gene in 20% (30 out of 150) cases. In 64% of all cases, OVA placed the correct disease gene in the top 10, compared to 51% by ExomeWalker. ExomeWalker ranked the correct variant outside the top100 in 42% of the cases, compared to only 10% in OVA rankings.

Prioritisation Result	ExomeWalker	Gene Ontology Analysis Tool
<b>Top 1</b>	16.46%	19.62%
<b>Top 10</b>	51.90%	60.76%
<b>Top 20</b>	55.06%	81.01%
<b>Top 50</b>	58.86%	89.87%
<b>Top 100</b>	58.86%	91.77%

**Table 8.** Variant prioritization comparison between OVA and ExomeWalker, summarized in **Figure 30**.

#### ***4.3.1.5 Prioritisation of novel disease gene variants***

Finally, in order to verify that these results are consistent with real cases of novel disease gene discovery, OVA was used to prioritize 20 recently published novel disease gene mutations not present in the OVA database. As before, disease causing variants were inserted into VCF files obtained from healthy individuals (**Dataset 6**) and prioritised using OVA. The results of the analysis are summarised in **Table 9**. The ranking is overall somewhat poorer than that observed using the other datasets, with 14 out of 20 genes ranked in the top 25. This may be in part due to the small size of the dataset; or, due to overestimated accuracy when testing using other datasets, as despite the best attempts to remove associations between test genes and diseases, some indirect circularity of knowledge nevertheless is likely to remain.

Ranking	Count	Percentage	Genes
<b>1st</b>	3	15%	CACNA1B, COQ4, WWOX
<b>Top 10</b>	8	40%	CACNA1B, COQ4, WWOX, KCNA2, NALCN, SEMA3D, SLC9A1, USP8
<b>Top 25</b>	14	70%	CACNA1B, COQ4, WWOX, KCNA2, NALCN, SEMA3D, SLC9A1, USP8, AFF4, DCDC2, DTNA, ETV6, KCNC1, PNKP
<b>Not in Top 25</b>	6	30%	CHCHD10, COL17A1, DDX58, PTRH2, SNRPB, CEP120

**Table 9.** The prioritization results of **Dataset 6**, consisting of 20 novel disease gene VCF.

#### **4.3.2 OVA Application**

All the methods described here are implemented and made available as a web-based application. The landing page allows the user to configure prioritization parameters and select the scoring method (from RF model, average and weighted average, with configurable weights) (**Figure 31A**).

Phenotype/disease selection screen requires the user to provide a set of input phenotype terms or disease(s) that describe their query. All queries are facilitated by a responsive, user-friendly auto-complete search (**Figure 31B**), which draws data directly from the OVA database to ensure only phenotypes/diseases present in the database are selected. Each query term is coloured by specificity, with green indicating more informative terms and red indicating very broad terms which should be avoided where possible.

Candidate genes can be supplied as a list (**Figure 31C**), a genomic region (**Figure 31D**) or a VCF file (**Figure 31E**). Gene and Pathway Ontology annotations identified as relevant to selected phenotype are displayed under the 'Review' screen (**Figure 31F**). Often, if the user has extensive knowledge of the disease

under study, it can be useful to prune identified annotations to retain only the most relevant ones. Additionally, a user may choose to remove all terms which are very broad descriptions (e.g. 'protein binding'), while retaining only most specific functions and pathways. For this purpose, a specificity filter is made available. Furthermore, additional functional and pathway annotations which were not picked up by the OVA algorithm can be added via this page (**Figure 31F**).

Prioritization results are provided as a sortable, interactive table (**Figure 31G**), which can be downloaded as a tab-delimited text file. If VCF files were provided as input, variant column in each row allows the user to view all filtered variants retained for each gene, as well as their effects on the protein sequence. Furthermore, gene functional annotations can also be viewed for each gene. The annotations are presented as a word cloud, with the size of the annotation scaling with how relevant it was found to query phenotype.

---

**Figure 30. OVA application user interface. A.** Start screen. **B.** Phenotype/Disease input screen. **C.** Gene list input screen. **D.** Genomic region selection screen. **E.** VCF file upload screen. **F-G.** Annotation review screen. This allows the user to query the back-end database and provide additional pathway and functional terms for prioritisation that might not have been picked up by the OVA algorithm (or, were unwantably removed in the previous screen). **H.** Results screen. Prioritisation results are displayed in an interactive, sortable table. Individual annotations and sequence variants can be viewed for each gene.

## OVA: Ontology Variant Analysis Tool

Start Website DEMO

Configure Parameters

Start Prioritisation

### About

Current high-throughput sequencing methods used for disease gene discovery or clinical diagnostics can generate very large volumes of data. While the extraction of non-synonymous, potentially deleterious variants can be easily automated, this often results in the identification of thousands of candidate disease genes. Since the experimental verification of an individual gene can be both difficult and time consuming, some method to prioritise the order in which such verification is sought is often employed.

Ontology Variant Analysis Tool is an online variant filtering and prioritisation application. Ontology Variant Analysis Tool can filter your VCF files on a wide array of criteria. Remaining genes are prioritised based on their functional and phenotypic profile similarity to a user supplied phenotype.

### Usage

Ontology Variant Analysis Tool is a single page application. Please use the buttons provided on the page to navigate instead of using your browser's 'Back' and 'Forward' buttons. Ontology Variant Analysis Tool requires JavaScript to function. If you are unable to proceed past this page, check your browser has JavaScript enabled. Finally, the website is best viewed on modern browsers as we cannot guarantee correct functionality on legacy browsers.

If this is your first time using our website, we recommend viewing a demonstration which will walk you through a common usage scenario. This can be accessed by clicking 'Start Website DEMO' button at the top of this page (temporarily unavailable due to technical issues). Alternatively, begin using Ontology Variant Analysis Tool by clicking 'Start Prioritisation' at the top of the page.

### Contact

If you have any questions, comments or suggestions, please send an email to :

umaan@leeds.ac.uk

## Phenotype Selection

Search by:  Phenotype Term  OMIM Disease

- Text (e.g. Hearing impairment)
- HPO Accession (e.g. HPO:0000612)

Supply the affected phenotype. This can be as OMIM /ORPHANET disease(s), or terms describing the phenotype. To begin, type in your query and select from the results. Phenotype terms are coloured by specificity, with red indicating very broad terms and green very specific. Aim to include more specific terms in your query.

Back

Proceed

Search Results				
Phenotype	Accession	Spec	Add	
Hearing abnormality	HPO:0000364		▲	
Hearing impairment	HPO:0000365		▲	
Mixed hearing impairment	HPO:0000410		▲	
Progressive sensorineural hearing impairment	HPO:0000408		▲	
Sensorineural hearing impairment	HPO:0000407		▲	

Current Selection	
Name	Id
No phenotypes currently selected.	

Clear All

# C



## Gene Selection

- Input as list (e.g. TMEM137, ENSG00000207608, NM\_001011720, 83734)
- Input genomic region(s)
- VCF file upload

OLF14, IL2, SEPP1

Supply candidate genes for prioritisation. Genomic region(s) may be specified, or a gene list supplied (most common delimiter characters accepted). Only valid genes will be prioritised. Ambiguous genes will also be discarded unless one of the identifiers presented is selected.

Submit

Back

Proceed

# D

## Gene Selection

- Input as list (e.g. TMEM137, ENSG00000207608, NM\_001011720, 83734)
- Input genomic region(s)
- VCF file upload

Chromosome:  Start:  End:

Search for protein coding genes only

Add to selection

Supply candidate genes for prioritisation. Genomic region(s) may be specified, or a gene list supplied (most common delimiter characters accepted). Only valid genes will be prioritised. Ambiguous genes will also be discarded unless one of the identifiers presented is selected.

Current Genomic Region Selection				
Clear All	Chromosome	Start	End	Genes in Region
X	1	5000000	10000000	110

Back Proceed

## E (i)

### Choose File:

No file chosen

### Filter Variants By Quality:

Minimum Quality Score

### Genotype options

- Heterozygous
- Homozygous
- Compound Heterozygous

### Exclude the following variant classes:

- Variants in non-protein coding genes
- Known variants ( with an rs identifier)
- Synonymous substitutions
- Variants in untranslated regions
- Intronic
- In Frame Insertions/Deletions ( >>[Click here to define maximum length](#)<<)
- Splice Sites (>>[Click here to define](#) <<)

## E (ii)

### Born in Bradford :

Filter by allele frequency in Born in Bradford exome sequencing data

Maximum allowed heterozygous

Maximum allowed homozygous

Variants in selected intervals only (e.g autozygous regions)

Chromosome:  Start:  End:

#### Current Genomic Region Selection

Chromosome	Start	End
------------	-------	-----

# F

[Begin](#) ✓ [Phenotype Selection](#) ✓ [Gene Input](#) ✓ [Review Annotations](#) ✓ [Results](#) ✗

## Review and Submit

This page shows pathway and functional ontology terms that are associated with the supplied phenotype. You can add or remove terms to further refine your query. Similarly to phenotype terms, Gene Ontology terms are coloured by specificity. We suggest removing very broad terms which appear in red from your query. Expand (+) to view all terms within category.

[Back](#) [Proceed](#)

[Review Terms](#) [Search For Additional Terms](#)

All Annotations						
Accession	Term	Spec	Domain	Remove		
+ reproduction				X		
- nuclear division				X		
GO:0045835	negative regulation of meiosis		Biological Process	X		
GO:0007143	female meiosis		Biological Process	X		
+ metabolic process				X		
+ cell growth				X		
+ immune system development				X		

View terms:

- Pathways
- All GO Terms
- Biological Process Domain Terms
- Cellular Component Domain Terms
- Molecular Function Domain Terms

Remove terms by specificity:

0 out of total 225 terms are selected for removal.

[Remove](#)

Term Categories:

[Expand All](#) [Collapse All](#)

# G



## Review and Submit

This page shows pathway and functional ontology terms that are associated with the supplied phenotype. You can add or remove terms to further refine your query. Similarly to phenotype terms, Gene Ontology terms are coloured by specificity. We suggest removing very broad terms which appear in red from your query. Expand (+) to view all terms within category.

[Back](#) [Proceed](#)

[Review Terms](#) [Search For Additional Terms](#)

**Search for more terms :**  **Search by:**

- All Gene Ontology Domains
- Biological Process
- Cellular Component
- Molecular Function
- Pathway Terms

Accession	Term	Spec	Domain	Add
GO:0001947	heart looping		Biological Process	▼
GO:0001985	negative regulation of heart rate involved in baroreceptor response to increased systemic arterial blood pressure		Biological Process	▼
GO:0001986	negative regulation of the force of heart contraction involved in baroreceptor response to increased systemic arterial blood pressure		Biological Process	▼

## Results

Prioritisation results are displayed below. Clicking on a column header will sort the results table. Some columns contain external links - for example, clicking on Ensembl ID will take you to Ensembl browser page for that gene. Annotations for each gene can be viewed by clicking 'View Annotations' button. Text size indicates how well a term matches query. Biological process domain terms are in blue, molecular function terms in green and cellular component terms are in red.

Top 200 results are displayed below. Full results can be downloaded as a [textfile here](#).

Select columns to view :

Rank     Name     Description     Ensembl ID     Related Diseases     Summary     BP Score, human     CC Score, human     MF Score, human  
 BP Score, mouse     CC Score, mouse     MF Score, mouse     Mouse Phenotype Score     Zebrafish Phenotype Score     Pathway Score     crossLinkScore     Human Phenotype Score     Annotations

Overall Rank	Name	Gene Summary	BP Score, Human	CC Score, Human	MF Score, Human	View GO Annotations	View Variants
166	<a href="#">CTD-2140B24.4</a>	Unavailable	0.5321	0.2645	0.2192	<input type="button" value="View"/>	<input type="button" value="View"/>
29	<a href="#">CENPU</a>	The centromere is a specialized chromatin domain, present throughout the cell cycle, that acts as a platform on which the transient assembly of the kinetochore occurs during mitosis. All active centromeres are characterized by the presence of long arrays of nucleosomes in which CENPA (MIM 117139) replaces histone H3 (see MIM 601128), MLF1IP, or CENPU, is an additional factor required for centromere assembly (Foltz et al., 2006 [PubMed 16622419]).[supplied by OMIM, Mar 2008]	0.3573	0.7655	0.6059	<input type="button" value="View"/>	<input type="button" value="View"/>

### **4.3.3 Discussion**

The development of open biomedical ontologies has exploded in the last decade and alongside it the coverage and accuracy of annotations. The work presented here takes advantage of this rich resource to bring together ontologies from across multiple domains to produce OVA, a knowledge-based gene and variant prioritization tool. OVA utilizes human and model organism phenotypes, functional annotations, curated pathways, cellular localizations and anatomical terms to find genes most relevant to a query phenotype using semantic similarity.

While gene similarity has been classically compared using sequence similarity (an evolutionary measure), the strength of semantic similarity is that the comparison is driven by the meaning of the descriptions pertaining to each entity. For example, simple lexical comparison of the words '*foal*' and '*horse*' would classify them as unrelated. Similarly, sequence-based comparisons will tell us nothing about the similarity of two genes which differ in sequence greatly, but perform key functions in the same biological process or pathway.

OVA exploits The Human Phenotype Ontology and Uberpheno structure and annotations to facilitate comparisons between human diseases and animal models of human diseases. Terms pertaining to model organism phenotypes (e.g. '*Abnormal snout morphology*' (MP:0000443)) are bridged to human phenotype terms (e.g. '*Abnormality of the nose*' (HP:0000366)) by using Uberpheno, allowing the quantification of similarities between them from the ontology graph. As numerous large-scale model organism phenotyping efforts are currently under way, such as those undertaken by The International Mouse Phenotype Consortium (Skarnes *et al.*, 2011), utilizing model organism phenotype data in a generalized gene prioritization approach is becoming more viable as coverage increases.

Gene ontology annotations have proved to be one of the most frequently utilized sources of gene functional knowledge in computational biology, with numerous applications taking advantage of this structured and highly curated resource. GO has also been heavily utilized as a data source for various candidate disease gene

prioritization applications. However, while candidate prioritization methods using Gene Ontology semantic similarity measures have generally been demonstrated to be effective, there are a number of drawbacks that this type of methodology suffers from that can detract from usability, accuracy and utility.

The majority of tools catalogued by the Gene Prioritization Portal (Bornigen et al., 2012) require the user to supply ‘seed’ genes – genes already known to be associated with the disease – and score candidates based on similarity to these. This is a major limitation of this approach, as in the case of rare or novel phenotypes, any prioritization based on similarity to known disease genes is impossible. Furthermore, the quality of available annotations of the supplied genes largely determines the success of this type of approach, while also allowing for little functional heterogeneity among disease genes.

Here, this approach is supplemented by building links across multiple ontologies. This allows enhancing the functional profile against which all candidate genes are scored by reasoning directly from a disease phenotype, as well as known genes. Consequently, this approach eliminates the requirement for the user to supply known ‘seed’ genes and reduces the reliance on quality seed gene annotations.

One of the major hurdles to overcome in a knowledge-based approach to candidate gene prioritization is the inconsistency of the level and quality of annotations across the genome. While the better studied genes are more likely to have high quality annotations, less well characterized yet more relevant genes can be overlooked simply because information available about them is incomplete. This issue is further addressed in OVA by the use of gene and phenotype annotations from model organism (mouse, rat and fish) orthologs to both support the human data and to compensate where data for human genes remains incomplete.

The integration of data from across multiple ontologies can supplement knowledge where it may be incomplete or inadequate in a particular domain. While most human genes now have associated Gene Ontology annotations available,

'shallow' annotations – that is, low information content terms– are still prevalent. Similarly, there are a number of terms that while they may not be considered uninformative, are not meaningful for candidate gene prioritization without further context. For instance, two genes annotated with the term '*regulation of transcription, DNA-templated*' (GO:0006355) would be considered highly functionally similar, and yet could participate in regulation of entirely different pathways. Accordingly, pathway ontology annotations can serve to fill in this knowledge gap, helping to decide whether a gene is truly relevant to the query phenotype and thus improving prioritization accuracy.

Here, novel gene discovery is simulated in well and poorly characterized diseases in order to demonstrate that the method presented here is capable of meaningful candidate gene prioritization even when direct functional knowledge about the disease is lacking. By inferring new gene-disease associations through phenotype semantic similarity search and cross-ontology bridges, OVA attempts to deduce missing annotations, enabling gene prioritization for new and rare human diseases while supplementing the functional profile of better characterized phenotypes. Furthermore, this work shows that OVA RF model distinguishes relevant genes accurately and, coupled with a variant filtering approach, performs better than another recently published variant prioritization tool, ExomeWalker.

Knowledge-based candidate gene/variant prioritization methods have been known to perform worse than reported when predicting novel disease genes (Bornigen *et al.*, 2012). However, large scale assessment using novel disease genes is not feasible. Cross-validation-based methods of individually removing direct disease-gene associations can serve to simulate novel gene discovery by ensuring that the test gene does not contribute to the query annotation profile. However, there is a degree of knowledge circularity in literature, and thus ultimately in ontology annotations that is difficult to account for. By prioritizing VCF files containing 20 newly reported mutations in novel disease genes, this work shows that there is agreement between these results and those obtained from a larger,

simulated dataset, although the novel variant dataset prioritization was somewhat less accurate.

Thus, the author maintains that the results based on previously described disease genes represent a reasonable approximation of the true accuracy of OVA.

Through an interactive and intuitive web interface, OVA allows the user to control many aspects of the prioritization process. OVA employs The Human Phenotype Ontology to facilitate detailed phenotypes queries in addition to previously described diseases, enabling prioritization for novel diseases which may not yet have been described in frequently-used databases such as OMIM.

## ***5. N6-Methyl Adenosine Sequencing***

### ***5.1 Introduction***

Since the early years of RNA research it has been known that RNA can be subject to numerous, chemically distinct post-transcriptional modifications. Starting with the discovery of the ‘fifth’ nucleotide – pseudouridine (Davis and Allen 1957; Cohn and Volkin 1951), over a hundred RNA modifications have been characterised to date (Machnicka et al. 2013), collectively termed the RNA ‘epigenome’ or ‘epitranscriptome’. RNA modifications can be seen as analogous to those of DNA and histones, adding a yet to be fully understood layer of intricate regulation to the transcriptome.

As with any complex regulatory network, there is potential for things to go wrong. Numerous studies implicate RNA modifications in disease, including diabetes (Vasan et al. 2014), obesity (Fawcett and Barroso 2010), infertility (Zheng et al. 2013), dyskeratosis congenital (Heiss et al. 1998), mitochondrial myopathy (Bykhovskaya et al. 2004) and various cancers (Chen et al. 2015b; Steinman et al. 2013). It is thus important to understand the dynamics of RNA modifications and the effects varying physiological conditions have on the epitranscriptomic landscape. However, until recently, methods for transcriptome-wide profiling of RNA modifications have proved elusive.

The advent of high-throughput sequencing technologies has facilitated the development of several approaches for the global characterisation of the epitranscriptome, including N6-methyl adenosine (Dominissini et al. 2013; Meyer et al. 2012), pseudouridine (Lovejoy et al. 2014; Carlile et al. 2015), 5-methylcytosine (Schaefer et al. 2009) and A-to-I RNA editing (Bahn et al. 2012). ‘Next-generation’ sequencing, while classically used to identify deleterious DNA variants or aberrant gene expression changes, has now opened up new avenues for understanding the contribution of RNA modifications to human disease.

However, bioinformatics efforts have struggled to keep up with the rapid developments in epitranscriptome sequencing. While numerous algorithms have been developed for DNA epigenome sequence data analysis, few yet exist for epitranscriptomics data. Where dedicated methods for epitranscriptomics data analysis have been lacking, DNA analysis software – if not entirely appropriate for the task- had been adapted (Dominissini et al. 2013). In light of this, this chapter introduces novel bioinformatics approaches for the analysis of RNA N6-methyl adenosine sequencing data.

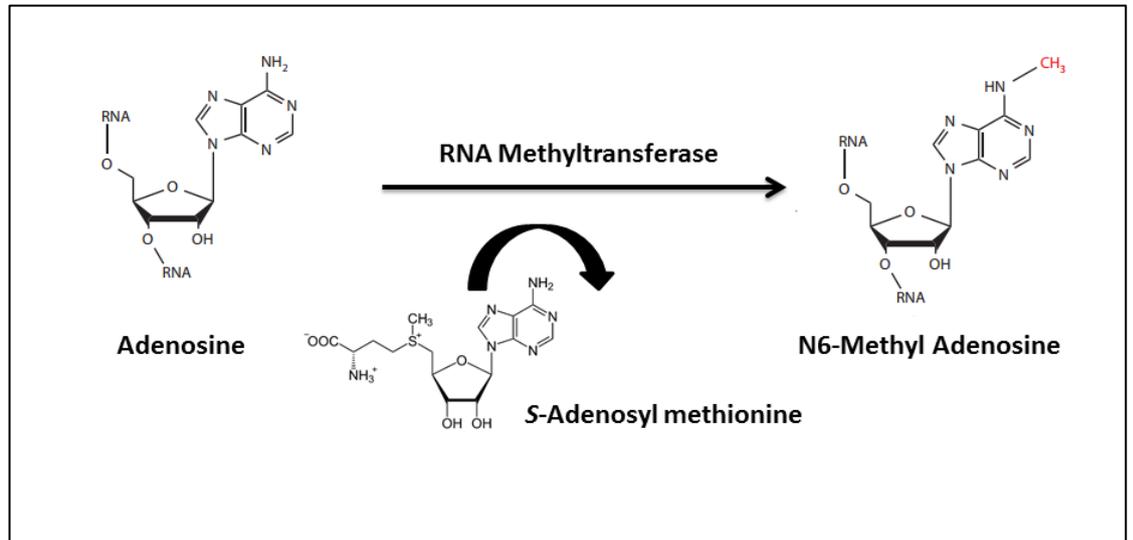
The chapter is structured as follows. First, a review of the current body of knowledge on the molecular biology of N6-methyl adenosine, as well as its physiological roles, is presented. Means for the detection of N6-methyl adenosine are summarised, with the emphasis on discussion of available computational methods for data analysis of high-throughput epitranscriptomic sequencing data. The remainder of the chapter discusses the development of novel software and methodology pertaining to N6-methyl adenosine sequencing data analysis. The methods presented here have been implemented as a stand-alone, GUI-driven desktop software application – m6aViewer. The last section of the chapter discusses the details of the software implementation.

## ***5.1.2 N6-Methyl Adenosine Molecular Biology***

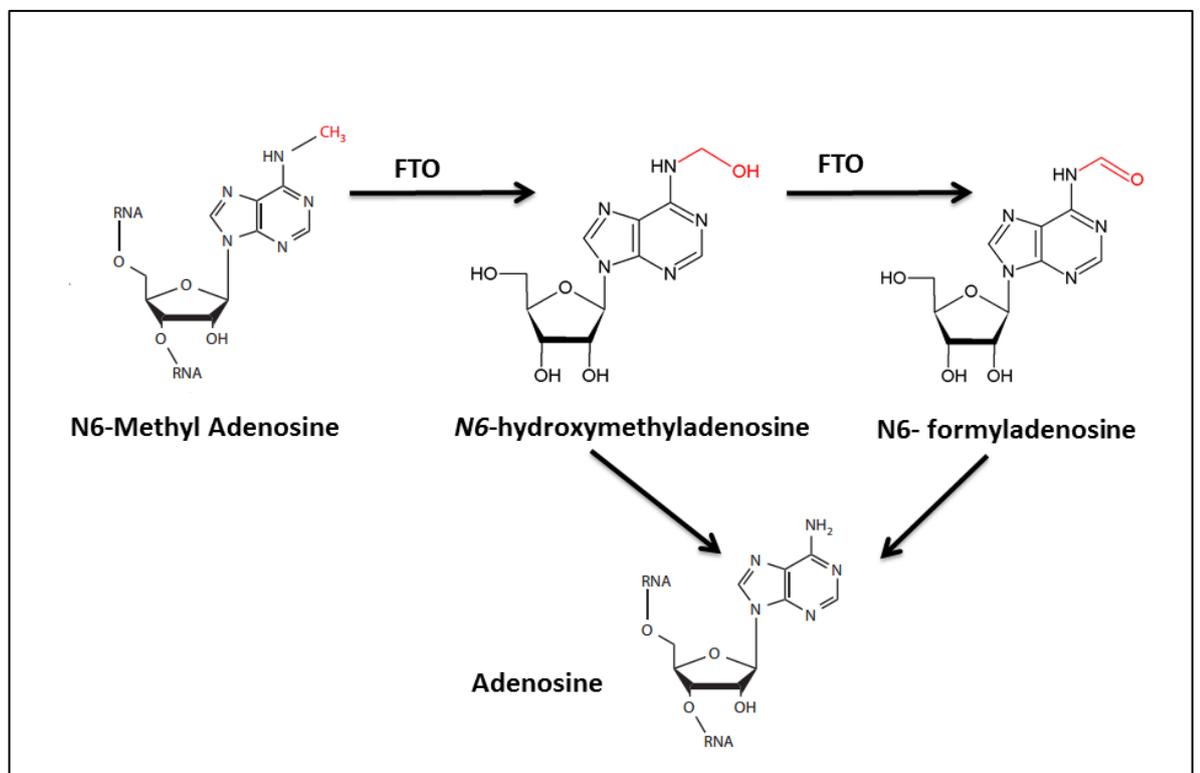
### ***5.1.2.1 Overview***

N6-methyl adenosine (henceforth referred to as m<sup>6</sup>A) was first discovered in the 1970s and was found to be one of the most highly abundant RNA modifications (Desrosiers et al. 1974; Lavi and Shatkin 1975; Wei et al. 1975). The addition of a methyl group to adenosine is catalysed by RNA methyltransferases in the presence of co-substrate S-adenosyl methionine (SAM) (**Figure 32A**). While initially m<sup>6</sup>A was thought to be a static modification, the discovery of the first RNA demethylase - FTO (fat mass and obesity-associated protein) - has indicated that the m<sup>6</sup>A modification is reversible and dynamic (Jia et al. 2011). Further investigation into the molecular dynamics of FTO-mediated RNA demethylation has uncovered two stable, intermediate products – hydroxymethylated and formylated adenosines (Fu et al. 2013)

(**Figure 32B**). These intermediates could have a role in regulating m<sup>6</sup>A dynamics – or even have functional roles of their own; however, little is yet known about them.



**Figure 32A.** The enzymatic conversion of adenosine into N6-methyl adenosine by RNA methyltransferase in the presence of a co-substrate, S-Adenosyl methionine.



**Figure 32B.** Enzymatic demethylation of N6-methyl adenosine by RNA demethylase FTO creates two stable intermediates.

The discovery of FTO as an RNA demethylase was largely responsible for the rekindled interest in m<sup>6</sup>A research, which had dwindled since the early studies of the modification. Several RNA methyltransferases, demethylases and mediator “reader” proteins have been discovered and m<sup>6</sup>A landscape has emerged as both dynamic and highly conserved (Batista et al. 2014; Dominissini et al. 2012; Meyer et al. 2012), hinting at its crucial biological functions. To date, the precise biological roles of m<sup>6</sup>A remain poorly characterised. There is emerging - and often conflicting - evidence that implicates m<sup>6</sup>A in RNA nuclear export (Zheng et al. 2013; Camper et al. 1984), degradation (Wang et al. 2014a), splicing (Zhong et al. 2008; Dominissini et al. 2012) and translation (Zhou et al. 2015; Meyer et al. 2015), as well as miRNA dynamics (Alarcón et al. 2015b; Chen et al. 2015b; Berulava et al. 2015; Ke et al. 2015).

Indeed, it is not infeasible that m<sup>6</sup>A may have multiple distinct functional roles within the cell, mediated by different “reader” proteins or via structural changes within the methylated RNA molecule. This is further evidenced by the apparent non-random distribution of m<sup>6</sup>A within mRNA molecules – adenosine methylation preferentially occurs within 3' and 5' untranslated regions (UTR) and long and/or alternatively spliced exons (Meyer et al. 2015; Schwartz et al. 2014; Ke et al. 2015; Meyer et al. 2012), suggesting that distinct functional roles may exist for different classes of m<sup>6</sup>A.

Thus far, m<sup>6</sup>A has been shown to occur in mRNA (Meyer et al. 2012), rRNA (Iwanami and Brown 1968), tRNA (Saneyoshi et al. 1969), snRNA (Bringmann and Lührmann 1987), miRNA (Berulava et al. 2015) and lncRNA (Meyer et al. 2012; Dominissini et al. 2012) and has been found in transcripts from diverse organisms, including mammals (Dominissini et al. 2012), plants (Luo et al. 2014; Li et al. 2014b), yeast (Schwartz et al. 2013), bacteria (Deng et al. 2015) and viruses (Lichinchi et al. 2016). It has been implicated in a range of different diseases including obesity and diabetes (Vasan et al. 2014; Fawcett and Barroso 2010); various forms of cancer (Lin et al. 2016; Zhang et al. 2016); depression (Du et al. 2015); infertility and asthenozoospermia (Yang et al. 2016);

Zheng et al. 2013); and may play a role in the dynamics of HIV infection (Lichinchi et al. 2016).

Despite the renewed interest in m<sup>6</sup>A, the precise roles this modification plays in RNA metabolism and the physiological consequences these may have are yet to be elucidated. Recent research has raised as many new questions as it has answered. How does the position of m<sup>6</sup>A within mRNA influence its function? Cancer cells are intrinsically associated with methionine metabolism (Leach and Tuck 2001; Tuck et al. 1996) – what are the roles of m<sup>6</sup>A in cell transformation? What physiological outcomes does the disruption of m<sup>6</sup>A manifest in? What role do differentially expressed RNA methyltransferases play during development (McGraw et al. 2007; Meyer et al. 2012)? What roles do tissue-specific RNA demethylases have? How does m<sup>6</sup>A interact with other aspects of RNA lifecycle? In order to identify the functions of m<sup>6</sup>A, it is imperative to characterise these not yet fully explored aspects of RNA biology.

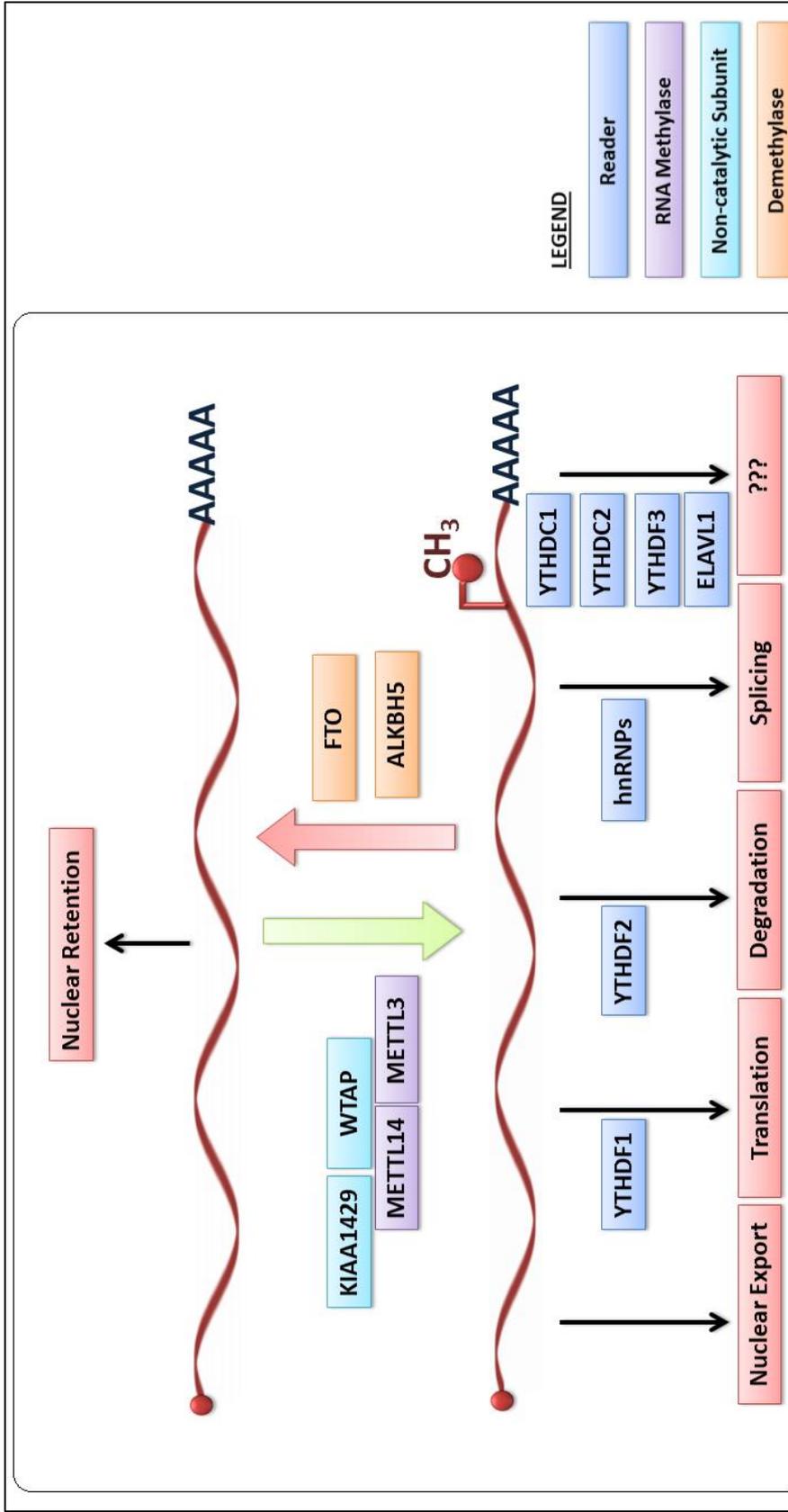
#### ***5.1.2.2 Dynamic Epitranscriptome – readers, writers and erasers***

m<sup>6</sup>A is a dynamic modification, with its changing landscape shaped by diverse groups of proteins that can be classified into the broad roles of ‘writers’, ‘erasers’ and ‘readers’. A number of RNA methyltransferases, RNA demethylases and effector “reader” proteins have been identified, with many more still likely to be discovered. **Figure 33** summarises the roles these proteins have been attributed in shaping the RNA methylome.

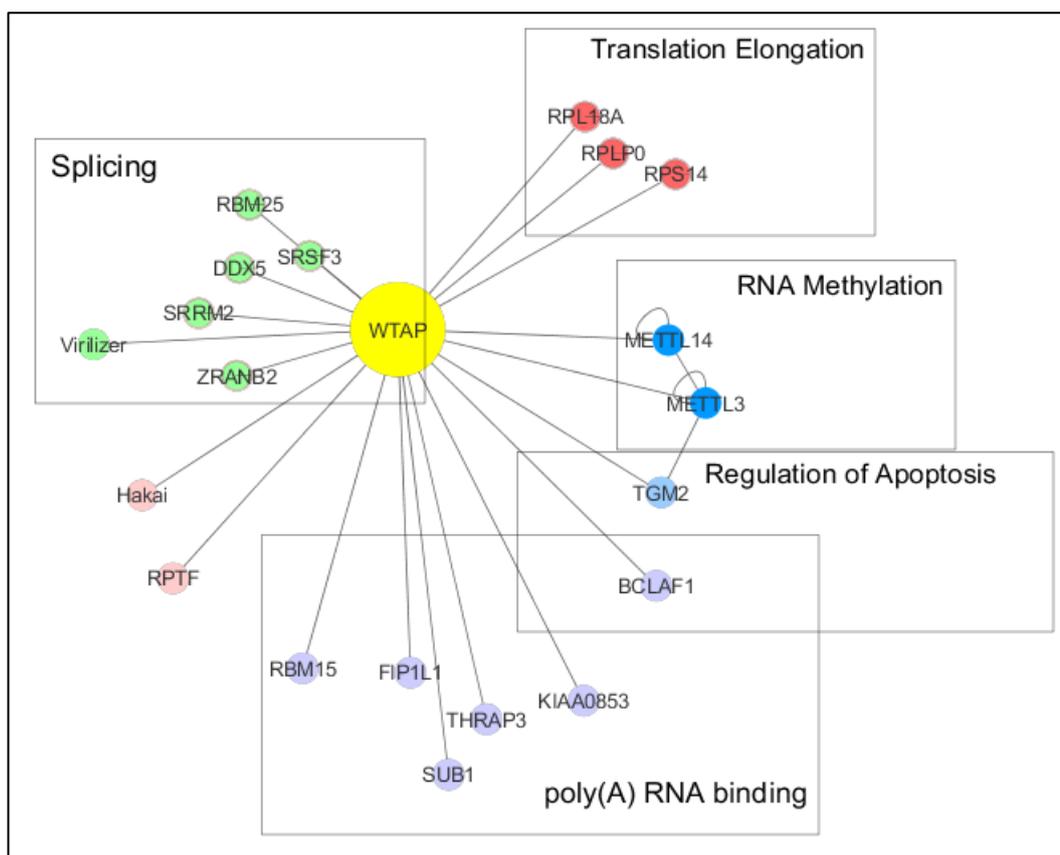
##### ***5.1.2.2.1 RNA Methyltransferases***

Although METTL3 was the first identified mammalian RNA methyltransferase, it was known that it belonged to a much larger, 200 kDa protein complex (Bokar et al. 1997). To date, three additional components of the mammalian RNA methyltransferase complex have been identified – METTL14, WTAP and KIAA1429 (Schwartz et al. 2014; Ping et al. 2014). METTL3 and METTL14 both possess RNA methyltransferase activity, and while WTAP and KIAA1429 are not catalytic, both have been shown to interact with METTL3 and/or METTL14 and to be required for RNA methylation (Ping et al. 2014; Schwartz et al. 2014). WTAP has been shown to be required for METTL3/METTL14 localisation to nuclear speckles (Ping et al. 2014). A number of other proteins were found to

physically interact with the RNA methyltransferase complex through proteomics screens, although some of these associations could be spurious (Schwartz et al. 2014; Horiuchi et al. 2013).



**Figure 33.** Summary of key proteins in mammalian m<sup>6</sup>A methylation and the proposed roles for m<sup>6</sup>A ‘reader’ proteins. RNA methylases METTL14 and METTL3 act together with non-catalytic sub-units WTAP and KIAA1429 to methylate mRNAs, while demethylases FTO and ALKBH5 are capable of reversing this modification. Effects of RNA methylation are mediated via ‘reader’ proteins that are capable of directly or indirectly recognise m<sup>6</sup>A residue and promote processes such as nuclear export, translation, degradation and splicing.



**Figure 34.** Proteins shown to interact with WTAP, identified through proteomic screens (Schwartz et al. 2014; Horiuchi et al. 2013). Functions ascribed to these proteins here are derived from their respective Gene Ontology annotations using the following procedure. Experimental RNA methyltransferase component protein interaction data was downloaded from the publishers' websites as supplementary data (Schwartz et al. 2014; Horiuchi et al. 2013), common contaminants were filtered out using CRAPOME database (Mellacheruvu et al. 2013). Protein interactions were visualised using Cytoscape (Shannon et al. 2003). Gene Ontology annotations were obtained through Cytoscape plug-ins BinGO (Maere et al. 2005) and GOrize (Garcia et al. 2007).

Interestingly, WTAP has been shown to interact with a number of different RNA binding proteins, as well as RNA Methyltransferases METTL3 and METTL14. Further investigation into this data revealed that notable sub-groups of these proteins are involved in splicing and translation, confirming m<sup>6</sup>A as a key player in these processes (**Figure 34**). Of particular note are BCLAF1 and TGM2 – proteins implicated in regulation of apoptosis (Lee et al. 2012; Hsieh et al. 2013), although it is unclear what role these may play in m<sup>6</sup>A metabolism.

Consistent with this observation, METT3 or WTAP silencing in *D. rerio* embryos causes early developmental defects and increased apoptosis (Ping et al. 2014). That the *in vitro* interaction between BCLAF1 and WTAP may not be spurious is further evidenced by the observation that BCLAF1 is also enriched in nuclear speckles *in vivo*, and directly impacts WTAP localisation therein (Horiuchi et al. 2013).

Before its characterisation as an RNA methyltransferase complex component, WTAP was heavily implicated as a component of the splicing machinery, but without a defined function. Its interaction with several proteins involved in splicing (**Figure 33**) also supports this role. However, the part m<sup>6</sup>A plays in splicing is less clear. Transcriptome-wide PAR-CLIP data from HeLa cells show that the majority of RNA methyltransferase METTL3 binding sites are intronic or intergenic (Liu et al. 2014) – a feature not reported in similar experiments by Ping *et al* (Ping et al. 2014); however this may be due to a data analysis approach that considers mature transcripts only. Indeed, in contrast to transcriptome-wide m<sup>6</sup>A screens (Meyer et al. 2012; Dominissini et al. 2012), early experiments indicated that m<sup>6</sup>A methylation is also prevalent in nascent RNA intronic regions (Carroll et al. 1990), suggesting a possible role in RNA splicing. Depletion of METTL3 or WTAP leads to a general disruption of splicing processes, including an alteration of splice isoform ratios. Indeed, transcriptome-wide m<sup>6</sup>A analyses indicate that adenosine is preferentially methylated in exons which are involved in alternative, rather than canonical splicing. Taken together with co-localisation of RNA methyltransferase complex and splicing machinery to nuclear speckles, this data suggests a tight integration between splicing and RNA methylation processes, although the precise role of m<sup>6</sup>A is as yet unknown.

It is worth noting that while a small number of intronic m<sup>6</sup>A sites were identified through transcriptome-wide m<sup>6</sup>A profiling, current RNA sequencing based approaches are not best suited for an unbiased characterisation of nascent RNA transcripts. Poly-(A) selection for mRNA sequencing, for example, excludes these transcripts. Likewise, in total RNA libraries, it can be difficult to

distinguish genuine intronic RNA sequencing reads from reads arising from DNA contamination.

RNA methylation machinery is highly conserved. In *S.cerevisiae*, RNA methylation is catalysed by a METTL3 homolog Ime4, in complex with a WTAP-like protein Mum2. Slz1, which has no mammalian counterpart, is also known to interact with yeast RNA methyltransferase complex by guiding the methylation machinery to the nucleolus. FIP37, a plant WTAP homolog, has been shown to be required for methylation in *A. thaliana* (Zhong et al. 2008).

In *E. coli* – an early system used to study RNA methylation – a number of methyltransferases have been identified. trmB6 is responsible for A37 tRNA methylation, rlmF has been shown to methylate A1618 in 23S rRNA (Sergiev et al. 2008) while RlmJ is specific to A2030 in 23S rRNA (Golovina et al. 2012). Interestingly, while the loss of these methyltransferases is not lethal, trmB6 mutants show an impaired ability to grow under stress, while rlmF null mutant cells exhibit growth retardation compared to wild type strains (Golovina et al. 2012; Sergiev et al. 2008).

Recently, it has been shown that bacterial mRNA, as well as tRNA and rRNA, undergo methylation, with over 100 m<sup>6</sup>A residues detected in *P. aeruginosa* mRNA and over 200 detected in *E.coli* mRNA (Deng et al. 2015). The methyltransferase(s) responsible for bacterial mRNA methylation remain unidentified, however, as rlmJ and rlmF null mutants do not significantly alter the m<sup>6</sup>A/A ratio found in bacterial RNA (Deng et al. 2015), suggesting these enzymes are rRNA-specific. It is possible that this elusive mRNA methyltransferase is also specific to certain bacteria, as in the species studied thus far, only gram-negative bacterial mRNAs have been shown to be methylated (Deng et al. 2015).

RNA methyltransferases have been shown to recognise several related RNA sequence motifs, mostly notably the mammalian 'DRACH', although only a fraction of these are methylated, indicating that additional factors are required for RNA methyltransferase binding (Csepany et al. 1990; Narayan et al. 1994; Ke et al. 2015; Dominissini et al. 2012). It has been proposed that secondary

RNA structure plays a key role in guiding RNA methyltransferases. Early studies showed that m<sup>6</sup>A formation was impaired in double-stranded RNA constructs (Narayan et al. 1994), while computational predictions of RNA structure around detected m<sup>6</sup>A residues have indicated a correlation with a more relaxed secondary structure (Zhou et al. 2016b). However, other studies failed to find any overlap between secondary RNA structure and m<sup>6</sup>A residues (Dominissini et al. 2012). Conversely, transcriptome-wide PARS (parallel analysis of RNA structure)(Kertesz et al. 2010) data from GM12878 cells suggests that the bases preceding m<sup>6</sup>A exhibit a strong tendency to be unpaired, while the methyl-adenosine itself shows no such enrichment (Roost et al. 2015). This suggests that RNA secondary structure may indeed play a role in RNA recognition by RNA methyltransferase machinery and highlights the poor precision inherent in computational RNA secondary structure prediction.

In addition to secondary RNA structure, miRNAs have been implicated in RNA methyltransferase binding. Chen *et al* ( 2015b) show that in mouse embryonic stem cells, formation of m<sup>6</sup>A can be modulated by miRNAs through sequence pairing. Strong correlation was observed between global m<sup>6</sup>A levels and the over- and under-expression of Dicer (a key enzyme in miRNA maturation pathway), while RNA methyltransferase or RNA demethylase levels were unaffected. Similarly, individual m<sup>6</sup>A sites could be thus manipulated by over- or under-expressing their corresponding miRNAs. Interestingly, the depletion of Argonaute proteins - main mediators of miRNA binding to target mRNAs - did not affect global m<sup>6</sup>A levels, suggesting a different mechanism for miRNA targeting to methylation sites.

While miRNAs have been shown to aid RNA methyltransferase binding, miRNAs themselves can be methylated (Berulava et al. 2015; Alarcón et al. 2015b). The m<sup>6</sup>A modification in primary miRNAs has been recently shown to be the means by which the miRNA microprocessor complex targets primary miRNA stem-loops, conferring specificity (Alarcón et al. 2015b). This dual relationship between m<sup>6</sup>A and miRNAs suggests potentially coupled regulation dynamics might exist between RNA methylation and miRNA maturation, where one is required for the other.

#### **5.1.2.2.2 RNA Demethylases**

To date, two mammalian RNA demethylases - FTO and ALKBH5 - have been discovered (Jia et al. 2011; Zheng et al. 2013). FTO is ubiquitously expressed, with higher expression levels in the brain, and localises either to the cytoplasm, or to nuclear speckles. The localisation to nuclear speckles suggests a level of interaction between methylation and demethylation processes, while the presence of FTO in the cytoplasm indicates that RNA could be demethylated both during and after processing and export from the nucleus. On the other hand, export to cytoplasm could be a way of regulating FTO activity via compartmentalisation. How/whether this dynamic corresponds to distinct m<sup>6</sup>A functions remains to be established.

In contrast to FTO, ALKBH5 has been shown to be expressed exclusively in the testes and is not present in the cytoplasm. ALKBH5 depletion results in increased export of poly-(A) RNA from the nucleus, suggesting a role for m<sup>6</sup>A as a regulator of gene expression through dynamic RNA nuclear export and retention.

The tissue specificity of ALKBH5 points to distinct physiological roles of RNA demethylases; indeed, a number of tissue- or condition-specific demethylases may yet be revealed. Recently, a method for selective inhibition of FTO and ALKBH5 demethylases has been developed (Huang et al. 2015), which will likely aid in elucidating any functional and physiological differences these two RNA demethylases may have.

Finally, while FTO and ALKBH5 are highly conserved across eukaryotes, there is no evidence for RNA demethylase existence in bacteria, suggesting that bacterial RNA methylomes are largely static; or are regulated by other mechanisms, such as RNA decay.

#### **5.1.2.2.3 m<sup>6</sup>A 'readers'**

The characterisation of m<sup>6</sup>A 'reader' proteins is key to elucidating the functional roles of m<sup>6</sup>A, as they are likely to be the foremost mediators of m<sup>6</sup>A roles. The first m<sup>6</sup>A readers to be described were YT521-B homology (YTH) domain proteins, encompassing five human paralogs (YTHDF1, YTHDF2 and

YTHDF3, YTHDC1, YTHDC2) (Luo and Tong 2014; Dominissini et al. 2012; Wang et al. 2014a; Schwartz et al. 2013), all of which - with the exception of YTHDC2 - have been shown to bind m<sup>6</sup>A *in vitro*.

Crystal and solution structures have provided insights into the basis of m<sup>6</sup>A recognition by the YTH domain. Methylated adenosine nestles into a hydrophobic binding pocket of the domain, and is stabilised by the formation of four hydrogen bonds, while the methyl group nests in a 'cage' formed by three aromatic side chains (Zhu et al. 2014; Theler et al. 2014; Li et al. 2014a). The YTH domain also interacts with the guanine adjacent to the methylated adenosine (Theler et al. 2014), suggesting that at least part of the observed RR-m<sup>6</sup>A-CH methylation consensus sequence also plays a role in m<sup>6</sup>A recognition by reader proteins. This may suggest that the degeneracy of the m<sup>6</sup>A consensus could be in part due to divergent m<sup>6</sup>A functions, with different consensus sequences surrounding the methylation site required to bind distinct m<sup>6</sup>A readers.

YTHDF2 is a cytoplasmic m<sup>6</sup>A reader, and has been found to be responsible for guiding methylated RNA to processing bodies for degradation (Wang et al. 2014a). This role is highly conserved. The YTHDF2 homolog Pho92 also binds m<sup>6</sup>A in *S.cerevisiae*, with individual mRNA stability inversely correlated with the number of YTHDF2 binding sites it harbours. Depletion of Pho92 in *S.cerevisiae* increases the half-life and abundance of its target mRNAs. It is interesting to note that the *S.pombe* YTHDF2 homolog, Mmi1, also participates in meiotic mRNA decay, in spite of the absence of RNA adenosine methylation in this species (Chen et al. 2011a).

In contrast to the negative regulatory role of YTHDF2, the cytoplasmic YTHDF1 has been shown to promote mRNA translation. Recent findings suggest that YTHDF1 increases translation efficiency of its target transcripts in an m<sup>6</sup>A-dependent manner by 'loading' the mRNA onto the ribosomes via interactions with transcription initiation factors. YTHDF1 depletion abolishes this effect and results in reduced ribosome occupancy of target transcripts (Wang et al. 2015b). The distinct roles of YTHDF1 and YTHDF2 suggest divergent regulatory functions of m<sup>6</sup>A that may be transcript- and context-dependent.

However, YTHDF1 and YTHDF2 share approximately half of their target transcripts, thus indicating that they may act in concert with each other, regulating gene expression via dynamic, perhaps enzymatic concentration dependent determination of methylated mRNA fate.

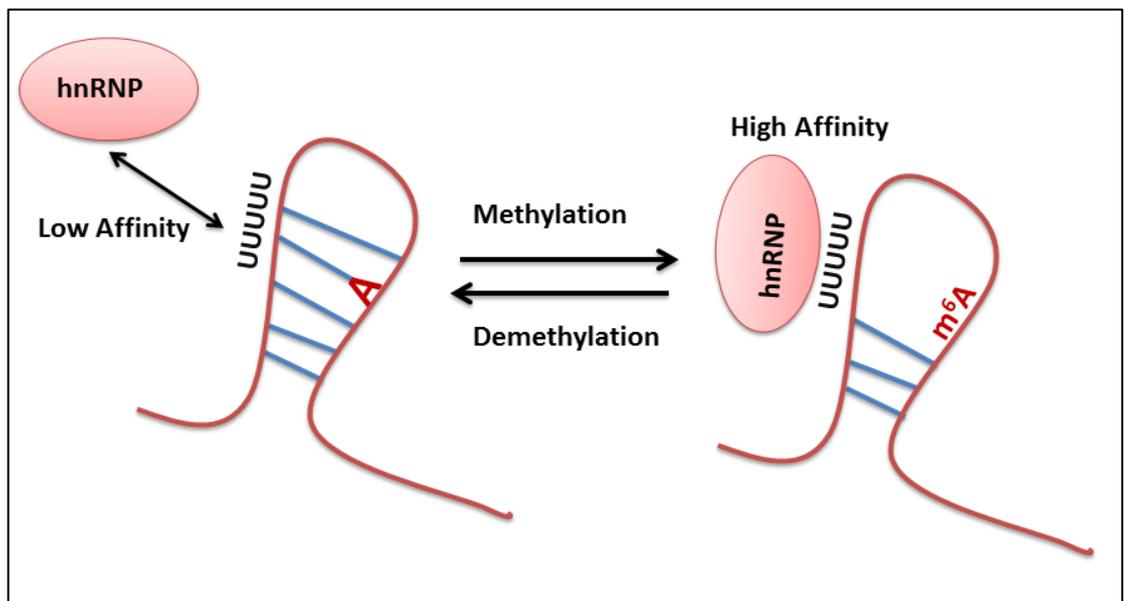
In addition to YTHDF1-3, the YTH domain (Zhang et al. 2010) family also includes YTHDC1 and YTHDC2. Initially predicted to bind m<sup>6</sup>A, akin to the confirmed m<sup>6</sup>A readers YTHDF1-3, YTHDC1 has recently been shown to regulate mRNA splicing through its recognition of methyl-adenosine (Xiao et al. 2016). YTHDC1 promotes the exclusion of its targeted exons– in line with the observation that m<sup>6</sup>A modification is frequently found on alternatively spliced transcripts.

While YTHDF3 has been shown to bind m<sup>6</sup>A (Dominissini et al. 2012) its precise function remains unclear. Similarly to YTHDF3, little is known about YTHDC2, beyond its structural similarities to the YTH domain proteins, and thus further investigation into its role as a putative m<sup>6</sup>A reader is required.

In addition to YTH domain proteins, heterogeneous nuclear ribonucleoproteins (hnRNPs) have been implicated as m<sup>6</sup>A readers (Alarcón et al. 2015a; Liu et al. 2015; Sparmann 2015). hnRNPs belong to a class of predominantly nuclear RNA binding proteins involved in the regulation of various aspects of the life cycle of RNA, including nascent RNA processing, splicing and trafficking (Yeap et al. 2014; Martinez-Contreras et al. 2007; Singh 2001; Dreyfuss et al. 1993). The basis of m<sup>6</sup>A recognition by hnRNPs has thus far been revealed as two-fold - either direct methyl-adenosine recognition (Alarcón et al. 2015a) or indirect, structure-mediated binding (Liu et al. 2015, 2013).

In a RNA homo- or hetero-duplex, adenosine will base pair with uracil, however, when methylated, the strength of this base pairing is reduced. Therefore, methylation of an adenosine may affect the stability of any secondary structure the residue is part of (Kierzek and Kierzek 2003). Consequently, it has been proposed that adenosine methylation and demethylation in RNA can regulate the binding affinity of RNA binding proteins

by inducing structural changes within RNA (Liu et al. 2015). Liu *et al* (2015) showed that the binding of heterogeneous nuclear ribonucleoprotein C (hnRNPC) – a protein involved in RNA splicing (Rajagopalan et al. 1998; McCloskey et al. 2012; König et al. 2010) - to RNA is greatly increased in the presence of m<sup>6</sup>A near the site of the uridine-track recognised by hnRNPC (Cieniková et al. 2014) and proposed the m<sup>6</sup>A ‘switch model’ (**Figure 35**), whereby the accessibility of the uridine track to bind hnRNPC is dependent on the local secondary structure, which in turn is dependent the presence of m<sup>6</sup>A. Indeed, mutations of key m<sup>6</sup>A residues thought to base pair with or close to hnRNPC binding sites result in greatly reduced binding of hnRNPC, thereby inducing a reduction in the abundance of alternatively spliced transcripts of its target RNAs (Liu et al. 2015). Transcriptome-wide screening revealed several thousand putative m<sup>6</sup>A ‘switches’, while investigations into structural changes induced by m<sup>6</sup>A in the lncRNA metastasis associated lung adenocarcinoma transcript 1 (MALAT1) provides further evidence for this mechanism (Zhou et al. 2016a; Liu et al. 2013, 2015).



**Figure 35.** The m<sup>6</sup>A hairpin switch regulates the binding of hnRNPs to RNA, modulating their functions.

In contrast to the hnRNPC, hnRNPA2B1 has been shown to bind to m<sup>6</sup>A sites directly, implicating it as another m<sup>6</sup>A reader (Alarcón et al. 2015a). hnRNPA2B1 has been shown to participate in a wide array of biological processes (He and Smith 2009), including miRNA-related pathways (Villarroya-Beltri et al. 2013) and splicing (Berson et al. 2012). Alarcón *et al* propose that hnRNPA2B1 may be the mediator of m<sup>6</sup>A-dependant alternative splicing, acting downstream of METTL3, as hnRNPA2B1 depletion replicates alternative splicing patterns observed in METTL3 depleted cells (Alarcón et al. 2015a). Consistent with reports implicating hnRNPA2B1 in miRNA synthesis, a significant number of hnRNPA2B1 binding sites were found to closely overlap with m<sup>6</sup>A sites in primary miRNAs; hnRNPA2B1 depletion consistently affected the levels of most of these miRNAs (Alarcón et al. 2015a).

Finally, ELAVL1 was also identified as a putative m<sup>6</sup>A binding protein through its statistically significant association with an m<sup>6</sup>A bait in RNA affinity chromatography (Dominissini et al. 2012); however the basis for this association and its functions relating to m<sup>6</sup>A remain unclear. Indeed, aligning ELAVL1 binding sites with m<sup>6</sup>A positions indicates that the majority of ELAVL1 binding sites are more than 100bp away from the nearest m<sup>6</sup>A site (Chen et al. 2015a). This suggests that either the initial association found between m<sup>6</sup>A and ELAVL1 was a false positive result; ELAVL1 acts in an indirect manner, similarly to the m<sup>6</sup>A 'switches' described earlier; or ELAVL1 is specific to a small proportion of m<sup>6</sup>A residues - perhaps recognising m<sup>6</sup>A in a context-dependent manner- and thus transcriptome-wide analysis is unlikely to reveal significant associations. Further analysis by Wang *et al* indicated that ELAVL1 preferentially targets methylated RNA only if its binding site is next to an m<sup>6</sup>A position, confirming the *in vitro* results by Dominissini *et al* (2012). However, if the ELAVL1 binding site is located further away (12-nt were used for this test), then ELAVL1 shows strong binding preference for unmethylated RNA and in fact, this preference for demethylated RNA has been observed in total mRNA extracted from mESCs (Wang et al. 2014b). As such, further investigations are needed to confirm ELAVL1's status as an m<sup>6</sup>A reader.

### 5.1.2.3 m<sup>6</sup>A physiological roles

Renewed investigations in the dynamics of RNA methylation have attributed a number of distinct functions for this epitranscriptomic mark. The widespread nature of m<sup>6</sup>A across several eukaryotic RNA species indicates that methylation could indeed play a number of diverse roles in the RNA life cycle.

While more recent efforts have focused on mRNA, miRNA and lncRNA methylation, early studies of m<sup>6</sup>A molecular functions investigated the more abundant tRNAs and rRNAs. In *E.coli*, tRNA<sup>1</sup>Val is methylated at A37, a position postulated to be necessary for stabilising the anticodon loop structure. Curiously, the knockout of the methyltransferase responsible for A37 tRNA methylation does not have detrimental effects on growth under normal conditions, but rather impairs survival of *E.coli* cells under osmotic and oxidative stress (Sergiev et al. 2008). In rRNA, methylation of A2058 in 23S has been linked to antibiotic resistance (Skinner et al. 1983). These modifications, along with other methyl-adenosines in *E.coli*, are thought to be static, serving a structural role, as no RNA demethylases have yet been found in bacterial species. As m<sup>6</sup>A appears to be involved in regulating alternative splicing, nuclear export and translation in eukaryotes – processes absent or distinctly different in prokaryotes – it is likely that the dynamic nature of this modification evolved in eukaryotes. Indeed, while *S.pombe* RNA is not methylated, m<sup>6</sup>A has been reported to be present in all other eukaryotes studied to date. On the other hand, *S.pombe* is just as likely to have lost m<sup>6</sup>A machinery – *S.pombe* retains the conserved YTH domain proteins, although these have been shown incapable of binding m<sup>6</sup>A (Wang et al. 2016).

m<sup>6</sup>A is emerging as an important, multi-faceted regulator of the RNA life cycle. It is not surprising then, that there is increasing evidence linking the disruption of m<sup>6</sup>A regulation with adverse physiological effects. m<sup>6</sup>A metabolism has been strongly implicated in several human diseases, including obesity and cancer (Zhang et al. 2015; Lin et al. 2016; Zhang et al. 2016). In fact, the rekindled interest in m<sup>6</sup>A methylation research in recent years has been largely attributed to the discovery that FTO - a gene implicated in obesity - is an RNA demethylase (Jia et al. 2011). The genetics of obesity have been a popular

research avenue, as 40-70% of all variation in BMI may be explained by genetic factors (Maes et al. 1997). In 2007, FTO was identified as the first gene to be significantly associated with obesity in several independent genome-wide association studies (GWAS) across multiple population groups (Dina et al. 2007; Hinney et al. 2007; Scuteri et al. 2007; Frayling et al. 2007). As the identified variants were all found to be located within the first intron of FTO, it has been suggested that the association with obesity may be due to the distal, regulatory effects the locus may exert on the expression of other genes (Claussnitzer et al. 2015); the variants were found to have no impact on FTO gene expression itself. However, overexpression of catalytically active FTO in pre-adipocytes enhances adipogenesis (Zhang et al. 2015), suggesting that the FTO protein itself contributes to obesity. That this effect may be mediated through its m<sup>6</sup>A demethylation activity is further evidenced by the mirrored outcomes of METTL3 knockdown in adipocytes (Zhao et al. 2014; Wang et al. 2015c).

FTO-deficient mouse models have shown inconsistency in phenotypes - germline FTO loss manifests with high perinatal lethality and reduction in lean and fat mass (Fischer et al. 2009), while, paradoxically, adult onset FTO depletion led to an increase in fat mass (McMurray et al. 2013). The mechanism of FTO action could partially explain these phenotypic discrepancies. It has been proposed that FTO regulates splicing through m<sup>6</sup>A demethylation - a well-documented role of m<sup>6</sup>A - in a number of transcripts involved in sterol metabolism. It is conceivable that different isoform ratios of these transcripts could have significant effects on adipogenesis. Congruently, FTO has been shown to regulate the splicing of RUNX1T1 (Zhao et al. 2014), which has two alternatively spliced isoforms with antagonistic effects on adipogenesis. Thus, substantial evidence implicates FTO-mediated m<sup>6</sup>A demethylation in obesity; however, how this process can be disrupted by the presence of intronic FTO variants is less clear.

FTO is heavily expressed in the brain, thus leading to the speculation that disruption of m<sup>6</sup>A methylation patterns could contribute to the manifestation of neurological disorders. Indeed, while the main focus of homozygous FTO variants was in the context of obesity, they are also associated with other

phenotypes, including attention deficit disorder (Choudhry et al. 2013), major depressive disorder (Rivera et al. 2012; Milaneschi et al. 2014) and reduced brain volume in otherwise healthy, elderly subjects (Ho et al. 2010). FTO seemingly exerts these phenotypes through demethylation of mRNAs involved in dopaminergic pathways (Hess et al. 2013). Indeed, a number of FTO variants have been linked to dysregulation of D2/3R-signaling (Sevgi et al. 2015), suggesting that alterations in reward response processing could be another way in which FTO influences obesity. This notion is further supported by the observed association between certain FTO variants and susceptibility to addictive behaviours, including alcohol dependence (Wang et al. 2013).

Interestingly, while FTO and ALKBH5 demethylases have been attributed very distinct physiological functions, possibly because FTO is ubiquitously expressed and ALKBH5 is limited to the testes, polymorphisms in ALKBH5, like FTO, have also been associated with major depressive disorder (Du et al. 2015). The mechanism of ALKBH5 role in depression is less clear, as the gene is expressed only in the testes. This association has not yet been confirmed in independent studies, thus it is possible that the finding is a false positive - an unfortunately common failing in complex disease genome-wide association studies (Hirschhorn et al. 2002; Sullivan 2007).

Mutations in ALKBH5, consistent with its limited expression in the testes, have been recently revealed to contribute to male infertility in a cohort of 77 men undergoing infertility treatment (Landfors et al. 2016). Interestingly, the same report found significant associations between infertility and mutations in FTO. An independent study also found that increased m<sup>6</sup>A methylation was a risk factor for asthenozoospermia (Yang et al. 2016), consistent with the findings that ALKBH5 deficiency impairs fertility in male mice (Zheng et al. 2013).

Besides the genetic associations in human disease, research into physiological effects of components of the RNA methylation pathway has yielded interesting observations. m<sup>6</sup>A methylation, in general, seems to be required for the viability of many organisms. In *D. melanogaster*, METTL3 homolog Ime4 is required for gametogenesis and its homozygous deletion is

lethal, with death occurring in larval and pupal stages (Hongay and Orr-Weaver 2011). The methyltransferase is also required for viability in *A. thaliana*, with death in knockouts occurring during early developmental stages (Zhong et al. 2008), similarly to *D. melanogaster*. This embryo-lethal phenotype is also observed in the WTAP homolog AtFIP37 null mutants of *A. thaliana* (Vespa et al. 2004). In *D. rerio*, MO knockdown of METTL3 or WTAP disrupts early development, increases apoptosis and leads to various physiological defects (Ping et al. 2014). In line with these findings, METTL3 and METTL14 knockdowns in mouse embryonic stem cells decreased the levels of many transcripts involved in pluripotency, while differentiation-specific mRNAs were not diminished (Chen et al. 2015b).

In *S. cerevisiae*, RNA methylation is required for meiosis, and knockdown of Ime4 results in a cell cycle arrest at G2 prophase – the stage that correlates with the highest levels of m<sup>6</sup>A accumulation in wild type yeast (Agarwala et al. 2012). Methylated RNA transcripts in yeast were reported to be enriched for meiosis-related functions (Schwartz et al. 2013), however it is difficult to tell whether this effect on m<sup>6</sup>A distribution is independent or is observed due to stage-specific gene expression.

Perhaps the most interesting role yet ascribed to m<sup>6</sup>A is the regulation of circadian rhythms. The loss of METTL3 elongates the circadian period by affecting the nuclear export and stability of clock gene mRNAs (Fustin et al. 2013) - an observation in line with reported global effects of m<sup>6</sup>A on RNA nuclear export. Furthermore, the putative m<sup>6</sup>A reader ELAVL1 has been previously reported to be involved in transcriptional circadian control (Lehmann et al. 2015; Keller et al. 2009).

### **5.1.3 m<sup>6</sup>A detection**

Until recently, scalable and reliable methods for transcriptome-wide detection of novel m<sup>6</sup>A residues have eluded researchers. N<sup>6</sup>-methyl adenosine does not disrupt normal Watson-Crick base pairing, and therefore cannot be readily detected from RNA sequencing data using mutation detection methods that are, for example, used to detect A-to-I deamination (Chepelev 2012) or RNA/DNA cytosine-5 methylation using bisulphite sequencing techniques

(Schaefer et al. 2009). Likewise, m<sup>6</sup>A also cannot be easily chemically converted into a residue that would terminate reverse transcription – a method used for transcriptome-wide profiling of pseudo-uridylation (Lovejoy et al. 2014).

Early methods for m<sup>6</sup>A detection relied heavily on mass spectrometry (Kowalak et al. 1993) or thin layer chromatography (Kane and Beemon 1985). The first anti-m<sup>6</sup>A antibody was described in 1987 (Bringmann and Lührmann 1987), which subsequently enabled the use of immunoblotting-based techniques.

SCARLET (site-specific cleavage and radioactive labelling followed by ligation-assisted extraction and thin-layer chromatography) for m<sup>6</sup>A detection and quantification was proposed by Liu *et al* (2013). This very recent technique can be used to assess the m<sup>6</sup>A status of potentially any base within the transcriptome and obtain reliable measures of stoichiometry; however it relies on thin-layer chromatography, which precludes any transcriptome-wide analysis.

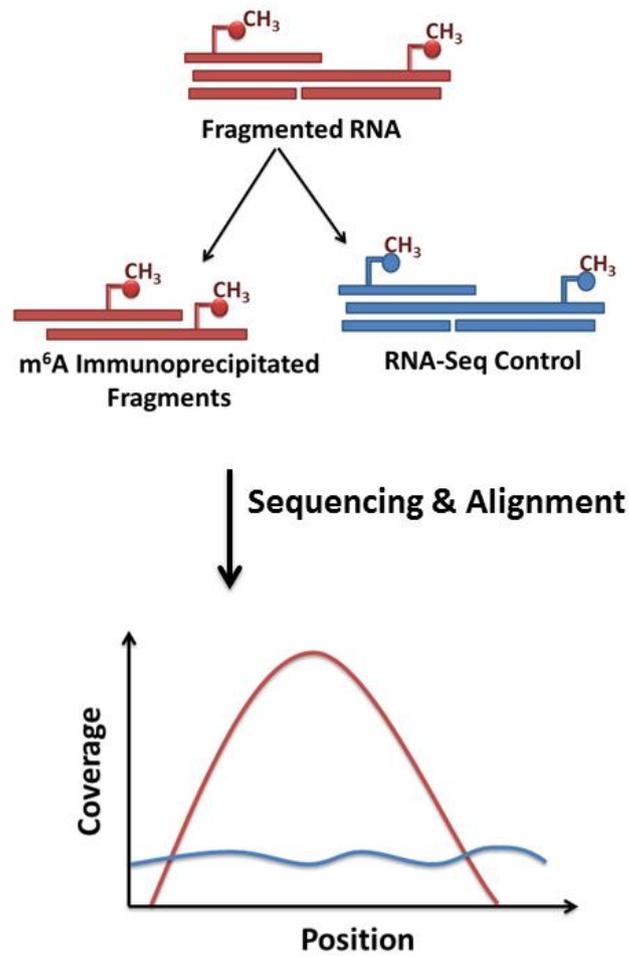
While m<sup>6</sup>A does not disrupt Watson-Crick base pairing, it is still likely to have an effect on the physical properties of RNA, such as base-stacking interactions. On this basis, Golovina *et al* proposed a method for monitoring individual m<sup>6</sup>A residues using high resolution melting analysis; however this method requires the precise position of the methylated adenosine to be known (Golovina et al. 2014).

In 2007, Dai *et al* proposed a ligation-based technique for pseudouridine and methyl-adenosine detection and quantification (Dai et al. 2007) that could potentially be adapted for use with microarrays. In brief, the technique exploits non-Watson-Crick base-pairing between adenosine and guanine – the N6-methyl group, if present, sterically clashes with the phosphate backbone in this non-canonical purine-purine pair. This greatly affects ligation efficiency, thus guanine can be used as a reporter residue to detect and quantify the presence of m<sup>6</sup>A. However, while theoretically scalable for high-throughput use, this approach screens for the presence of modifications at defined positions only, and therefore has proven hard to adapt for novel m<sup>6</sup>A site detection.

Harcourt *et al* noted that *T.thermophilus* DNA polymerase I (which can act as a reverse transcriptase in the presence of Mn<sup>2+</sup>) displayed substantial selectivity against m<sup>6</sup>A in reverse transcription reactions and thus could be used for m<sup>6</sup>A detection (Harcourt et al. 2013) in a similar manner to the approach described by Dai *et al*. However, this approach also suffers from similar drawbacks.

In light of these challenges, two independent groups proposed an immunoprecipitation based method for transcriptome-wide m<sup>6</sup>A detection (Meyer et al. 2012; Dominissini et al. 2012). In essence, the method is a marriage between ChIP-Seq and RNA-Seq (**Figure 36**). RNA is fragmented into approximately 100bp length fragments, and fragments bearing the m<sup>6</sup>A modification are recovered using an anti-m<sup>6</sup>A antibody. Following a standard RNA library preparation protocol, these fragments are then sequenced together with a normal RNA-Seq control library. In a manner similar to ChIP-Seq, aligned read coverage from the immunoprecipitated fraction is expected to form a detectable peak, about twice the sequenced fragment length, indicating the region wherein a methylated adenosine lies. However, unlike DNA immunoprecipitation, sequenced read coverage for any position is also heavily dependent on the underlying gene expression and is also subject to various sequencing biases – thus, an RNA-Seq control is required to detect regions which are genuinely enriched in immunoprecipitated fragment reads. This method for m<sup>6</sup>A detection was termed m<sup>6</sup>A-seq, or Me-RIP (hereby referred to as m<sup>6</sup>A-seq only).

Further improvements to m<sup>6</sup>A-seq were proposed by Schwartz *et al*, who used smaller RNA fragments, thereby allowing the detected regions harbouring m<sup>6</sup>A to be further refined (Schwartz et al. 2014). Despite these advances, m<sup>6</sup>A-seq remains a low-resolution technique rife with difficulties.



**Figure 36.** Immunoprecipitation-based method for transcriptome wide m<sup>6</sup>A detection. Initially, RNA is fragmented into short fragments and an anti-m<sup>6</sup>A antibody is used to enrich for m<sup>6</sup>A-modified molecules prior to library preparation and sequencing. An RNA-Seq input control is sequenced together with immunoprecipitated fragments and m<sup>6</sup>A positions can then be detected by identifying read pile-up peaks in the transcript coverage distribution.

Analysis of m<sup>6</sup>A-seq data from RNA methyltransferase knockdowns indicates that a substantial proportion of all detected m<sup>6</sup>A peaks are false positives (Schwartz et al. 2014). These could potentially arise from non-specific antibody binding, DNA contamination during sample preparation and/or sequencing read alignment errors, for example due to low-complexity sequence regions.

Furthermore, it has been noted that the method struggles to accurately capture the stoichiometry of m<sup>6</sup>A, although no direct comparisons with more accurate methods for m<sup>6</sup>A stoichiometry quantification (such as SCARLET) have yet been performed to the knowledge of the author.

The approach suffers from low resolution – ideally, a single m<sup>6</sup>A residue would generate a peak approximately 200nt wide at its base, with the peak summit indicating the position of the residue. In practice, a combination of non-specific antibody binding, immunoprecipitation specificity, alignment and amplification errors and the potential for several modified bases to be in close proximity can result in regions enriched in immunoprecipitated reads that can span several kilobases. Furthermore, Linder *et al* note that the summits of the m<sup>6</sup>A-seq peaks in their analysis only rarely precisely corresponded to the m<sup>6</sup>A position detected at single nucleotide resolution (Linder et al. 2015). The presence of a m<sup>6</sup>A consensus sequence within the enriched region may indicate the position of the methylated residue; however, the RRACH motif is degenerate, and several consensus sequences can appear within enriched regions due to chance.

Finally, while the work described herein was underway, a mutation-based method for transcriptome-wide detection of m<sup>6</sup>A at single nucleotide resolution was described (Linder et al. 2015), which substantially improves on some of the shortcomings of m<sup>6</sup>A-seq. Similar to m<sup>6</sup>A-seq, the authors propose the use of anti-m<sup>6</sup>A antibodies, whereby UV cross-linking of antibody to RNA induces signature mutations which can be detected in sequencing data. While greatly improving the resolution at which m<sup>6</sup>A can be detected, this approach still suffers from all the concerns associated with antibody use.

#### ***5.1.4 Computational methods for m<sup>6</sup>A-seq data analysis***

Initially proposed in 2012, m<sup>6</sup>A-seq has been quickly embraced by the research community, with a sizable body of work already reporting applications of the technique (Dominissini et al. 2012; Schwartz et al. 2014; Meyer et al. 2012; Meyer and Jaffrey 2014; Meyer et al. 2015; Hess et al. 2013; Berulava et al. 2015; Alarcón et al. 2015b; Chen et al. 2015b). Accordingly, RNA methylome sequence data deposited online has also increased substantially (Kolesnikov et

al. 2015; Barrett et al. 2013) – however, bioinformatics efforts have struggled to keep up with this rapidly advancing field. The lack of dedicated software for m<sup>6</sup>A-seq data analysis has been particularly telling, with one popular protocol (Dominissini et al. 2013) suggesting adapting the ChIP-Seq peak-calling software MACS (Zhang et al. 2008) for the task. Indeed, both Dominissini *et al* (2012) and Meyer *et al* (2012) used in-house scripts for m<sup>6</sup>A-seq data analysis.

#### **5.1.4.1 m<sup>6</sup>A-Seq analysis software**

An early dedicated m<sup>6</sup>A-seq analysis pipeline was described and implemented in Perl by Li *et al* (Li et al. 2013); however it is no longer accessible via the published URL. In brief, sequenced reads from the immunoprecipitated and control samples are aligned to the reference genome using BWA read alignment software (Li and Durbin 2009) and uniquely mapped read coverage across the reference sequence is computed using SAMtools (Li et al. 2009) and BEDtools (Quinlan and Hall 2010). The reference sequence is subdivided into small, 25nt width windows and each window from immunoprecipitated fraction is compared to the control using Fisher's Exact test, in order to detect statistically significant enrichment. This approach allows different library sizes between the immunoprecipitated and the control samples to be taken into account. Adjacent regions enriched in the immunoprecipitated sample are then concatenated. This part of the pipeline represents a near-faithful implementation of the original methodology described by Meyer *et al* (2012), with one key difference. Meyer *et al* (2012) artificially extend sequencing reads in the 5'-to-3' direction up to 100bp in their analysis in order to account for the difference between the RNA library insert size (sheared and size-selected to an average of 100bp) and sequencing read length, which at the time was still largely limited to 36bp.

In the original report by Dominissini *et al* (2012), the authors take a conceptually similar approach for m<sup>6</sup>A-seq data analysis. Similarly to Meyer *et al* (2012), uniquely aligned sequenced reads are extended in the 3' direction and per-nucleotide reference coverage is computed. The reference sequence is scanned using partially overlapping 100 bp windows, in contrast to the smaller, non-overlapping windows used by Li *et al* (2013) and Meyer *et al* (2012). As an

alternative to Fisher's Exact method for detecting significantly enriched regions, Dominissini *et al* (2012) compute a unique 'window score':

$$\begin{aligned} & \textit{Window Score} \\ & = \log_2\left(\frac{\textit{Mean Window Coverage IP} / \textit{Median Gene Coverage IP}}{\textit{Mean Window Coverage Control} / \textit{Median Gene Coverage Control}}\right) \end{aligned}$$

This method does not inherently compute a statistical significance level for detected regions. Thus, the authors empirically estimate the false discovery rate of this method by simply reversing the immunoprecipitated and control samples. Finally, another key difference between Meyer *et al* (2012) and Li *et al*'s (2013) application of statistical testing and the window score computed by Dominissini *et al* (2012) is the scaling – while Meyer *et al* (2012) take into account total sequencing library sizes, the window score accounts for background differences by only considering local, median gene coverage in the immunoprecipitated and control fractions.

Following this established principle of 'binning' reference coverage data, Meng *et al* developed an R and MATLAB package 'exomePeak' for m<sup>6</sup>A-seq data analysis (Meng *et al.* 2013, 2014). In contrast to the previously discussed approaches, rather than extending the reads in the 3' direction, the authors instead shift the reads by half the fragment length. Under sufficient coverage, there should be little practical differences between the two approaches; however, the shifting method does not accurately represent the sequenced fragment, only the central position of its alignment, and the coverage peaks are artificially 'slimmer' as a result.

Using the sliding window approach, significantly enriched regions are identified by modelling the read coverage in the immunoprecipitated and control fractions as a Poisson distribution. The conditional probability that reads are enriched in the immunoprecipitated fraction is estimated using the Przyborowski and Wilenski's C-test, which compares the means of two Poisson distributions (Przyborowski and Wilenski 1940). While Dominissini *et al* (2012) used gene-level and Meyer *et al* (2012) transcriptome-level background, Meng *et al* (2013) opt to test for both. ExomePeak combines two significance values to derive the final p-value (Fisher 1925).

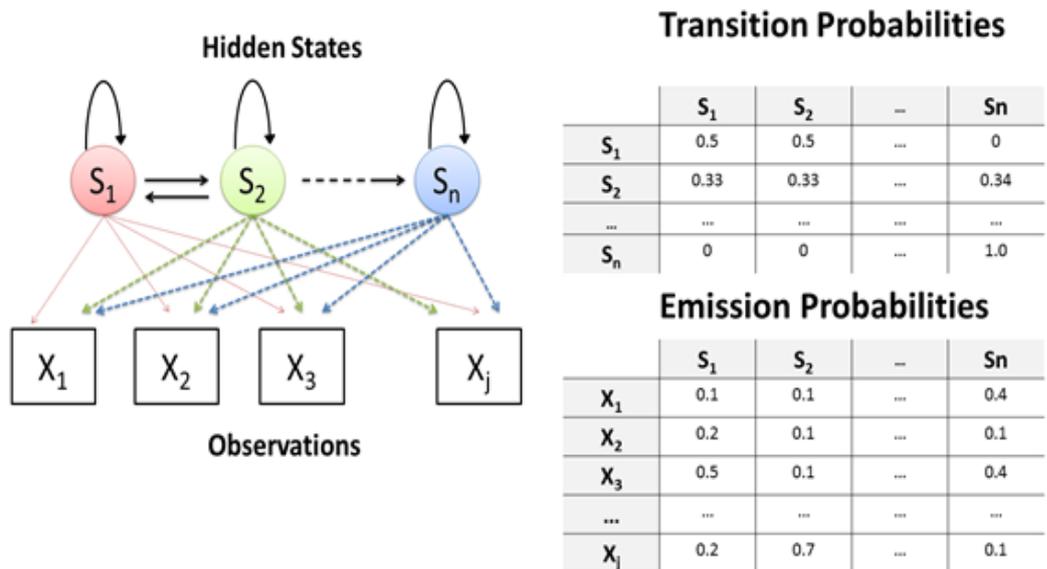
Meng *et al*/ also highlight the advantages of exome-based analysis - the approach that is implicit in previously discussed work - in circumventing the difficulties arising from transcriptome heterogeneity. However, as peak discovery in this approach is directed by the location of known exons, it crucially fails to detect intronic and intergenic peaks that may be detected from pre-mRNA or in annotated transcripts. For similar reasons, the exomePeak package cannot be readily used to detect RNA methylation in organisms with poorly annotated transcriptomes. Data generated by m<sup>6</sup>A-seq protocols which sequence poly-A selected RNA should not, in theory, generate intronic reads – and any that are present are often attributed to DNA contamination, rather than the presence of nascent mRNA. However, alternative m<sup>6</sup>A-seq protocols that do not exclude non-poly-adenylated RNA are also popular, and indeed, have provided insights into the methylation status of other RNA species, such as lncRNA and miRNA.

Following the development of the exome-based ‘exomePeak’ algorithm, the same group recently introduced the R package HEPeak, which marginally improves upon both sensitivity and specificity of ‘exomePeak’ m<sup>6</sup>A peak-calling in tests on simulated data (Cui *et al.* 2015). As an extension to the sliding window – or ‘binning’ – approach described previously, the authors model the sequence read coverage distributions using a hidden Markov model (see **Box 1**). In this approach, each genomic window can be considered to have a binary, ‘hidden’ methylation status to be determined, with consecutive windows forming a 1<sup>st</sup> order Markov chain (**Figure 37A**). Unlike previously described approaches, which incorrectly assume independence between each window tested for enrichment, a Markov model can capture the sequential nature of the data. This allows for a more intuitive identification of enriched regions, wherein the outer ‘slopes’ of each peak can be included in the called region (**Figure 37B**). The authors suggest that this approach also permits disregarding of some degree of noise in the data, whereas an independent testing approach would identify small, local spikes as significantly enriched regions.

## Box 1. Hidden Markov models

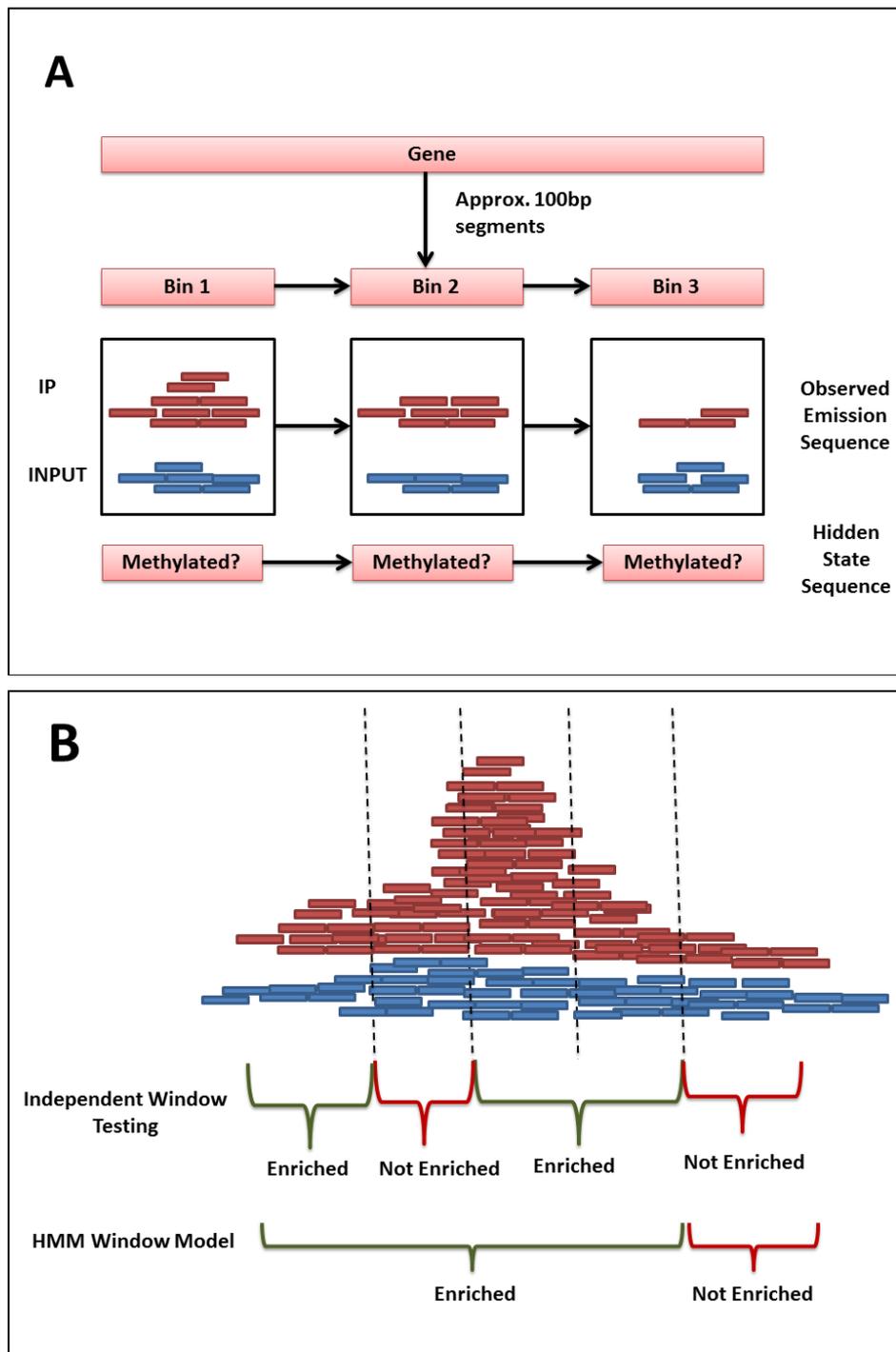
A hidden Markov model (HMM) is a ubiquitous tool for modelling probability distributions for sequential, periodical or time series data (Baum and Petrie 1966; Eddy 2004). Often used for signal processing problems, such as voice recognition, it can also be naturally applied to problems in the biological domain e.g. for modelling DNA or protein sequences.

HMM represents a series of 'hidden' states,  $S$ , and observations,  $X$ , that can be emitted by each state with differing probabilities:



The observations are assumed to be generated by some stochastic process that satisfies the Markov property – that is, at any step in the process, the state  $S$  at step  $t$  is independent of all states prior to  $t - k$ , where  $k$  is the order of the Markov process modelled. The model can be effectively represented as two matrices of transition and emission probabilities.

HMM is a probabilistic model, thus for any sequence of observations, given the transition and emission probabilities, the most likely sequence of hidden states to have generated said observations can be computed using standard Bayesian principles. That is, the probability that a HMM generates any given hidden sequence with respective series of observations is the product of the corresponding emission and transition probabilities. In many problems, emission and transition probabilities can be directly obtained from the frequencies observed in labelled training data.



**Figure 37. A.** HEPeak Hidden Markov Model representation of  $m^6A$ -seq coverage data. **B.** A schematic comparison between enriched regions (indicated by dashed lines) called using independent testing, and a first order hidden Markov model which models dependency between consecutive windows.

In HEPeak, initial emission probabilities are estimated using a binomial approximation of the ratios between mean read counts in the immunoprecipitated and control fractions. Expectation maximisation (**Box 2**) is used to estimate the model parameters given the observed data and the Viterbi algorithm (**Box 3**) is used to obtain the final solution of likeliest methylation state of each interval. The statistical confidence level for each region  $X$  is estimated using log odds ratios of posterior probabilities:

$$Score = \log \frac{P(Methylated | X)}{P(Unmethylated | X)}$$

These scores are transformed into standard z-scores and the p-value is estimated as a one-tailed probability from the resulting Gaussian distribution.

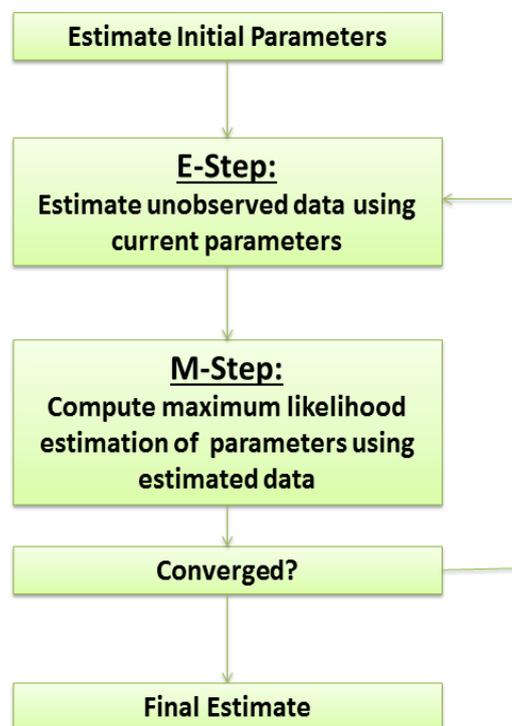
To date, to the best knowledge of the author, ExomePeak and HEPeak R packages remain the only accessible, dedicated software for the detection of m<sup>6</sup>A peaks from m<sup>6</sup>A-seq data. While the authors of these programs claim high sensitivity and specificity of these tools in simulated tests, an objective evaluation has not been performed due to the lack of testing data sets – that is, there is no m<sup>6</sup>A-seq dataset wherein all the m<sup>6</sup>A positions are known and independently verified. The exome-based approach used by both HEPeak and ExomePeak, while it avoids ambiguities that may arise from non-canonically spliced transcripts, precludes detection of peaks within intronic regions and unannotated transcripts.

However, perhaps the major drawback of the HEPeak and ExomePeak packages is the low resolution of detected regions. Enriched regions can span several kilobases, often encompassing peaks arising from several m<sup>6</sup>A residues. Indeed, using the HEPeak approach will result in larger detected regions than those called by ExomePeak in some cases. This makes it difficult to verify the m<sup>6</sup>A status of individual sites, as well as precluding any accurate comparisons between multiple samples.

## Box 2. Expectation Maximisation algorithm

Expectation Maximisation (EM) is an iterative method for approximating the maximum likelihood and is often used to estimate the parameters of probabilistic models, where data may be incomplete (Do and Batzoglou 2008; Dempster et al. 1977). HMMs (see **Box1**) are problems of this form, as they contain unobserved, or ‘hidden’, states; however EM algorithm has a wide variety of applications, such as clustering problems and natural language processing.

The EM algorithm alternates between two steps: estimating the probability distribution of the incomplete data and re-estimating the model parameters:

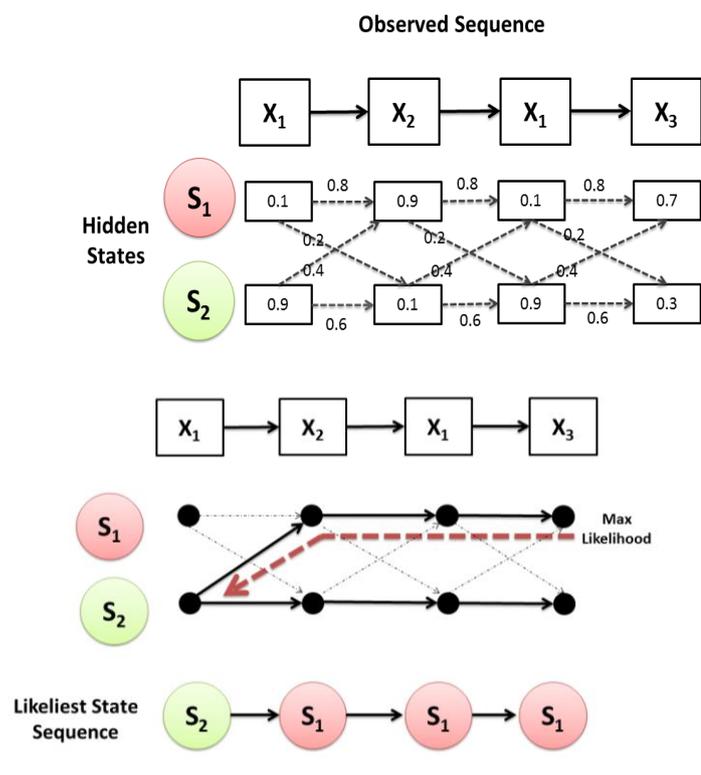


In essence, EM attempts to iteratively find parameters that maximise the probability of the observed data by reducing the problem into simpler sub-problems. Given an initial guess at the model parameters, the algorithm computes the probability distribution of all possible completions of the missing data in the E-Step. During the M-Step, the model parameters are re-estimated using the weighted training examples provided by the probability distribution of completions obtained in the E-Step. The algorithm iterates for a fixed number of steps, or until convergence (as with each iteration, the model likelihood increases, at a diminishing rate).

### Box 3. Viterbi Algorithm

Viterbi algorithm is often used to find efficient solutions to HMM problems. Given a sequence of observations in a HMM, a naïve solution to finding the likeliest sequence of hidden states that generated the observations would be to compute the likelihood of all the possible solutions. However, given  $n$  states and a sequence of length  $l$ , this would require  $n^l$  probability computations – an impractical number for all but small problems. Viterbi algorithm reduces the number of calculations required.

In a HMM sequence of states, at any given step, we can compute the likeliest path (state sequence) to that particular state:



Therefore, when computing a transition between one state and the next, instead of calculating the probabilities of all the possible paths in the HMM, we need only consider the likeliest path. At the end of the sequence, we have computed the likelihood of the most probable solution, and the likeliest sequence of hidden states can be obtained by backtracking through the trellis.

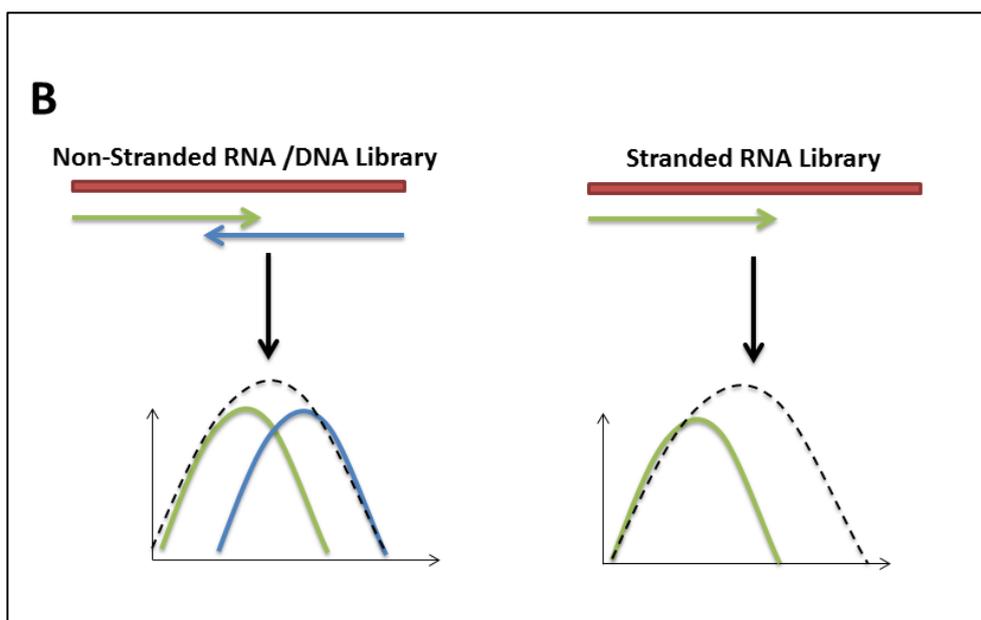
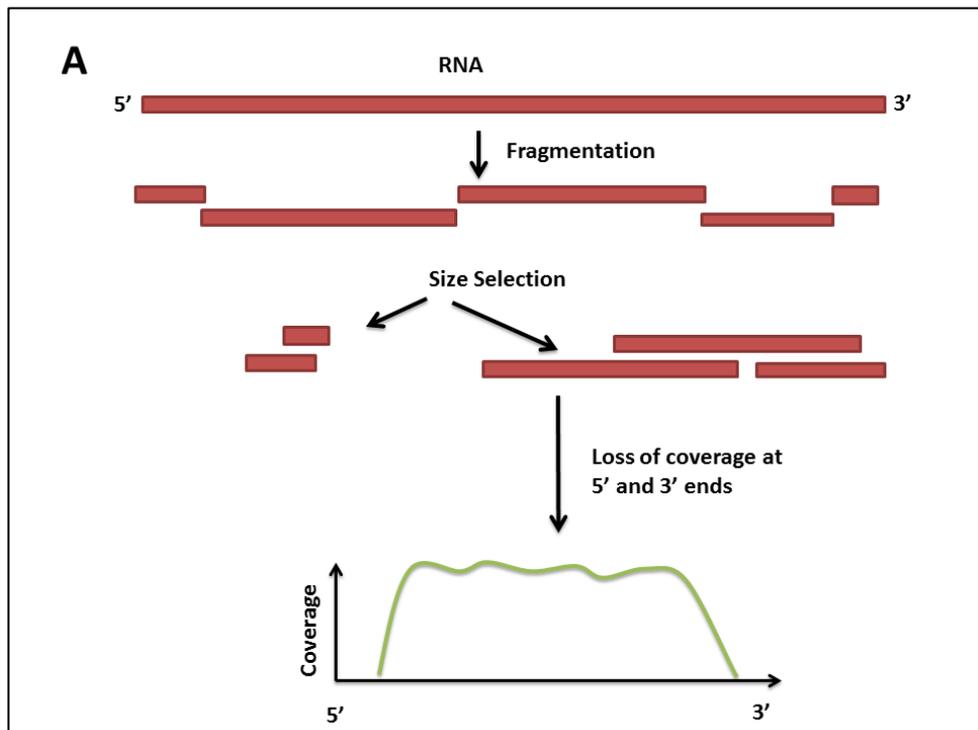
#### **5.1.4.2 Use of ChIP-Seq Peak callers for m<sup>6</sup>A-Seq data**

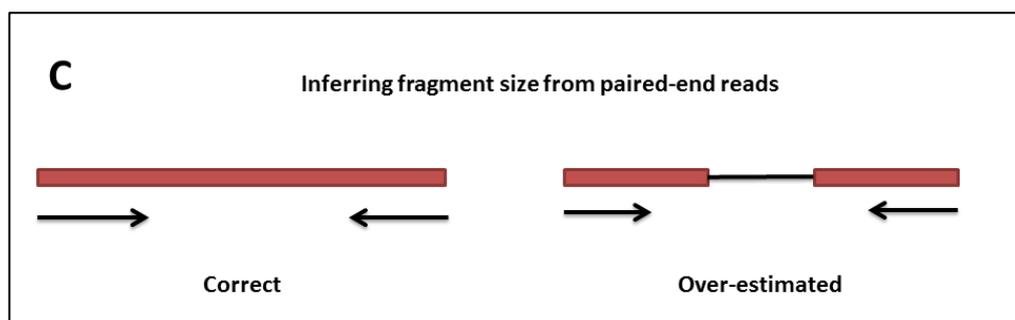
While m<sup>6</sup>A-seq data differs from ChIP-seq data in several key aspects, it has been suggested that ChIP-Seq peak-calling software could also be used for the analysis of m<sup>6</sup>A-Seq data. Thus far, one such protocol has been proposed (Dominissini et al. 2013), using peak-calling software MACS (Zhang et al. 2008). Similarly to previously discussed methods, using a sliding window approach, MACS detects significantly enriched windows by considering the sequenced read coverage as a Poisson distribution, not unlike the approach employed in ExomePeak. In order to account for local fluctuations and biases in the read distribution, MACS considers the surrounding read distribution in the immunoprecipitated sample at 1kb, 5 kb and 10kb resolution, and uses the maximum coverage to determine significant p-value cut-offs. Alternatively, this background distribution can be estimated from the control sample, if such is available. This feature makes MACS somewhat compatible with m<sup>6</sup>A-seq data, which heavily relies on the RNA-Seq control. In contrast to the long, enriched regions detected by tools like ExomePeak, MACS attempts to predict the protein-DNA (or, in this case, m<sup>6</sup>A to antibody) binding sites by reporting the location with the highest fragment pile-up within detected regions.

The use of ChIP-Seq peak-callers for m<sup>6</sup>A-seq data, nevertheless, is not entirely appropriate. There are several key differences between these two types of data that inevitably arise from the differences between RNA and DNA. While ChIP-Seq data also exhibits some regional coverage variation due to mapping biases, chromatin structure and copy number variations, the background read coverage is generally assumed to be fairly uniform in DNA sequence data models. On the other hand, the major determinant of regional coverage in m<sup>6</sup>A-Seq data is the level of individual gene expression, which yields very varied regional coverage.

Additionally, due to the fragmentation and size selection steps, the 5' and/or 3' ends of RNA transcripts are frequently lost, resulting in the bias in coverage at the ends of the transcript (**Figure 38A**) – an issue DNA sequence data is not subject to. As ChIP-Seq peak-callers typically estimate background read coverage from the immunoprecipitated data, this can lead to overestimation of

background coverage at the 5' and 3' ends of the transcript and increased false negative rate in those regions. Conversely, due to read depletion at the ends of transcripts, the overall background coverage for internal regions with higher coverage may be underestimated, leading to an increase in false positive peak calls. These concerns preclude the use of ChIP-Seq peak-calling software that does not facilitate the inclusion of non-immunoprecipitated background control.





**Figure 38. A.** The RNA fragmentation step in m<sup>6</sup>A-seq can induce the loss of coverage at 5' or 3' ends of the transcript, depending on library preparation protocol, the bias not seen in sequenced fragments from DNA immunoprecipitation experiments. **B.** Where sequencing read length is shorter than immunoprecipitated fragment length, non-stranded library preparation protocols produce a bimodal coverage distribution, whereas stranded RNA sequencing can result in a shifted coverage distribution. The read coverage distributions are shown as blue (all reads aligning to forward strand) and green (all reads aligning to reverse strand) lines, whereas the expected fragment coverage distribution is illustrated as a black dashed line. **C.** Paired-end sequencing allows to accurately infer sequenced fragment length in DNA sequencing (left), however, ChIP-Seq algorithms will consistently over-estimate fragment length if the paired reads spans an intron (thin black line).

While high-throughput sequencing technology has considerably improved in recent years, with sequenced reads commonly covering 100 bases or more, many ChIP-Seq algorithms (Zhang et al. 2008; Boeva et al. 2012; Fejes et al. 2008; Valouev et al. 2008) make provisions for cases where the DNA library insert length is longer than the sequenced read. In DNA (and non-stranded RNA) library preparation protocols, aligned reads generate a bimodal coverage distribution (**Figure 38B, left**). The mid-point between the two peaks generated by forward and reverse strand reads is often used to estimate the sequenced fragment length and shift reads accordingly to correct for this, thus improving the accuracy of called binding sites in ChIP-Seq data (Zhang et al. 2008;

Valouev et al. 2008). However, this becomes impossible for m<sup>6</sup>A-seq data generated with stranded library preparation methods (**Figure 38B, right**), and will result in consistently displaced peak calls, thus making many ChIP-Seq peak-callers, including MACS, unsuitable for this type of data.

Due to advances in sequencing technology, paired-end sequencing – wherein each fragment is sequenced from both ends – has become commonplace. While paired-end data greatly improves the accuracy of read alignment, in particular in repetitive sequence regions (Chen et al. 2012), perhaps the main benefit of paired-end data for ChIP-Seq and m<sup>6</sup>A-Seq is the ability to accurately infer each individual sequenced fragment length (**Figure 37C, left**), foregoing the need to shift or artificially extend aligned reads. However, RNA of eukaryotes invariably contains introns, complicating the direct inference of sequenced fragment boundaries from the ends of read mate pairs. As MACS and other ChIP-Seq peak-callers were not designed to work with large gapped alignments, read mate pairs mapping potentially several kilobases apart over intronic regions can lead to wildly inappropriate results.

Furthermore, split-read mapping across intron-exon boundaries is a key feature in RNA-Seq data, with several dedicated RNA sequence aligners available (Dobin et al. 2013; Trapnell et al. 2009). However, such alignments are rare in DNA sequence data (or may even be discarded by DNA sequence alignment algorithms) and are primarily exploited for the detection of DNA structural variants (Rausch et al. 2012). Consequently, even single-end RNA sequence data is problematic for ChIP-Seq peak detection algorithms, as additional provisions must be made for gapped alignments. Indeed, an m<sup>6</sup>A site close to an intron-exon boundary would be identified as two separate peaks by MACS or similar ChIP-Seq peak-callers.

In summary, it is thus apparent that currently available algorithms for the analysis of m<sup>6</sup>A-Seq data are inadequate. Dedicated m<sup>6</sup>A-seq analysis software is limited to the R environment, which requires knowledge of this statistical programming language to use. Furthermore, while these methods properly take intron-exon boundaries into account, they are unsuitable for m<sup>6</sup>A detection in poorly annotated transcriptomes. On the other hand, as ChIP-Seq is an older

technology, methods for peak-calling from DNA sequence data are more advanced and can offer increased sensitivity, specificity and resolution – nonetheless, these fail to capture the key biases and intricacies present in RNA sequence data.

### **5.1.5 Computational prediction of m<sup>6</sup>A sites**

As m<sup>6</sup>A-seq has become more widespread in use, an increasing amount of sequence data has become available. This has spawned a new branch of research for computational prediction of new m<sup>6</sup>A sites from primary sequence. In addition to the benefits the predictive power of these models provides, e.g. for hypothesis generation, feature-based models of m<sup>6</sup>A sites can also provide new insights into the biological context and roles of this modification.

Chen *et al* developed the first m<sup>6</sup>A prediction web server - m6Apred - based on a support vector machine (SVM) model (Chen et al. 2015d), followed by iRNA-Methyl, a predictor utilising a different set of features, but still utilising the same supervised learning algorithm (Chen et al. 2015c, Chang and Lin 2011). In iRNA-Methyl, the authors utilise the *S.cerevisiae* RNA methylome data (first published by Meyer *et al* (2012)) to identify a set of yeast m<sup>6</sup>A sites as positive training examples, and use randomly selected unmethylated positions containing the yeast methylation consensus GAC as negative training examples. Short (50 nt) RNA sequences surrounding the methylation site are represented as feature vectors using a pseudo-component approach, originally developed for representing protein sequences (Chou 2001) and later further extended for nucleotide representation (Chen et al. 2015e; Guo et al. 2014). In this approach, instead of considering sequences as composed on four nucleotides, they can be 'encoded' using physiochemical properties instead. Sequences are represented as a measure of translational (rise, shift, slide) and angular (twist, tilt, roll) properties of adjacent RNA bases. kmer content information and sequence enthalpy, entropy and free energy based on dinucleotide composition are also included. Physiochemical properties of bases directly affect the structure of RNA, and therefore are likely to play a role in m<sup>6</sup>A deposition/recognition, and as such, yield more predictive power for m<sup>6</sup>A sites in yeast than RNA sequences alone.

m6Apred represents an extension of this work, instead comparing chemical properties, such as ring structure, functional groups and hydrogen bonding properties of each training sequence. Additionally, pRNAm-PC – a web server for predicting m<sup>6</sup>A sites from pseudo-dinucleotide composition using iRNA-Methyl training data - was established by the same co-authors (Liu et al. 2016). The differences between the approach used in iRNA-Methyl and pRNAm-PC are unclear.

The three classifiers have been shown to be able to distinguish true m<sup>6</sup>A sites in yeast from randomly selected, unmethylated consensus sites with 65-78% accuracy. These performance rates, while encouraging, highlight the difficulty in developing an accurate m<sup>6</sup>A site predictor. These difficulties may be due, in part, to the accuracy of training data sets. While the SVM algorithm used by Chen *et al* (2015) has been demonstrated to be tolerant of somewhat noisy training data (Glick et al. 2006; Kumar et al. 2011), RNA methyltransferase knockout experiments across various organisms have suggested that a large proportion of all enriched sites detected by m<sup>6</sup>A-seq could in fact, be false positives. The high proportion of false positive sites mislabelled as genuine m<sup>6</sup>A sites in the training dataset could heavily confound the results, and as performance was assessed using a cross-validation based approach, this would further skew the reported results.

Additionally, these classifiers consider only base composition and physical and/or chemical properties of sequences. While this allows for an unbiased classification, requiring no further information for making predictions than the RNA sequence of interest, additional data could serve to improve prediction accuracy. For example, it has been suggested that RNA secondary structure may play a role in m<sup>6</sup>A site formation and/or recognition by reader proteins, and thus could be used as a predictor feature for a classifier. Furthermore, it has been noted that the distribution of m<sup>6</sup>A residues in mRNA is non-random, with enrichment in UTR regions, and long and/or alternatively spliced exons. Thus, positional information could be used to further enhance the predictive power of such classifiers, although this may preclude its use for sequences which are poorly annotated.

Finally, SVM parameters can be difficult to tune, often requiring exhaustive and time-consuming grid searches to achieve optimal classification performance (Gaspar et al. 2012). It is thus possible that further m<sup>6</sup>A site prediction accuracy could be eked out by additional parameter optimisation.

m6a-pred and iRNA-Methyl classifiers have been trained on yeast data, and therefore are likely to perform worse if applied for mammalian sequence classification. To address this issue, Zhou *et al* developed a mammalian m<sup>6</sup>A prediction server SRAMP (sequence-based RNA adenosine methylation site predictor) (Zhou et al. 2016b). Similar to methods used by Chen *et al* (2015), SRAMP uses only sequenced-based features for classification derived from a mammalian (human and mouse) m<sup>6</sup>A training data sets. SRAMP encodes features based on positional sequence information with respect to the methylated adenosine and predicted secondary structure (using RNAfold tool (Lorenz et al. 2011)) information. Zhou *et al* trained a Random Forest (Breiman, 2001) classifier, which outperforms both m6a-pred and iRNA-Methyl on a mammalian testing dataset, but not an independent yeast one, suggesting there are crucial biological differences between methylated RNA in yeast and mammals that prevent the training of a universal, species-independent classifier.

Zhou *et al* (2016) used a largely unbalanced (1:10 positive to negative instance ratio) training data set in order to simulate the observation that only a small proportion of m<sup>6</sup>A consensus motifs are actually methylated. However, unbalanced training data sets are problematic for many supervised learning algorithms (Maimon and Rokach 2010) and can result in inaccurate predictions for the minority group, as the penalty for misclassifying minority instances is significantly lessened (i.e. 90% total classification accuracy on the training dataset can be achieved simply by classifying every instance as belonging to the majority group). This unbalance in the training data set can result in a misleading assessment of the overall classifier accuracy, if the precision with respect to individual groups is not reported.

Zhou et al's results indicate that there is strong positional nucleotide preference not just within the short consensus sequence surrounding the m<sup>6</sup>A

site, but also at more distal regions. However, positional binary encoding of nucleotide sequence is unable to fully capture the sequential nature of the data, such as sequence motif enrichment or periodicity around the m<sup>6</sup>A site.

Overall, current work in computational prediction of m<sup>6</sup>A sites is promising; however, these methods still suffer from fairly low classification accuracy, such that their use for real world problems is not yet practicable.

### **5.1.6 Summary**

Research into the RNA methylome has exploded in recent years, with the development of transcriptome-wide methods for m<sup>6</sup>A site detection. As a result, functional roles surrounding this modification are being slowly elucidated. Several RNA methyltransferases, demethylases and m<sup>6</sup>A binding proteins have been discovered, forming a dynamic, regulatory network. m<sup>6</sup>A is emerging as a key regulator of RNA fate, with strong evidence to suggest that adenosine methylation exert control over processes as diverse as RNA splicing, nuclear export, translation and degradation. Disruption of this delicate balance of m<sup>6</sup>A modifications within the cell has been shown to result in diverse phenotypes. Whether through knockdown experiments, or genetic associations, RNA methylation is implicated in complex human disease, including obesity, infertility and various neurological disorders.

RNA methylome sequencing data presents not only a unique analytical challenge, but also an unprecedented opportunity for gaining new insights into RNA biology. However, while wet-lab investigations have flourished, bioinformatics have floundered. Implementations of dedicated algorithms for m<sup>6</sup>A-seq data analysis have thus far been limited to R packages 'ExomePeak' and 'HEPeak'. Both of these algorithms take an exome-based approach for transcriptome analysis and therefore are unable to detect m<sup>6</sup>A sites in intronic regions, novel transcripts or transcriptomes of poorly annotated organisms.

Current methods for the analysis of m<sup>6</sup>A-seq data fail to address the plethora of problems inherent to this type of data. Detected m<sup>6</sup>A sites suffer from low resolution, high false positive rate and inability to detect m<sup>6</sup>A sites in

low expression transcripts. These issues could be at least partially addressed *in silico*.

Furthermore, computational prediction of m<sup>6</sup>A sites from primary RNA sequence could open additional avenues for m<sup>6</sup>A data analysis. In conjunction with m<sup>6</sup>A-seq data, it could be used to call m<sup>6</sup>A sites with increased resolution, sensitivity and specificity. However, this avenue remains largely unexplored. While the accuracy of m<sup>6</sup>A site prediction algorithms developed thus far has been too low for practical applications, there is scope for improvement. The use of alternative training datasets could be explored together with algorithmic improvements. The predictive power of many sequence features remains to be investigated, and could provide not only improvements in m<sup>6</sup>A prediction accuracy, but also yield insights into the biological context and dynamics of this modification.

## ***5.2. m6aViewer application methods***

### ***5.2.1 Overview***

The following section discusses how some of the issues in m<sup>6</sup>A-seq data analysis can be addressed. A new tool for m<sup>6</sup>A-seq data analysis – m6aViewer – is introduced for this purpose, implementing novel methods for m<sup>6</sup>A-seq data processing, peak-calling and visualisation. Developed using the programming language Java, m6aViewer is a cross-platform tool controlled through a graphical user interface. In contrast to exomePeak and HEPeak packages, it requires no programming skills to use.

The rest of this chapter is dedicated to the implementation of m6aViewer's major features and the m<sup>6</sup>A peak-calling and processing methodology applied.

### ***5.2.2 Sequence Read Processing***

m<sup>6</sup>A-seq sequence data alignment should be performed using a programme that implements a splicing-aware split-read mapping algorithm. All m<sup>6</sup>A-seq data analysed here has been aligned using the STAR aligner (Dobin et al. 2013). Introduced in 2013, STAR is a relatively new alignment algorithm, built exclusively for RNA sequence alignment problems, unlike several more

established tools such as Tophat (Trapnell et al. 2009; Kim et al. 2013), which represent the natural adaptation of DNA sequence aligners to the problem. The STAR aligner has been shown by its authors (Dobin et al. 2013), as well as independently (Engström et al. 2013), to compare favourably to other alignment algorithms in terms of specificity, sensitivity and novel splice junction discovery. Most importantly, the algorithm utilises uncompressed suffix arrays stored in local memory and thus, while it requires substantial computational resources to run, high quality alignments can be generated up to two orders of magnitude faster than with other popular RNA-Seq data aligners (Engström et al. 2013).

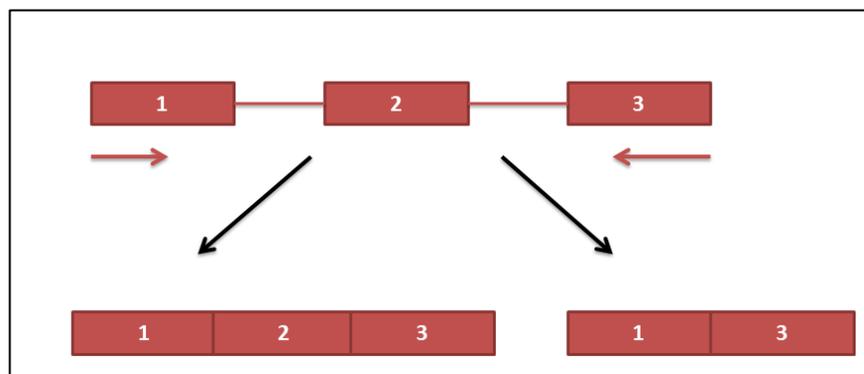
Aligned RNA sequence reads are typically stored in a SAM or BAM formatted file (Li et al. 2009), a widely accepted file format for storing sequence alignment data. While recently a number of alternative and/or improved ways of storing sequence alignment data have emerged (Cochrane et al. 2012; Hsi-Yang Fritz et al. 2011), the timeliness of SAM/BAM file format has thus far ensured its dominance. Due to its universal adoption, sorted and indexed BAM format files are used as m6aViewer input and as a starting point for all the subsequent analyses described herein.

The SAM file format consists of the header and alignment sections, where each alignment is stored on a single line containing 11 mandatory fields describing the read, reference and the alignment (**Table 10**), while the header stores meta-data that mainly describes the reference sequence to which the data was aligned to.

**Table 10** summarises the types of data stored in the alignment line. This file format stores key information required for computing sequenced RNA fragment coverage across the reference that is an integral part of m<sup>6</sup>A-seq data analysis. Read depth can easily be inferred from each read's starting coordinate and length, simply by tallying up the coverage at any given position. There are several concerns, however, that make this a less trivial task than could be initially supposed.

Field	Description
1	QNAME Name of read
2	FLAG A bitwise flag field, storing binary information on the read, including primary/secondary alignment status, PCR/optical duplicate status and paired/unpaired read flags
3	RNAME Name of the reference sequence
4	POS Left-most mapping position of the read on the reference
5	MAPQ Quality score of the alignment
6	CIGAR CIGAR string, encoding alignment matches, mismatches, insertions/deletions, clipping, etc. with respect to reference sequence
7	RNEXT Name of reference sequence of the next read or mate pair read
8	PNEXT Position of the next read or mate pair read
9	TLEN Length of the read
10	SEQ Sequence of the read
11	QUAL ASCII encoded quality score

**Table 10.** Summary of mandatory read alignment information stored in BAM/SAM file format.



**Figure 39.** In paired-end RNA sequencing, if the mates are aligned to non-consecutive exons, the sequenced fragment cannot be easily inferred, as potentially several differently spliced transcripts could give rise to the observed paired-end read alignment.

In RNA sequencing, reads spanning intron-exon boundaries are commonplace; such split-read alignments should be properly accounted for, else the obtained coverage will be inaccurate. While this is a seemingly straightforward adjustment in coverage calculations, it can be more complicated in paired-end sequence data, however. Consider the case where mates within a read pair are aligned to non-consecutive exons – potentially, the read pair could have been generated by an RNA fragment including any sequential permutation of skipped exons (**Figure 39**). This issue is perhaps decidedly less frequent in m<sup>6</sup>A-seq data than in other forms of RNA sequencing due to current m<sup>6</sup>A-seq protocols necessitating a size selection of RNA fragments, averaging 100 nt. A reasonable estimation, based on the mean expected fragment size, can be made in cases where the insert size is smaller - that is, only small exons can be wholly spanned by the read pair while still conforming to the assumption of a limited fragment size.

Additionally, in paired-end sequence data, some conformations of mate pair alignment can be indicative of mapping errors – for example, where both mates are aligned to the same strand. While these types of arrangement can be indicative of genuine RNA transcripts that can arise from genomic duplications, rearrangements, chimeric transcripts, or even circular RNAs (Qu et al. 2015), the fragment reference coverage becomes impossible to deduce. It is important to identify these scenarios, as these cases can cause program runtime errors due to violated assumptions; or worse, result in inaccurate coverage estimations without any indication of a problem.

Other considerations must also be taken into account that could result in errors when computing coverage. Read base calling and alignment quality scores are encoded within the SAM file and are indicative of the level of confidence one should place in the data. Typically, sequence data quality control steps are performed at fastq as well as post-alignment stages. Should additional quality control steps be included at runtime to discard poor quality reads and alignments? Blindly including every read regardless of mapping quality can lead to an increase in noise in the coverage data; on the other hand, quality score-based checks are bound to exclude some correctly aligned reads,

thus resulting in a decrease in coverage, which can be problematic for detecting m<sup>6</sup>A residues in poorly expressed transcripts. Furthermore, additional quality checks, while individually trivial, are likely to substantially increase computation time, as they will need to be performed on millions of reads in a typical m<sup>6</sup>A-seq experiment.

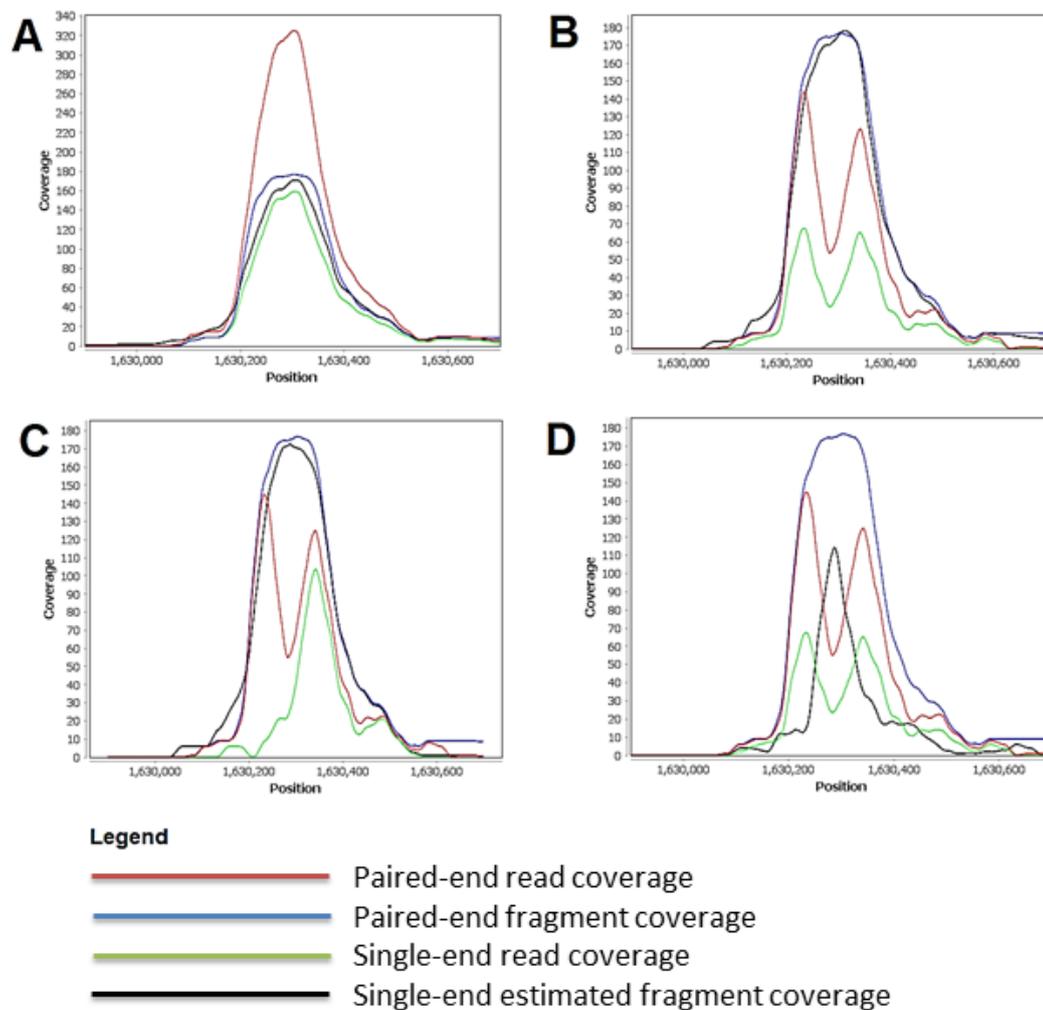
There is also the matter of duplicate reads. In DNA sequencing, duplicate reads are typically assumed to be PCR amplification artefacts and their filtering is recommended (Dozmorov et al. 2015). These will naturally arise during the amplification step, as some fragments become preferentially enriched due to smaller size or low GC content, and therefore are more likely to be sequenced. Sample availability can be another major contributor – low input amounts will require more PCR cycles to achieve the concentrations required for sequencing and result in further reduction of library complexity. Some fragment duplication, however, can also be expected to arise due to sampling coincidence from fragmentation in RNA sequencing for genes with very high expression. Thus, removal of duplicate reads is only fully justified when the sequencing depth is low, and sampling coincidence is unlikely (Zhou et al. 2014). This is problematic for RNA sequencing data, as coverage across the transcriptome is extremely variable - filtering duplicates on the assumption that they have arisen because of PCR amplification would be beneficial for genes with low expression, but pose problems for genes which are highly expressed. For a peak-calling strategy that is more concerned with limiting the false positive rate than the false negative rate of called peaks, removal of duplicate reads using software such as SAMtools (Li et al. 2009) can be beneficial. If limiting the false negative peak calling rate is an issue, then an alternative strategy may be required.

As touched upon previously, sequence reads are often not an entirely accurate reflection of the RNA fragments. Most current sequencers can only read short fragments, as errors accumulate with increased read length, resulting in a substantial drop in sequencing quality. This often results in reads which are shorter than the fragment/library insert size. Paired-end data can more accurately represent the sequenced fragment; however while paired-end

sequencing is becoming the norm, single-end reads still constitute a large share of m<sup>6</sup>A-seq data that is currently publicly available.

For a single enriched region that constitutes an m<sup>6</sup>A site, the observed coverage distribution can be remarkably different under varying methods of sequencing and coverage computations. Consider the case where averaged fragmented RNA size is 100 nt and sequencing read length is also 100 bp, obtained using non-stranded library preparation (**Figure 40A**). The difference between counting fragment coverage and read coverage is slight. Paired-end data results in a slightly wider peak, suggesting that some information is lost at the end of the read in case of single-end data. If the read length is shorter than the insert size, the difference becomes more pronounced, and a bimodal read coverage distribution can be observed (**Figure 40B**). In paired-end data, this can be corrected by including the gap between mates, as well as the reads themselves, in the coverage. For single-end data, extending the reads up to 100bp in length also corrects this bimodal pattern. If the sequencing data is stranded, instead of the bimodal coverage pattern, the bias manifests as a shift in the distribution for single-end reads (**Figure 40C**) that can also be corrected by read extension.

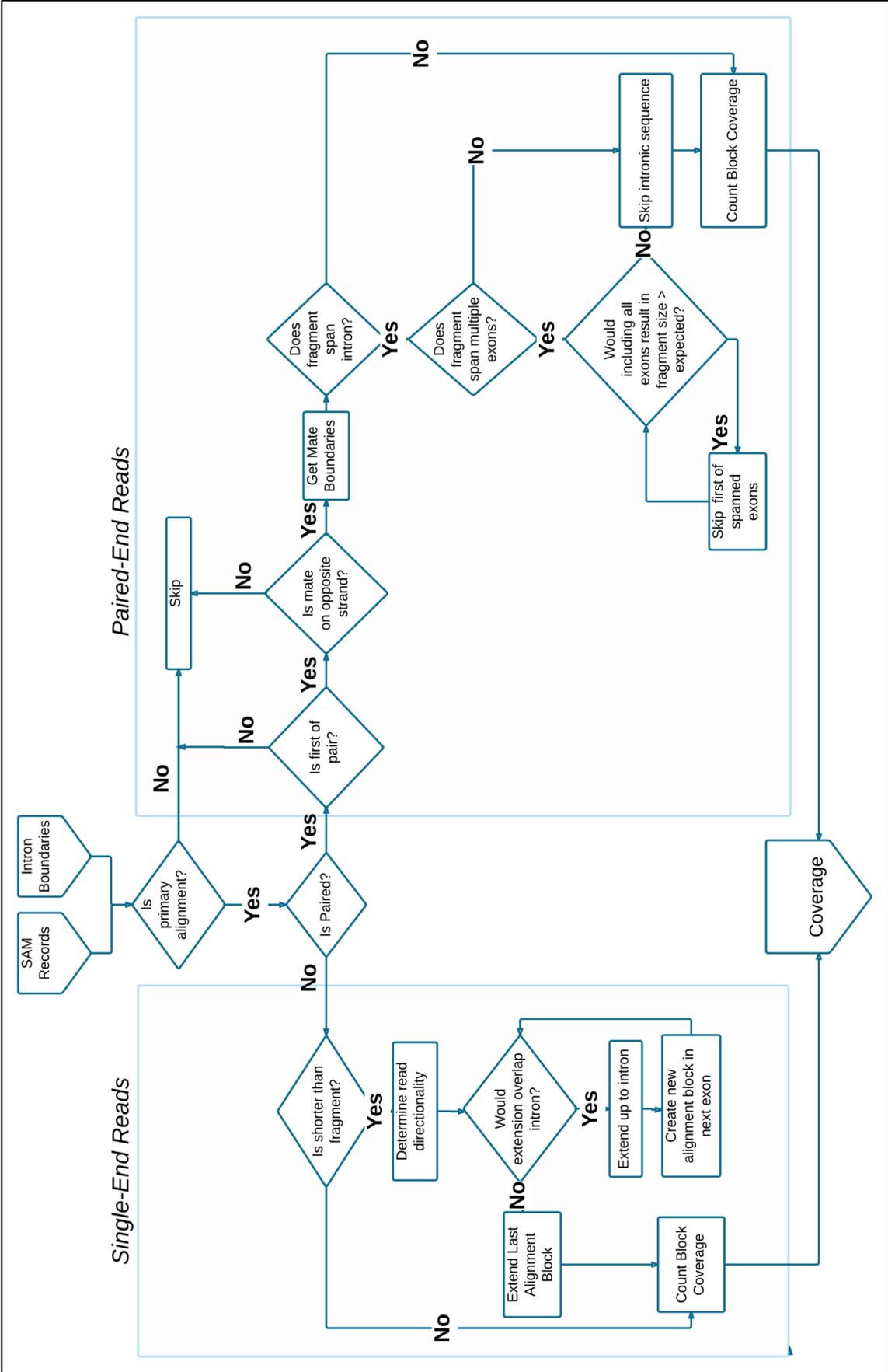
Extending single-end reads up to the average sequenced fragment size appears to be a valid strategy, although some small discrepancies remain. Indeed, modelling the length of read extension based on a real distribution of fragment lengths would yield more accurate corrections; however this can be highly variable between different samples and often this information is not available, so that modelling read size extension as a random variable is not practical.



**Figure 40.** Simulated read coverage distributions. Simulations were performed using m6A-seq data with average 100bp fragment size, sequenced using non-stranded library protocol with 100bp paired-end reads. Non-stranded single-end reads were simulated by filtering out the second mate. Stranded single-end reads were simulated by filtering out the mate mapped to position the coding strand. Coverage data presented here is smoothed to aid clarity. **A.** 100bp unstranded reads, average 100bp fragment size **B.** Artificially trimmed (after alignment, to discount aligner effects) to 50bp, unstranded reads, average 100bp fragment size. **C.** Artificially trimmed to 50bp, stranded reads, average 100bp fragment size. **D.** Artificially trimmed (after alignment, to discount aligner effects) to 50bp, unstranded reads, average 100bp fragment size. Single-end read fragment coverage illustrates the distribution obtained using the read shifting, rather than extension strategy.

An alternative to single-end read extension towards the 3' end is fragment shifting, a strategy frequently used in ChIP-seq peak callers (Zhang et al. 2008), as well as the m<sup>6</sup>A peak caller exomePeak (Meng et al. 2013). **Figure 40D** illustrates the effects of this approach on single-end read coverage distribution. While the coverage distribution accurately captures the centre of the peak and corrects the bimodal distribution observed in unadjusted data, the resulting peak does not accurately capture the real fragment coverage, which can be inferred from paired-end reads. The peak is 'slimmer', but most importantly, coverage at the summit is lower, which has a direct impact on peak detection, potentially resulting in an increased false negative rate.

Based on these considerations, a robust approach to obtaining an accurate representation of the true RNA fragment coverage distribution can be formulated, that is a compromise between stringency and computational complexity. This is summarised in **Figure 41**. Aligned sequence input data is assumed to have undergone desired quality control steps, however, m6aViewer implements optional, user-configurable (disabled by default) filtering of poor quality alignments, as well as reads flagged as PCR/optical duplicates. Only two quality checks are performed by m6aViewer by default – read secondary alignment and paired-end read proper pair checks, as these types of reads are not routinely filtered out from aligned data and can directly impact the m6aViewer algorithm. In order to correctly infer fragment boundaries from paired-end data, both mates must be mapped, and must be mapped in a correct orientation. m6aViewer performs this check and discards any pairs violating this requirement, as these likely arise due to alignment errors. Such alignments can also arise due to circular transcripts, genomic duplications or trans-splicing/chimeric transcripts; however, the detection of m<sup>6</sup>A in these rare RNA species is beyond the scope of this work.



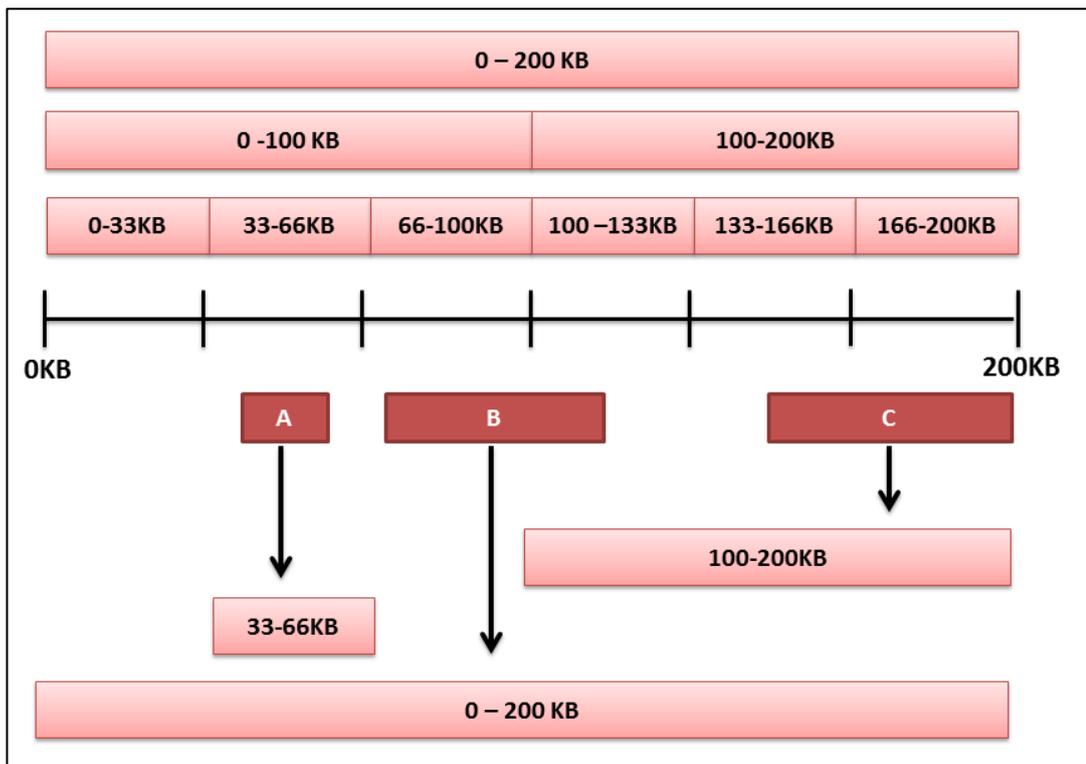
**Figure 41.** Outline of the strategy adopted for processing single- and paired-end sequencing reads to obtain an accurate representation of the true RNA fragment coverage distribution. The term 'block' is used to refer to a single continuous block of aligned sequence with respect to reference and will most commonly consist of a single read or the sequence between the boundaries of paired-end mates; however, in the case of split-reads, a single read may consist of several alignment blocks. This is also the case where alignment blocks would be artificially extended past an exon boundary into intronic sequence, or (presumed) alternatively spliced exon.

After filtering, for a given set of sequence reads, coverage is estimated iteratively by determining the boundaries of alignment blocks generated by each read (or pair of reads) and incrementing covered positions in a depth of coverage array.

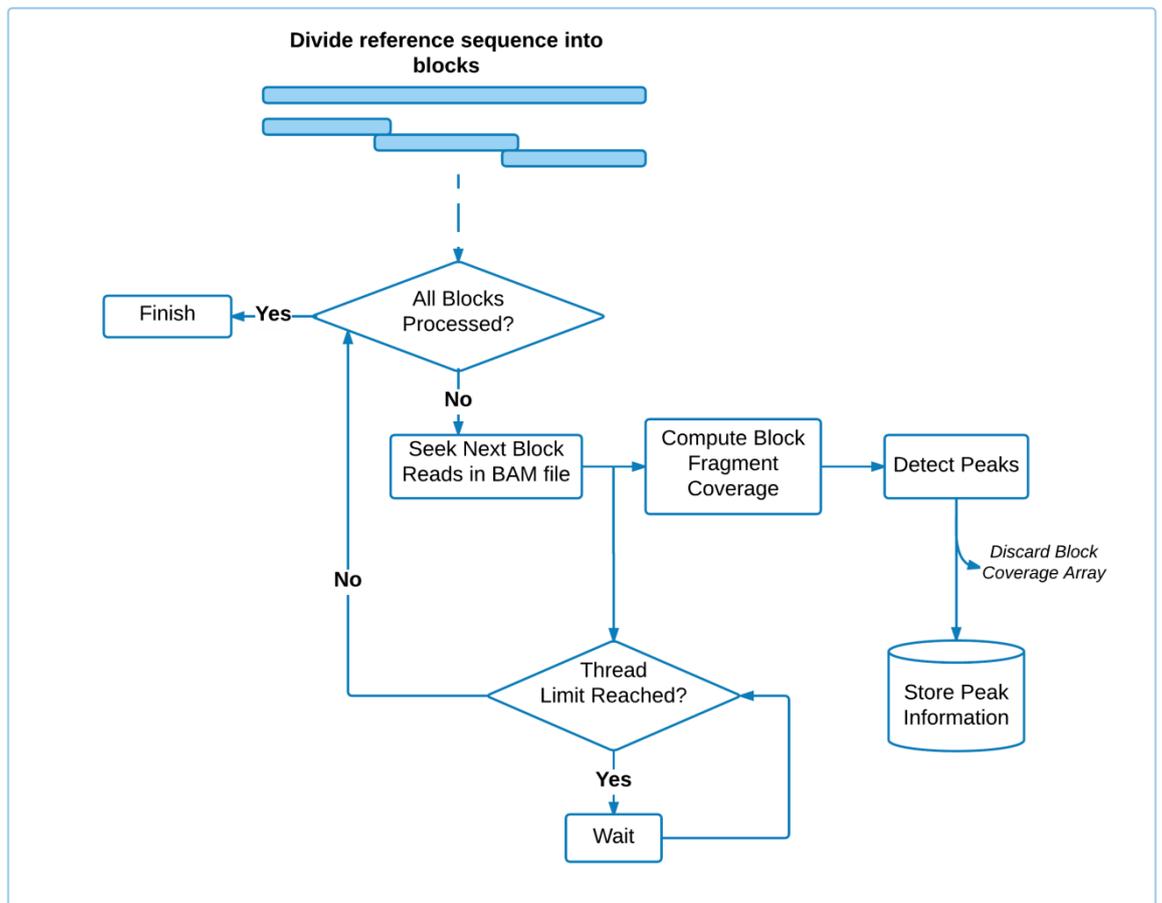
This approach requires additional considerations for management of computational resources. Processing the entire dataset at once is infeasible for all but the smallest genomes. The human genome, for example, consists of approximately 3 billion base pairs and as such, in order to process a single m<sup>6</sup>A-seq sample, one would need to keep track of some 6 billion integers (for both immunoprecipitated and control coverage). This would require a substantial amount of RAM (~4GB per billion 32 bit integers), which is not readily available in most desktop computers. Furthermore, in Java (and many other programming languages) integers are used to keep track of array indexes, thus providing an upper limit to the length of any single array (2,147,483,648). While no single human chromosome is that big, this is certainly not true for all other organisms (Pellicer et al. 2010). Memory requirements can be managed by processing the data in smaller chunks. As the alignment file is read sequentially, this will require the reads to be sorted; otherwise the SAM file would need to be read more than once.

This sequential block design, however, does not easily allow taking advantage of modern multi-core processors. Here, BAM format files are used to facilitate parallel data processing, BAM files are more compact than their SAM counterparts and have been widely adopted by researchers as the definitive

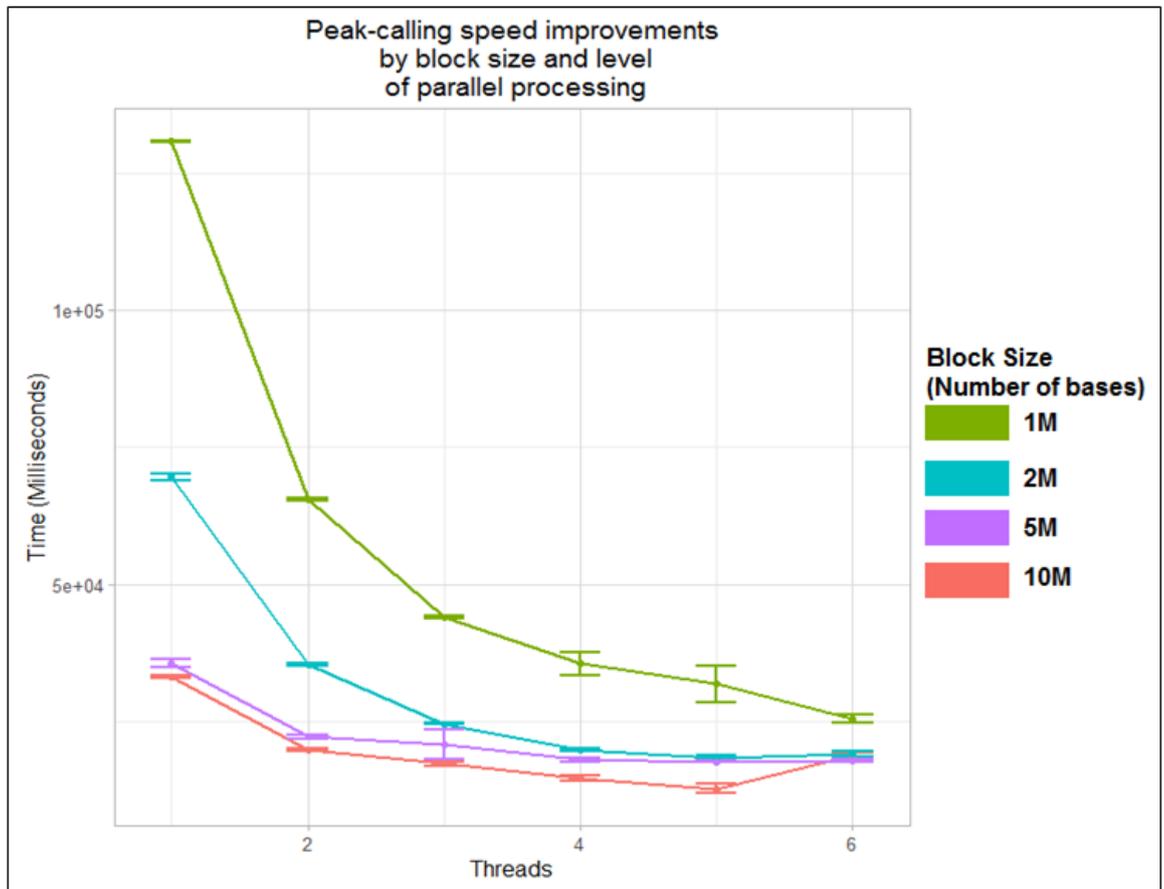
sequence alignment file format. As BAM files are the compressed counterpart to SAM files, they use BGZF block compression format, such that each compressed block is no bigger than 64 kilobytes. Blocks can be indexed by storing file offsets in a BAI file format and thus can be used for random access. In a sorted and indexed BAM file, a binning strategy is employed to allow easy identification of data blocks which contain reads overlapping a specific region in the reference sequence. This approach, adopted from the database access optimisation used by the UCSC Genome Browser (Karolchik et al. 2004), utilises an interval tree type data structure. The reference is subdivided into smaller blocks in a hierarchical manner and each read alignment is placed in the block which can contain it in its entirety (**Figure 42**). In order to retrieve all alignments in the region of interest, one needs to retrieve overlapped blocks. While introduction of a 'jagged' data structure would improve read binning, typically almost all reads will still be assigned to the smallest bins, and therefore examining bins at the top of the hierarchy will not be a costly operation. This indexing structure allows for fast random access of reads for any given region in the reference, overcoming the inherent limitations of traditional file reading methods, which are limited to streaming data in a sequential manner. Utilising BAM file random access, the reference is subdivided into smaller blocks and each can be processed in parallel, thus limiting memory use by not processing the whole dataset at once and improving speed by taking advantage of parallel processing (**Figure 43**). A similar strategy is used for retrieval of local coverage for real-time data visualisation.



**Figure 42.** An illustrative example of the aligned read binning scheme used to sort and index BAM files. Given a genomic region of 200KB, a hierarchical binning scheme can be devised. Each aligned read is placed in the smallest bin which can wholly contain it – in the example here, read **A** can be contained in its entirety in the 33-66KB bin, read **B** must be placed in the top tier bin, while read **C** is placed in the 100-200KB bin.



**Figure 43A.** Computational resources are managed by subdividing the reference into smaller blocks in order to limit memory use. Peak calling for several blocks can be performed in parallel to take advantage of multi-core processing.



**Figure 43B.** Increasing both block size and parallelisation can increase peak-calling speed by an order of magnitude. Similar gains in speed can be achieved by initially doubling the available memory (block size) or the number of parallel processing threads (1 Thread, 1M Block: mean= 130772; 2 Threads, 1M Block: mean=65427; 1 Thread, 2M Block: mean= 69560.6). Both increasing memory use (block size) and the number of processing threads have diminishing returns.

### 5.2.3 m<sup>6</sup>A peak-calling

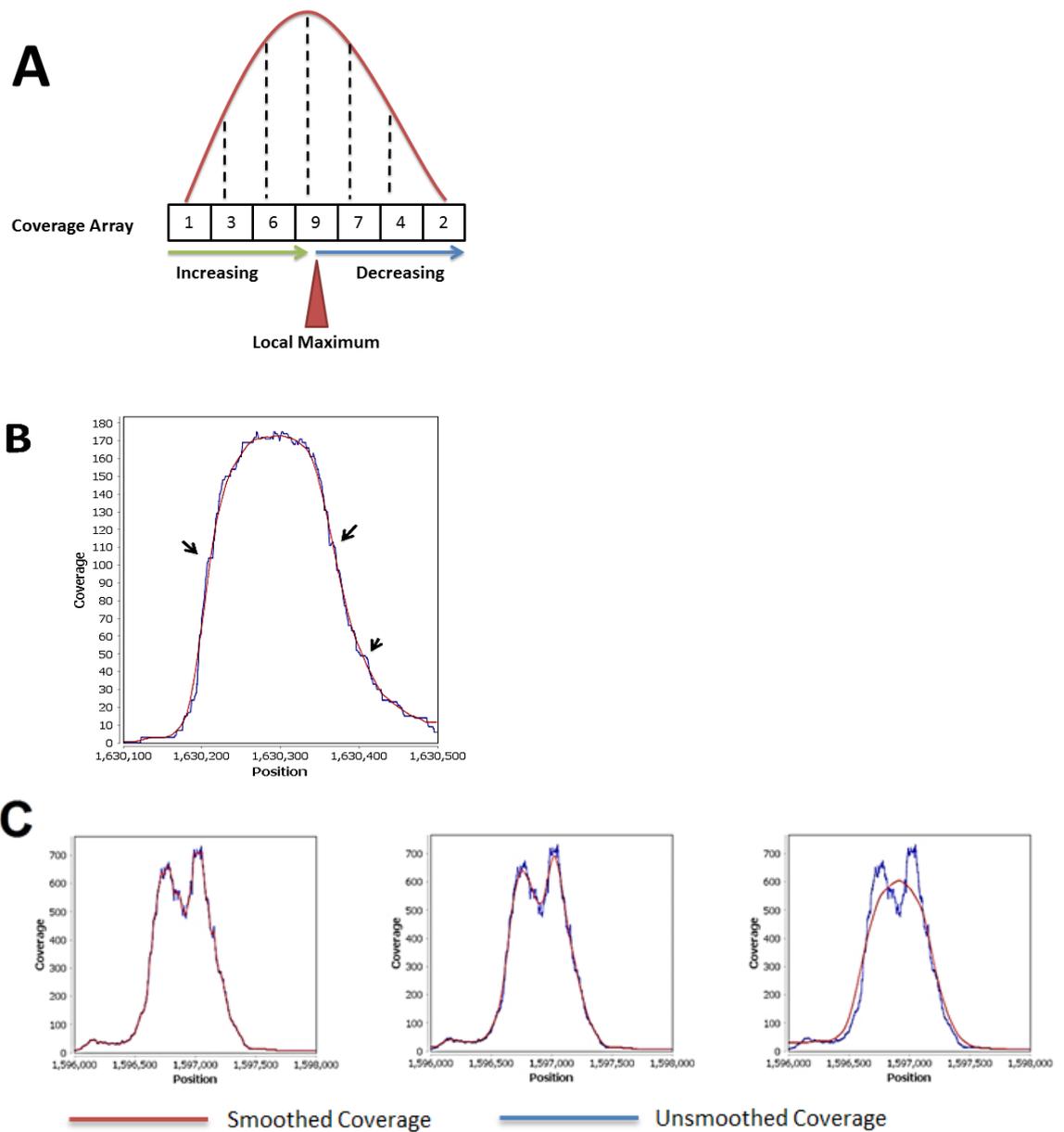
A common strategy for m<sup>6</sup>A peak calling uses binning, where, in general, the reference is subdivided into small blocks and each block is tested for statistically significant coverage enrichment over control. Here, an alternative to the binning method is proposed that relies directly on the distinct 'peak' shape of the distribution to identify m<sup>6</sup>A peaks.

Initial candidate peak positions are identified by finding all the local maxima in the coverage distribution. This is done by scanning the coverage array and detecting a change in gradient (**Figure 44A**). In practice, however, the coverage data is noisy, with small irregularities blurring the bell-shaped distribution signal and introducing small local maxima (**Figure 44B**). Thus, coverage data is first smoothed in order to remove small confounding signals that can affect true peak detection. Data smoothing methods have been widely studied. The simplest and perhaps oldest method is the mean sliding window approach, where each point,  $C_x$ , is smoothed out using  $n$  data points preceding it:

$$C_x = \frac{1}{n} \sum_{i=0}^{n-1} C_{x-i}$$

Or, alternatively, surrounding it:

$$C_x = \frac{1}{n} \sum_{i=-(n-1)/2}^{(n-1)/2} C_{x-i}$$



**Figure 44.** **A.** Peaks can be detected from the coverage distribution by determining local maxima as the point of change between increasing and decreasing gradient in the coverage array. **B.** A peak from chromosome 1. Blue line indicates the raw coverage distribution; red shows the smoothed coverage distribution. Black arrows indicate where small local maxima are present in the unsmoothed data and would confound true peak detection. **C.** The effects of increasing smoothing window size (from left to right: 10nt window size, 50nt window size; 200nt window size) on the data. Small window sizes fail to eliminate small, confounding local maxima, while window sizes that are too large can result in genuine signal being lost.

This approach has several advantages - firstly, it is computationally simple, which is crucial in this case, as due to the size of typical datasets, even trivial computations, when applied to a whole transcriptome-wide dataset, can result in significantly increased running times. Secondly, while the decision is largely arbitrary, the level of smoothing required can be easily achieved by adjusting the bandwidth parameter  $n$ . It is important that a smoothing method is able to reduce the impact of noise in the data, without losing the signal. **Figure 44C** shows how varying the bandwidth can affect the smoothing of the coverage data. While a small  $n$  size may not be enough to smooth out the noise in the coverage data, the coverage data becomes increasingly flattened as the bandwidth grows. At the extremes, this can result in the loss of genuine peak signal (**Figure 44C, right**).

Here, a mean smoothing window size of 20 nt is used – one empirically determined to be sufficient for smoothing out small local variation, while preserving the overall peak signal. In some cases, however, this is not sufficient to remove all confounding signal from the coverage distribution. Consequently, a look-behind mechanism is implemented to identify and merge local maxima which have been detected in very close proximity.

Local maxima (peaks) and local minima (valleys between overlapping peaks) are subsequently identified from the smoothed coverage by searching the coverage array for gradient inversion events. A potential peak is thus initially defined as a position  $i$  in the smoothed coverage array  $C$ , where  $C_i - C_{i-1}, \dots, i-n > 0$  and  $C_i - C_{i+1}, \dots, i+m > 0$ , where  $n$  and  $m$  are either the last prior (or first subsequent) gradient change event position detected (if greater than  $1/10^{\text{th}}$  of the expected peak width, to account for overlapping peaks but also prevent detection of small irregularities in coverage distribution), half the expected peak width or last (or next) position with 0 read coverage, whichever occurs closest. For cases where peak summits are flat – i.e. local maxima spans multiple bases - the putative peak position is defined as the central point.

Each local maximum identified is tested against the null hypothesis that the read distribution in the immunoprecipitated sample is not higher than that in the control using Fisher's Exact test. The total number of fragments for the peak

region in IP (immunoprecipitated) and INPUT (control) samples are counted as the number of fragments aligning to (but not necessarily wholly contained within) the peak region. The peak region is defined as the region encompassing the number of bases equal to the sequenced fragment length to each side of the detected maximum; in cases of peak overlap, the region boundary to the overlapping side(s) of the peak is defined as a mid-point between the two peaks. Respective contingency tables are thus computed from the total IP and INPUT fragments at the putative peak position and the total IP and INPUT library size. The p-value is then computed as:

$$p = \frac{(R_{IP} + L_{IP})! (R_{INPUT} + L_{INPUT})! (R_{IP} + R_{INPUT})! (L_{IP} + L_{INPUT})!}{R_{IP}! R_{INPUT}! L_{IP}! L_{INPUT}! (R_{IP} + R_{INPUT} + L_{IP} + L_{INPUT})!}$$

where  $R$  is the number of reads at a putative m<sup>6</sup>A site and  $L$  is all other aligned reads in the library. Alternatively, local background can be used in this calculation instead, using fragment counts at peak position and total reads aligning to the respective transcript.

The data is then subjected to the Benjamini-Hochberg (Benjamini and Hochberg 1995) correction, to account for the multiple testing bias. Multiple testing corrections aim to recalculate the probabilities obtained from repeated, independent applications of statistical tests. One of the earliest - and still commonly used- multiple testing corrections is the Bonferroni (Armstrong 2014) correction. However, Bonferroni correction is particularly stringent, and while it does effectively control type I errors, it can result in an increase of false negative calls. Here, Benjamini-Hochberg is used as an alternative, less conservative multiple-testing correction, which aims to strike a balance between limiting type I and type II errors. In a list containing  $n$  sorted (smallest to largest) p-values, the adjusted probability  $P(x)$  is computed as:

$$P(x) = \frac{P(x)*n}{i}, \quad i = n, n - 1, \dots 1;$$

Where  $i$  is the rank of the value in the list. Unlike in the Bonferroni correction, where all p-values are adjusted universally, here the most significant hits are corrected more conservatively than less significant hits.

Alternatively, FDR can be estimated and controlled by m6aViewer by treating the INPUT sample as IP and performing peak detection in order to obtain an empirical p-value distribution from the switched IP and INPUT samples. The FDR of peak p-values can then be estimated from the obtained distribution and represents the chance of seeing an equivalent read enrichment in the RNA-Seq control data. While this is not an ideal measure, in practice it provides a cut-off that is less stringent than Bonferroni correction, but more stringent than Benjamini-Hotchberg.

The peak-calling strategy described here has several advantages. Firstly, detecting peaks based on the shape of the distribution results in increased peak calling resolution. Each peak can be identified as a single nucleotide position representing the peak summit, whereas binning-based approaches will result in significant regions which can span several kilobases in length. This poses problems for both downstream comparative data analysis and wet lab peak validation.

As discussed by Cui *et al* (2015), the binning strategy models enrichment within individual segments as independent, an assumption which is not generally correct. Working with the entire coverage distribution avoids this problem posed by segmentation entirely. Indeed, this approach can be seen as the reverse to that adopted by ChIP-Seq peak caller MACS – MACS detects significantly enriched regions and then attempts to refine these by finding peak summits, whereas here peaks are detected in the immunoprecipitated sample first, and tested for statistically significant enrichment afterwards. This type of approach allows one to differentiate (to some degree) multiple peaks in close proximity - a phenomenon that has been shown to be fairly common (Linder *et al*. 2015). MACS attempts to refine peak summits to a single nucleotide resolution call by finding the highest coverage position within identified significantly enriched region. However, in the case of overlapping peaks, this in general will only identify the peak summit position with the highest coverage, disregarding any putative presence of multiple m<sup>6</sup>A residues in the region.

## 5.2.4 m<sup>6</sup>A peak deconvolution

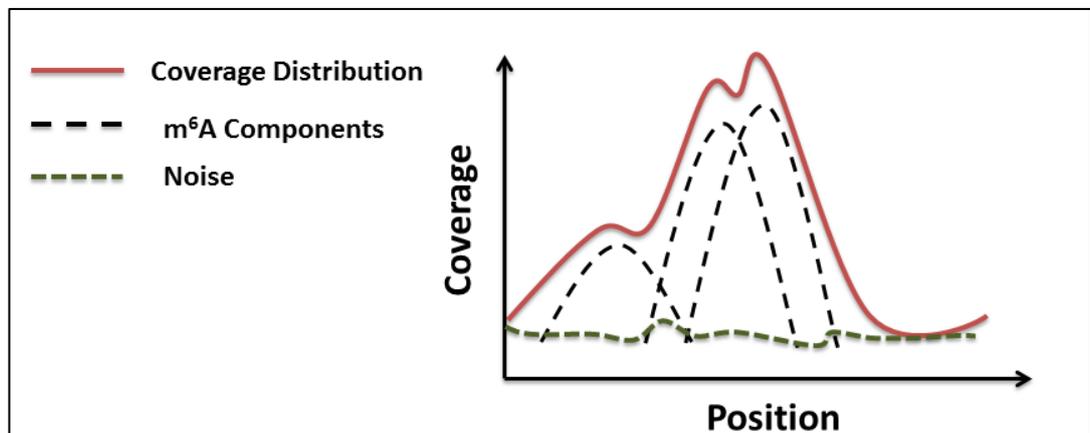
### 5.2.4.1 Probabilistic aligned read modelling using expectation maximisation

Peaks in RNA fragment coverage distribution arising from multiple m<sup>6</sup>A sites in close proximity can often be visually indistinguishable from single sites. Indeed, Linder *et al* ( 2015) observed that multiple methylated adenosines are often present in close proximity. Conversely, the summit of detected single m<sup>6</sup>A peaks rarely corresponds precisely to the site of methylation (Linder *et al.* 2015).

Improving peak-calling resolution is important for downstream validation of m<sup>6</sup>A sites via PCR as well as additional experiments, such as identification of m<sup>6</sup>A reader proteins. Furthermore, it is important to separate individual m<sup>6</sup>A sites in experiments which aim to detect methylation changes across different conditions. A window-based approach, such as that employed by tools like exomePeak (Meng *et al.* 2013), could fail to identify methylation changes in windows which encompass multiple m<sup>6</sup>A residues if these changes are not uniform across all sites. Here, an approach is described that attempts to improve m<sup>6</sup>A peak-calling resolution, as well as deconvolute several m<sup>6</sup>A residues in close proximity that results in overlapping peaks. This method is implemented in m6aViewer software as an alternative peak-calling mode that increases peak-calling accuracy at the cost of increased software running time.

When considering the aggregated RNA fragment coverage alone, positional information on individual RNA fragments is lost, which could be used to inform m<sup>6</sup>A residue positions. Consider a region enriched in immunoprecipitated RNA fragment coverage, such as shown in an illustrative example in **Figure 45**. Reads aligning to this region are a product of one of several possible cases – antibody binding to one or more m<sup>6</sup>A residues; antibody binding to RNA non-specifically; free RNA or DNA fragment contamination that can arise from RNA “sticking” to beads or other surfaces; or erroneous read alignment (due to poor read quality, low complexity regions, etc.). The latter cases can be considered noise, and with the possible exception of non-specific antibody binding, should constitute a minor fraction of all the reads aligning to an enriched region. Thus, the observed coverage distribution can be seen as a mixture of noise and one

or more m<sup>6</sup>A components. If these components could be deconvoluted, multiple m<sup>6</sup>A sites in close proximity, as well as individual m<sup>6</sup>A sites, could be identified more accurately.

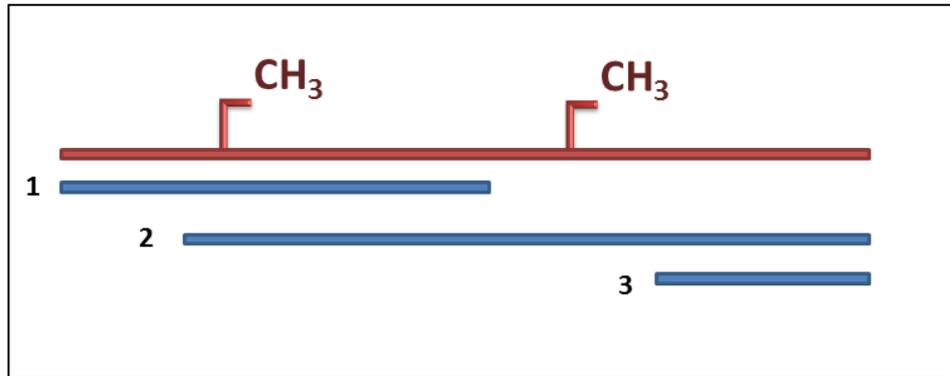


**Figure 45.** Diagram representation of an enriched IP region as a mixture of several m<sup>6</sup>A components and noise reads.

Given multiple m<sup>6</sup>A sites in close proximity, it is possible then to assign each aligned fragment a probability that represents how likely it is that we see it as a result of antibody binding to each possible m<sup>6</sup>A position or as a result of noise. Thus, if m<sup>6</sup>A positions were known, a probabilistic approach could be used to assign each data point (fragment) to each m<sup>6</sup>A site. The reverse is also true - given the probability distribution of all fragments, unknown m<sup>6</sup>A positions could be inferred. Thus, for a given region modelled in such a probabilistic way, m<sup>6</sup>A positions can be more accurately called by finding the combination of adenosines in the reference sequence which gives rise to the highest overall model likelihood.

The problem can then be framed as one of maximum likelihood – what combination of components explains the observed RNA fragment distribution best? A naïve approach would be to consider all adenosines in a given region and compute the likelihood of all the possible combinations, choosing the highest. However, this would be prohibitively computationally expensive, resulting in factorial algorithmic complexity ( $O(n!)$ ) and certainly would not be scalable to whole transcriptomes – a 1kb region containing 250 adenosines and

up to four putative m<sup>6</sup>A sites, for example, would require 161,487,125 possible models to be assessed.



**Figure 46.** An example of immunoprecipitated RNA fragment (blue) alignment to a reference sequence (red) containing two m<sup>6</sup>A sites. Here, fragment 1 is likely to arise due to antibody binding to the leftmost m<sup>6</sup>A position; fragment 2 could be a result of antibody binding to either (or both) m<sup>6</sup>A position; while fragment 3 is likely to be noise, as it does not overlap any of the m<sup>6</sup>A sites.

Instead, Expectation Maximisation (EM) algorithm (Dempster et al. 1977; Do and Batzoglou 2008) can be used to find a solution in a reasonable time by iteratively computing the likeliest m<sup>6</sup>A sites and fitting the observed fragment distribution to the model, until the algorithm converges to a solution. EM is often used to find maximum likelihood-based approximations of parameters in probabilistic models, and here m<sup>6</sup>A positions can be seen as parameters to be estimated in a probabilistic RNA fragment distribution model.

For a given enriched region, let  $X = (x_0, \dots, x_{k-1})$  be a  $k$ -sized vector of sequenced RNA fragments aligned to the region, drawn from an unknown mixture of size  $n$  of  $D$  distributions, each representing either an m<sup>6</sup>A site or a noise read cluster. We wish to find a set of parameters  $\theta$  that maximise the log likelihood function:

$$L(\theta) = \ln P(X|\theta)$$

Where  $\theta$  consists of m<sup>6</sup>A positions and noise cluster  $C \in \{c_0, \dots, c_{n-1}\}$  and a corresponding prior probability vector  $S \in \{s_0, \dots, s_{n-1}\}$  where  $\sum_{i=0}^n S_i = 1$  and  $0 \leq$

$S_i \leq 1$ . Here, both the RNA region ‘sequencability’ and differing m<sup>6</sup>A stoichiometry (the proportion of methylated RNA molecules out of the total pool of RNAs) are accounted for by incorporation of prior probabilities, as both influence the observed peak height.

Expectation maximisation approach can iteratively estimate  $\theta$  while maximising  $L(\theta)$  and consists of two steps. During the expectation step, posterior probabilities of each mapped RNA fragment arising from each putative m<sup>6</sup>A position in C are computed, given the estimated parameters  $\theta$  at step  $t$ .

$$P(x_i \in C_j \mid \theta_t) = \frac{S_j \cdot P(x_i \mid D, C_j)}{\sum_{j=0}^n S_j \cdot P(x_i \mid D, C_j)}$$

The maximisation step then re-estimates parameters  $\theta_t$  from the posterior probabilities obtained during the expectation step. Each prior probability at step  $t$  is estimated as:

$$S_{j,t} = \sum_{i=0}^k P(x_i \mid S_{j,t-1}, C_{j,t-1}) / k$$

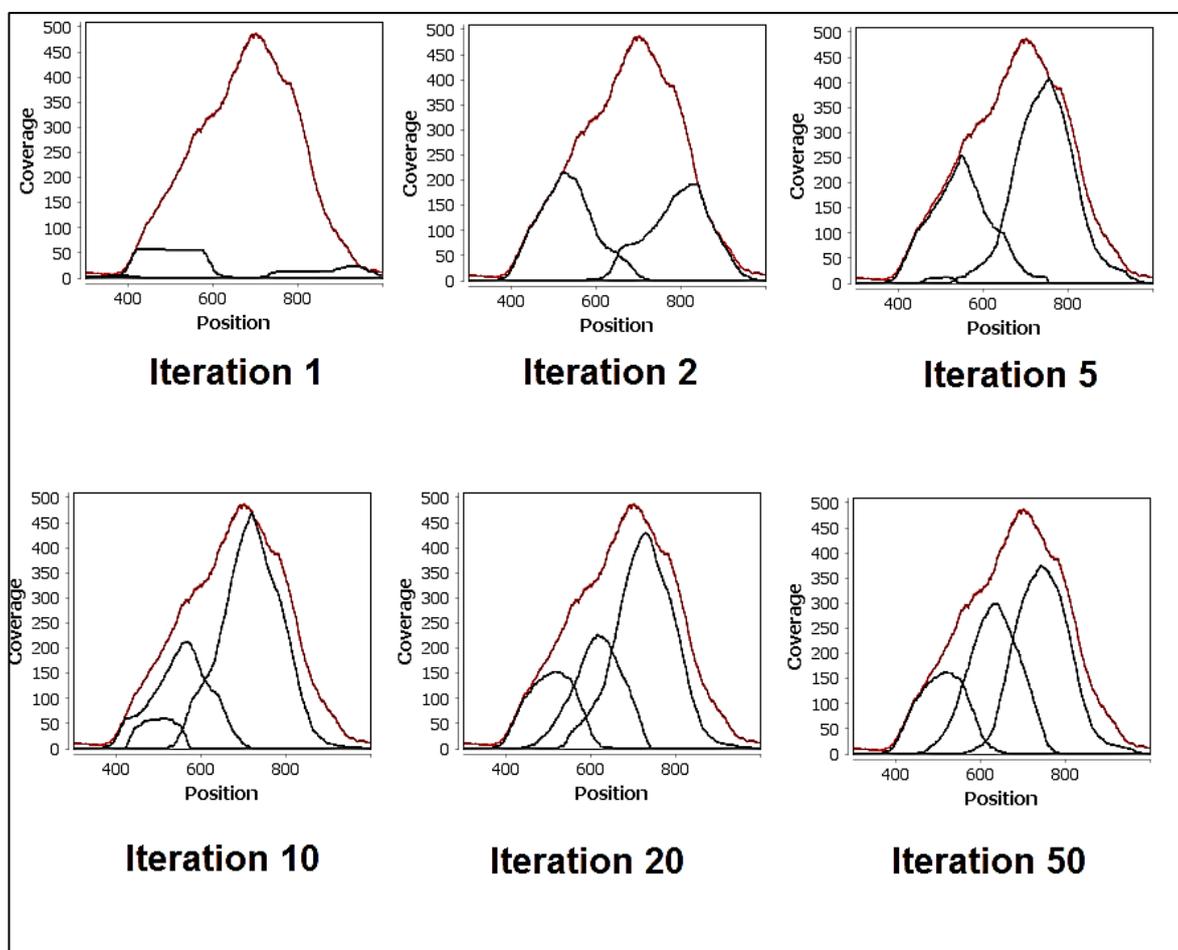
and each m<sup>6</sup>A position as:

$$C_{j,t} = \frac{\sum_{i=0}^k P(x_i \mid S_{j,t-1}, C_{j,t-1}) \cdot C_{j,t-1}}{\sum_{i=0}^k P(x_i \mid S_{j,t-1}, C_{j,t-1})}$$

The process repeats for a set number of iterations, or (in most cases) until the algorithm has converged when  $L(\theta_t) - L(\theta_{t-1}) < 0.01$ , where the likelihood at step  $t$  is computed as:

$$L(\theta_t) = \ln\left(\prod_{i=0}^k \max_{j=0; j < n} P(x_i \mid C_j, S_j)\right)$$

Where each sequenced read fragment is effectively assigned to either the likeliest m<sup>6</sup>A position cluster or a noise cluster. This iterative procedure is visualised in a simulated example in **Figure 47**.

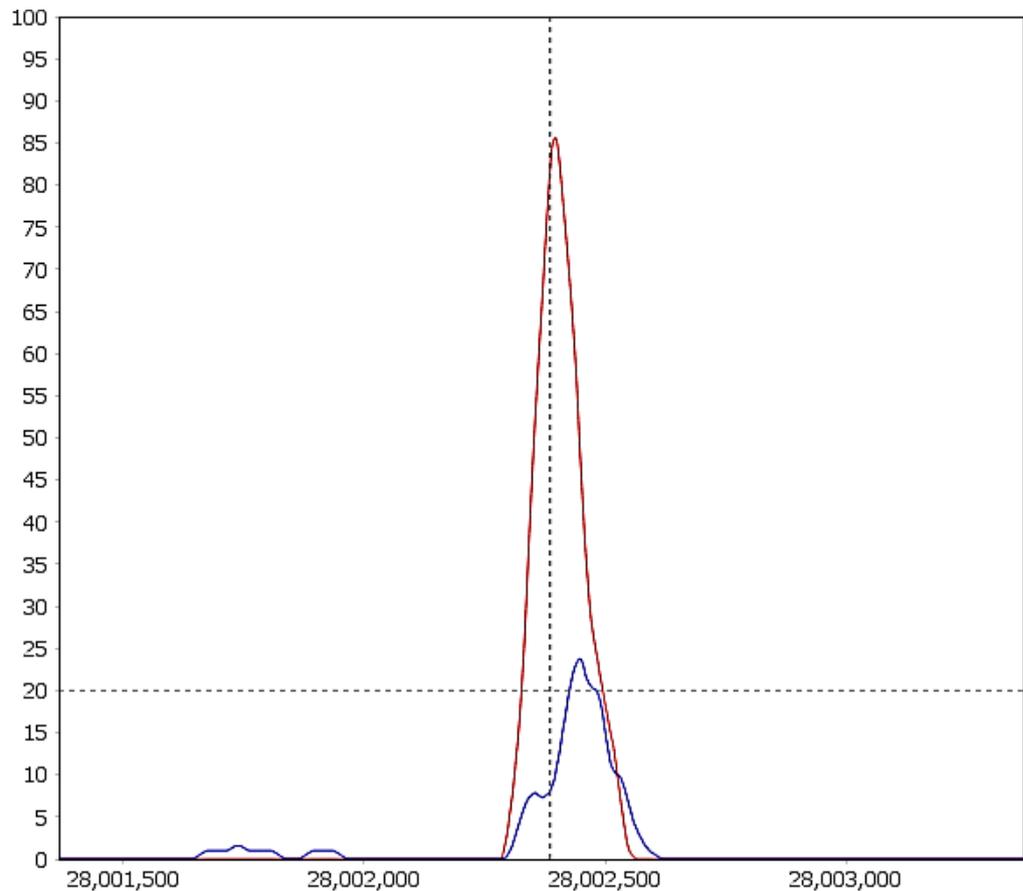


**Figure 47.** Iterative fitting of multiple peaks using EM algorithm on simulated data. 1000 RNA fragments were simulated to result from m<sup>6</sup>A residues at positions 500, 600 and 750 with 20%, 35% and 40% corresponding probabilities to simulate differing stoichiometry and 5% randomly placed noise reads. Fragment lengths were drawn randomly from a normal distribution to introduce variability. EM was initialised at 3 random positions with equal priors. Total simulated fragment coverage is shown in red, while peaks fitted during each iteration are shown in black. Noise read cluster is omitted to aid clarity.

#### 5.2.4.2 Estimating EM probabilities

In order to approach the problem of m<sup>6</sup>A peak deconvolution in a probabilistic way, a robust way of estimating the probability that a fragment is seen as a result of antibody recognition of a specific site is required. We can assess this in terms of how well each fragment supports the expected coverage distribution around the m<sup>6</sup>A site. In the case of an aligned fragment that does not overlap a putative m<sup>6</sup>A site, this probability is effectively zero. For all other fragments, this is more complicated – consider example 2 in **Figure 46**: given only one aligned fragment as evidence, under the (inaccurate) assumption of a random fragmentation and antibody binding process, both the putative overlapped positions are equally likely to be methylated. However, we know that the stochastic process that generates these observations is governed by a non-uniform latent variable –the actual m<sup>6</sup>A distribution. Thus, the posterior probability that the observed fragment was generated by antibody binding to a putative m<sup>6</sup>A position is non-uniform and can be modelled by including this prior.

In order to take into account any antibody-binding biases, an empirical distribution is obtained from the data by considering a set of training RNA fragments and how they fall in relation to known m<sup>6</sup>A positions. As ground truth, a set of high confidence m<sup>6</sup>A positions was selected from the data reported by Linder *et al*'s (Linder et al. 2015) single nucleotide resolution m<sup>6</sup>A map in HEK293T cells. The residues selected are not in proximity to other m<sup>6</sup>A sites which could confound the results, and form clear, distinct single peaks in a HEK293T paired-end m<sup>6</sup>A-seq dataset (Schwartz et al. 2014). All fragments overlapping these sites in m<sup>6</sup>A-seq dataset are selected, resulting in 10058 training fragments. An example training peak is shown in **Figure 48**.



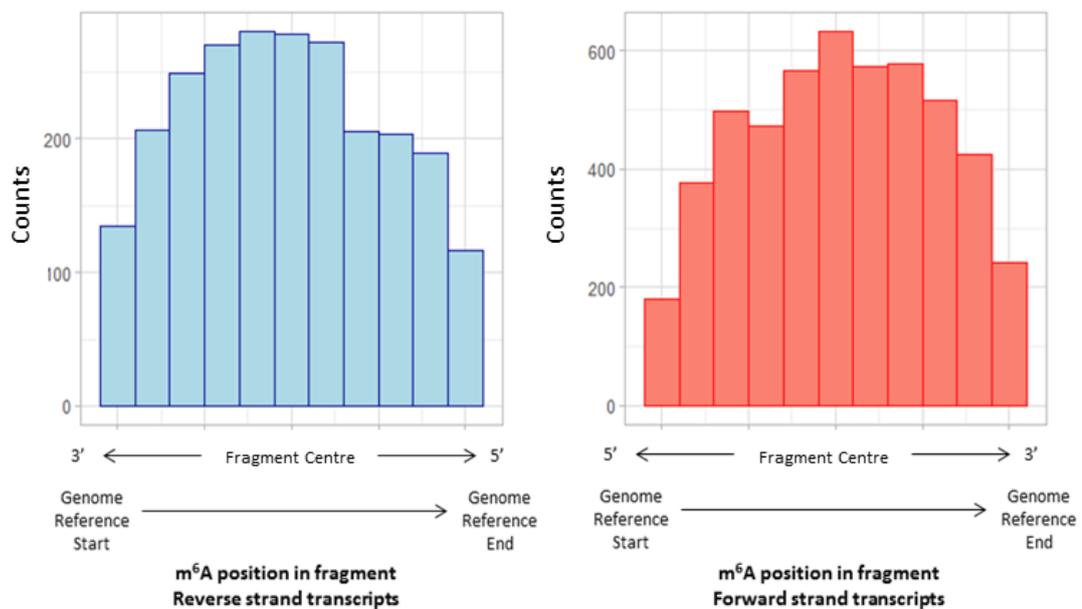
**Figure 48.** An example training peak from HEK293T cell line m<sup>6</sup>A-seq dataset also reported by Linder et al, 2015. Immunoprecipitated coverage is shown in red, while control RNA-seq coverage is in blue. Vertical dotted line indicates nearest RRACH consensus, while horizontal dotted line indicates the minimum coverage threshold requirement for peak selection.

Given the training fragments, probability density function estimation can be obtained from the frequency distribution and subsequently used to estimate probabilities for other RNA fragments.

Under the assumption of a random fragmentation and antibody binding process and no enhanced degradation of isolated, fragmented RNA, the position of m<sup>6</sup>A within each sequenced fragment should be random and uniformly distributed. However, this is not the case. **Figure 49** shows the distribution of m<sup>6</sup>A positions with respect to the sequenced fragments: m<sup>6</sup>A residues are depleted near fragment ends but there is a clear enrichment near the centre of the fragment, with a slight bias towards the 3' end. This bias is

evident when considering transcript strandedness – sense and anti-sense strand transcripts show mirrored distributions with respect to genomic reference. Thus, to account for this bias, each fragment probability is computed with respect to the transcript directionality, rather than uniformly across the reference.

There could be a number of reasons for this non-uniform distribution. A natural explanation for m<sup>6</sup>A depletion at sequenced fragment ends is simply due to some m<sup>6</sup>A residues being close to the end of the transcript. However, this does not appear to be the case - m<sup>6</sup>A positions near transcript ends show the same skewed distribution as m<sup>6</sup>A positions central to the transcript (**Figure 50**).



**Figure 49.** The distribution of m<sup>6</sup>A positions within sequenced RNA fragments in reverse strand transcripts (**left**) and forward strand transcripts (**right**).



**Figure 50.** Comparison of distributions of m<sup>6</sup>A positions within sequenced RNA fragments. 3'UTR and 5'UTR encompasses m<sup>6</sup>A positions within 200 base pairs of transcription start and end sites, while CDS encompasses m<sup>6</sup>A positions central within the transcript. Fragment centre is indicated as a dashed line.

The distribution could also arise due to technical bias, such as RNA fragmentation. While a number of approaches exist for DNA and RNA fragmentation, none fragment the molecule in a truly random fashion. Enzymatic fragmentation can result in sequence-specific cleavage biases - for example, RNase III specifically cleaves double stranded RNA. Chemical fragmentation – although requiring an additional end-repair step- is more homogenous across a transcript, and as m<sup>6</sup>A-Seq protocols described by Dominissini *et al* (2012) and Meyer *et al* (2012) (as well as the data used here) use zinc chloride fragmentation buffer, fragmentation bias should be minimal.

Alternatively, it is likely that the antibody shows some preference for m<sup>6</sup>A positions away from fragment ends. This leads to a concern that it is possible that different anti-m<sup>6</sup>A antibodies generate somewhat different fragment distribution profiles in relation to m<sup>6</sup>A position and this could lead to inaccuracies in the model. Furthermore, in the case of polyclonal antibodies, different batches may introduce significant variation that could be difficult to account for. The most commonly used antibody for m<sup>6</sup>A-seq is a rabbit polyclonal antibody provided by Synaptic Systems (Dominissini et al. 2012; Meyer et al. 2012; Batista et al. 2014; Luo et al. 2014) originally developed by Munns *et al* (1977), although alternatives exist. Zhou *et al* (2015) used the rabbit polyclonal antibody ABE572 from Millipore, while Meyer *et al* (2012) used a different rabbit polyclonal antibody developed at New England Biolabs (Kong et al. 2000). While commercial monoclonal m<sup>6</sup>A antibodies exist (for instance, ENZ-ABS301-0100 Enzo Life Sciences), to the best knowledge of the author, these have not yet been used in m<sup>6</sup>A-seq studies published at the time of the writing of this thesis. A comparison of fragment distributions generated by different antibodies would be particularly useful as it is a potential source of bias, however only Synaptic Systems antibody paired-end m<sup>6</sup>A-seq data was publicly available at the time of writing.

#### **5.2.4.3 Expectation Maximisation initialisation**

It is important to initialise EM calculations with good starting values – poor starting values in particular can result in the algorithm becoming ‘trapped’ at local maxima and therefore failing to converge to the global maximum (Wang and Zhang; Maitra 2009; Melnykov and Melnykov 2012). EM is frequently initialised randomly, while more robust approaches adopt multiple restart strategies (Melnykov and Melnykov 2012); however, this can be computationally expensive, which is a major concern when trying to apply this approach to whole transcriptome data. Consequently, a two-step data-guided approach is employed that does not require EM to be run multiple times, while ensuring good initialisation values, unlike the random start approach.

As m<sup>6</sup>A positions are more likely to occur in regions which have high fragment coverage, initially  $n$  positions in a region are chosen based on the

RNA fragment coverage values, where  $n$  is the number of peaks to be fitted. This selection is done iteratively, such that each read encompassing the previously selected positions is not counted towards the next position; this strategy prevents initialisation of multiple positions per peak. The initial selection is then refined based on the reference sequence, as methylation is more likely to occur at a RRACH motif and must occur at an adenosine. All reference positions for 'A', 'AC' and 'RRACH' are extracted, with more weight being given to 'RRACH' and 'AC' motifs than just adenosines. Motif weights ensure that a 'RRACH' motif, for example, 10 bases away will be prioritised over 'AC' 5 away, but if the motif is too far, the nearest 'AC' or 'A' is used instead. Optimal adjustment of initial coverage-based positions can then be formulated as the assignment problem, where each initial position needs to be matched to the closest motif position in a manner which minimises the total adjustment distance for all positions. Here, this is solved with the Hungarian algorithm (Munkres 2006) using a distance matrix constructed to represent the bipartite graph between coverage-based positions and sequence motif positions.

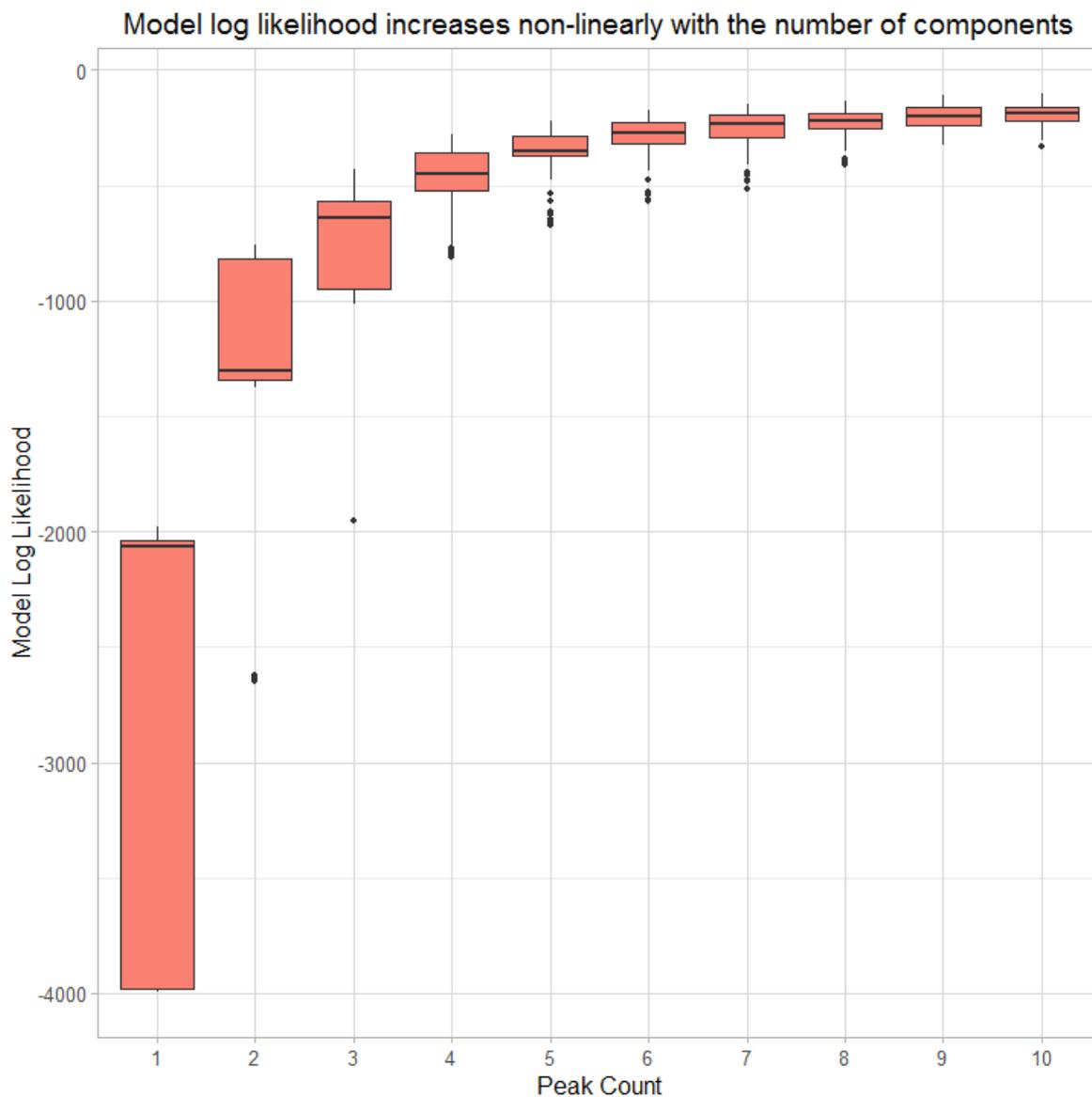
#### ***5.2.4.4 Expectation Maximisation – how many peaks?***

Lastly, the final parameter that needs to be estimated is the number of peaks to be fitted to any given region. For any given mixture model, model likelihood increases (non-linearly) with the number of mixture components and can result in over-fitting. This is illustrated in **Figure 51**, where the simulated data in **Figure 47** is initialised with increasing number of peaks, using 100 random starts at each increment.

In this case, for any given region, the maximum model likelihood could be achieved by fitting an m<sup>6</sup>A position for each base in the region, as this would maximise the probabilities for each individual data point. This is clearly a nonsensical result. Thus, a method is needed to select an optimal number of components that takes into account model complexity in addition to likelihood.

This problem has been widely studied, as it is applicable not only in the context of EM, but also any other unsupervised algorithms that could be used for a clustering task. Specifically, how many clusters are in the data is a fundamental problem and many different approaches have been described

(Pelleg and Moore 2000; Hamerly 2007; Fraley and Raftery; Gupta et al. 2010; Steele and Raftery 2010)



**Figure 51.** Model likelihood increases non-linearly with the number of components introduced. Simulated data was used as in **Figure 47**, with 3 peaks being the 'ground truth'. The boxplots show the distribution of log likelihood obtained by initialising each iteration with 100 random starts.

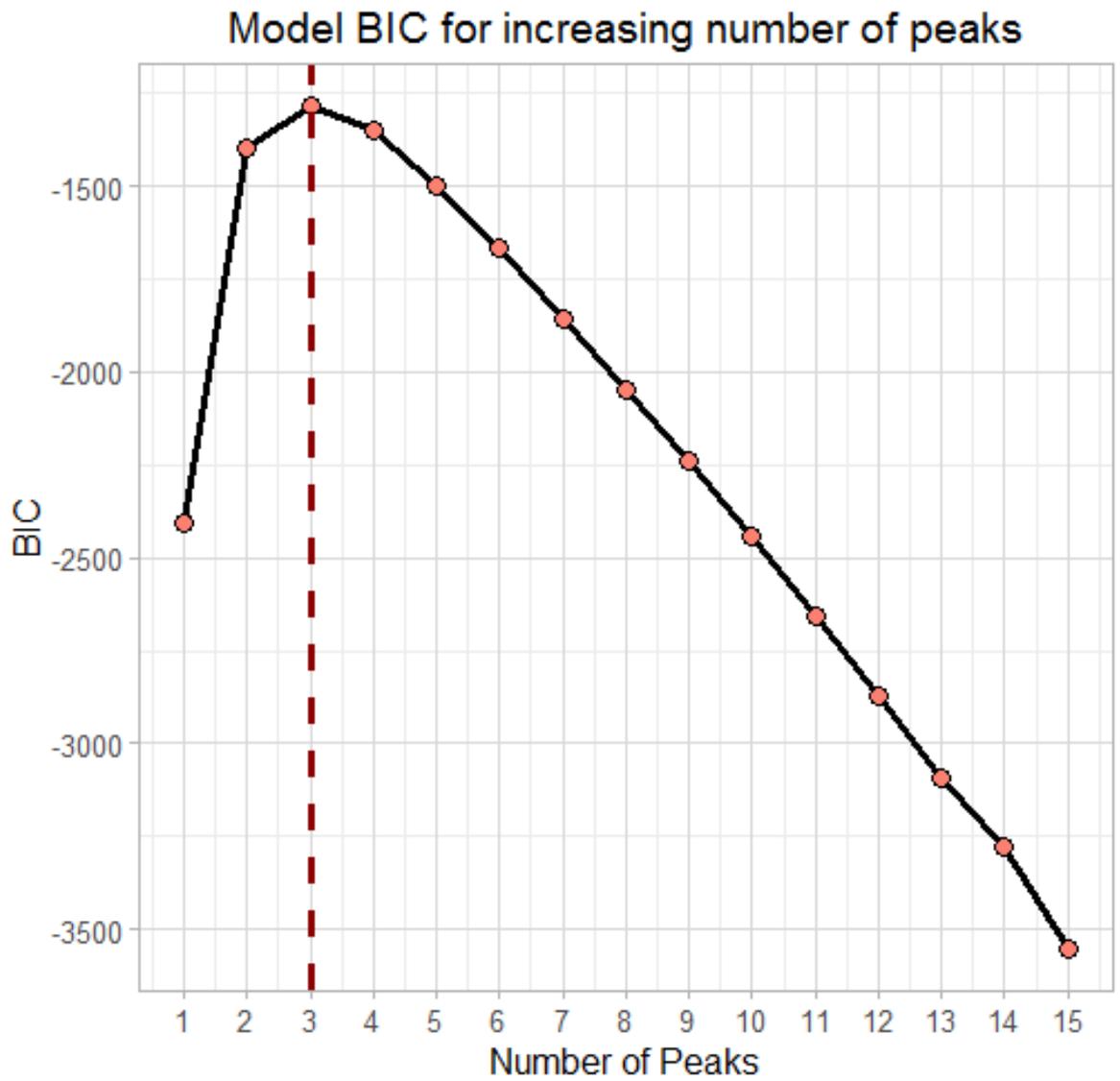
Here, Bayesian information criteria (Schwarz 1978) (BIC) is used to adjust model likelihood scores, as it has been previously shown to perform well in the

context of EM (Hirose et al. 2011; Steele and Raftery 2010). Effectively, BIC introduces a penalty for increasing the number of parameters in the model and can be calculated as:

$$BIC = -\ln(LL) + k \cdot 2\ln(n)$$

Where LL is the likelihood of the model, k is the number of parameters and n is the number of data points. Dagsupta and Raftery (Dasgupta and Raftery 1998) suggested computing BIC for the multiple possible models with increasing cluster counts and choosing the one that corresponds to the first BIC maximum. This is illustrated in **Figure 51**, where the first maximum is found at 3 peaks, as expected. Effectively, this identifies the point where the increase in model likelihood no longer compensates for the increase in model complexity penalty.

While Dagsupta and Raftery (Dasgupta and Raftery 1998) suggest estimating a maximum of likely amount of true clusters for the data and computing BIC for all the resulting models, here models are computed iteratively only until a maximum is detected in order to save computation times. That is, in example in **Figure 52**, only 4 models would be computed instead of the 15 shown.



**Figure 52.** The first maximum BIC indicates the optimum trade-off between model complexity and model likelihood and can be used as a selection criteria for the number of peaks to be fitted to any given region. BIC was calculated from the maximum likelihood model out of 100 random starts at each number of peaks shown in **Figure 51**.

## **5.2.5 Technical False Positive m<sup>6</sup>A peak identification**

### **5.2.5.1 Motivation**

False positive peaks can constitute a large proportion of all detected peaks and are problematic in the analysis of m<sup>6</sup>A experiments. These can be expected to arise from technical variation, such as non-specific antibody binding; a major concern for any antibody-based technique. It has been postulated that for m<sup>6</sup>A-seq experiments, the false positive rate may be particularly high (Schwartz et al. 2014). Schwartz *et al* (2014) had noted that upon RNA methyltransferase knockdown, despite global depletion of m<sup>6</sup>A levels seen via TLC, m<sup>6</sup>A-seq data did not show the expected universal reduction in detected peaks. A subset of peaks which did not exhibit any reduction in immunoprecipitated read enrichment was identified, and was found to be enriched in degenerate, purine-rich motifs instead of the RRACH motif expected at m<sup>6</sup>A sites, indicating that a high proportion of these sites were likely to comprise non-specific antibody binding sites. It is concerning that such a large proportion of all reported m<sup>6</sup>A residues in literature may actually be false positives; however, it may be possible to differentiate these sites from real methylated adenosines using computational approaches.

### **5.2.5.2 Training Data**

In order to create a model that would allow identification of false positive peaks, a definite set of examples from both real m<sup>6</sup>A sites and false positive peaks is required. However, no such dataset is available – even m<sup>6</sup>A positions identified by Linder *et al* (2015) cannot be considered true positives, as these were also obtained using an antibody-based technique and therefore are likely to contain a subset of sites arising due to non-specific antibody binding. Thus, a set of training sites were obtained from the RNA methyltransferase complex knockout data by Schwartz *et al* (2014), as explained below, with the understanding that the training data is likely to contain some mislabelled instances.

RNA methyltransferase knockdown and matched control m<sup>6</sup>A-seq data from HEK293T cells, A549 cells and mouse fibroblast cells were downloaded from ArrayExpress (Kolesnikov et al. 2015), aligned using the STAR aligner (Dobin et

al. 2013) to either human hg19 or mouse mm10 reference genomes, and sorted and indexed using Samtools (Li et al. 2009). Peak-calling was performed by m6aViewer running in default mode, as the model-based peak-calling involves a sequence-based initialisation step which may bias the results. Matched knockdown and control sample sites were intersected, and peaks labelled as true positive m<sup>6</sup>A sites if a comparable m<sup>6</sup>A peak was not detected in the knockdown sample and the gene expression level of the transcript has not decreased so as to prevent detection of the peak. Similarly, a peak was labelled as a technical false positive only if the change in peak enrichment levels between the knockdown and the control was less than 0.5 fold. For the purposes of intersecting the samples, peaks in different samples were considered to be the same site if located within 50 nt of each other. On the other hand, two sites were considered independent sites if they were detected further than 200 nt apart in the matched samples. Peaks between 50 and 200 nt apart in two samples were considered ambiguous and therefore excluded from the dataset in order to obtain the highest quality training set. Using this approach, a high confidence HEK293T cell line dataset was created, comprising 2098 peaks; of which 1030 are false positive instances and 1068 true positive instances. The datasets obtained from A549 and mouse fibroblast cells was reserved solely for independent testing.

### ***5.2.5.3 Sequence-based model***

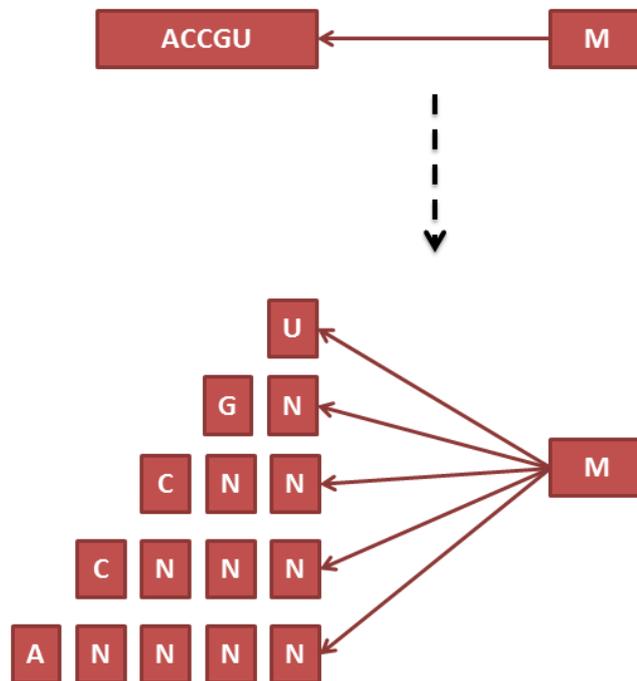
In order to ascertain whether true m<sup>6</sup>A sites could be differentiated from false positives using only unbiased features that are independent of external data/annotations, an RNA sequence-only model was initially considered. A sequence-only model is attractive in that it can be applied universally in an unbiased manner, requiring only the knowledge of the transcriptome sequence. As such, for each peak in the training dataset, the 400 base pair RNA sequence surrounding the peak was obtained with each training sequence represented as a combination of characters A, G, C, U and M, where M represents the putative methylated adenosine position in the sequence. Rare occurrences of sequences with ambiguous bases were excluded from the training data. If false positive peak and genuine m<sup>6</sup>A site sequences exhibit intrinsically different

sequence features, such nucleotide composition, periodicity or sequence motifs, these differences could be captured using a Markov model.

Markov chains are sequences of a random variable  $X$ , where the probability distribution for  $X$  depends only on  $X_{t-1}, X_{t-2}, \dots, X_{t-n}$ , and as such are ideally suited for representing RNA sequences. In this case, a 1<sup>st</sup> order Markov chain would model the probability of observing any given base with reference to only the preceding base, whereas a 5<sup>th</sup> order Markov chain would consider a preceding 5 base sequence (5-mer) instead. Higher order dependencies are thus computationally difficult to model, as the number of required parameters grows exponentially with increasing order and even relatively low order chains can become impossible to compute. There is also the matter of training data requirements – in order to accurately estimate the transition probability matrix, all kmers must be observed in the training data set at a sufficient frequency. As the kmer length increases, however, the probability of observing it in any given sequence decreases dramatically, and in some genomes/transcriptomes many kmers occur only rarely.

A number of algorithms have been described that can help alleviate these issues. An interpolation approach has often been applied to the gene finding problem (Salzberg et al. 1999, 1998), where rather than considering a fixed order Markov chain, the kmer length is variable and dependent on the training data. This largely overcomes the uncertainty of estimating the transition probability matrix values for most large kmers, while still capturing sequences which are genuinely overrepresented in the data. On the other hand, this approach does little to alleviate the computational requirements of the model, as the number of model parameters (i.e. the transition probability matrix) that require estimation is still very large. Mixture Transition Distribution (Berchtold and Raftery 2002) (MTD) models have been utilised in sequence modelling of DNA methylation sites (Seifert et al. 2012) and provide an alternative way of estimating the transition probabilities for high order Markov chains. Rather than computing the probability of observing a particular base after a specific kmer, which cannot be estimated accurately when the kmer is large, the probability is estimated as a combination of different 'lag' probabilities (**Figure 53**). This is a

computationally attractive model, as the transition probability matrix required grows linearly with increasing Markov chain order, rather than exponentially.



**Figure 53.** Mixture transition distribution model lag dependency schematic diagram. Instead of the probability of each base (M in this example) depending only the preceding kmer,  $P(M | ACCGU)$  is instead estimated as a combination of  $P(M | U)$ ,  $P(M | GN)$ ,  $P(M |, CNN)$ ,  $P(M |CNNN)$  ,  $P( M | ANNNN)$ .

Formally, given all possible bases  $X = (X_0 = A, X_1 = C, X_2 = G, X_3 = U, X_4 = M)$ , an MTD model can be defined as:

$$P(X_t | X_{t-1}, \dots, X_{t-n}) = \sum_{l=1}^n w_l P(X_t | X_{t-l})$$

Where  $X_t$  is the observed base at step  $t$ ,  $n$  is the order of the model and  $w$  is a lag weight, where the lag weights are constrained as follows, to ensure the results follow a probability distribution:

$$\sum_{l=1}^n w_l = 1; w_l \geq 0$$

For a  $k^{\text{th}}$  order model,  $k$  transition probability matrices are thus computed, storing the probability distribution of each lag separately and requiring only  $k |X| (|X| - 1)$  values and increasing the order of the model requires only a single additional matrix.

In this case, representing RNA sequences as a Markov process may not be strictly accurate, as this approach models sequence dependencies only in the 5' -> 3' direction. However, spatial sequence information may be more informative as we are not trying to model a transcription/translation process where this directionality is important. Consider a 2<sup>nd</sup> order Markov model for m<sup>6</sup>A sequences – given a sequence GG-M-CU, where methylation 'M' occurs within a known consensus RRACH, the probability of M would be estimated using only the preceding 'GG' – however, the following bases 'CT' are also important predictors. Similarly, given the putative roles of m<sup>6</sup>A as RBP 'switches', m<sup>6</sup>A may be surrounded by regions which have high degree of sequence complementarity and thus it is important to capture this. As such, a small extension is introduced where instead of defining the probability of observing a base  $X$  at  $t$  only as being dependent on the preceding bases  $P(X_t | X_{t-1}, \dots, X_{t-n})$ , surrounding bases are  $P(X_t | X_{t-1}, \dots, X_{t-n}, X_{t+1}, \dots, X_{t+n})$  are considered instead.

In order to obtain the model that best describes the training data, model parameters need to be optimised such that the log-likelihood of the model is maximised. The log likelihood can be defined as the sum of the log likelihood of all training sequences, where the log likelihood of a sequence of length  $m$  is:

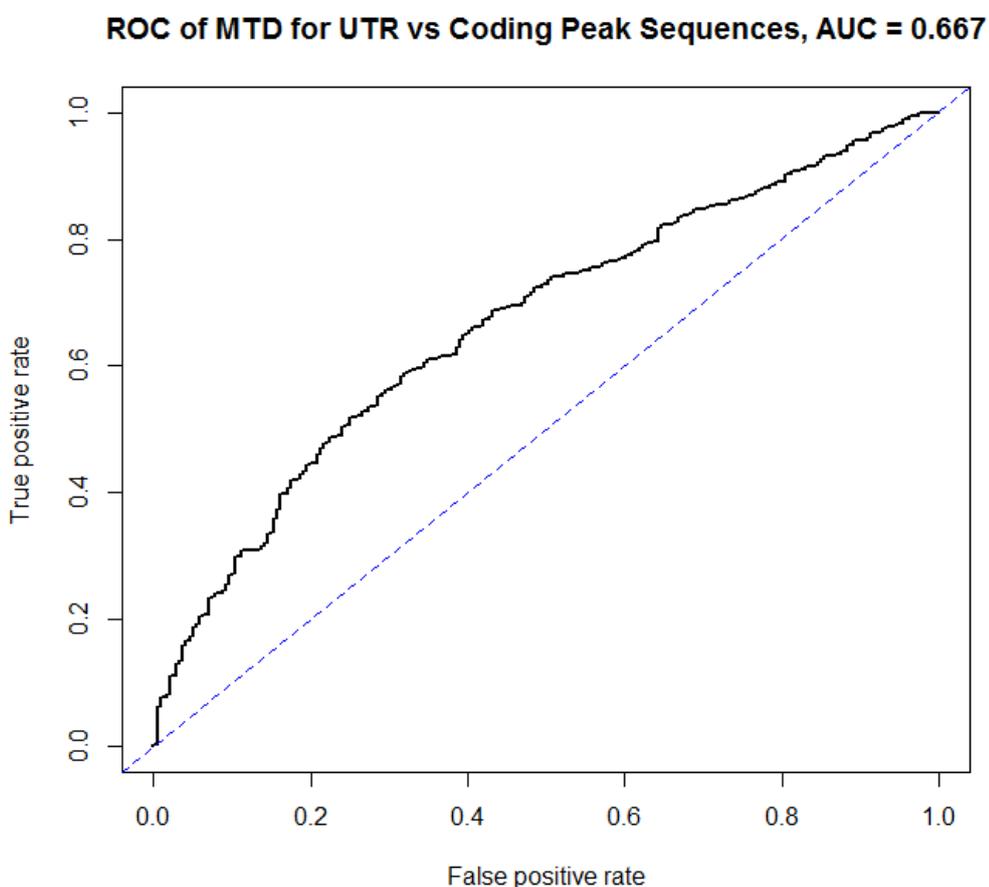
$$LL = \sum_{t=1}^m \log \left( \sum_{l=1}^n w_l P(X_t | X_{t-l}) \right)$$

While a number of parameter estimation methods for MTD models have been described, here the original approach described by Berchtold (2001) is used to optimise the lag weights and transition probability matrices. As the lag weights and transition probability matrix rows must sum to 1 to satisfy the probability constraints, the increase in one parameter must be balanced by the decrease in another. An iterative procedure is thus used that at each step modifies two elements of the lag weight vector and two elements of each row in the transition matrix (increasing one and decreasing the other). The procedure iterates until the increase in model log-likelihood between iterations becomes negligible.

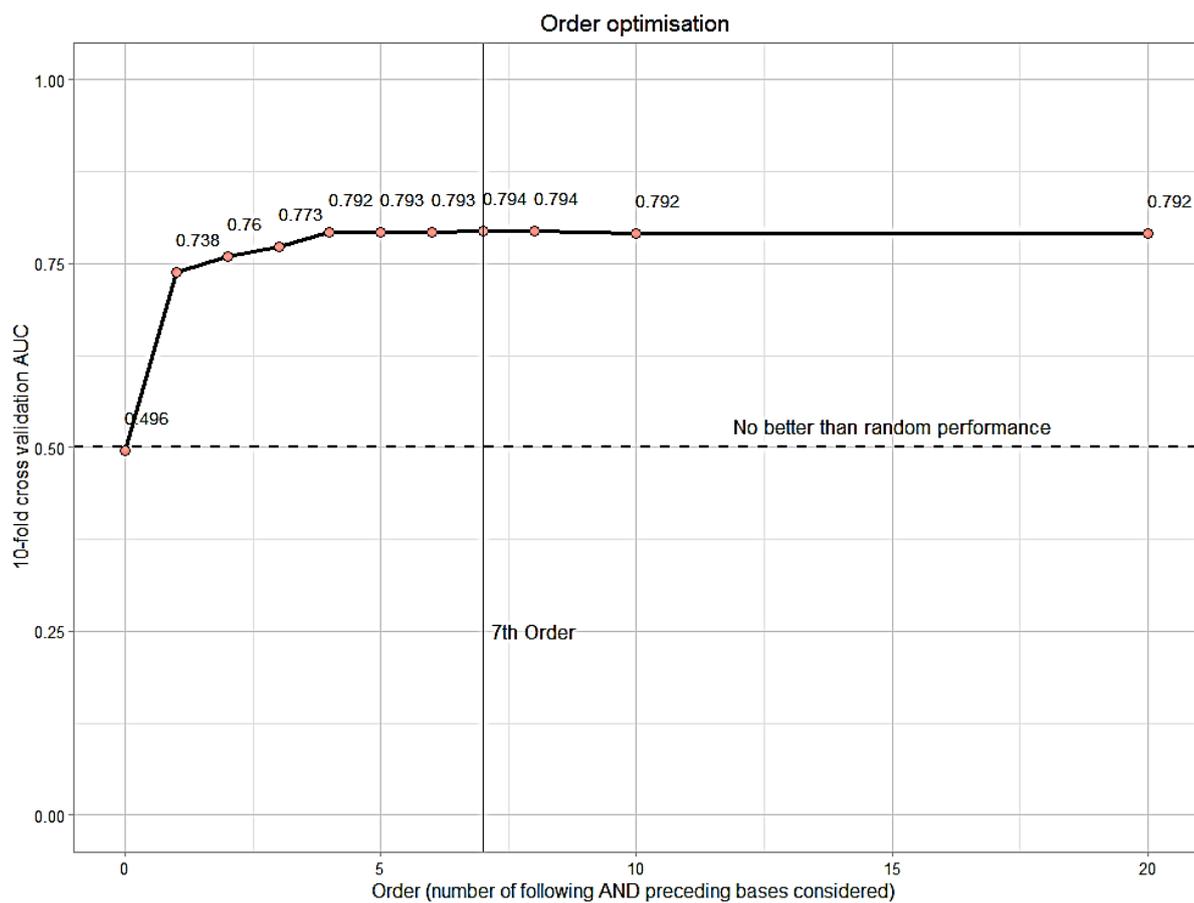
As m<sup>6</sup>A is more likely to occur in UTR regions, there is a danger that using only two classes (true m<sup>6</sup>A and false positive m<sup>6</sup>A) would result in a model which effectively classifies sequences into coding and non-coding, rather than real m<sup>6</sup>A and non-specific binding sites because of this bias. This can be illustrated in **Figure 54**, where a 3<sup>rd</sup> order MTD model was trained using the true positive dataset with peak sequences additionally separated into coding and non-coding subsets. Using 10-fold cross-validation, coding and non-coding peak sequences could be classified with better than random accuracy (AUC=0.667), which suggests that controlling for the type of peak sequence is important to avoid biases. As such, training sequences were further separated into coding, intronic, 5'UTR and 3'UTR groups in order to avoid building a model that merely captures the differences between coding and non-coding regions. However, while this helps avoid biases towards non-coding sequences, it also reduces the available training data which may have a negative impact on the accuracy of the final model.

Using the approach outlined above, multiple increasing order models were trained to ascertain the highest-performing classifier. Each iteration was assessed using a 10-fold cross-validation approach and the results are shown in **Figure 55**. The highest performance was achieved by a 7<sup>th</sup> order sequence model, with the area under the ROC curve at 0.794, where the probability of

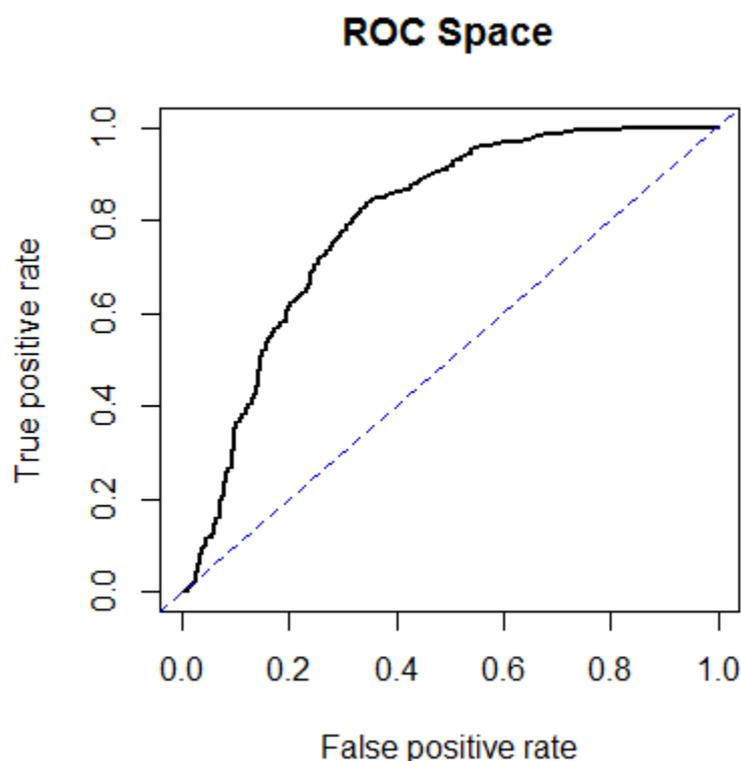
observing each base in sequence is dependent on the 7 preceding and 7 following bases. It is interesting to note that the biggest gain is achieved by the jump from 0-order sequence model, where the probability of each base in the sequence is effectively the frequency at which it is observed in the training data, to the 1<sup>st</sup> order sequence model. The AUC of the 0-order model (0.496), indicating classification performance that is no better than random, suggests that there is no sequence composition bias at single base level in the training sequences.



**Figure 54.** The ROC curve of MTD of coding vs non-coding m<sup>6</sup>A peak sequences. The dashed blue line is at AUC=0.5, indicating classifier performance that is no better than random.



**Figure 55.** The area under the ROC curve (AUC) achieved by sequence models of increasing order, assessed using a 10-fold cross validation.



**Figure 56.** The ROC curve of the 7<sup>th</sup> order MTD sequence model, AUC = 0.7936126. The dashed blue line is at AUC=0.5, indicating classifier performance that is no better than random.

The ROC curve of the 7<sup>th</sup> order model is shown in **Figure 56**. AUC of 0.794 indicates that the sequence model here can differentiate between technical false positive peak sequences and m<sup>6</sup>A peak sequences in a high proportion of all cases. However, the cost-benefit trade-off was found to be unacceptable for practical applications. Thus, to be applicable for real data m<sup>6</sup>A-seq data, the classifier requires further improvement.

#### ***5.2.5.4 Supplementing RNA sequence model with secondary RNA-structure predictions***

It has been suggested that RNA secondary structure might be important for RNA methyltransferase binding or RNA recognition by m<sup>6</sup>A reader proteins, such as in the case of m<sup>6</sup>A 'switches'. Thus, secondary structure might be an important predictor that could be used to differentiate true m<sup>6</sup>A sites from peaks which arise from non-specific antibody binding. On the other hand, the basis of non-specific antibody binding to non-methylated RNA is unknown and similar

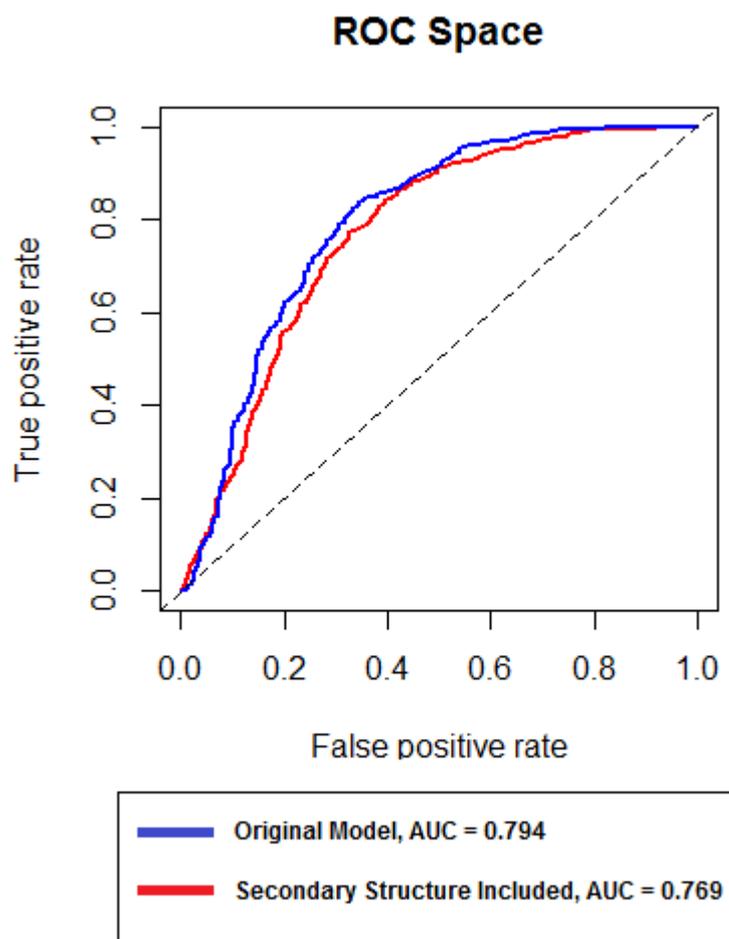
RNA secondary structure to that of actual m<sup>6</sup>A sites could play a role in it, thus negating any predictive power.

While it would be difficult to fully incorporate secondary RNA structure, the sequence-based MTD model can be easily extended to integrate the information on whether the individual bases are unpaired or paired (but not which bases they are paired with). This additional data could be important, as adenosines have been shown previously to be preferentially methylated in RNA regions of low secondary structure complexity and to be unlikely to occur in double-stranded RNA. In order to incorporate secondary sequence information into the sequence model, instead of using a dictionary which captures only the bases of RNA (A, C, G, U and M), the training sequences can instead be represented as additionally having a paired or unpaired status (Ass, Ads, Css, Cds, Gss, Gds, Uss, Uds, Mss, Mds, where ss= single stranded and ds = double stranded), thus requiring 10 different possible states to represent the sequence instead of 5.

For each training sequence, RNA secondary structure was predicted using Vienna RNA software (Lorenz et al. 2011). As predictions can be greatly affected by sequence length, the entire transcript sequence, rather than just the RNA sequence surrounding the peaks was used to obtain the secondary structure predictions. Predictions made by Vienna RNA were parsed from the 'dot-bracket' format and the new training sequences were used to retrain the sequence model.

The 10-fold cross validation results (**Figure 57**) show that incorporating secondary RNA structure predictions into the sequence model does not enhance performance. In fact, the AUC value of the more complex model is marginally lower (0.794 vs 0.769) than the sequence-only model. This result is counter-intuitive, in particular with the expectation that m<sup>6</sup>A should favour single-stranded RNA regions. The poorer predictive power when using RNA secondary structure may arise from significant errors in the predicted structures used in the analysis. Additionally, only the paired or unpaired status of bases is captured here, although additional information that is excluded from the model could be important, for example the presence of loop structures or base-pairing

information. On the other hand, introducing additional states to the model increases the spread of the training data, so that each state probability is estimated from fewer data points, which can have an additional negative impact on the performance of the model.



**Figure 57.** The comparison of performance between a sequence-only model and a model incorporating predicted RNA secondary structure information. Area under the ROC curve of 0.5 ( $x=y$ ) indicates no better than random performance, whereas area under the ROC =1 indicates perfect classification. The performance of sequence-only model was marginally better than that of sequence + predicted RNA secondary structure model.

#### 5.2.5.5 Feature-based ensemble model

As the sequence-based model could not be further improved by incorporating secondary structure based information, a feature-based model was considered

next. There are a number of features in addition to sequence information and RNA secondary structure which could be predictive of m<sup>6</sup>A status. The following features were obtained for all training sequences:

<b>Feature type</b>	<b>Possible values</b>
<b>MTD sequence model scores</b>	5'UTR, 3'UTR, CDS and Intronic sequences respectively
<b>Transcript information</b>	Coding, non-coding, length, 5'UTR length, 3'UTR length, CDS length, intron lengths, exon lengths
<b>Sequence composition</b>	2-mer, 3-mer and 4-mer frequencies
<b>Peak information</b>	Distance to nearest consensus, distance to nearest AC, peak enrichment score, peak height, peak width, total peaks in transcript.
<b>Conservation information</b>	phastCons and phyloP conservation scores for all surrounding bases from 100 vertebrate multiple alignments
<b>miRNA information</b>	mirBase data: distance to nearest miRNA site, site score, miRNA type(only for mouse and human data). Distance to miRNA seed sequence from custom miRNA list.
<b>Secondary structure information</b>	Nearest ViennaRNA-predicted hairpin, bulge and stem paired and unpaired base information.

**Table 11.** List of transcript features that could be predictive of an m<sup>6</sup>A methylation site.

A multi-class random forest learner is chosen as the model, where each training sequence is again labelled as 5'UTR, 3'UTR, CDS, Intronic or Other Non-Coding in order to avoid biases discussed earlier, resulting in 10 total classes to be predicted (m<sup>6</sup>A 5'UTR, false positive 5'UTR...etc.). A random forest model is a suitable approach to the task, as it is robust to over-fitting and remains one of the highest performing supervised learning algorithms for a wide variety of tasks.

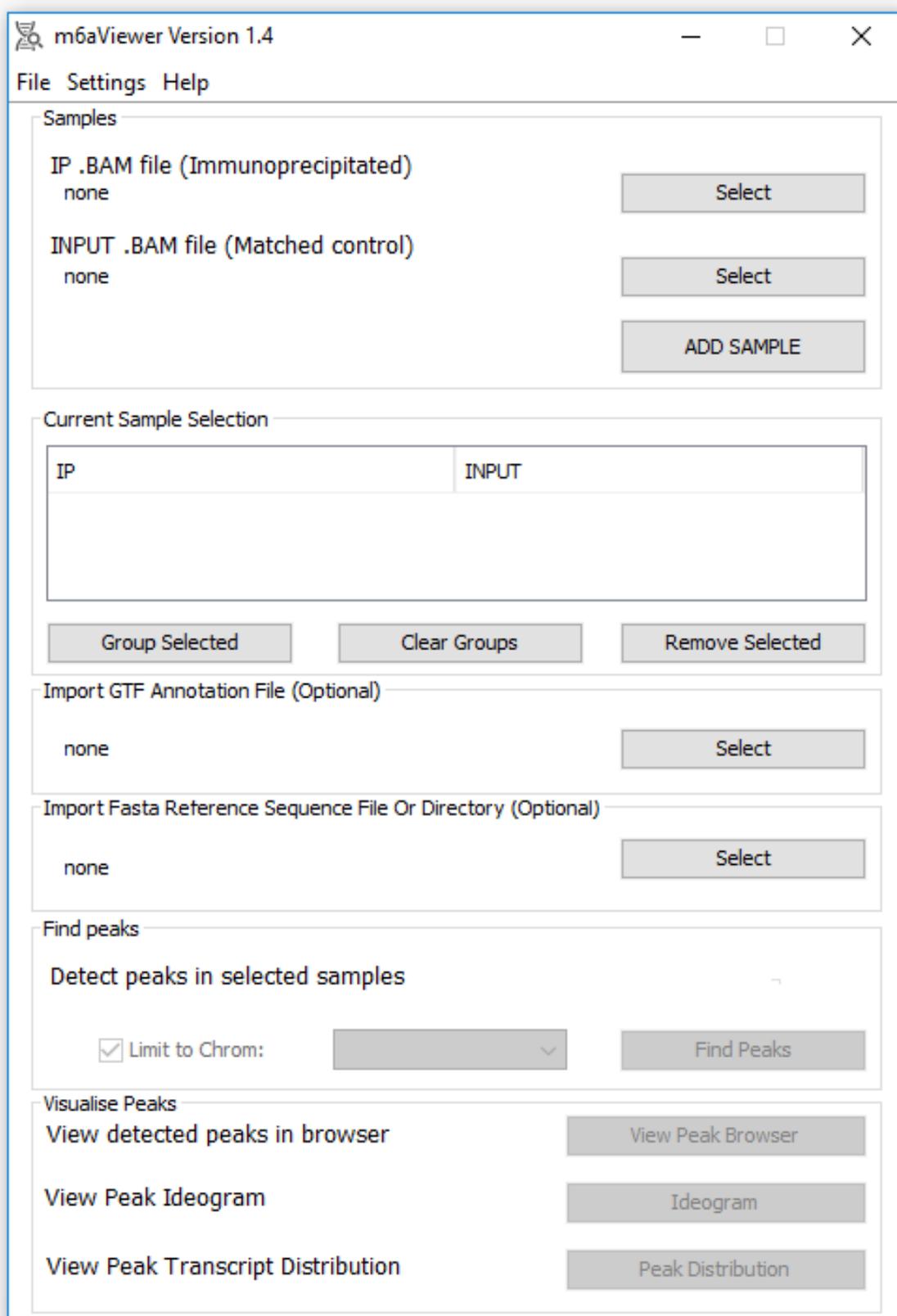
A lot of the features listed above are likely to be uninformative or redundant and therefore are likely to negatively affect the performance of the classifier. For example, more than 1000 kmer frequencies are encoded as features, while only a few of these are likely to be informative. Thus, in order to identify only useful data, feature selection was first performed. A greedy stepwise feature search was performed, where at each step a random forest classifier consisting of 100 trees was trained using a subset of all features in **Table 11** and evaluated using 10-fold cross validation. At each step, another most informative feature was added, until the addition of extra features resulted in a decrease in the overall performance.

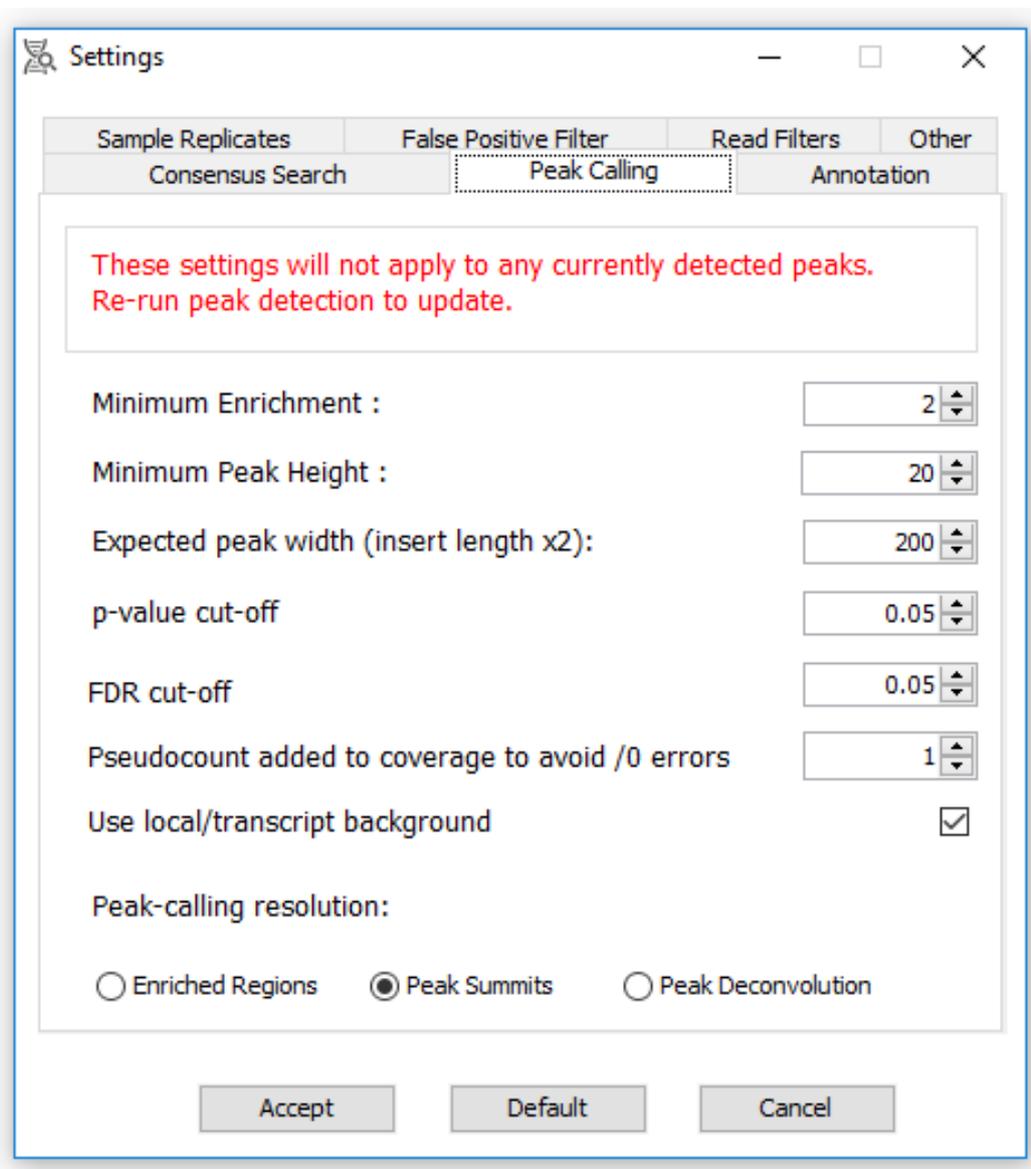
Using the final selected features, a random forest model is trained using 1000 random decision trees, each considering 7 random features.

## ***5.3 Results and Discussion***

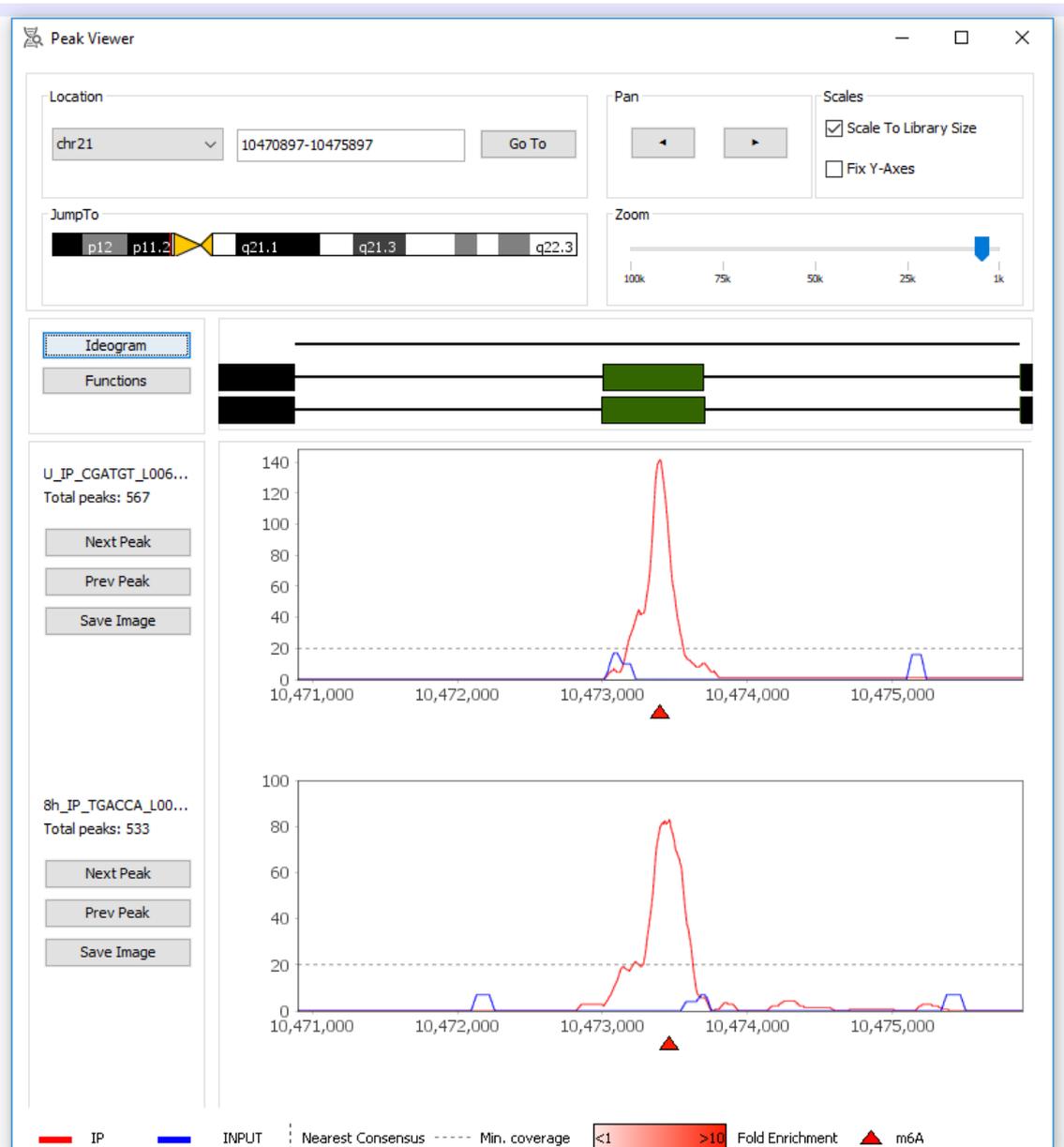
### ***5.3.1 m6aViewer implementation***

While the methods section of this chapter focused on the algorithmic details of m<sup>6</sup>A peak calling, the implementation can be equally important. In addition to the algorithms and models described, a number of utility functions for visualisation and data analysis are also implemented as part of m6aViewer software.





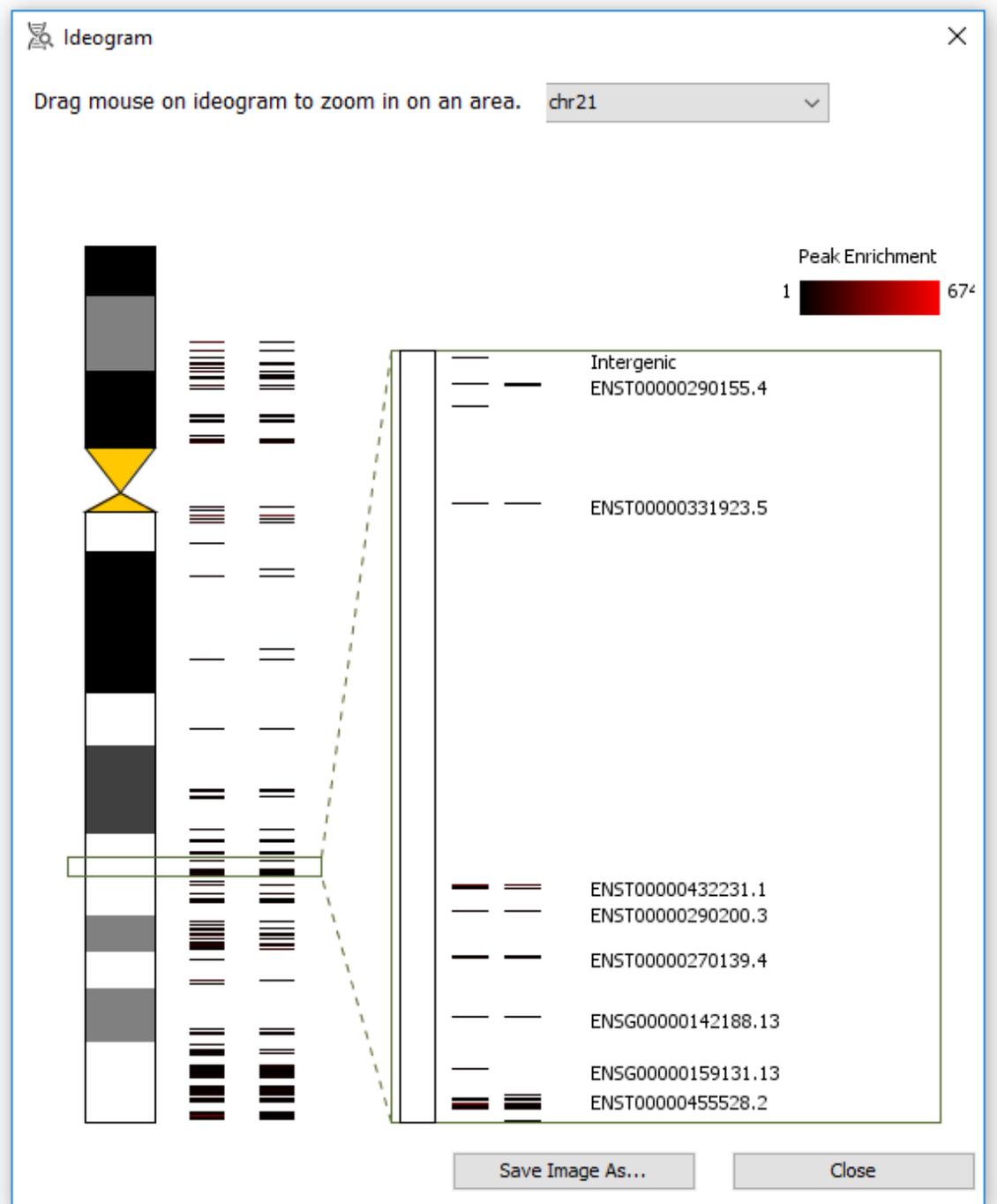
**Figure 58. A:** The main user interface window of m6aViewer. For basic usage, users are required to provide a minimum of two BAM files – one from immunoprecipitated sample (IP) and another one from a control (INPUT). Further options can be accessed from ‘File’ and ‘Settings’ menus. **B:** m6aViewer settings menu provides a large number of configurable options. A number of parameters can be configured here, including sensitivity/specificity of peak-calling and alternative peak-calling.



**Figure 58C.** m6aViewer’s peak browser interface allows the user to quickly visualise m<sup>6</sup>A peaks across multiple samples and easily jump between peaks using ‘Next Peak’ and ‘Last Peak’ buttons. The main panels display normalised control and immunoprecipitated coverage; the nearest m<sup>6</sup>A consensus site (configurable) to each peak is drawn as a vertical dashed line. Peak mouse-over tooltips are available for each peak and provide additional information. Genes are drawn above the track, provided a gene annotation file was provided or in-built annotation was selected.

m6aViewer is implemented via a graphical user interface in Java 1.7. The user interface is divided into three main views, consisting of the main control window (**Figure 58A**), an extensive options menu (**Figure 58B**) and a genome browser-style peak browser (**Figure 58C**). The main application window provides an interface for data input and peak calling functions. The minimal requirement for m<sup>6</sup>A peak-calling is two indexed BAM files, files one containing IP derived aligned data and the other the matched aligned RNA-seq control data. Transcript level information is supplied as GTF annotation files and is required for peak annotation and false positive filtering tasks.

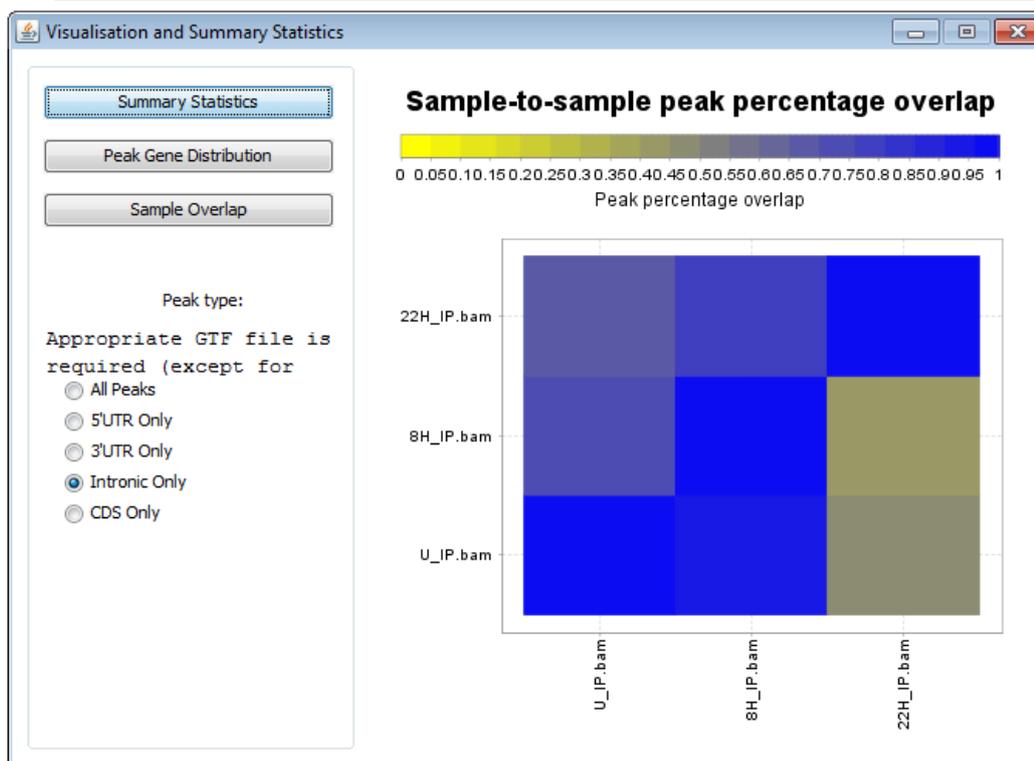
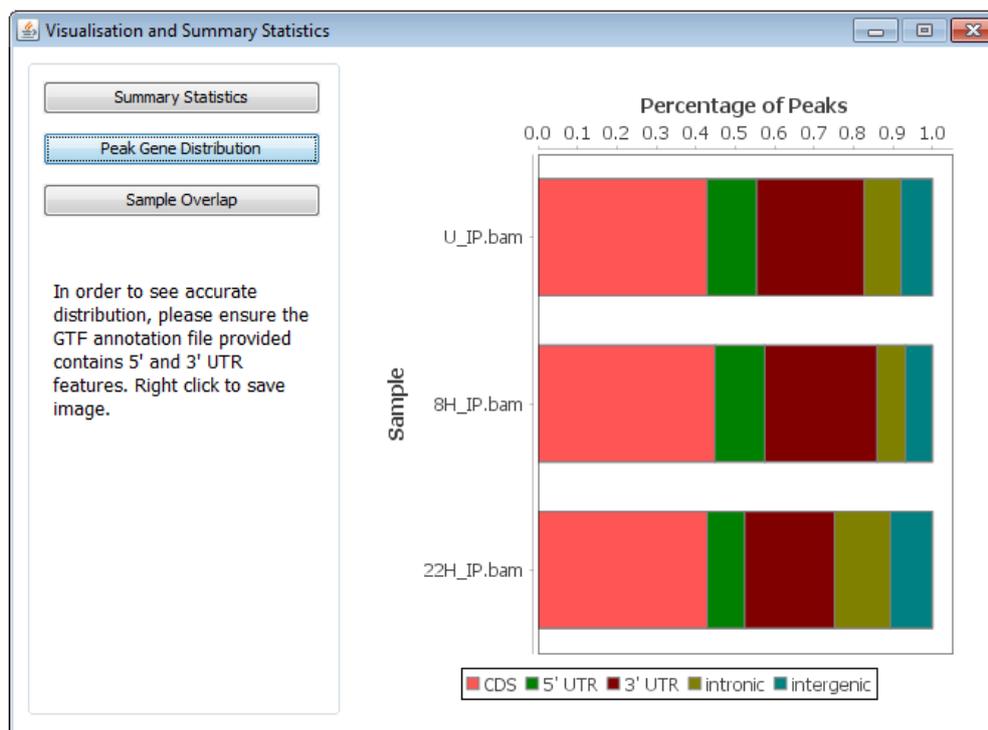
As with sequence data repositories, such as UCSC (Karolchik et al. 2004), Ensembl (Cunningham et al. 2014) or Genbank (Benson et al. 2005), the input fasta files are expected to contain the reference sequence(s) as a series of fixed length lines. As commonly studied organisms (e.g. human, mouse) have very large genomic sequences, it is computationally impractical to parse the entire reference sequence in order to annotate m<sup>6</sup>A peaks. Here, random access is instead used to retrieve only the sequences immediately surrounding each called m<sup>6</sup>A peak. For a given reference position, the corresponding file byte position is calculated by scanning the start of the fasta file to determine the length of the header line, the number of bases stored per line, the type of end of line character used and the type of character encoding used by the file. The availability of fasta files enables a number of other features. Individual peak sequences can be extracted and saved to a multi-fasta format file where they can be used for downstream analyses, such as for example, novel consensus motif detection via software such as the MEME suite (Bailey et al. 2009). Peaks may be filtered based on the peak distance to the nearest consensus site, as the presence of a consensus sequence may confer high confidence to the peaks. Nearest consensus sites are also annotated in text output and visualised in the peak browser (**Figure 58C**). Appropriate fasta files are also required in order to run EM-based peak calling, as sequence information is required for EM initiation step.



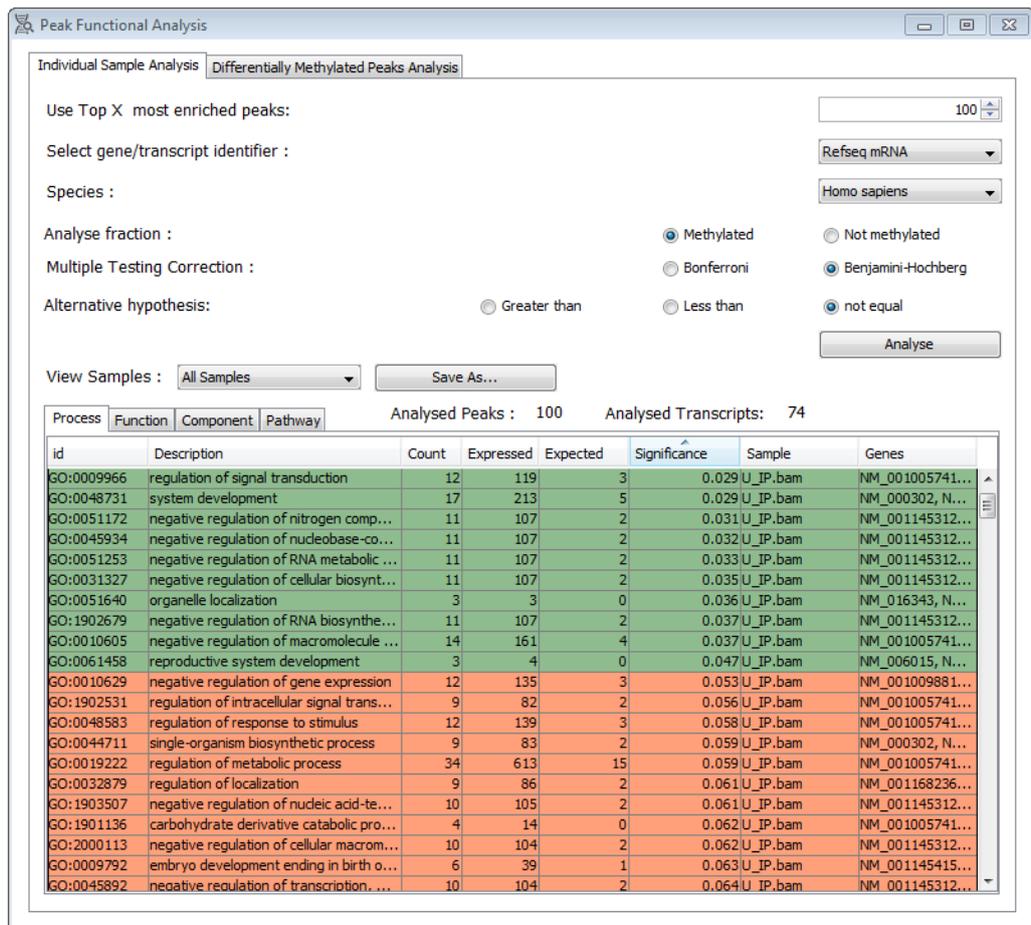
**Figure 58D.** Peak ideogram window in m6aViewer allows for a global overview of all m<sup>6</sup>A methylation sites called across all samples at chromosome level. Each peak position is marked as a single line, shaded by peak enrichment. Multiple samples are stacked side by side. The ideogram window is interactive and any part of the chromosome can be zoomed in on by dragging a window using the mouse.

Similarly to the peak sequence annotation, the peak browser utilises random access of BAM files to enable real time coverage data browsing without directly storing coverage data in memory or on disk. The peak browser window displays the immunoprecipitated and control read coverage across all samples, the position of the nearest m<sup>6</sup>A consensus sequence relative to the peak and gene annotation, and provides context-dependant peak information, such as enrichment scores and p-values, as a mouse-over tooltip. Peaks may also be visualised using the ideogram view (**Figure 58D**), while the summary of peak distribution within transcripts is also provided (**Figure 58E**).

It can often be of interest to know whether transcripts harbouring m<sup>6</sup>A sites are enriched for any particular group of functions in a given experiment. m6aViewer can obtain Gene Ontology functional annotations (Ashburner et al. 2000) and Reactome (Fabregat et al. 2016) pathway annotations directly from Ensembl (Cunningham et al. 2014) MySQL servers and perform an enrichment calculation, identifying categories of transcripts which are over- or under-represented in the methylated transcripts. This is performed against the background of all expressed genes identified in the RNA-Seq control sample, in contrast to popular web gene enrichment analysis web services such as DAVID (Huang et al. 2008) or Panther (Mi et al. 2017), where category enrichment is tested against whole transcriptome background and can be unsuitable for epitranscriptomic data. **Figure 58F** shows m6aViewer's the peak functional enrichment analysis menu.



**Figure 58E.** m6aViewer implements an additional visualisation window which provides a summary overview of all samples. This includes peak distribution charts, statistics and sample-to-sample peak overlap heatmaps.



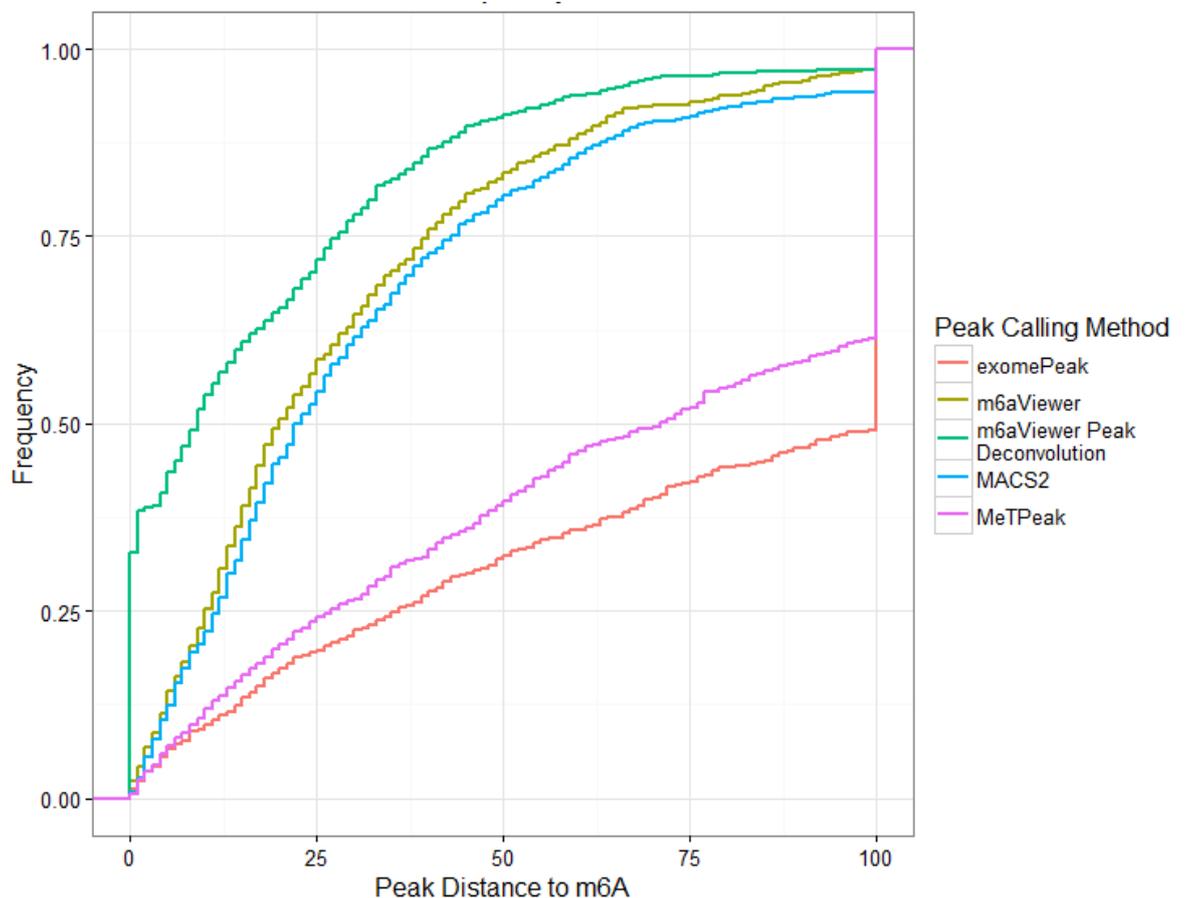
**Figure 58F.** Peak functional annotation enrichment analysis window. The results panel at the bottom displays all Gene Ontology and Reactome pathway annotations for analysed methylated transcripts. Significant hits based on a user selected alternative hypothesis (Enriched in transcript set, under-represented in transcript set or either) are coloured in green, while non-significant hits are shown in red. In order to make the significance values more transparent and understandable, the table also shows the number of methylated transcripts that are annotated with a particular term ('Counts' column); the number of transcripts annotated with the term that are expressed in the data set ('Expressed'); and the number of methylated transcripts annotated with the term expected to occur by chance in the analysis set ('Expected').

### 5.3.2 Evaluation of m6aViewer's peak calling performance

While the work described in this thesis was in progress, Linder *et al* (2015) described a mutation-based technique to obtain a single-nucleotide resolution map of m<sup>6</sup>A methylation sites in HEK293T cells. Here, this data is used to benchmark m<sup>6</sup>A peak-calling resolution. No matched m<sup>6</sup>A-seq dataset is available, however in order to facilitate a comparison, a different HEK293T m<sup>6</sup>A-seq dataset from Schwartz *et al*, 2014 (Schwartz *et al*. 2014) was used. The m<sup>6</sup>A-seq dataset was aligned to human hg19 reference genome using the STAR aligner (Dobin *et al*. 2013) and the alignments were sorted and indexed using Samtools (Li *et al*. 2009). Peak-calling was performed using both m6aViewer's default running mode, and the peak deconvolution modes, where peaks with p-value < 0.05 and > 2-fold enrichment were retained. In order to facilitate a comparison to other available m<sup>6</sup>A peak-calling software, peaks were also called using MACS2 (Zhang *et al*. 2008), exomePeak (Meng *et al*. 2013) and MeTPeak (Cui *et al*. 2016) - an undated and renamed version of HEPeak (Cui *et al*. 2015) (Huang.Y, personal communication, 11/07/2016). Peak-calling parameters were set to mirror those used by m6aViewer where possible. MACS2 was used with an additional command line option ('—call-summits') and data was treated as single-end, rather than paired end, as when MACS2 peak-calling was performed on paired-end data, very few and largely erroneous peaks were called due to vast overestimation of sequenced fragment sizes and expected peak widths when fragments spanned exon splice sites. For regions detected by exomePeak and MeTPeak, the centre of the region was computed in order to obtain a single peak position.

The 1000 highest confidence residues were identified in the Linder *et al* (2015) single-nucleotide resolution mutation map that also corresponded to an enriched region in the m<sup>6</sup>A-seq dataset. Each peak identified in these enriched regions by m6aViewer, MACS2, exomePeak and MeTPeak was compared to the 'ground truth' peak positions from Linder *et al* (2015). For each called peak, the distance to the nearest peak in the Linder *et al* (2015) dataset was computed and the distances visualised as cumulative frequency distribution (**Figure 59**).

These results highlight that the called m<sup>6</sup>A-seq peak summits very rarely precisely correspond to the actual site of methylation, a pattern also noted by Linder *et al* (2015). This difficulty can be further compounded by the presence of multiple methylated sites in close proximity, which blurs the expected peak signal. **Figure 59** shows that it is possible to improve the precision with which m<sup>6</sup>A residues are called by modelling each region as a mixture of fragment coverage distributions. The model-based peak deconvolution approach correctly identified the precise position of a methylated residue in 34% of cases, compared to 1-3% by methods (including the default m6aViewer summit calling approach) considering peak summits alone.



**Figure 59.** Cumulative frequency distribution of peak distance to the nearest m<sup>6</sup>A residue identified by Linder et al, 2015 at single nucleotide resolution level.

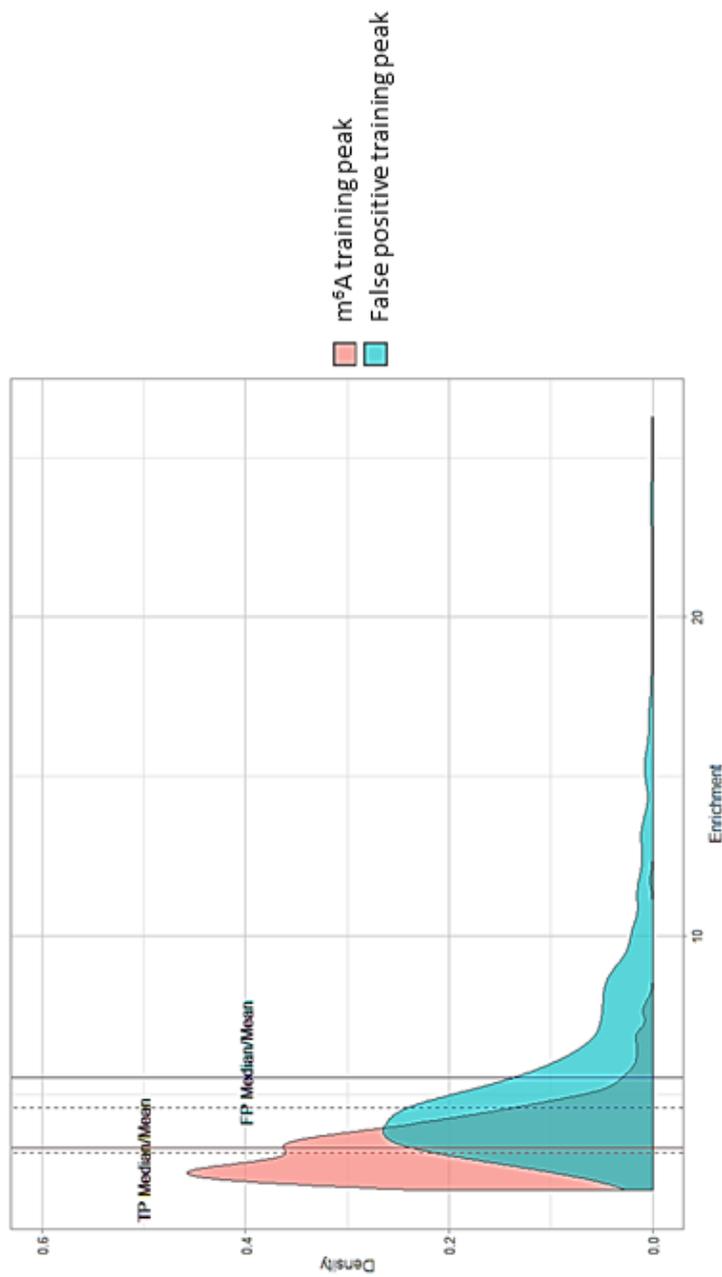
It is worth noting that, as expected, methods which attempt to identify the peak summit, such as MACS2 or m6aViewer in standard running mode, perform much better in terms of peak-calling precision than ‘binning’ based methods if the centre of the called region is considered as the peak position. While this is a somewhat unfair comparison, it nonetheless serves to illustrate the vast discrepancy in peak-calling resolution achieved by different approaches. The general performance of MACS2 and m6aViewer in standard running mode seems to be largely on par, with the slightly inferior performance of MACS2 explained by MACS2 inability to process paired-end RNA sequence reads.

### ***5.3.3 Evaluation of m6aViewer’s False Positive Filter Performance***

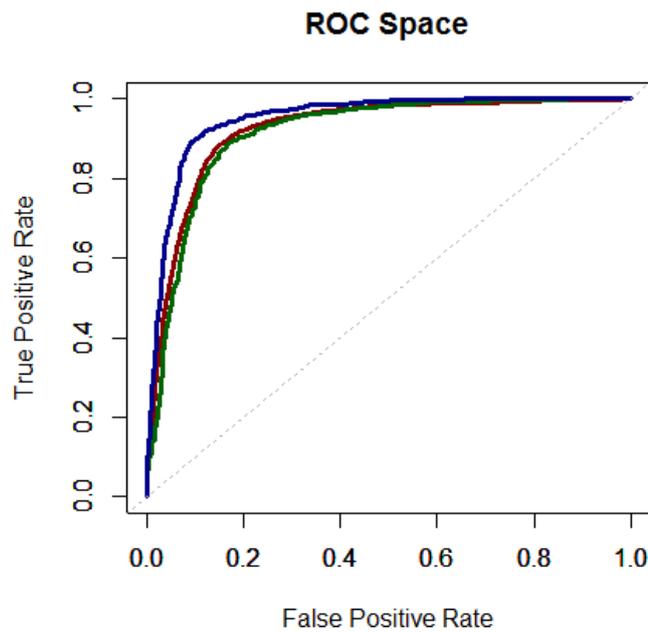
In addition to developing methods for improved peak-calling performance, this work investigates the possibility of identifying and classifying m6A peaks that conform to features of false positive sites that could arise due to non-specific antibody binding.

MTD RNA sequence and secondary structure models were investigated, before settling on an ensemble model encoding additional information due to increased performance. In keeping with the MTD sequence model, in the feature selection step of the ensemble model, the expanded RNA secondary structure features were not selected as informative for the final model, suggesting that either the error rate in RNA secondary structure prediction is too high, RNA secondary structure is not important for RNA adenosine methylation, or a similar RNA secondary structure is present at non-specific antibody binding sites to that of actual m<sup>6</sup>A sites, which could also be an antibody site recognition factor. miRNA binding information and conservation of nearest consensus motif and AC sites was found to be predictive. Out of peak information features considered, distance to nearest consensus and peak enrichment scores were found to be informative. When the latter was investigated, counter-intuitively slightly higher peak enrichment was found for false positive peaks than for real m<sup>6</sup>A sites (**Figure 60**). While this observation could arise due to some bias in training peak selection procedure, it could also be the case that the difference is due to antibody binding dynamics. That is, given all the RNA molecules of a particular

methylated transcript, only a proportion of them are expected to contain the m<sup>6</sup>A residue, thus the antibody can only bind to that proportion. On the other hand, if some other RNA property is causing non-specific antibody binding, it is likely that it is not subject to the same stoichiometry – thus, all/more of the molecules of that species are available for antibody binding, resulting in higher enrichment values upon sequencing.



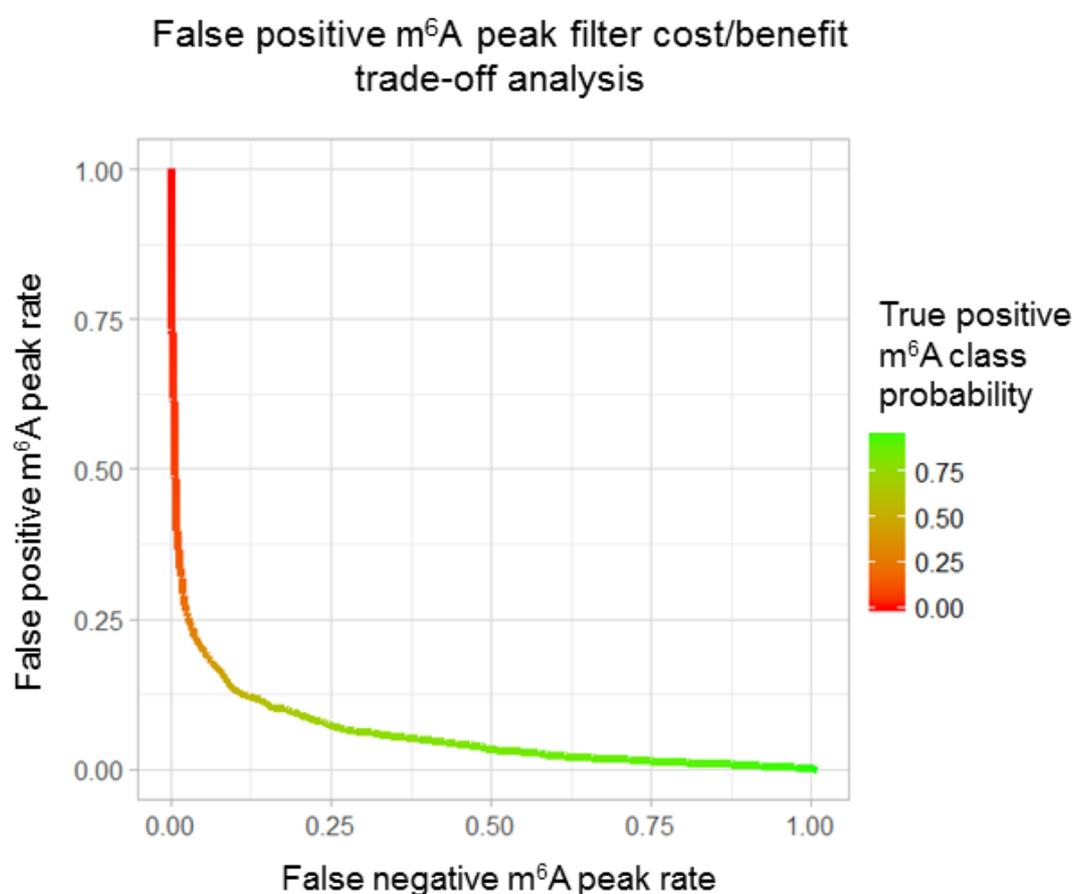
**Figure 60.** Distribution density of peak enrichment in false positive peak training sites (blue) and m<sup>6</sup>A training peaks (salmon).



**Figure 61.** Performance assessment of the random forest model, incorporating the MTD sequence model scores. **RED** = 10-fold cross validation using HEK293T cells, AUC= 0.923; **BLUE** = independent test set, A459 cells, AUC = 0.950; **GREEN** = independent test set, mouse fibroblasts, AUC =0.913.

In order to assess the performance of the final classifier, 10-fold cross-validation was performed using HEK293T cell line data. In addition, to assess how well the model generalises to different tissue types or even different species, two independent testing data sets were also used: the A459 cell line and mouse fibroblast m<sup>6</sup>A-seq data. The results are shown in **Figure 61**. The area under the ROC curve achieved by the combined random forest and MTD sequence model classifier is substantially higher than MTD alone (0.923 vs 0.794). These results suggest that this approach also generalises well to different cell/tissue types (AUC of 0.950), as well as different species (AUC of 0.913, mouse fibroblasts). While the results from mouse data are favourable, the model is far less likely to be accurate for more distantly related species. However, as publicly available RNA methyltransferase knockout m<sup>6</sup>A-seq datasets are limited to the cell types investigated here, the performance of this approach cannot yet be estimated for other species.

As random forests effectively use a voting system, where each decision tree decides whether a particular instance is likely to be a true m<sup>6</sup>A site or a false positive, the vote frequency can be used as an easily interpretable likelihood score. This allows implementing the classifier as a false positive filter, where the decision tree vote frequency can be used as a customisable cut-off. **Figure 62** shows the cost-benefit analysis of different cut-off values. At the default cut-off of greater than 0.5 (i.e. where more than half of all the random decision trees in the model have ‘voted’ for a true positive m<sup>6</sup>A site class), 86.02% of all false positive peaks in 10-fold cross validation test are identified correctly, at the cost of mislabelling 9.23% of genuine m<sup>6</sup>A sites as false positives. Increasing or decreasing this cut-off allows the filter to be easily skewed towards favouring precision or recall and can be easily customised towards different experiments and requirements.

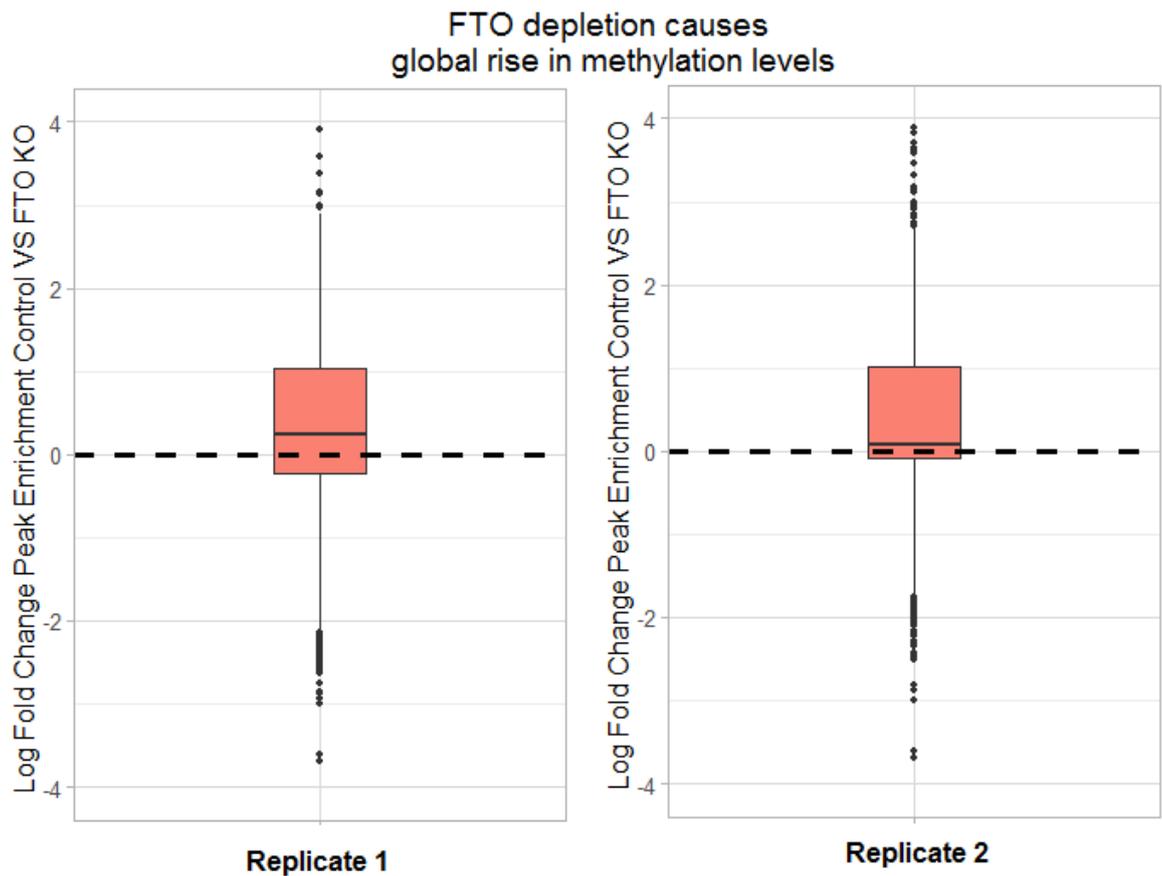


**Figure 62.** Cost-benefit analysis of false positive peak filter using 10-fold cross validation of the training data set.

It would be useful to further ascertain whether the peaks are indeed classified into specific and non-specific binding antibody binding sites, or whether this categorisation could be due to some other factor. Technical false positive peaks should, in general, be unaffected by biological influences which affect m<sup>6</sup>A levels – unless gene expression levels are also affected. One such factor that could be considered is RNA demethylation. On RNA demethylase knockdown, there should be a general increase in RNA methylation levels across all transcripts which are the targets of the demethylase. On the other hand, for all peaks which are observed due to technical noise such as non-specific antibody binding, the level of observed methylation should not change.

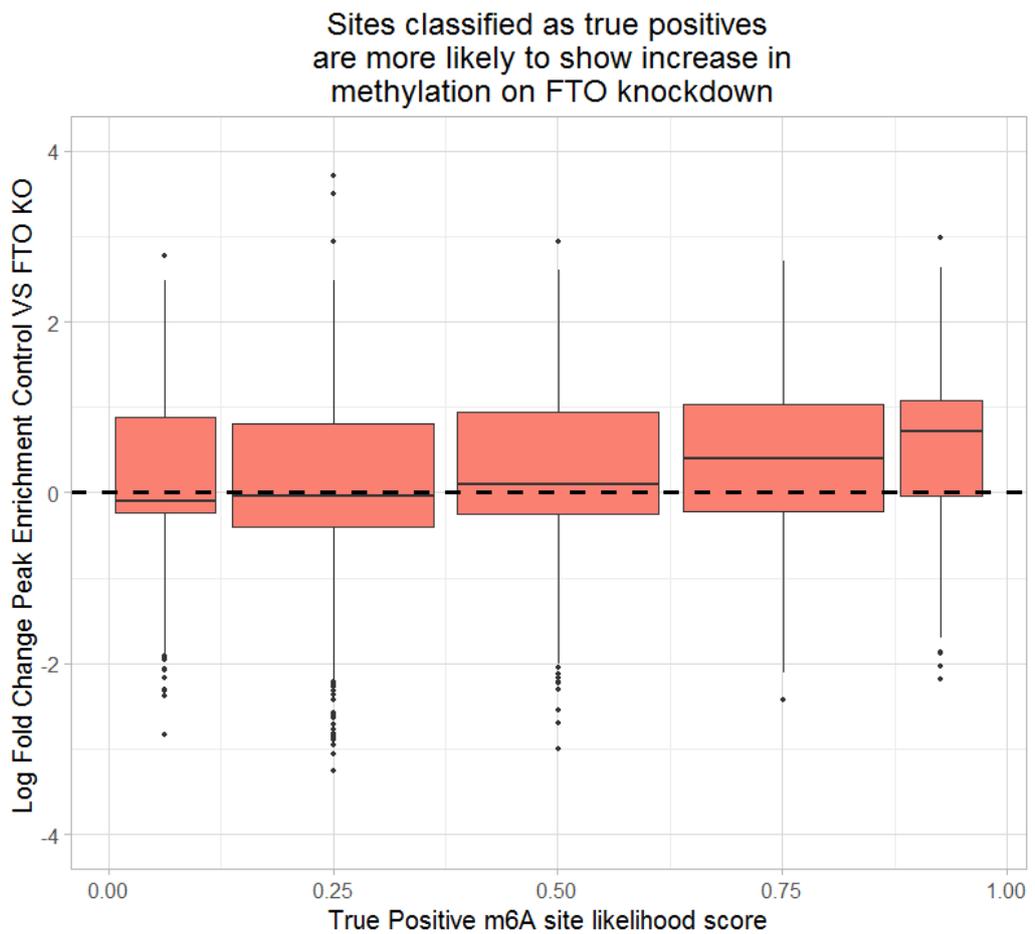
The general trends made by these assumptions can be tested. Zhao *et al*, 2014 (2014) performed a series of m<sup>6</sup>A-seq experiments on FTO-depleted mouse adipocytes and observed that overall, as expected, m<sup>6</sup>A levels increase upon FTO knockdown. The m<sup>6</sup>A-seq data of 2 matched control - FTO knockdown mouse adipocyte samples was downloaded from ArrayExpress, aligned to the mm10 mouse reference genome using the STAR aligner (Dobin *et al.* 2013) and m<sup>6</sup>A peaks were called using m6aViewer using default settings (> 2 fold enrichment, p-value <0.05) to form two experimental replicates. Each site was cross-referenced between matched FTO-knockout and control samples and log 2 fold-changes between two enrichment values were computed. As expected, in both replicates a larger proportion of individual peaks show increased enrichment in the FTO-depleted sample than in the control (**Figure 63**), although this bias is much less pronounced in the second replicate.

Next, all peaks across the two samples were pooled and annotated with the random forest model score and two subsets of peaks were selected – those that could be considered stable peaks (absolute log<sub>2</sub> fold-changes between the peak enrichment of FTO-KO and control less than 0.25) and upregulated peaks, i.e. those that responded to FTO-knockdown (log<sub>2</sub> fold change > 2).

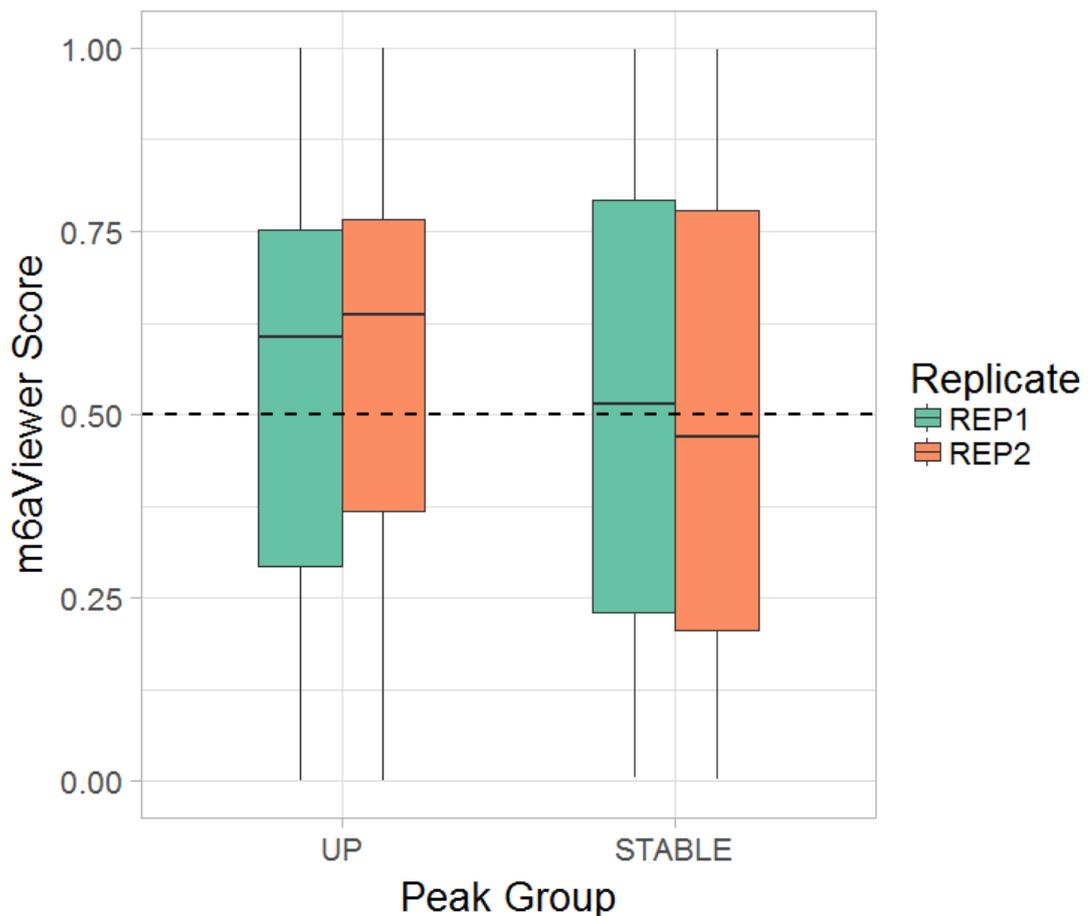


**Figure 63.** Depletion of RNA demethylase FTO causes a global rise in m<sup>6</sup>A methylation levels. Two biological replicates of FTO-depleted and control m6A-Seq datasets were considered separately.

**Figures 64A** shows the distribution of the random forest model score distribution in these groups. In replicate 1, 51.21% of stable peaks and 65.64% of upregulated peaks were scored 0.5 or higher; in replicate two, the difference between score distributions was more pronounced, with 47.87% of stable and 70.13% of upregulated peaks scoring 0.5 or higher. At a lower threshold of 0.4, in replicate 1, 71.26% of upregulated peaks and 58.78% of stable peaks are classified as true positives; in replicate 2, 54.28% of stable and 75.34% of upregulated peaks are classified as true positives (**Figure 64B**).



**Figure 64A.** Peaks which are classified as real m<sup>6</sup>A sites with greater likelihood are more likely to be hyper-methylated. Boxplots show the log fold change distributions when comparing m<sup>6</sup>A peak enrichment levels between control samples and FTO-depleted cells across a range of predicted 'true positive' m<sup>6</sup>A site scores. Peaks which are identified as more likely to be real m<sup>6</sup>A sites are more likely to respond to FTO depletion than sites which are scored lower.



**Figure 64B. Comparison of peak score distribution in FTO-depleted m<sup>6</sup>A-seq samples.** m6aViewer’s false positive filter classifier score distribution in peaks from two FTO-depleted and matched control sample replicates. Two subsets of peaks are compared, a “STABLE” (peaks which show no change in enrichment between FTO-depleted and control samples) subset and an “UP” (peaks which show at least a two fold increase in peak enrichment in FTO-depleted samples over control samples) subset. In replicate 1, 51.21% of stable peaks and 65.64% of upregulated peaks were scored 0.5 or higher; in replicate two, the difference between score distributions was more pronounced, with 47.87% of stable and 70.13% of upregulated peaks scoring 0.5 or higher. At a lower threshold of 0.4, in replicate 1, 71.26% of upregulated peaks and 58.78% of stable peaks are classified as true positives; in replicate 2, 54.28% of stable and 75.34% of upregulated peaks are classified as true positives. Two biological replicates of FTO-depleted and control m6A-Seq datasets were considered separately.

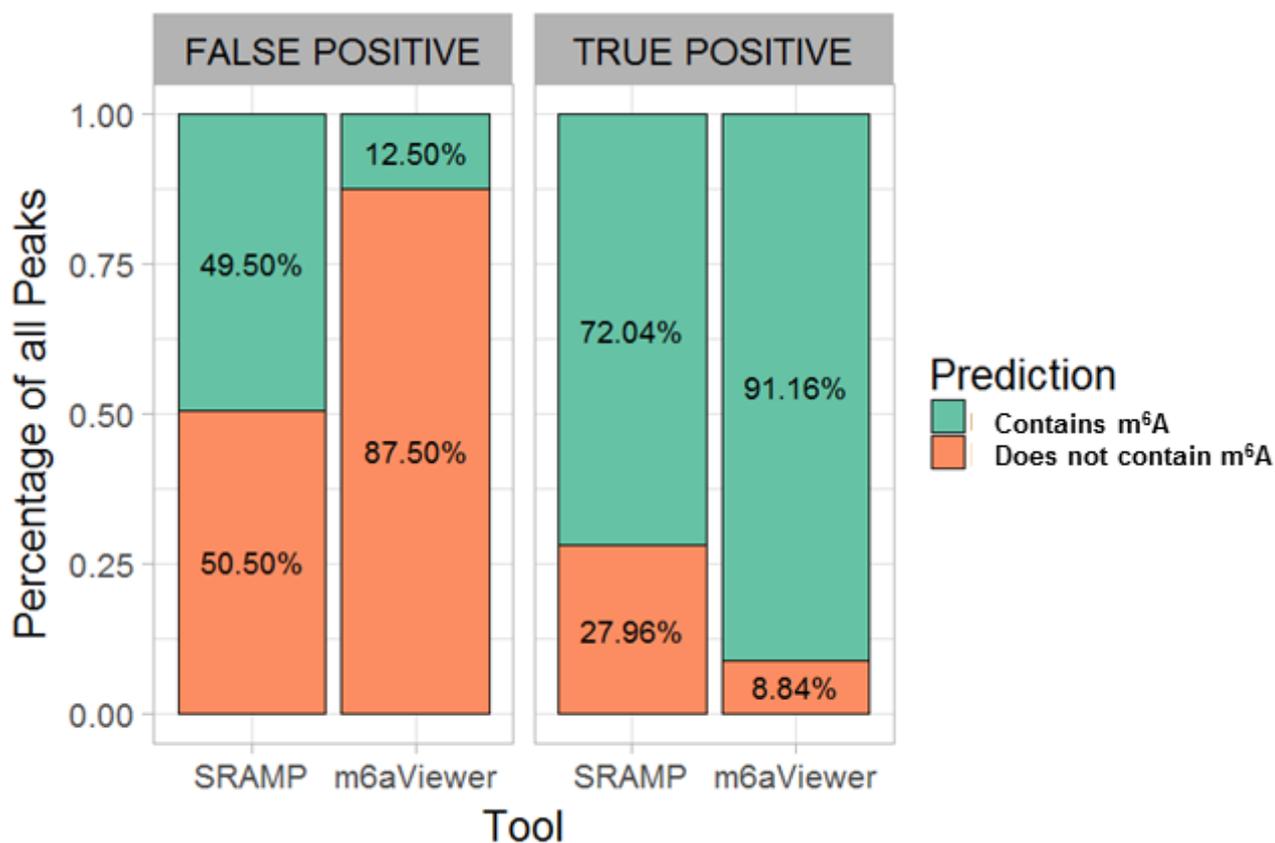
### 5.3.4 Comparison with m<sup>6</sup>A prediction algorithm SRAMP

Recently, a number of machine learning models for computational identification of m<sup>6</sup>A sites from sequence features were proposed. Zhou *et al.* (2016) implemented SRAMP, a web server for m<sup>6</sup>A site prediction from transcript sequence in a number of mammalian cell types. SRAMP combines multiple random forest predictors trained on sequence features from known m<sup>6</sup>A sites. A similar web tool implementing a support vector machine predictor has been developed for yeast data (Chen *et al.* 2015d).

In contrast to the methods proposed by Zhou *et al.* (2016) and Chen *et al.* (2015), false positive peaks are used as negative training examples, rather than randomly selected transcriptomic positions. Despite the best efforts to obtain a high-confidence training dataset, it is likely that some inaccuracies remain. While the training examples were obtained from cell-type matched m<sup>6</sup>A-seq datasets to minimise the effects of biological variation, it is likely that a proportion of training instances are mislabelled. Additionally, while generally robust, siRNA knockdown of methyltransferases does not abolish the presence of m<sup>6</sup>A methylation completely; this is likely to also contribute to mislabelled training instances. High classification accuracy, however, suggests that the approach described here is resistant to noise in the training data, with the positive and negative instances overall forming sufficiently biologically distinct groups.

To assess how well m6aViewer's false positive filter compares with sequence-based predictions of SRAMP, the SRAMP web-server was used to predict m<sup>6</sup>A positions within peak sequences (400nt surrounding the detected peak's position) from the A459 cell line testing dataset. The interrogation of the SRAMP web server was automated via HTTP POST queries by custom code. SRAMP predicted m<sup>6</sup>A residues to be present in 72.14% of all instances labelled as true positive m<sup>6</sup>A peak sequences while m6aViewer classified 91.14% of these sequences as such (**Figure 65**). In the false positive peak subset, SRAMP predicted m<sup>6</sup>A to be present in nearly half the sequences, while m6aViewer misclassified these in 12.5% of total instances. As discussed

previously, this discrepancy could be partially explained by inaccuracies in the dataset used for training m6aViewer's classifier, where genuine m<sup>6</sup>A sites could be mislabelled as false positives.



**Figure 65.** Comparison between m<sup>6</sup>A predictions made by web server SRAMP and m6aViewer. SRAMP predicted m<sup>6</sup>A within peak sequences from the true positive subset in 72.04% of cases, while m6aViewer classified these as true positives in 91.16% of cases. In the false positive peak subset, SRAMP predicted m<sup>6</sup>A residues in 49.5% of cases, while m6aViewer classified 12.5% of these as true positives. Sequences predicted to contain m<sup>6</sup>A by either algorithm are in green, while sequences predicted to not contain m<sup>6</sup>A are in orange.

Furthermore, SRAMP was run with default settings - a running mode which is faster, but does not consider predicted secondary RNA structure information, which reportedly enhances classification performance. These results, however, are not surprising, as SRAMP effectively detects the potential for m<sup>6</sup>A

methylation in a given sequence, whereas m6aViewer aims to frame these predictions in an experimental context by considering m<sup>6</sup>A-Seq data features. As such, some adenosine residues encompassed by the false positive peak sequences used here could be potentially methylated under certain conditions due to dynamic nature of m<sup>6</sup>A; thus, SRAMP's predictions would be correct within the designed scope of the software. Indeed, this potential for methylation may be a major contributing factor for the much lower discriminatory power of the sequence-only model established earlier.

### ***5.3.5 Integration with m6aViewer software***

Finally, in order to integrate the model with m6aViewer software and apply it to new data, a consistent way of annotating peaks to features is required. While the majority of the features considered can be obtained directly from the GTF annotation files or fasta sequence files, which are already required by other m6aViewer features, sequence conservation scores and miRNA binding site information is also required.

miRNA binding sites for hg19 human and mm10 mouse data were downloaded and stored in a small SQLite database, which can be packaged together with the main m6aViewer application executable. In order to apply the false positive filter to other organisms, or other human or mouse genome assemblies, a user may provide the data as a text file. Sequence conservation scores are obtained via direct connection to UCSC public MySQL database.

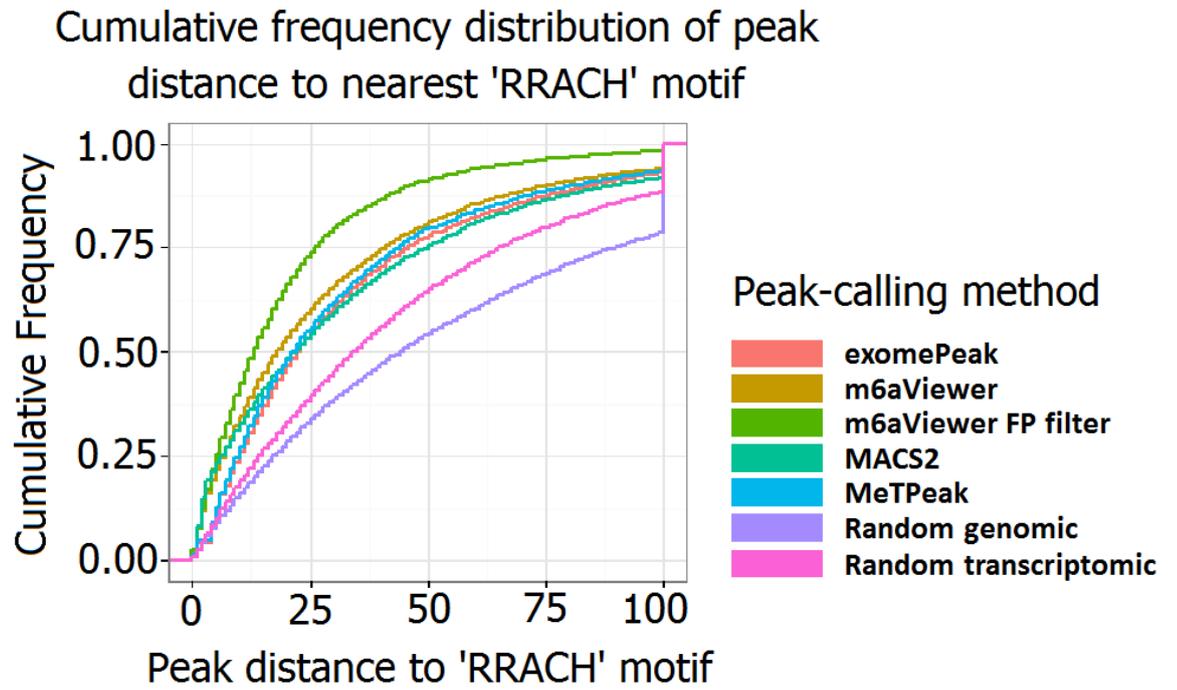
Alternatively, if the data is not available, the random forest classifier is able to impute some missing values, though naturally at the cost to accuracy. In the case of missing values in a peak instance the classifier is trying to score, the data can be estimated from the most similar instances in the dataset that was used to train the classifier.

### ***5.3.6 Distribution of identified peaks to nearest m6A 'RRACH' sequence motif***

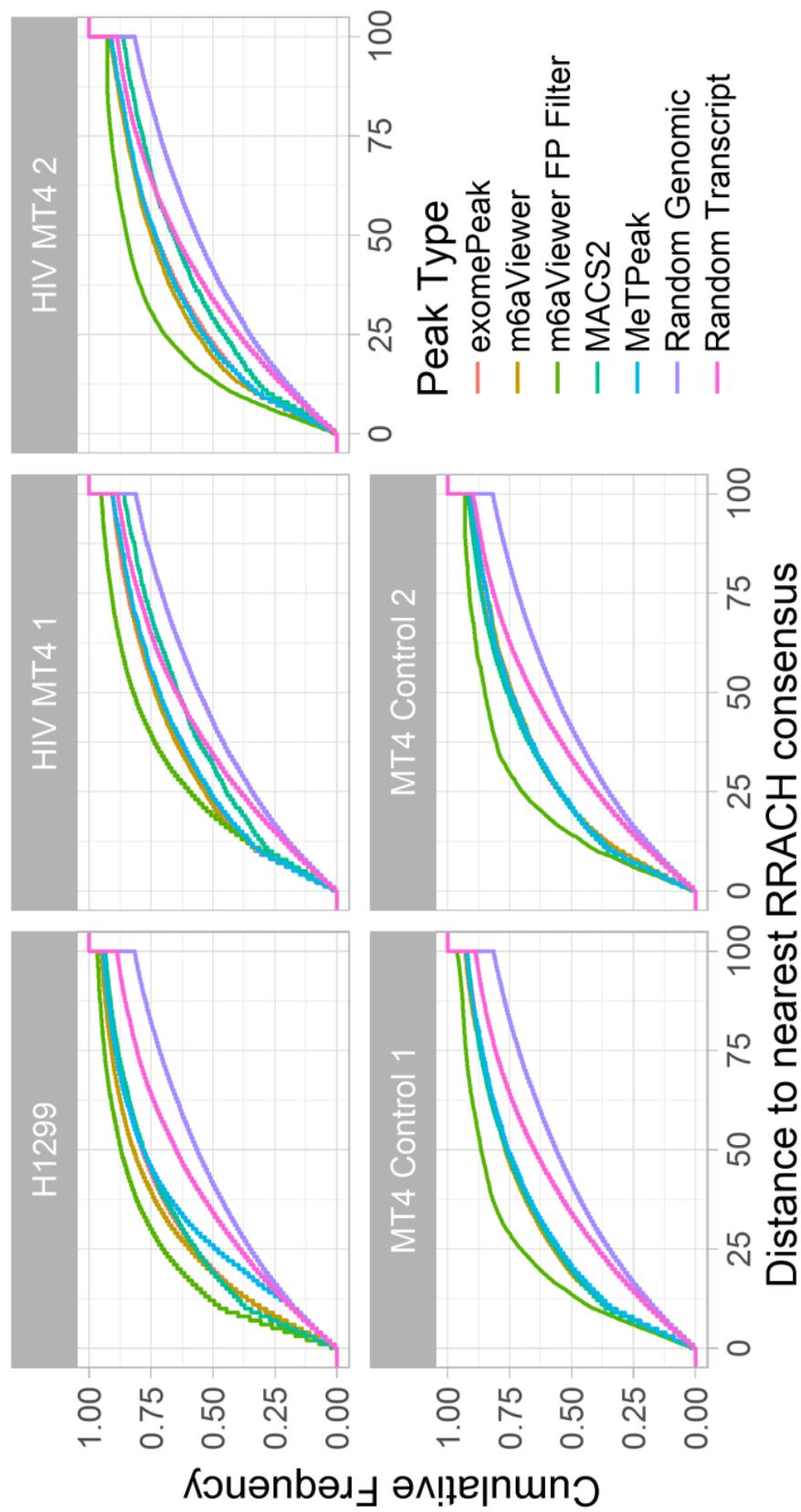
It is difficult to objectively evaluate the performance of any m<sup>6</sup>A scoring algorithm, since there is no m<sup>6</sup>A-seq testing dataset for which the locations of all m<sup>6</sup>A residues are known. Therefore, in order to further establish the validity of

peak-calling methodology implemented in m6aViewer, distance to the nearest m<sup>6</sup>A consensus sequence was used as a metric of performance. The tight co-localisation of the m<sup>6</sup>A 'RRACH' consensus motif with detected peak positions can confer confidence to the peak-calling method, and distance to the nearest m<sup>6</sup>A consensus has been previously used as an m<sup>6</sup>A peak-calling performance metric (Meng et al. 2013). Furthermore, assuming that the consensus sequence motifs are likely to coincide with the actual sites of the methylated residues, the distance to the nearest consensus can illustrate peak-calling precision.

**Figures 66** and **67** compare the peak to nearest m<sup>6</sup>A consensus distances between peaks detected by m6aViewer, exomePeak, MeTPeak, MACS2 and randomly selected transcriptomic or genomic control sites. m6aViewer was run in default peak-calling mode, as the deconvolution mode preferentially selects adenosine, 'AC' or 'RRACH' sequence positions during the EM initiation step, and thus the comparison would be extremely biased in favour of the EM algorithm and therefore largely meaningless. Peaks detected by m6aViewer show high levels of enrichment for previously reported consensus motifs, with known motifs appearing much more frequently near peak positions than near randomly selected transcriptomic positions. The significance of this observation was confirmed by performing a Kolmogorov–Smirnov test for the alternative hypothesis that the cumulative distribution function of m<sup>6</sup>A peak distance to nearest consensus lies above that obtained from randomly selected transcriptome positions ( $p < 2.2e^{-16}$ , statistic  $D = 0.2474$ ). Peaks detected by all algorithms tested are overall closer to a 'RRACH' consensus than the randomly selected control sites, with MACS2 and m6aViewer calling peaks closer to the m<sup>6</sup>A consensus than MeTPeak and exomePeak calls. As MeTPeak and exomePeak output significantly enriched regions, the centre of these intervals was used as the peak position; thus, while these points are not entirely comparable to MACS2 or m6aViewer peaks, it nonetheless serves to demonstrate the difference in peak-calling resolution achieved by these methods.



**Figure 66.** Cumulative frequency distribution of detected peak distance to nearest 'RRACH' consensus sequence motif in peaks called by different peak-calling software.



**Figure 67.** Cumulative frequency distribution of detected peak distance to nearest 'RRACH' consensus sequence motif in peaks called by different peak-calling software, assessed on m6A-Seq data from H1299 cells (ArrayExpress Accession: E-GEOD-76367), and HIV infected and control MT4 cells (ArrayExpress Accession: E-GEOD-74016).

### ***5.3.7 Discussion***

The results presented here suggest that a model-based approach can result in a sizable improvement in m<sup>6</sup>A peak-calling precision over binning or summit identification techniques. Nevertheless, more than a half of all m<sup>6</sup>A residues considered in the testing scenarios are still not identified precisely on target. This highlights the difficulty in trying to apply a single probabilistic approach to the whole transcriptome. The size of the data is limiting, too – a more complex model may be able to capture the underlying stochastic process which generates the sequenced reads better, but due to an increased computational complexity would not be usable in practice on whole transcriptome datasets. Indeed, the current model makes a number of assumptions which cannot hold for all cases. For example, the probability of each aligned read being a ‘noise’ read is modelled uniformly – a substantial oversimplification. Due to alignment errors in lower complexity regions, PCR amplification biases such as those in GC-rich regions or fragment selection biases such as those near the ends of transcripts, the contribution of ‘noise’ reads to the total immunoprecipitated read coverage can hardly be expected to be uniform. While enriched regions selected for deconvolution are relatively small (on average about a kilobase) and thus the level of variation in noise read distribution is expected to be less than within (or between) whole transcripts, nonetheless some variation will exist. This could be modelled for each position individually by considering the read distribution in the control RNA-Seq fraction and estimating a positional read ‘mappability’ or ‘sequencability’ factor that could then be included in the model. Additionally, all aligned reads are assumed to be generated by a single process estimated from the training peaks. This, however, may vary between different samples, protocols or batches of the polyclonal anti-m<sup>6</sup>A antibody used. While noise distribution oversimplification could be accounted for by increasing model complexity, any additional sources of variation are much harder to assess and address due to lack of data.

Due to these concerns, the probabilistic model-based peak calling approach is implemented as a supplementary, rather than the default peak-calling algorithm in m6aViewer software, despite showing substantial gains in performance in the testing data.

Another major feature implemented as part of m6aViewer software is the putative false positive peak detection via an ensemble model. The model was trained and assessed using RNA methyltransferase knockdown data and further independently assessed using RNA demethylase knockdown data to confirm that the opposite pattern can be observed.

Despite the best efforts to obtain a high-confidence training dataset, it is likely that some inaccuracies remain and therefore the performance values reported here are not wholly accurate. While the training examples were obtained from cell-type matched m<sup>6</sup>A-seq datasets to minimise the effects of biological variation, it is likely that some training instances are in fact mislabelled. Additionally, while generally robust, methyltransferase knockdown does not abolish the presence of m<sup>6</sup>A methylation completely; this is likely to also contribute to mislabelled training instances. However, high classification accuracy suggests that this approach is resilient to noise in the data, with the positive and negative instances overall forming sufficiently biologically distinct groups. It is feasible that these two groups are not in fact, non-specific binding sites and real m<sup>6</sup>A sites, but arise as a result of some other biological dissimilarity. The peaks which are unaffected by RNA methyltransferase knockdown, for example, could be methylated instead by some other, yet undiscovered RNA methyltransferase. Nevertheless, using a biologically and technically independent dataset where m<sup>6</sup>A-Seq was performed in a RNA demethylase FTO depleted system, a correlation between hyper-methylated peaks and high classifier score was observed. Again, it is feasible that this is due to some other biological factor – FTO-unresponsive sites could harbour genuine m<sup>6</sup>A methylation and are simply not targeted by FTO. However, taken together with RNA methyltransferase knockdown data, there is strong evidence to suggest that the classifier presented here has discriminatory power for specific and non-specific antibody binding sites in m<sup>6</sup>A-seq data. Nonetheless, wet lab experiments would be required to confirm these predictions and as such, are beyond the scope of this work.

## ***6. Characterisation of Kaposi's sarcoma Herpesvirus-8 m<sup>6</sup>A methylome and identification of a putative novel m<sup>6</sup>A 'reader' protein***

### ***6.1 Motivation***

m6aViewer software described in **Chapter 3** was developed alongside the investigation into the m<sup>6</sup>A methylome of Kaposi's Sarcoma Herpesvirus to meet the NGS data analysis needs. The following chapter, therefore, presents the motivations and aims of the project, analyses of NGS data as well as potential future research avenues.

NGS data analyses and outcomes are presented in two parts. The first part details the investigation into KSHV and host cell m<sup>6</sup>A methylomes. The second part focuses on the characterisation of a novel putative m<sup>6</sup>A modification 'reader' protein, SND1, here identified through its interactions with a key methylated KSHV transcript.

### ***6.2 Kaposi's Sarcoma Herpesvirus-8***

Kaposi's Sarcoma Herpes Virus-8 (KSHV) is a large, oncogenic double stranded DNA virus. Originally discovered in 1994 (Chang et al. 1994), the virus has since been identified as the main cause of Kaposi's sarcoma, Primary Effusion Lymphoma and Multicentric Castleman's Disease (Chang et al. 1994; Soulier et al. 1995; Cesarman et al. 1995; Schalling et al. 1995). Similar to other herpes viruses, KSHV is transmitted via saliva, blood or sexual contact, though transmission from pregnant women to the fetus is rare (Martin et al. 1998; Kedes et al. 1996). While the prevalence of KSHV infection in Western populations is less than 10%, in other areas, such as sub-Saharan Africa, it is estimated to be as high as 50% (Parkin et al. 2008). KSHV infection is typically well controlled by the host's immune system via cytotoxic T-cell recognition of viral epitopes (Robey et al. 2010; Stebbing et al. 2003). Irrespective of geography, its prevalence is higher in patients also infected with HIV (Wabinga

et al. 1993), who along with other immune-compromised patients, show a greater incidence of KSHV-related cancers.

Similarly to other herpesviruses, KSHV usually infects lymphoid cells; however it can also infect other cell types of endothelial lineage, as well as monocytes. B lymphocytes are the primary site for latent infection that allows the virus to establish a long-term latent viral reservoir (Blackbourn et al. 2000; Monini et al. 1999; Caselli et al. 2005). After entry, the virus generally establishes a latent infection phase, expressing a limited subset of key genes required to maintain infection. These include LANA (latency associated nuclear antigen), a phosphoprotein that acts as a transcriptional regulator and inhibits TGF-beta and p53 signalling pathways, which result in impaired apoptosis and increased cell proliferation (Nabel et al.; Si and Robertson 2006); vCyclin promotes cell cycle progression (Zhi et al. 2015; Godden-Kent et al. 1997); while vFLIP expression results in NF-kB activation by interfering with FAS associated death domain and caspase-8, thus promoting cell survival (Bagn ris et al. 2008; Liu et al. 2002). During the lytic phase, which KSHV rarely enters *in vivo*, the entire viral genome is expressed resulting in viral replication and assembly, and release of virions from the infected cell which is destroyed in the process.

Transition from latent to lytic phase is largely controlled by the RTA (regulator of transcriptional activation) DNA-binding viral protein, a transcriptional activator that triggers the lytic cascade of viral replication. Consisting of 2 exons in the ORF50 region, it's one of the few spliced KSHV mRNAs (Arias et al. 2014; Yu et al. 2007; Cohen et al. 2006).

KSHV lytic replication is essential for the development and maintenance of KSHV-associated tumours. It has been shown that drugs targeting lytic replication can lead to regression of KSHV-associated tumours. Consequently, it is important to characterise the transcriptional and post-transcriptional control mechanisms that act as mediators of the latent-lytic switch in KSHV infection, which could ultimately lead to the development of anti-cancer therapies.

It has been shown that the transcription of KSHV lytic genes is controlled via histone and DNA modifications; however, it is possible that post-transcriptional

control of viral transcript expression also plays a key role in the KSHV viral life cycle. While KSHV virus does not itself encode m<sup>6</sup>A methylation machinery, nevertheless the manipulation of host m<sup>6</sup>A methylation pathways could be required for viral cycle progression. As the roles of m<sup>6</sup>A in the cell have been shown to be very diverse, KSHV could potentially utilise the host m<sup>6</sup>A machinery for several purposes, including control of host and viral splicing, RNA stability, translational efficiency and nuclear export. For instance, efficient splicing could be required to drive the lytic pathway, as RTA transcript typically undergoes splicing to produce the mature transcript. Furthermore, the organisation of the KSHV viral genome is compact, and latent and lytic genes are often transcribed as large open reading frames which are further spliced into individual transcript products; m<sup>6</sup>A could potentially modulate this process. m<sup>6</sup>A has also been shown to enhance transcript stability, nuclear export and translation; hijacking these key processes could be a key mechanism for rapid viral gene expression and modification of the host's expression profile to promote virion assembly. Consequently, here the potential for the KSHV transcriptome to be methylated is investigated using m<sup>6</sup>A-Seq.

## **6.3 Methods**

### **6.3.1 Sequence Data Generation**

#### **6.3.1.1 m<sup>6</sup>A-Seq**

Sample preparation, RNA extraction, NGS library preparation and sequencing described in this section were carried out by collaborators (Whitehouse group, FBS) in accordance with the protocol described by Dominissini *et al* (2012). Briefly, RNA was extracted from HHV-8 infected TReX BCBL-1-RTA cells, at 0, 8 and 20 hours post induction of the lytic pathway via doxycycline-induced expression of RTA. In each case, two separate biological replicates were obtained. RNA was fragmented and an aliquot of each sample was retained to produce the 'INPUT' RNA-Seq control libraries. Affinity purified anti-m<sup>6</sup>A rabbit polyclonal antibody (Synaptic Systems, cat. no. 202 003) was used to select for m<sup>6</sup>A-containing RNA by immunoprecipitation of the fragmented RNA. Immunoprecipitated and the input control RNA were used for

NGS library production using Illumina's TruSeq Total RNA kit. The first set of libraries were sequenced on a HiSeq 2500 101 bp paired end lane, while later libraries were sequenced on a HiSeq 3000, 151 bp paired end lane.

#### **6.3.1.2 *SND1 fRIP-Seq***

SND1-fRIP-Seq was carried out by collaborators (Whitehouse group, FBS) in accordance to the protocol described by Hendrickson *et al*, (2016). Briefly, formaldehyde cross-linking was performed by incubating TReX BCBL-1-RTA cells with 0.1% formaldehyde. Cells were sonicated to lyse and fragment the RNA; a fraction of lysate was set aside to create the control (INPUT) NGS libraries, while the remainder was subjected to immunoprecipitation with anti-SND1 antibody. RNA was extracted and purified from both the immunoprecipitated and control fractions, and used to make the NGS libraries using the TruSeq Stranded Total RNA library production kit. As before, 3 time points (0H, 8H and 20H post-RTA induction) were sequenced on two 151 bp paired end lanes on the HiSeq 3000 instrument.

#### **6.3.1.3 *RNA lifetime profiling***

RNA lifetime profiling was carried out by collaborators (Whitehouse group, FBS), in accordance to the protocol described by Wang *et al* (2014). Briefly, BCBL-1 cells were transfected with SND1 siRNA or control siRNA and after 48 hours, actinomycin D was added to transfected cells at 5 µg/mL at 6 hours, 3 hours, and 0 hours before trypsinisation collection. RNA was extracted and purified from two biological replicates of latent and lytic BCBL-1 cells, and used to make the NGS libraries using TruSeq Stranded Total RNA library preparation kit. ERCC spike-in mix 1 (Jiang *et al*. 2011) was added proportional to total RNA prior to NGS library preparation. As before, the libraries were sequenced on 151 bp paired-end lanes on the HiSeq 3000 instrument.

#### **6.3.2 *Publicly available sequencing data***

The following data was downloaded from the ENCODE project (ENCODE Project Consortium *et al*. 2012) as raw fastq files:

- Hep2G cell line SND1 eCLIP data, two replicates

- Hep2G cell line SND1 knockdown and control RNA-Seq data, two replicates

The following data was obtained from GEO database (Barrett et al. 2013) as raw fastq files:

- HeLa cell line m<sup>6</sup>A-Seq data, two replicates (Accessions: GSM1135030, GSM1135031, GSM1135032, GSM1135033)
- HeLa cell line YTHDF2 knockdown and control RNA life time profiling data, two replicates (Accessions: GSM1197622, GSM1197623, GSM1197624, GSM1197625, GSM1197626, GSM1197627, GSM1197628, GSM1197629, GSM1197630, GSM1197631, GSM1197632, GSM1197633)
- HeLa cell line YTHDF2 PAR-CLIP data, three replicates (Accessions: GSM1197605, GSM1197606, GSM1197607)

The following data was obtained from DNA Data Bank of Japan (Mashima et al. 2016):

- HeLa cell line of RNA life time profiling of RNAs under normal conditions (Accessions: DRA000345, DRA000346, DRA000347, DRA000348 and DRA000350)

### ***6.3.3 Sequence data analysis***

#### ***6.3.3.1 Processing of raw sequence data***

All m<sup>6</sup>A-Seq, fRIP-Seq and RNA lifetime profiling data were generated at Leeds NGS Facility and extracted and de-multiplexed using bcl2fastq software, which exports a matched pair (read 1 and read 2) of compressed fastq files per sample. All further analysis was performed by the author.

Quality control of all sequence data, including publicly available datasets, was carried out using FastQC software (Andrews 2010), which allowed the identification of sequence adapter contamination, overrepresented sequences, estimation of the PCR/Optical duplicate rate and overall sequencing quality.

All raw sequence data were then processed using Cutadapt (Martin 2011) software, in order to remove poor quality bases (quality score less than 20) as

well as Illumina universal sequencing adapter sequence (AGATCGGAAGAG) from the 3' end of reads.

The KSHV reference genome sequence was downloaded in fasta format from the NCBI website (Benson et al. 2005), while a GTF file containing genomic feature coordinates (ORFs, genes, exons, UTRs) was assembled manually using data from the KSHV 2.0 annotation dataset created by Arias et al (2014). The human hg38 reference genome sequence was downloaded from the UCSC FTP site in fasta format. The human hg38 genome annotation was downloaded using the UCSC Table Browser Tool (Karolchik et al. 2004). KSHV data were manually added to the human reference fasta and GTF files as an additional chromosome.

The genome sequences in the merged fasta file were indexed for alignment using STAR software (Dobin et al. 2013). Paired-end sequence data was subsequently aligned to this index using the splice-aware read aligner STAR, in paired-end, two-pass mode. Aligned reads in BAM format were sorted by coordinate and indexed using Samtools (Li et al. 2009) and PCR and optical duplicates flagged (but not removed) using Picard software.

Data downloaded from public repositories were aligned as above, except without the addition of KSHV reference genome.

#### ***6.3.3.2 m<sup>6</sup>A-Seq data analysis***

m6aViewer was used to identify the m<sup>6</sup>A peaks in the m<sup>6</sup>A-Seq data, as described in **Chapter 5** with the peaks exported to text files for subsequent analysis.

Peak motif analysis was performed by exporting the flanking 100 base of RNA sequence surrounding peaks in KSHV methylome to a fasta file. Sequences containing repetitive viral sequence were removed. The remaining data was then used for enriched sequence motif detection using the MEME (Bailey et al. 2009) software, with scrambled sequences used as a control. KSHV methylome maps were produced using custom Java code.

### **6.3.3.3 fRIP-Seq data analysis**

SND1 binding sites were initially identified at transcript-level resolution by counting reads in the SND1 IP (immunoprecipitated) and INPUT (control) sample data that mapped to each RefSeq and KSHV transcripts. rSubread R package (Liao et al. 2013) was used to obtain raw read counts as follows. Each uniquely mapping read pair was counted towards the total transcript count for each sample, while multi-mapping reads were counted as partial reads, based on the number of mapped positions. Since the library preparation protocol preserved the strand of the original RNA molecule, only 'correctly' stranded read pairs were counted for each transcript. This allows to differentiate more accurately between the expression of sense and anti-sense transcripts.

Read counts were normalised using TMM (trimmed mean of MA values) normalisation (Robinson et al. 2010) and DESeq2 R package (Love et al. 2014) was then used to identify transcripts that showed a significant increase in the coverage of the normalised IP samples when compared to the INPUT controls.

In order to increase the resolution of the SND1-bound regions, custom Java code was used that identified transcriptome regions that were enriched in the IP data when compared to the control data. Initially, the application segmented regions into intronic or exonic sequences: a region was classified as exonic if the sequence was present in at least one mature transcript. The per base read coverage was determined for both intronic and exonic sequences and normalised using the parameters determined from the TMM normalisation step to account for library compositions and sizes. Using a sliding window approach, regions which showed enrichment in the IP data (at least 1.5 fold enrichment in the IP data compared to the INPUT data and at least 20 reads in the IP fraction) were identified. All data from the different samples in the analysis were then merged to generate a single dataset that contained read depth data for all consensus enriched and unenriched regions (both intronic and exonic). Each segmented region was subsequently treated as an individual gene for re-analysis using DESeq2, in order to identify regions with a significant increase in IP over INPUT signal.

### **6.3.3.4 RNA lifetime profiling data analysis**

As before, raw read count data was normalised between libraries using TMM normalisation. Normalised read counts were then scaled by linear fitting of the ERCC RNA spike-in transcripts, as described by Wang *et al* (2014). The RNA degradation rate for each transcript was computed by fitting a non-linear regression model to the data, with the formula:

$$\text{Expression at 0 Hours} + \text{Decay Rate} * \text{Time}$$

The RNA half-life was then computed for each transcript as:

$$\text{Half-life} = \ln(2)/\text{-Decay Rate}$$

Transcripts with poor quality fits (97.5% and 2.5% confidence intervals > 0.2) were filtered out from subsequent analysis.

#### **6.3.3.5 Other Analyses**

Differential expression analyses were performed in R using Deseq2 package (Love et al. 2014). YTHDF2 and SND1 binding sites in public PAR-CLIP and eCLIP data were detected using PARalyzer software (Corcoran et al. 2011). Functional enrichment analyses were performed using the clusterProfiler R package (Yu et al. 2012). Alternative splicing events were detected using Spladder software (Kahles et al. 2016). Read coverage and splicing graphs were visualised using IGV data browser (Robinson et al. 2011).

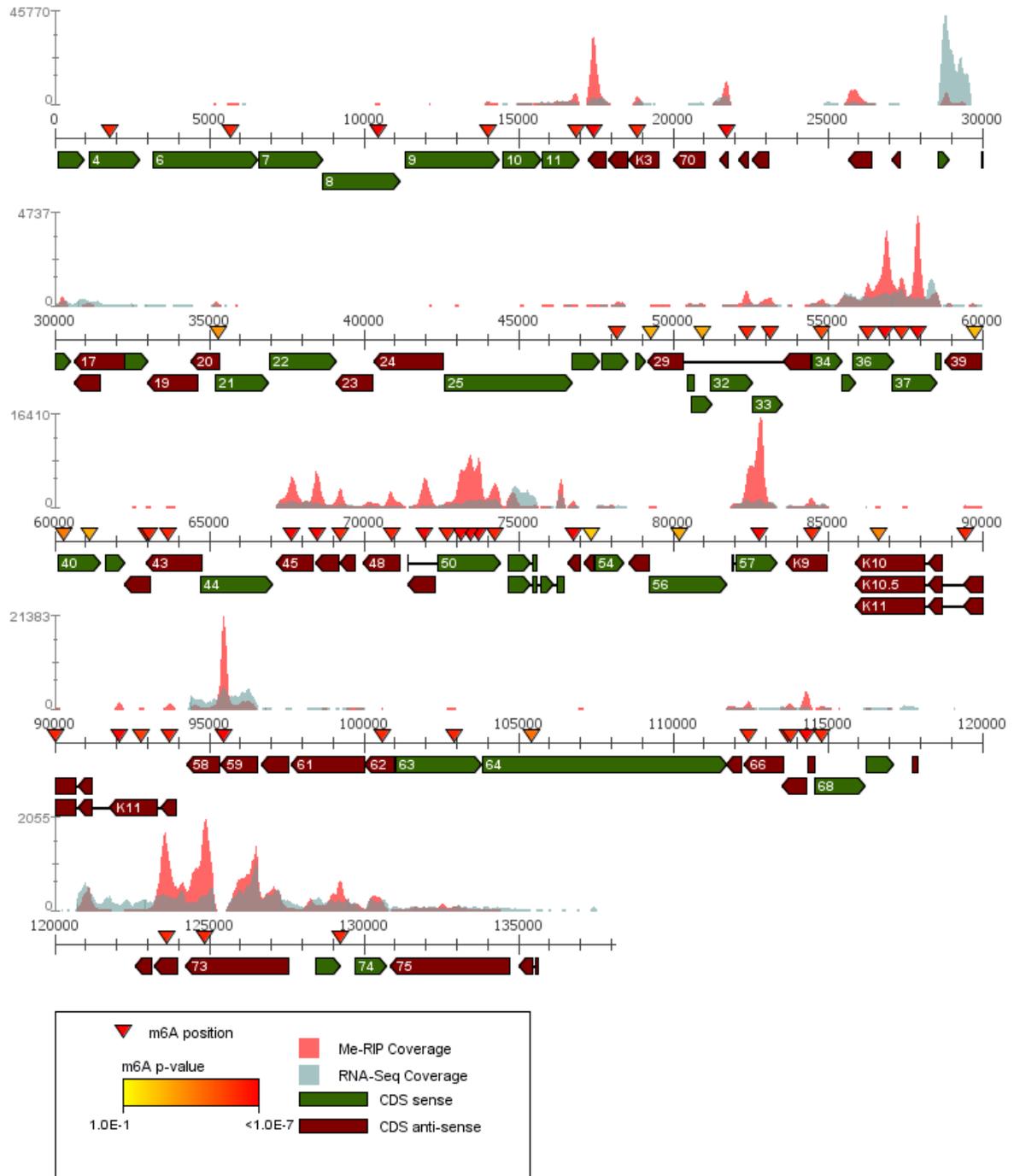
## **6.4 Results and Discussion**

### **6.4.1 KSHV Methylome**

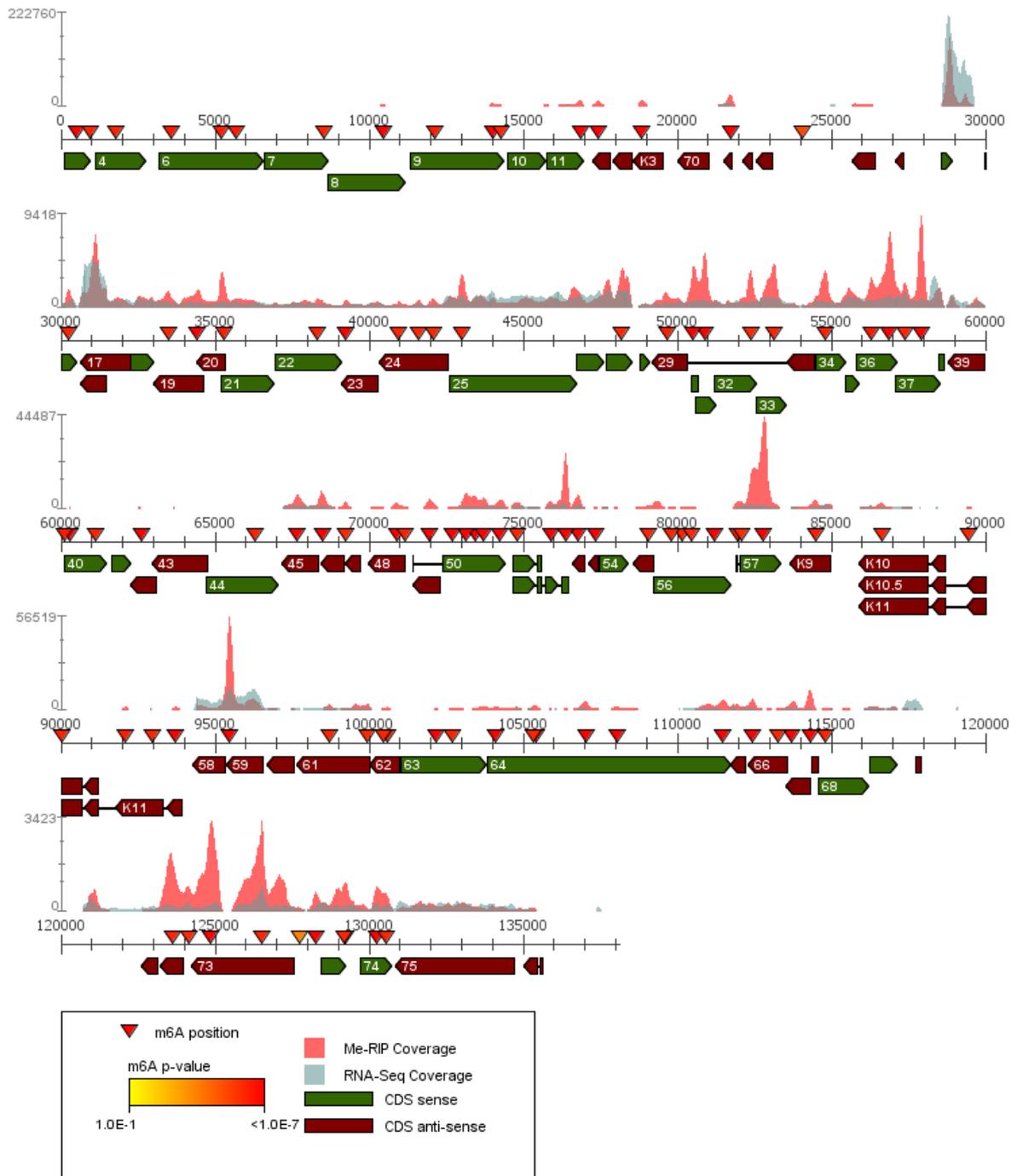
m<sup>6</sup>A modification was found to be widespread across both viral latent and lytic transcripts in the KSHV transcriptome. **Figures 68A, 68B and 68C** show the KSHV methylome maps at 0, 8 and 24 hours post-RTA activation respectively. KSHV m<sup>6</sup>A methylation was found to be largely consistent between biological replicates in KSHV transcriptome (**Figure 69**). In total, using a high confidence cut-off of at least 2-fold IP change over control, 34 viral m<sup>6</sup>A peaks were identified in latent cell transcriptomes; 57 were found at 8 hours post-reactivation; and 101 were detected at 24 hours post-reactivation.



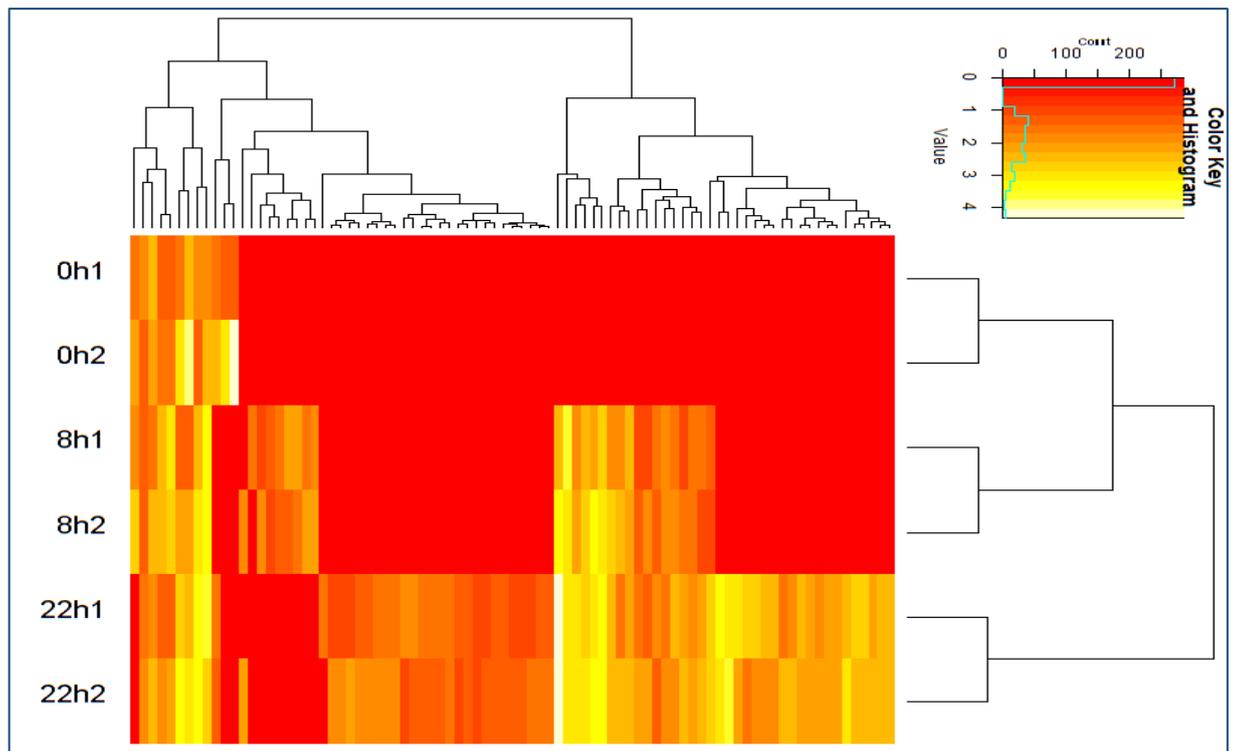
**Figure 68A.** m<sup>6</sup>A methylation of KSHV transcriptome at 0 Hour time point, prior to switch to lytic phase. Coverage derived from biological replicate 2. Note that each individual coverage data track is scaled to a different maximum, due to very variable viral gene expression levels.



**Figure 68B.**  $m^6A$  methylation of KSHV transcriptome 8 hours post-activation. Coverage data derived from biological replicate 2. Note that each individual coverage data track is scaled to a different maximum, due to very variable viral gene expression levels.



**Figure 68C.** m<sup>6</sup>A methylation of KSHV transcriptome 24 hours post-activation. Coverage data derived from biological replicate 2. Note that each individual coverage data track is scaled to a different maximum, due to very variable viral gene expression levels.



**Figure 69.** Heatmap showing m<sup>6</sup>A peak enrichment distribution in KSHV transcriptome at latent (0H) and early lytic (8 H) and late lytic (20H) stages. Peak calls from multiple samples were aggregated by identifying overlapping sites. The resulting peak enrichment over control matrix rows(samples) and columns (peaks) were clustered (hierarchical clustering, complete linkage) and enrichment values visualised as a heatmap. Replicate samples cluster together, indicating high m<sup>6</sup>A site reproducibility in viral RNAs. The majority of m<sup>6</sup>A sites are gained throughout the course of infection, though this corresponds to viral transcription activation.

For an m<sup>6</sup>A residue to be detected using m<sup>6</sup>A-Seq, the modified transcript has to be expressed above the detection cut off value. Consequently, only the genes used to maintain latent infection by KSHV can be reliably analysed at the 0 hour time point. This largely limits the analysis to the locus that encodes LANA (ORF73), a key viral latency maintenance transcript and other early latent genes (including vCyclin/ORF72): this locus was found to be heavily methylated at all time points. However, following activation of RTA and the commensurate switch from latent to lytic, KSHV begins to heavily transcribe the rest of the viral transcriptome, enabling m<sup>6</sup>A detection across other viral transcripts. Thus,

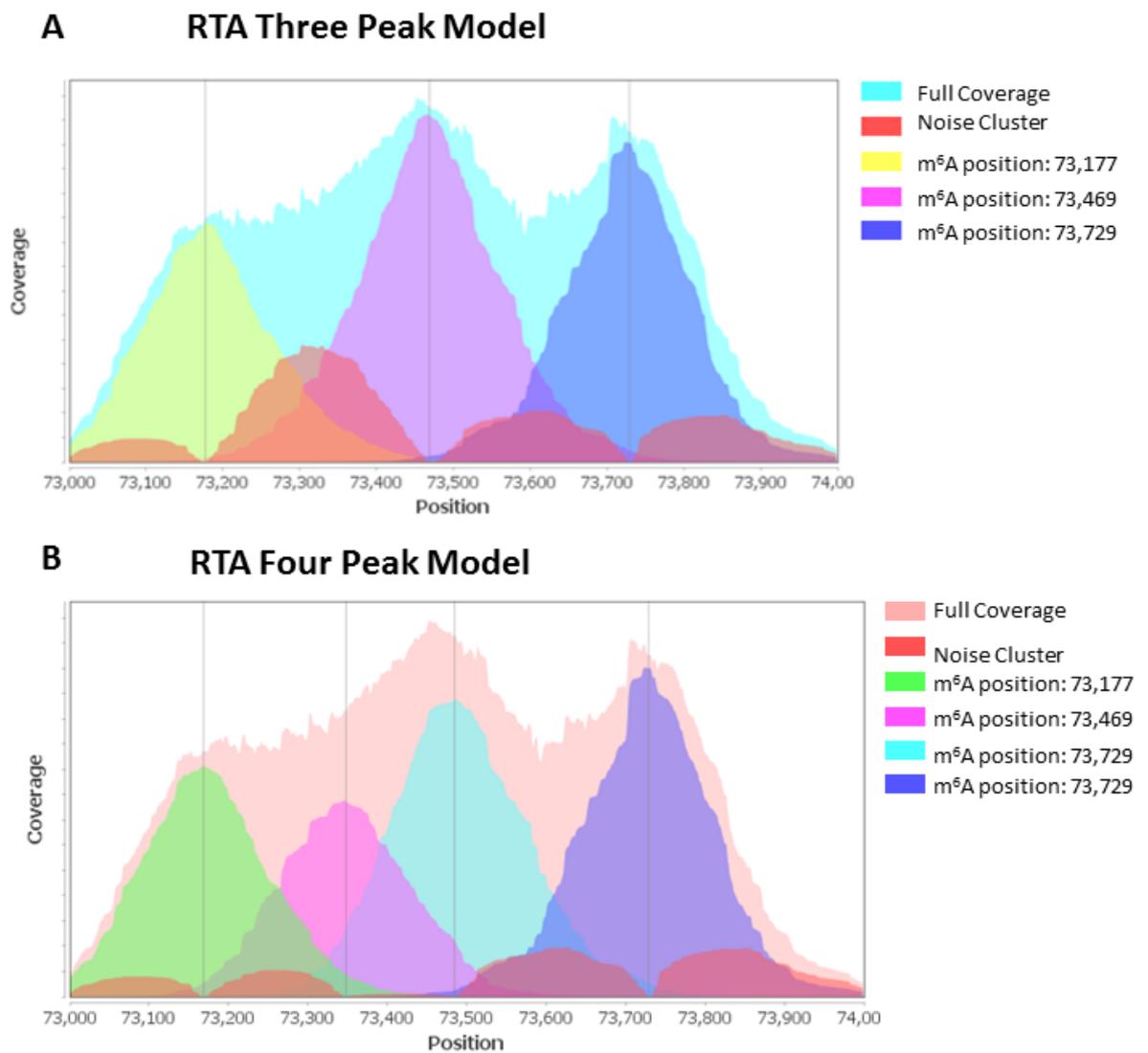
analysis of later time points shows that most of the early and late lytic transcripts harbour one or more methylation sites. It is important to note that a small proportion of (experimentally) un-activated latent cells undergo spontaneous lytic reactivation; consequently, it is not surprising that in latent cells both transcript expression and methylation was also detected in a number of lytic KSHV genes, including RTA, vIL6 (K2), vIRF-2 (K11) and ORF75.

RTA itself was found to contain multiple m<sup>6</sup>A sites that remain unchanged across all time points. Immunoprecipitated read coverage data distribution initially suggested that RTA may contain three distinct methylation sites; however, m<sup>6</sup>aViewer's model-based peak deconvolution suggests that the presence of four m<sup>6</sup>A residues is more likely (**Figure 70**).

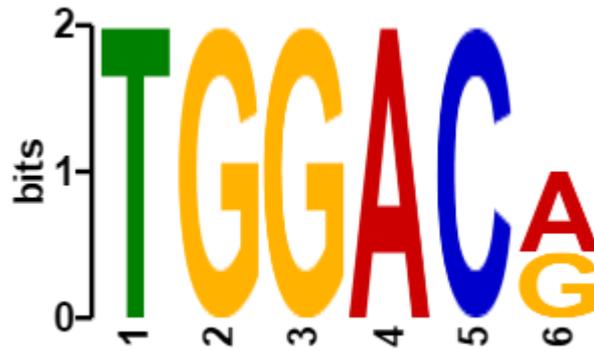
In order to ascertain whether viral m<sup>6</sup>A methylation sites also utilise the human methyltransferase 'RRACH' (typically GGACH) consensus sequence, motif analysis was carried out on 200 nt of sequences surrounding all detected m<sup>6</sup>A peaks. This analysis recapitulated a strong 'RRACH' motif, suggesting that the virus uses the host cell methyltransferase machinery (**Figure 71**). More specifically, the enriched motif consisted of a strong m<sup>6</sup>A motif preceded by thymine, and was found in the majority (56 out of 81) of queried KSHV m<sup>6</sup>A peak sequences.

Next, the location of detected viral m<sup>6</sup>As was compared against the distribution of m<sup>6</sup>A in cellular transcriptome. It was found that while in cellular transcripts m<sup>6</sup>A shows a preference for 3'UTR regions and coding regions; in the viral transcriptome non-coding m<sup>6</sup>A-modified bases constitute a much smaller proportion of the methylome (70-82% of all m<sup>6</sup>As detected in KSHV coding sequencing) (**Figure 72A and 72B**). This may be due to the much tighter organisation of the viral genome, with short UTR sequences and large open reading frames used to transcribe multiple genes at once. Furthermore, in KSHV, many open reading frames overlap and the same genomic coding strand sequence may act as both an ORF and a UTR for two different transcripts. Therefore, the prevalence of viral UTR methylation here may be difficult to estimate. Nonetheless, this highlights potential differences in m<sup>6</sup>A function between viral and host transcriptomes. 3' UTR-m<sup>6</sup>A mediated post-

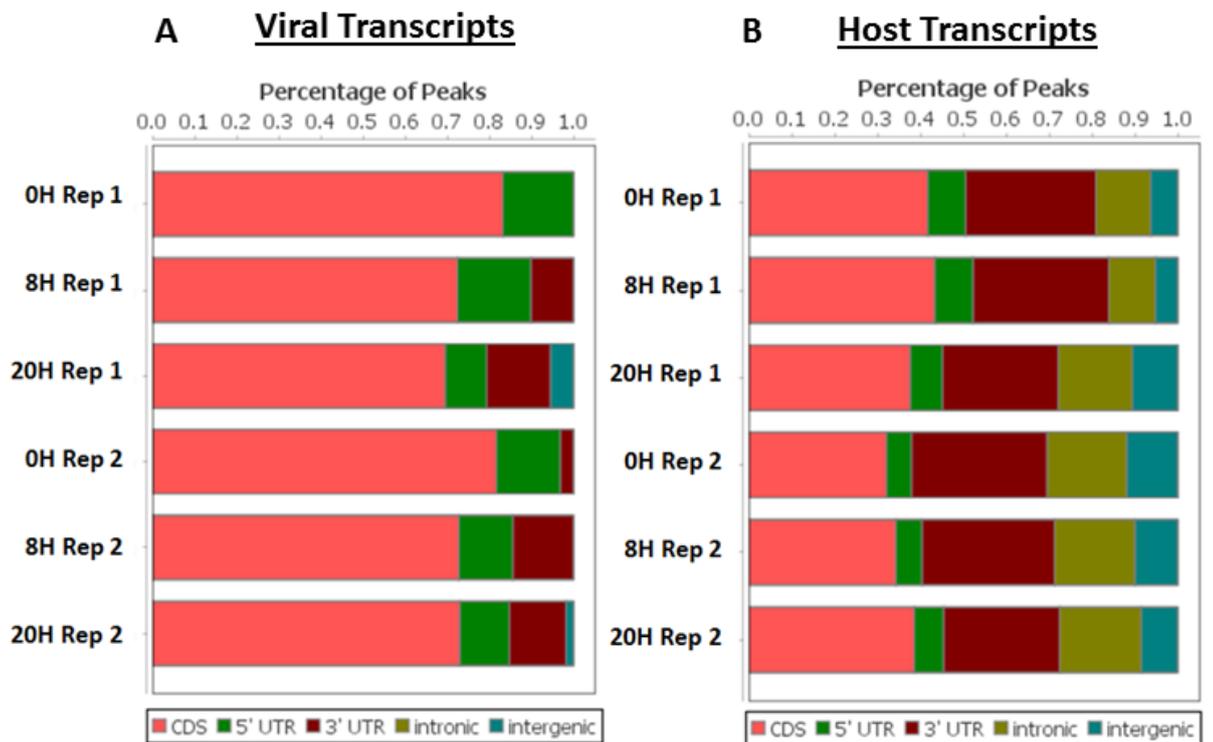
transcriptional regulation has previously been shown to regulate transcript stability via miRNA or reader protein recruitment and subsequent targeting to P-bodies for degradation. It is unlikely that hijacking these particular pathways would be beneficial to viral gene expression. Consequently, the different distribution of m<sup>6</sup>A residues in KSHV suggests that m<sup>6</sup>A methylation may serve a different set of functions in the virus when compared to the host cell.



**Figure 70.** Comparison between the three (**A**) and four (**B**) m<sup>6</sup>A residue model for IP read coverage in enriched region in RTA transcript. Fitting three peaks to the region such that the three m<sup>6</sup>A positions account for the majority of reads aligning to the region results in a cluster of reads between the first peak and the second peak that overlap neither peak position. On the other hand, a four peak model of the region fits an additional peak that accounts for these reads.

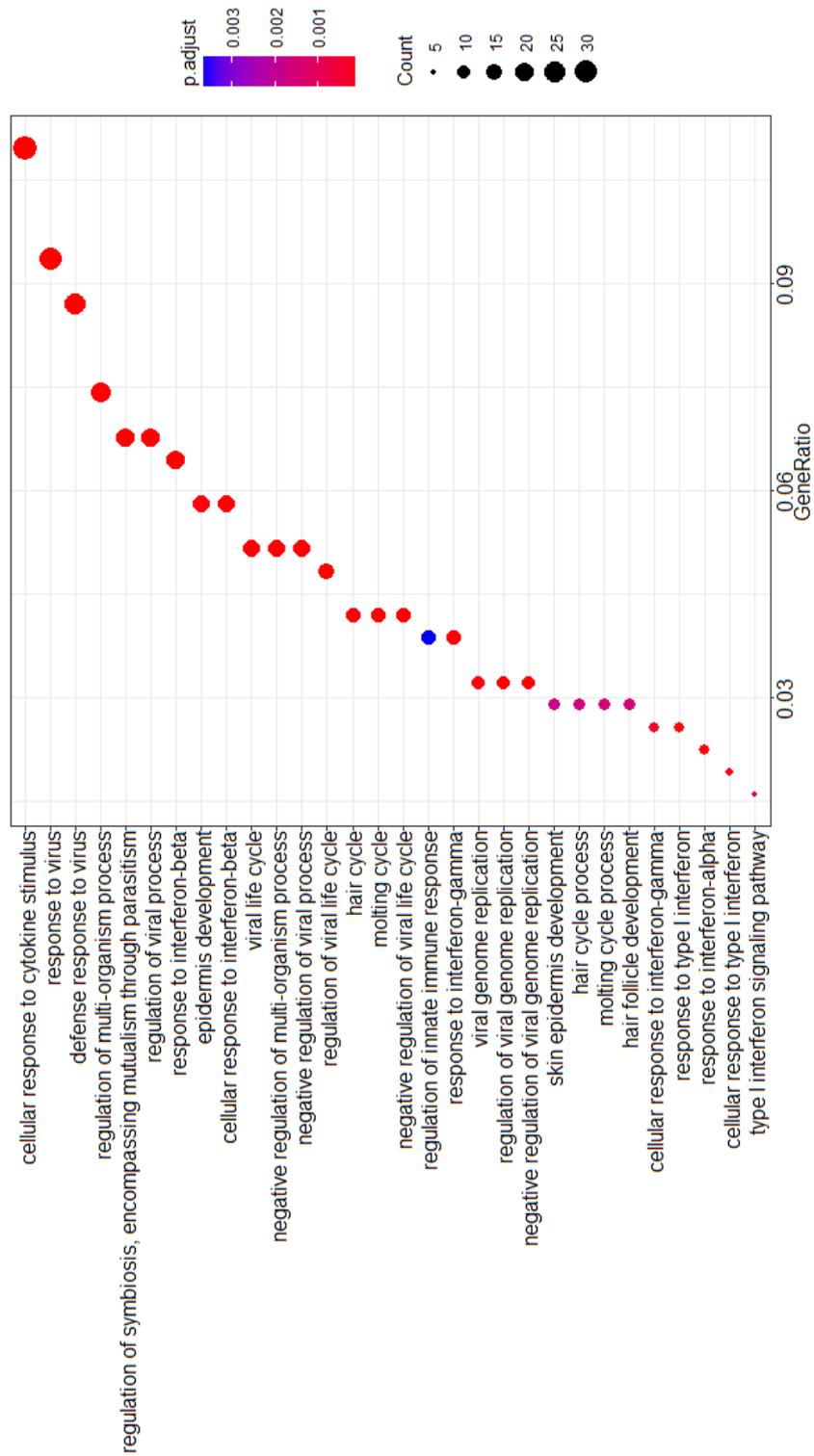


**Figure 71.** m<sup>6</sup>A consensus motif was recapitulated using KSHV peak sequences.



**Figure 72A.** m<sup>6</sup>A peak distribution in viral transcripts. **B.** m<sup>6</sup>A peak distribution in cellular transcripts.

While there were no differentially methylated m<sup>6</sup>A sites in the viral transcriptome (that could not be accounted for by m<sup>6</sup>A not being detected in transcripts that are also unexpressed), 903 total host m<sup>6</sup>A sites were found to be differentially methylated. These transcripts were found to be enriched for functions relating to virus defence mechanisms, stress responses and interferon response pathways (**Figure 73**). These results confirm that m<sup>6</sup>A modification is not static and suggest that differential m<sup>6</sup>A methylation exerts post-transcriptional control and modulates the host virus defence response. Whether this is a result of the host fighting the infection or the virus down regulating the anti-viral response is not yet known.



**Figure 73.** Top GO biological process functions enriched in differentially methylated transcripts across the experiment time course.

#### 6.4.2 Characterisation of a new putative m<sup>6</sup>A ‘reader’

In order to determine the precise function m<sup>6</sup>A modification has within the viral transcriptome, it is important to identify the reader proteins which recognise the modified sites. Here, comparative mass spectrometry analysis carried out by collaborators (Whitehouse group, FBS) identified a putative novel m<sup>6</sup>A reader protein, SND1 (Tudor staphylococcal nuclease), that preferentially binds a methylated oligo containing the flanking sequences of the RTA transcript’s first m<sup>6</sup>A peak position with greater affinity than other known m<sup>6</sup>A readers YTHDF1-3, YTHDC1 and hnRNPA2B1 (**Table 12**).

<i>Protein</i>	<i>ORF37</i>		<i>RTA 1st peak</i>		<i>RTA 4th peak</i>	
	<i>A oligo</i>	<i>m<sup>6</sup>A oligo</i>	<i>A oligo</i>	<i>m<sup>6</sup>A oligo</i>	<i>A oligo</i>	<i>m<sup>6</sup>A oligo</i>
<i>YTHDF1</i>	3	10	5	10	0	10
<i>YTHDF2</i>	2	9	4	9	0	9
<i>YTHDF3</i>	1	9	3	9	0	9
<i>YTHDC1</i>	0	2	0	8	0	2
<i>SND1</i>	0	1	2	27	0	0
<i>hnRNPA2B1</i>	19	18	16	15	16	15

**Table 12.** Mass spectrometry identified SND1 as a putative m<sup>6</sup>A reader protein. Known m<sup>6</sup>A readers are highlighted in blue. Putative novel reader SND1 is highlighted in red. The values are counts of unique identified peptides.

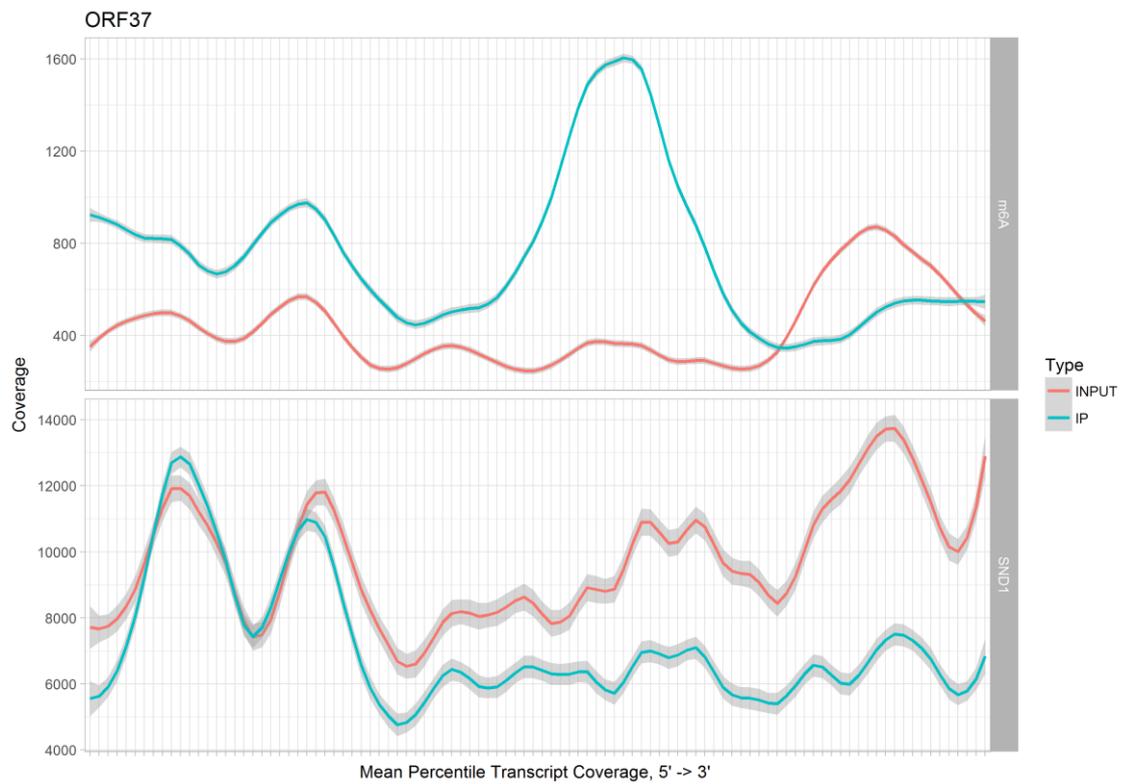
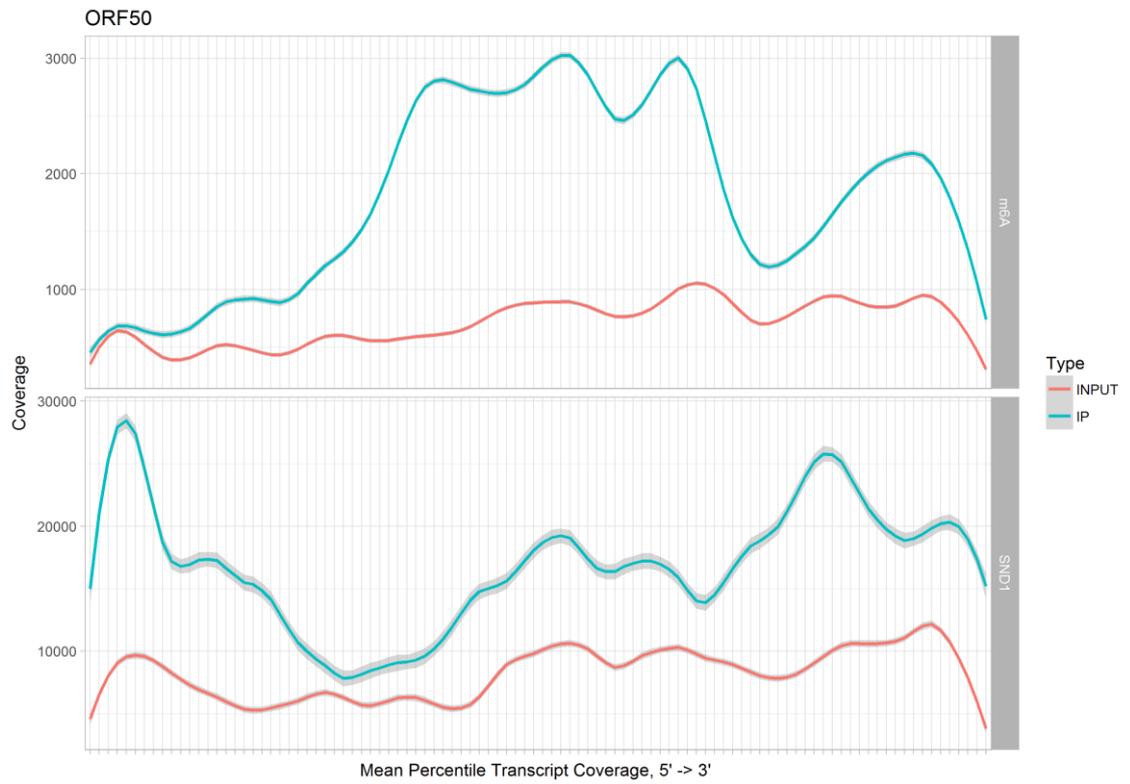
### **6.4.3 Identification of transcriptome-wide SND1 binding sites**

In order to further investigate the role of SND1 in the context of m<sup>6</sup>A and KSHV infection, transcriptome-wide SND1-fRIP-Seq was carried out. fRIP-Seq is in principle a very similar method to m<sup>6</sup>A-Seq. RNA-protein binding sites are identified by first formaldehyde cross-linking proteins to RNA in order to prevent post-lysis dissociation, and then isolation of protein-associated RNAs by immunoprecipitation. Similar to m<sup>6</sup>A-Seq, an RNA-Seq control is compared to the immunoprecipitated fraction in order to define high confidence protein binding sites. In contrast to m<sup>6</sup>A-Seq, fRIP-Seq is reportedly a much lower resolution technique, as cross-linked RNA cannot be easily sheared to a small fragment size. Nonetheless, while SND1 binding sites cannot be precisely narrowed down using this technique, it is possible to identify transcript-level enrichment.

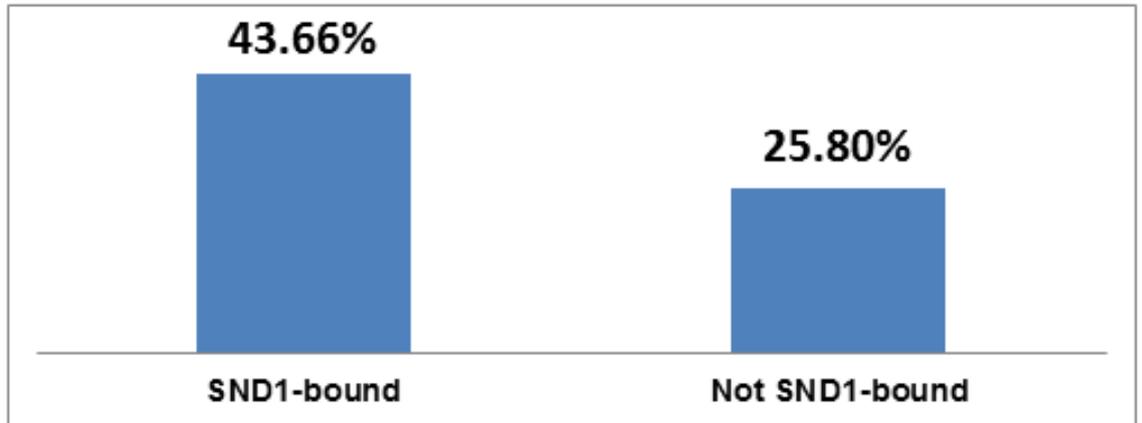
In parallel to m<sup>6</sup>A-Seq, SND1-fRIP-Seq was carried out by our collaborators (Whitehouse Group, FBS) at three time points, capturing latent and lytic KSHV infection phases. Data was then processed as described in the **Methods** section. 5082 transcripts were identified as being significantly enriched in IP over INPUT across all time points.

In confirmation of the mass spectrometry results, SND1 was found to be significantly enriched across the RTA/ORF50 transcript, but not ORF37 (**Figure 74**). In order to identify SND1 target transcripts where SND1 binding may be mediated by m<sup>6</sup>A, fRIP-Seq data was overlapped with previously identified m<sup>6</sup>A sites. To limit any bias from reduced sensitivity to detect either SND1-binding or m<sup>6</sup>A methylation in transcripts with low expression, transcripts which showed at least high expression (> 200 FPKM) in both m<sup>6</sup>A-Seq and SND1-fRIP-Seq INPUT controls were selected for further analysis. A subset of all high confidence SND1 targets (> 2 fold change in IP over INPUT) was selected, as well as a subset of high confidence transcripts not targeted by SND1 (> 2 fold increase in INPUT over IP). In keeping with the hypothesis that SND1 is a putative m<sup>6</sup>A reader, SND1 target transcripts were found to be 1.69 times (43.66% vs 25.80% SND1-m<sup>6</sup>A overlap) more likely to be m<sup>6</sup>A methylated than

transcripts with similar expression levels that are not targeted by SND1 (**Figure 75**).



**Figure 74.** Normalised percentile read coverage (5' to 3') across RTA/ORF50 and ORF37 transcript in m<sup>6</sup>A-seq and SND1-fRIP-Seq data. Both ORF50 and ORF37 show m<sup>6</sup>A enrichment, but only ORF50 is enriched in SND1-fRIP-Seq.

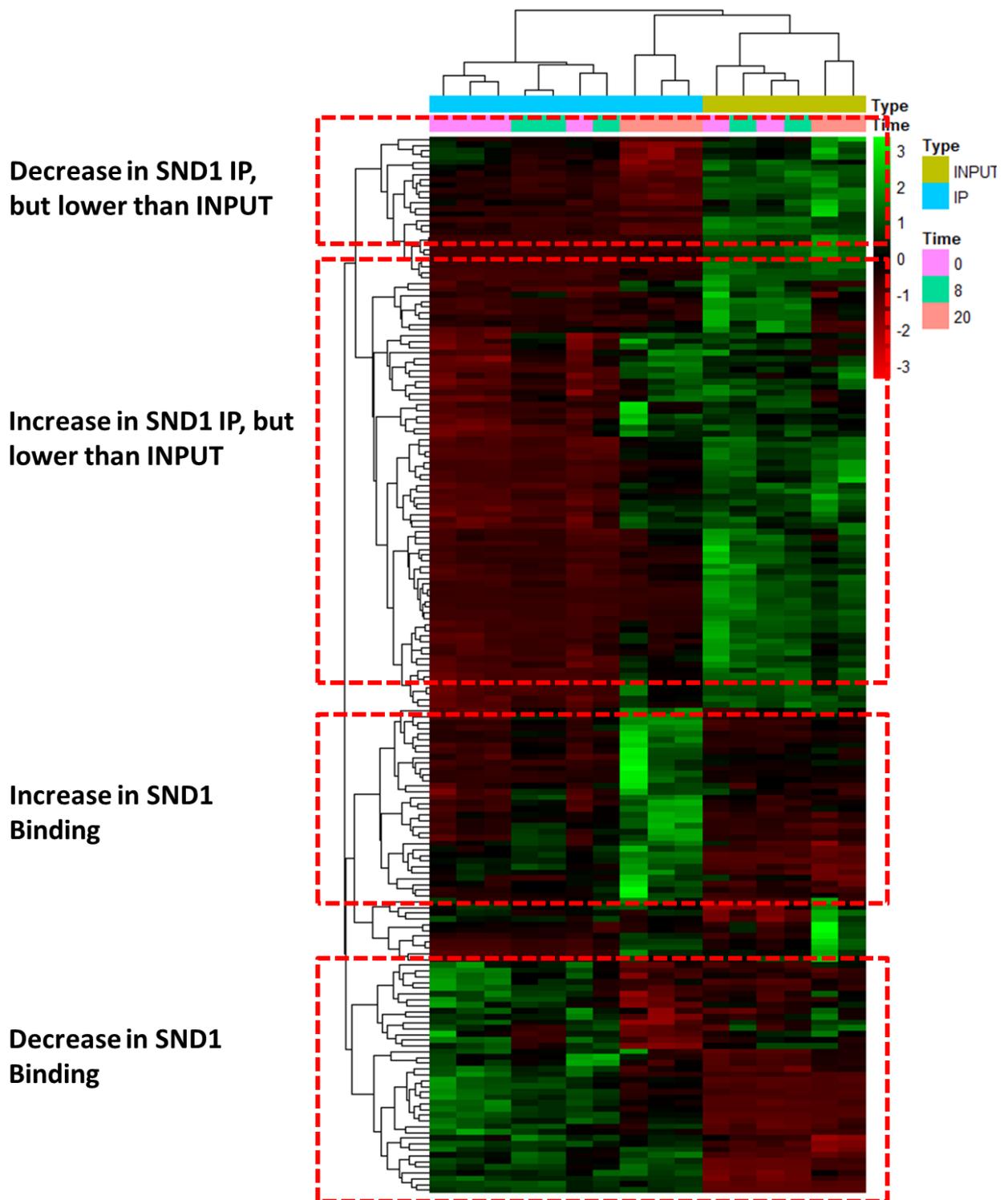


**Figure 75.** Percentage of m<sup>6</sup>A overlap with highly expressed transcripts identified as targets of SND1 (> 2 fold change IP/INPUT) or not targets of SND1 (> 2 fold change INPUT/IP) transcripts.

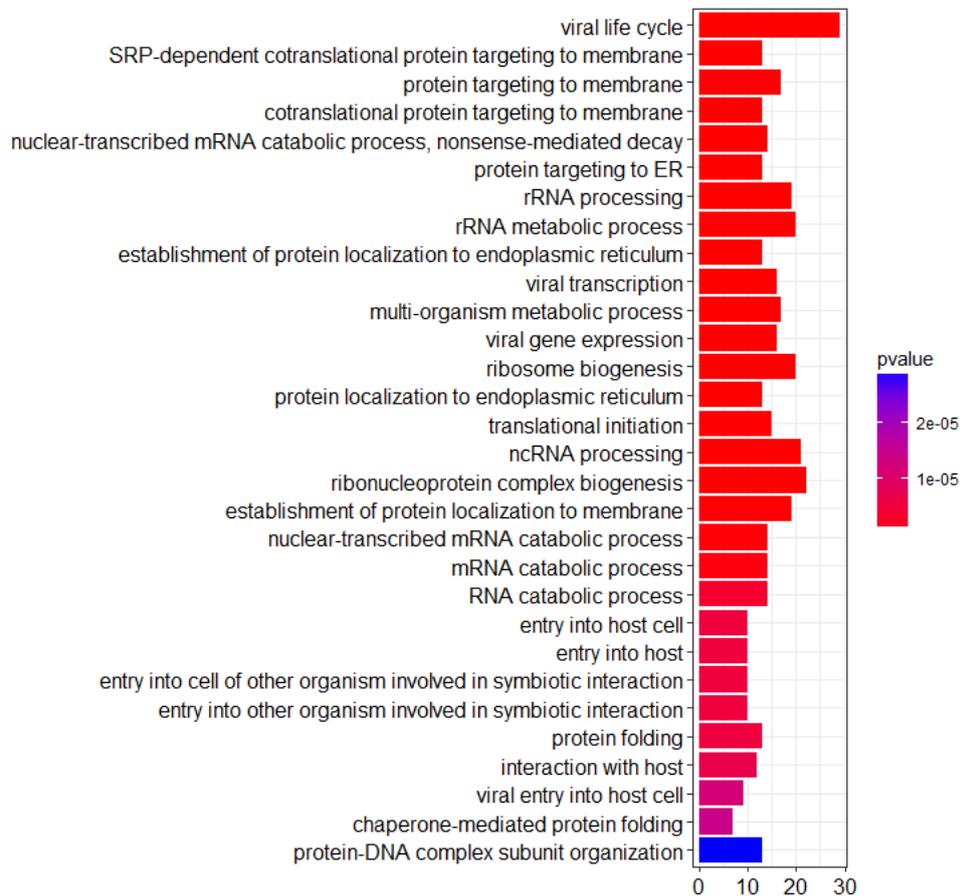
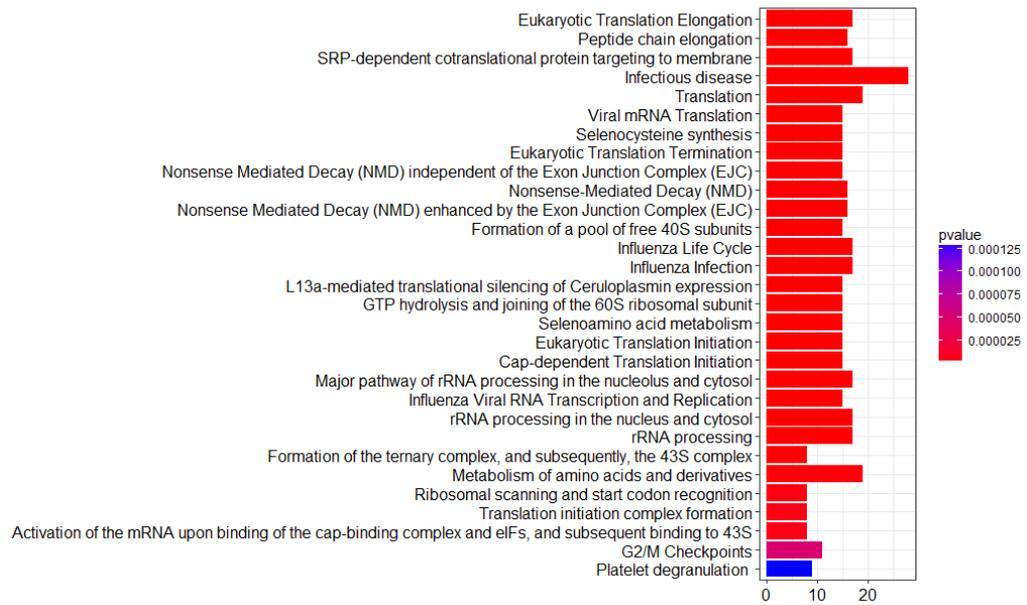
Next, 294 transcripts were identified as significantly differentially IP-enriched across the 0H, 8H and 20H time course (**Figure 76**). These were, unsurprisingly, enriched for gene functions relating to virus-host interactions, suggesting that SND1 plays a crucial role in the regulating these transcripts. Additionally, this subset was highly enriched for genes participating in translation initiation, elongation and termination pathways, as well as co-translational protein targeting to ER (**Figure 77**), reflecting the state of the cell during viral lytic replication, where KSHV genes are rapidly transcribed and translated for virion assembly and release.

The analysis highlighted four separate clusters of transcripts (**Figure 76**): a small group of transcripts exhibiting mostly stable expression levels, but showing an increase in SND1 binding over the time course; a larger group of mostly stably expressed transcripts which show a loss of SND1 binding over time; a group of transcripts with somewhat variable expression, where SND1 binding strongly increases over time at a greater rate than the increase in

expression levels; and finally, the largest group where transcript expression is largely stable and SND1 binding appears to increase over the time course – however, the transcript-level SND1-IP coverage in this group does not exceed that of the INPUTs. This could be due to a number of factors, such as inaccurate library size normalisation or, localised SND1 enrichment which would suggest that SND1-fRIP-Seq technique may have higher than transcript-level resolution.



**Figure 76.** Transcripts identified as significantly varying in IP/INPUT enrichment over the experiment time course. The heatmap shows a number of transcript clusters showing differential SND1 binding across KSHV infection time course. Genes can be grouped into increasing or decreasing binding clusters, as well as higher and lower IP vs INPUT groups.



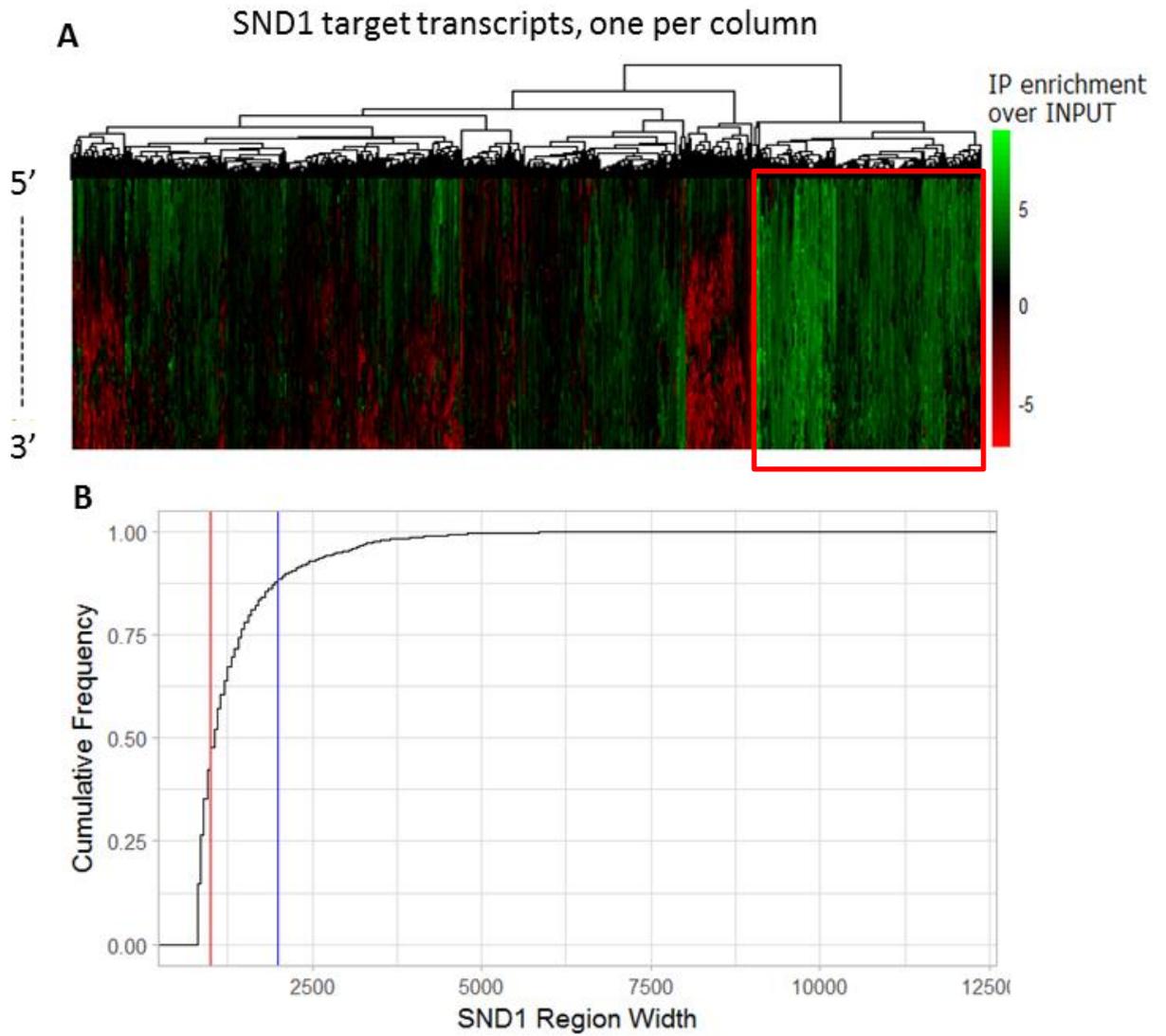
**Figure 78.** Top GO biological process functions (top) and Reactome pathways (bottom) enriched in differentially SND1-bound transcripts across the experiment time course.

In order to investigate whether higher resolution identification of binding sites from SND1-fRIP-Seq data was possible, 5' to 3' percentile coverage was computed for each transcript identified as enriched across all time points. Transcripts were then clustered using hierarchical clustering. **Figure 79A** shows that a number of distinct transcript coverage distribution clusters exist, representing distinct read distribution patterns. While some of the identified transcripts show enrichment across the whole transcript, others show a strong preference for 5' of the transcript.

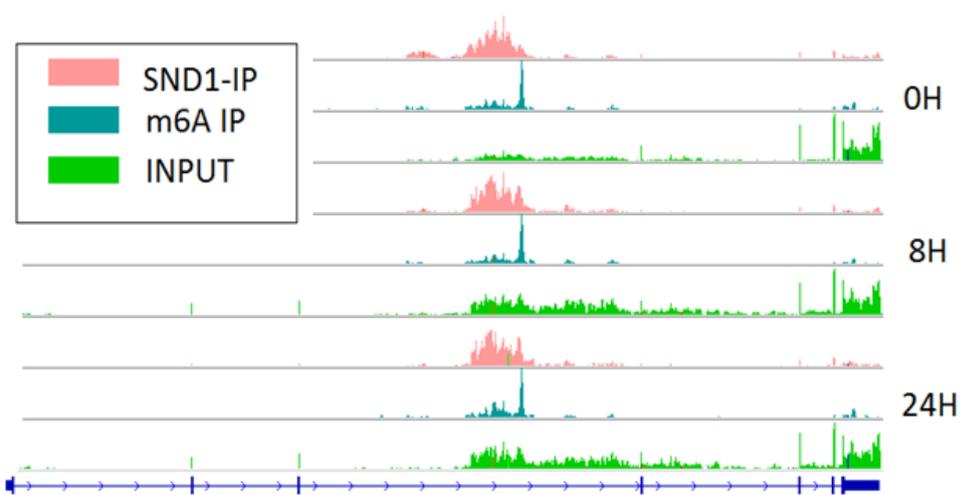
This prompted further investigation into the resolution of the SND1-fRIP-Seq data. In order to narrow down SND1 binding sites, the transcriptome was segmented using a sliding window approach into regions which could be considered IP-enriched and those that are not enriched. Introns and spliced transcripts were treated separately to preserve the distinction between nascent and mature RNA transcripts. This resulted in the segmentation of the transcriptome into 741,170 distinct regions, out of which 107,374 could be considered statistically significantly enriched in SND1 IP fraction.

These were found to be mostly less than 2Kb in size, with just under half of all significantly enriched regions being under 1Kb in size (**Figure 79B**). Thus, while the resolution of fRIP-Seq is much poorer than that of m<sup>6</sup>A-Seq (200-300bp) and no distinct 'peaks' can be detected, SND1 binding sites can be narrowed down to approximately 1-2kb resolution.

SND1-binding sites may therefore appear more defined in the introns of nascent transcripts, since intronic regions are generally much longer than exons. Thus, intronic regions were investigated in more detail. It could be observed, that as expected, in larger introns SND1 binding sites could be identified as distinct 1-2kb read clusters enriched in IP fraction. Remarkably, many of these clusters also directly overlapped m<sup>6</sup>A residues identified by m<sup>6</sup>A-seq – for example, the third intron of DUSP22 transcript (**Figure 80**).



**Figure 79A.** Percentile-based coverage of identified SND1-target transcripts. Transcript cluster showing whole transcript enrichment is highlighted in red. **B.** SND1 IP-enriched fRIP-Seq region size cumulative distribution. Red marker = 1kb; blue marker = 2kb.



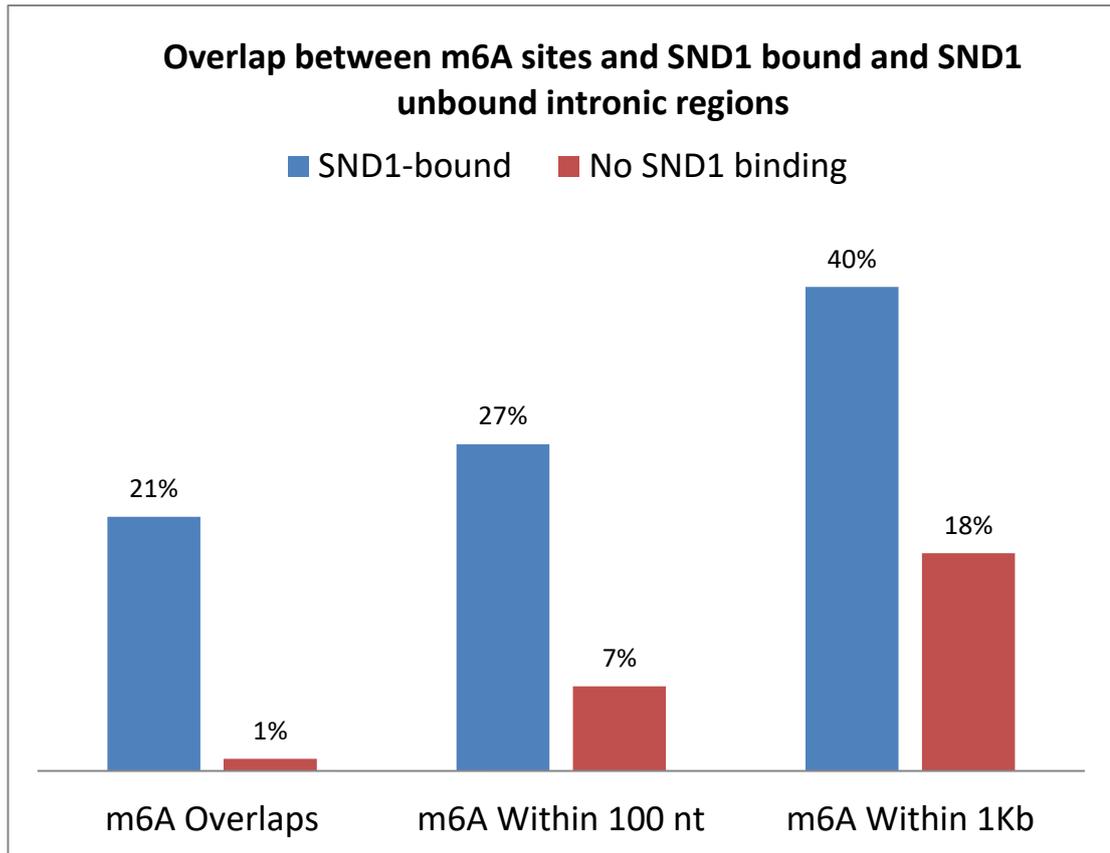
**Figure 80.** DUSP22 gene read coverage as visualised in IGV. Localised SND1 enrichment can be easily observed in longer introns, such as in the third intron of DUSP22 gene. m<sup>6</sup>A is often co-localised with SND1 intronic enrichment, as is the case here.

In order to investigate whether SND1 binding in these intronic regions is likely to be mediated by the presence of m<sup>6</sup>A, intronic SND1 regions (5406 total) were intersected with previously detected m<sup>6</sup>A peak positions. As a control, all intronic regions (spanning total of 17469 unique sites in the genome) which showed at least moderate expression (more than 50 reads per region) and were depleted for SND1 IP reads ( $< -1 \log_2$  fold change) were selected. 21% of all intronic SND1-bound regions were identified to directly overlap an m<sup>6</sup>A site – compared with just 1% of all control regions (**Figure 81**). This suggests that m<sup>6</sup>A modification could be co-regulated together with SND1, or m<sup>6</sup>A is facilitating the binding of SND1.

In order to investigate whether SND1 may be an indirect m<sup>6</sup>A reader, not binding the modification directly, but rather upstream or downstream of it, SND1-bound and control regions were extended either side by 100 nt to investigate short range effects and by 1Kb to investigate putative longer-range effects. Reasoning that if SND1 acts downstream/upstream of m<sup>6</sup>A, there should be an increase in m<sup>6</sup>A overlap in these larger regions when compared to control regions. **Figure 81** shows that, as expected, when the SND1 region is expanded by including some upstream and downstream regions, the amount of

overlap with previously identified m<sup>6</sup>A positions increases. This increase is consistent between SND1-bound intronic regions and control regions, suggesting that m<sup>6</sup>A is not more likely to appear upstream/downstream of SND1-bound regions, but rather directly overlap it.

While 21% overlap with m<sup>6</sup>A sites may seem somewhat small, SND1 is a multi-functional protein with diverse roles, and as such, not all of these functions are likely to be mediated via m<sup>6</sup>A recognition. It is also worth noting that this overlap is considerably higher than has been reported for other m<sup>6</sup>A readers, such as hnRNPA2B1. Alacorn *et al* (2015) found that out of 39,737 hnRNPA2B1 binding sites identified by PAR-CLiP, only 2096 (5.2%) overlapped m<sup>6</sup>A residues.

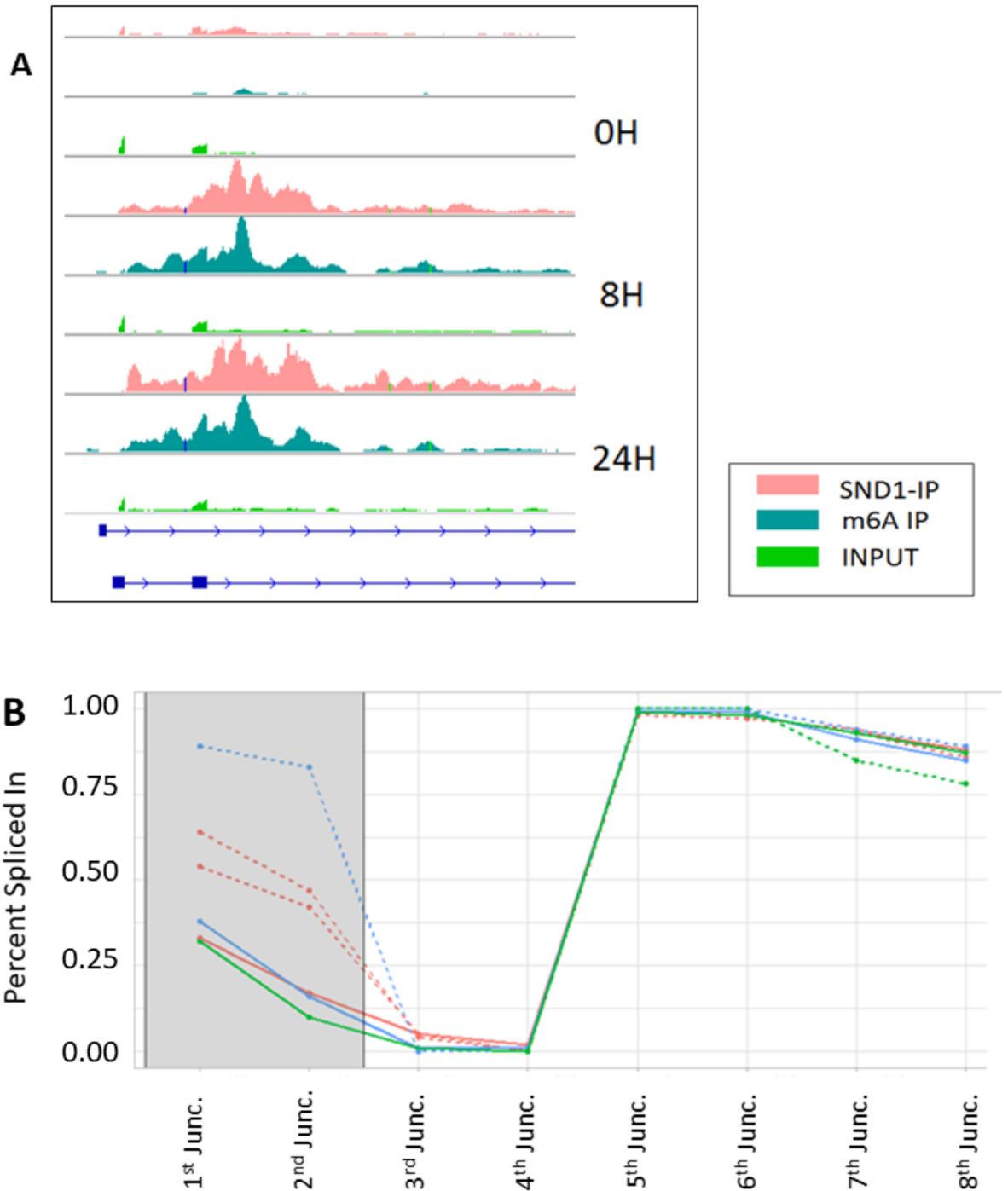


**Figure 81.** Overlap between intronic SND1-enriched/unenriched regions and m<sup>6</sup>A. Intronic SND1-bound regions were overlapped with corresponding regions in m<sup>6</sup>A-seq dataset and 21% of all sites were found to directly overlap identified m<sup>6</sup>A residues; m<sup>6</sup>A was detected with 100bp of 27% of SND1-bound regions and within 1Kb in 40% of SND1-bound regions. As a control, unbound but expressed intronic regions were selected, in each case displaying lower levels of overlap with m<sup>6</sup>A residues.



oligo, suggesting that these proteins may be recruited by and/or interacting with SND1.

In line with this hypothesis, some interesting cases have emerged. For example, GUK1 gene is not differentially expressed, but was found to gain an intronic SND1 binding site over the time course (**Figure 83A**). This gain of SND1 binding is also mirrored in a gain of an m<sup>6</sup>A peak overlapping the binding site. Interestingly, this also coincides with an increase of split-reads supporting the inclusion of the second cassette exon, but only in the SND1 IP fraction – splicing distribution remains largely unchanged in the INPUT fraction (**Figure 83B**). This would suggest, that fRIP-Seq is capturing and enriching for SND1 binding events on nascent transcripts during the RNA transcription/processing in the nucleus; and these events are ‘diluted’ when considering total cell RNA, which mostly consists of the cytoplasmic fraction.



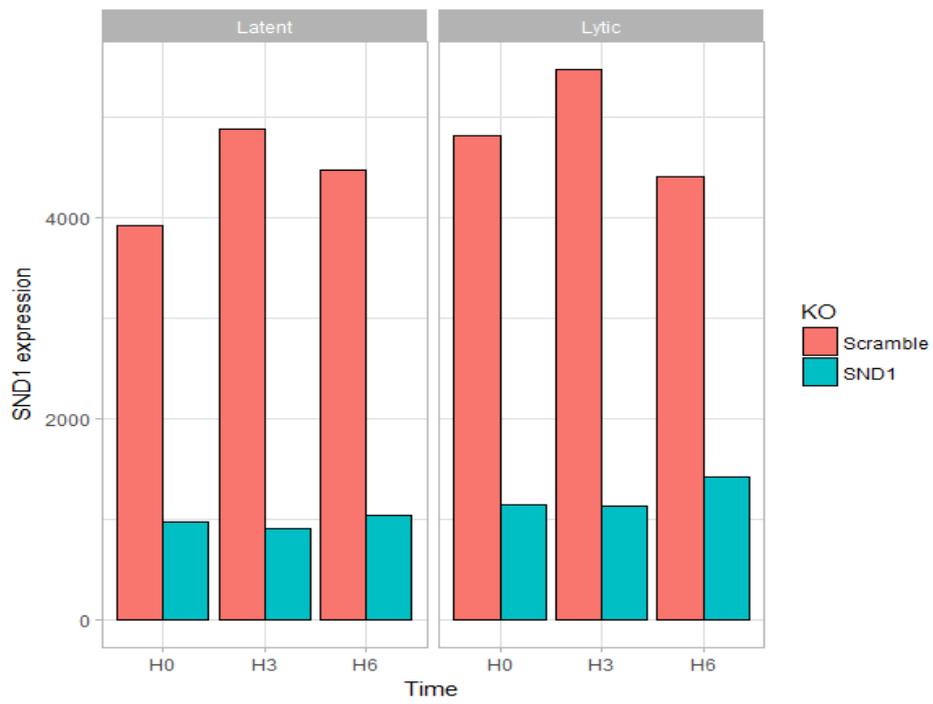
**Figure 83A.** Gain of SND1 binding at 8/24H in the intronic region of GUK1 is mirrored by the gain in m<sup>6</sup>A methylation in the same region. **B.** Split-read distribution as percentage spliced in (PSI) of alternative GUK1 exons in IP and INPUT. 20H IP time point not shown for all junctions due to lack of coverage. Differential junctions highlighted in grey. IP samples are shown as dotted lines, control INPUTs as solid lines. 0 hour time point is shown in red, 8 hour in blue and 20 hour in green.

#### ***6.4.4 RNA gene expression, lifetime profiling and alternative splicing analysis in SND1-depleted BCBL-1 cells***

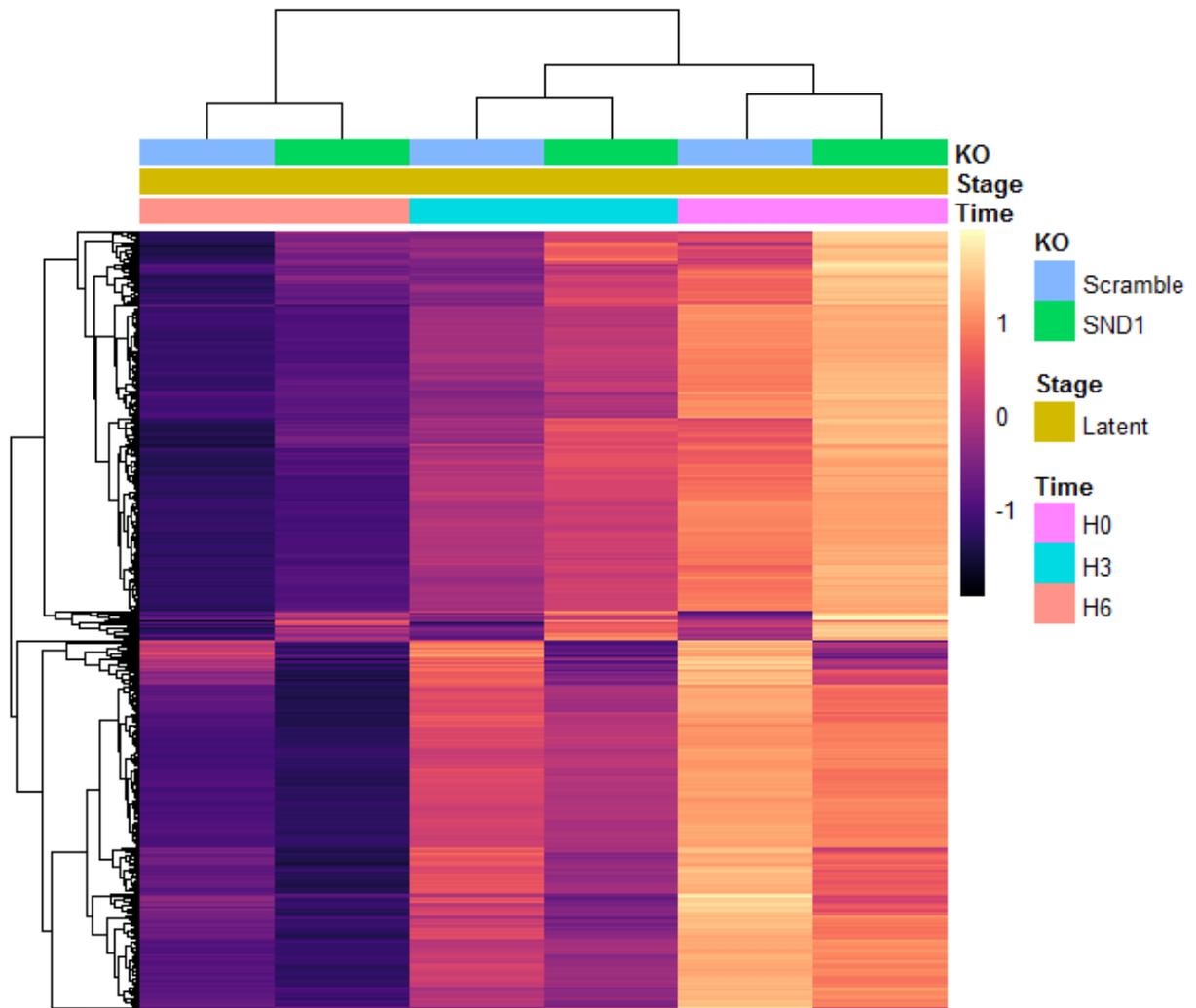
In order to further investigate the role SND1 may have in regulating its target transcripts, SND1 was knocked down in KSHV-infected latent and lytic BCBL-1 cells and RNA lifetime profiling assay was carried out by collaborators (Whitehouse Group, FBS). The experimental set up is summarised in **Table 13**, and data was processed as described in **Methods** section.

Initially, SND1 was confirmed to be depleted in knockdown samples in RNA sequencing data (**Figure 84**). SND1 depletion directly impacts the expression of 1643 transcripts in latent cells (**Figure 85**), resulting in consistent up or down regulation across most samples sequenced. The majority of differentially expressed transcripts identified were not targeted by SND1 in the SND1-fRIP-Seq data (**Figure 86**), suggesting that the immediate effects on gene expression of SND1 knockdown are likely due to the transcriptional, rather than post-transcriptional regulation by SND1. Interestingly, transcripts which were found to be upregulated on SND1 knockdown were more likely to be SND1 target RNAs than down regulated or unaffected transcripts. As SND1 has been previously described as a transcriptional activator, these results highlight a putative dual role of this protein. Down regulated RNAs were found to be less likely to be directly targeted by SND1, thus SND1 may exert its regulatory role on these molecules via transcriptional control, whereas up regulated RNAs were more likely to be directly targeted by SND1, thus suggesting that at least some of these RNAs may be regulated post-transcriptionally.

In order to further investigate the effects of SND1 knockdown on cellular transcripts beyond transcriptional expression changes, alternative splicing events were detected using Spladder software. In total, 3932 single exon skipping events, 808 multiple exon skipping events, 1671 intron retention events, 1587 alternative 3'UTR usage events and 2013 5'UTR usage events were detected across all samples in the data. No alternative splicing events were found to be statistically significantly different between SND1 knockdown and control samples.



**Figure 84.** SND1 mRNA was confirmed to be depleted to approximately 25% of scramble control in the knockdown samples.



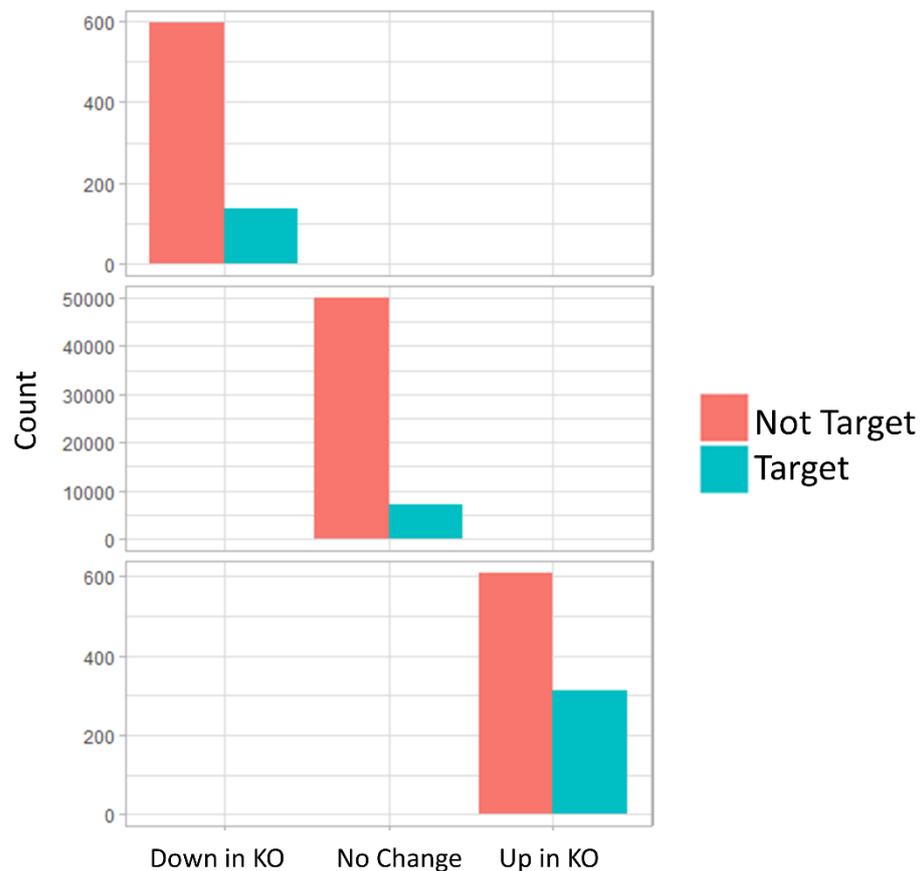
**Figure 85.** Heatmap showing cellular transcripts in latent BCBL-1 cells that were identified as significantly differentially expressed upon SND1 knockdown.

Latent Cells		Lytic Cells	
Time Point	Knock down	Time Point	Knock down
0H	SND1	0H	SND1
0H	Scramble	0H	Scramble
3H	SND1	3H	SND1
3H	Scramble	3H	Scramble
6H	SND1	6H	SND1
6H	Scramble	6H	Scramble

**Table 13.** Experimental sample set up of SND1 knockdown and control RNA life time profiling in latent and lytic cells. The experiment was repeated across two biological replicates.

The analysis was repeated using publicly available SND1 knockdown data from the ENCODE project in Hep2G cells, to the same result (1335 alternative 3'UTR, 1629 alternative 5'UTR, 3307 single exon skip, 600 multiple exon skip and 1196 intron retention events across all samples, with no significantly differential events detected). The CD44 gene was previously reported to be alternatively spliced on SND1 depletion in prostate cancer cells, however the gene is not expressed in BCBL-1 cells or in Hep2G cells, and thus this event could not be confirmed in available RNA sequencing data.

Next, the effects of SND1 knockdown on stability of its target transcripts were investigated using RNA lifetime profiling. RNA half-lives for each transcript were computed as described in the **Methods** section using a time course experiment of actinomycin D treated cells, to allow the measurement of the rate of degradation of each individual transcript.



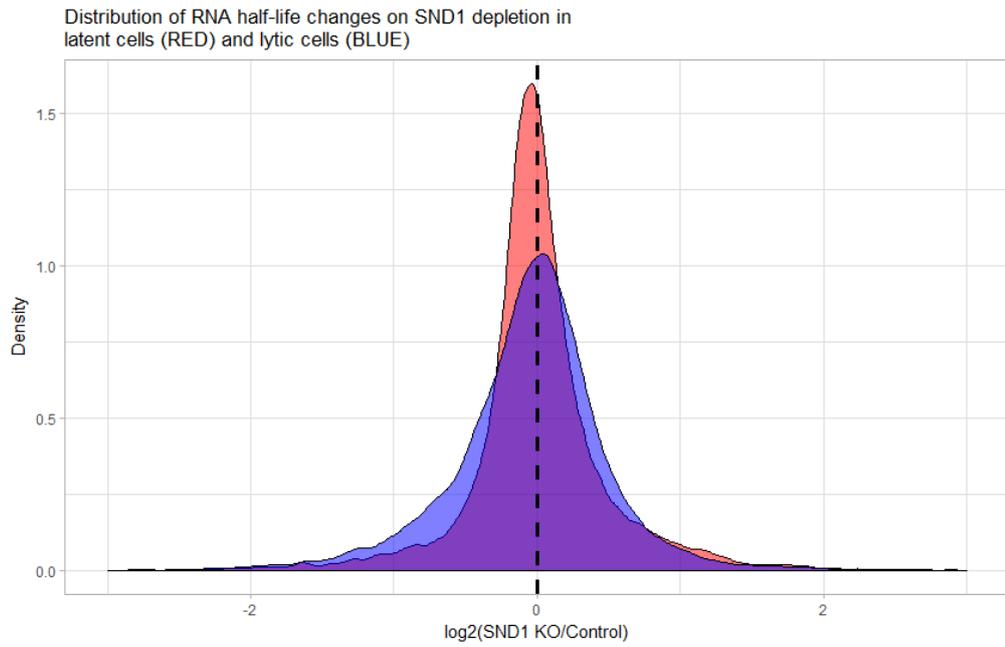
**Figure 86.** Summary of transcripts which are down regulated on SND1 depletion, transcripts which show no change in expression on SND1 depletion and transcripts which are up regulated on SND1 depletion and the proportion of each group that was found to be a target RNA for SND1 protein.

The rate of degradation of the majority of cellular transcripts in both lytic and latent cells were unaffected by SND1 depletion, although more variation was observed in lytic cells (**Figure 87**). Interestingly, when examining the effects of SND1 depletion on SND1 target transcripts only, as opposed to all RNAs shown in **Figure 87**, different patterns emerged. In latent cells, SND1 target transcripts are more likely to be stabilised or destabilised by SND1 depletion than non-target transcripts, suggesting a dual mode of SND1 action on RNA stability (**Figure 88A**). In lytic cells, SND1 target transcripts show a tendency towards decreased stability on SND1 depletion (**Figure 88B**); this effect was confirmed to be significant (Mann-Whitney test p-value: < 2.2e-16). Interestingly, this effect

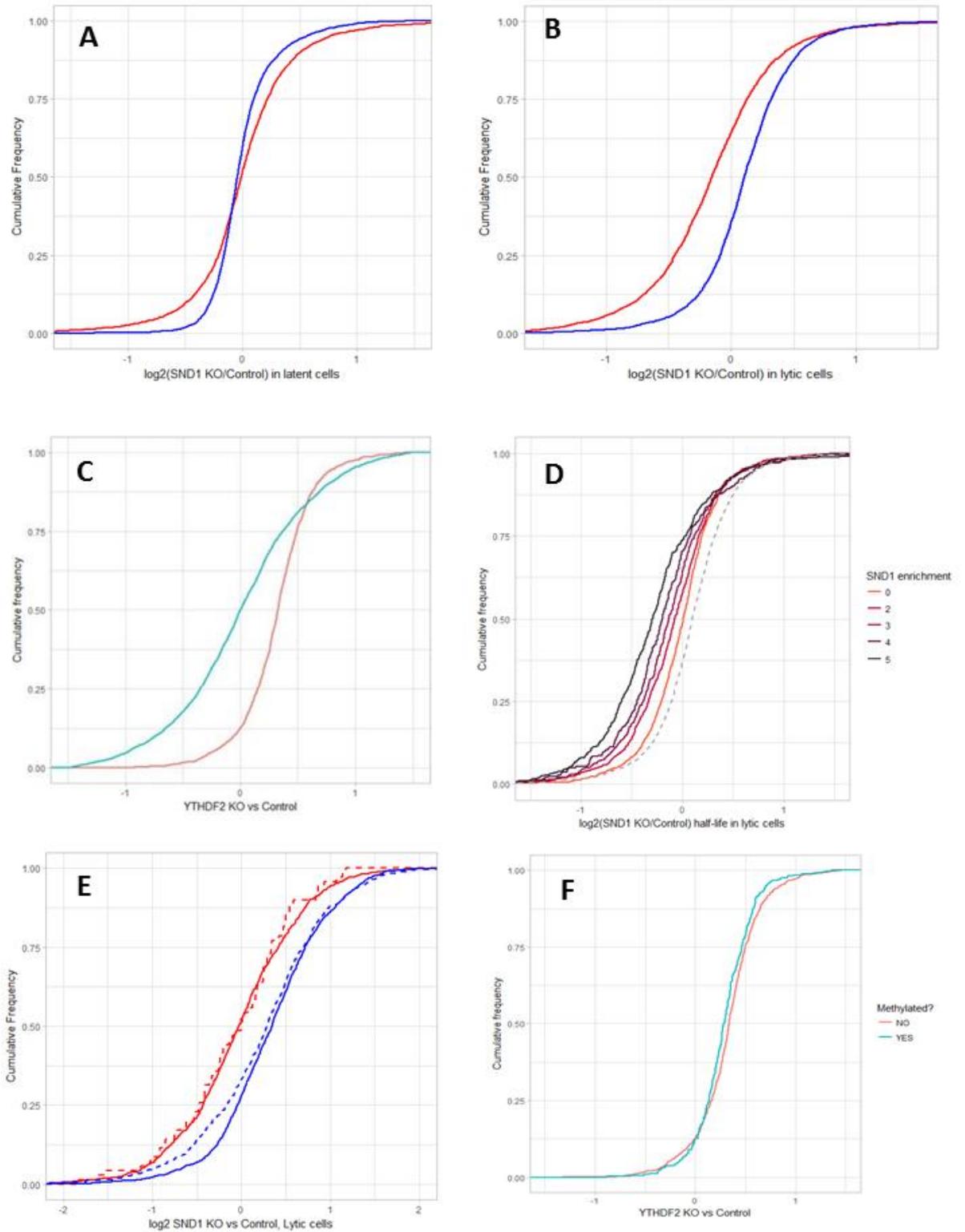
appears to be precisely the reverse of that reported for the YTHDF2 m<sup>6</sup>A reader protein, where YTHDF2 depletion stabilises its target RNAs (**Figure 88C**).

These observations are in line with reported roles of SND1 as a regulator of transcript stability by increasing RNA degradation when localised in P-bodies and promoting RNA stability when localised within stress granules. The dramatic shift of SND1 target stability in lytic cells that is not observed in latent cells in particular supports SND1 role as a component of stress granules, sequestering key RNAs and promoting their stability during periods of cellular stress. This is further evidenced by the observation that increased enrichment of SND1 target transcripts in SND1-fRIP-Seq data, which could be interpreted as stronger SND1-RNA associations, correlates with an increased effect on the stability of these transcripts (**Figure 88D**).

Whether this effect on RNA stability is mediated by SND1 recognition of m<sup>6</sup>A methylation is less clear. No significant difference in stability was observed in SND1 target transcripts which were identified as methylated or not methylated in m<sup>6</sup>A-Seq experiment (**Figure 88E**). However, a substantial amount of technical noise and biological variability is expected to accumulate when comparing data across multiple sequencing experiments, carried out over the course of several years. It is interesting to note the difference between stability of methylated and unmethylated YTHDF2 target transcripts in the sequencing data produced by the He group (Wang *et al*, 2014) was overall very small, and not significant at p-value < 0.01 (Mann-Whitney test, p-value: 0.01356) (**Figure 88F**), although the effects could be validated by qPCR in multiple key transcripts.

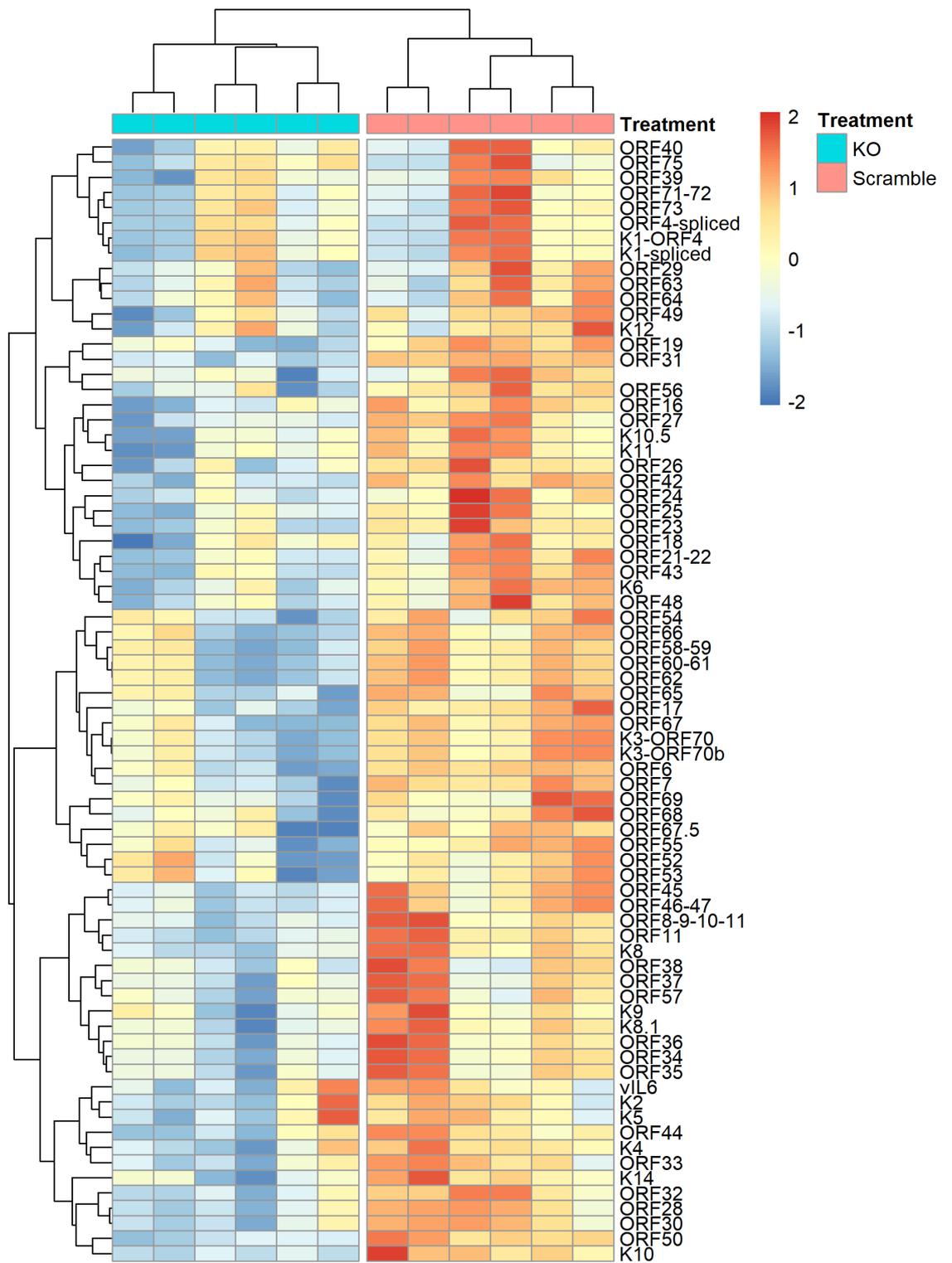


**Figure 87.** The distribution of log<sub>2</sub> fold changes of SND1 knockdown versus scramble control RNA half-lives in latent and lytic cells.



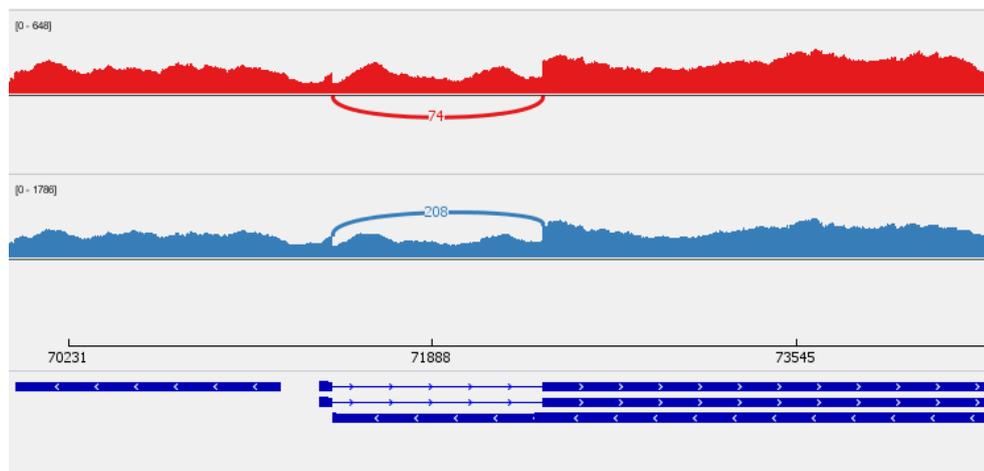
**Figure 88A.** Cumulative frequency distribution of RNA half-life changes on SND1 knockdown in latent cells. SND1 target transcripts are in red, while non-targets are in blue. Mann-Whitney p-value: 0.01016. **B.** Cumulative frequency distribution of RNA half-life changes on SND1 knockdown in lytic cells. SND1 target transcripts are in red, while non-targets are in blue. Mann-Whitney p-value:  $< 2.2e-16$ . **C.** Cumulative frequency distribution of RNA half-life changes

on YTHDF2 reader knockdown in HeLa cells. Figure obtained from reanalysis of raw data from Wang *et al*, (2014). YTHDF2 target transcripts are in red, while non-targets are in blue. Mann-Whitney p-value:  $< 2.2e^{16}$ . **D.** Cumulative frequency distribution of RNA half-life changes on SND1 knockdown in lytic cells. SND1 target transcripts are drawn in a solid line, separated by IP enrichment over control, while non-target distribution is shown as a dotted line. **E.** Cumulative frequency distribution of RNA half-life changes on SND1 knockdown in lytic cells. SND1 target transcripts are in red, while non-targets are in blue. Distribution of unmethylated transcripts is shown as solid lines, while methylated transcripts are shown as dashed lines. Mann-Whitney p-value methylated vs unmethylated SND1 targets: 0.56; Mann-Whitney p-value methylated vs unmethylated SND1 non-targets: 0.18. **F.** Cumulative frequency distribution of RNA half-life changes on YTHDF2 reader knockdown in HeLa cells. Figure obtained from reanalysis of raw data from Wang *et al*, (2014). YTHDF2 target transcripts that are methylated are in blue, while unmethylated targets are in red. Mann-Whitney p-value: 0.01356.

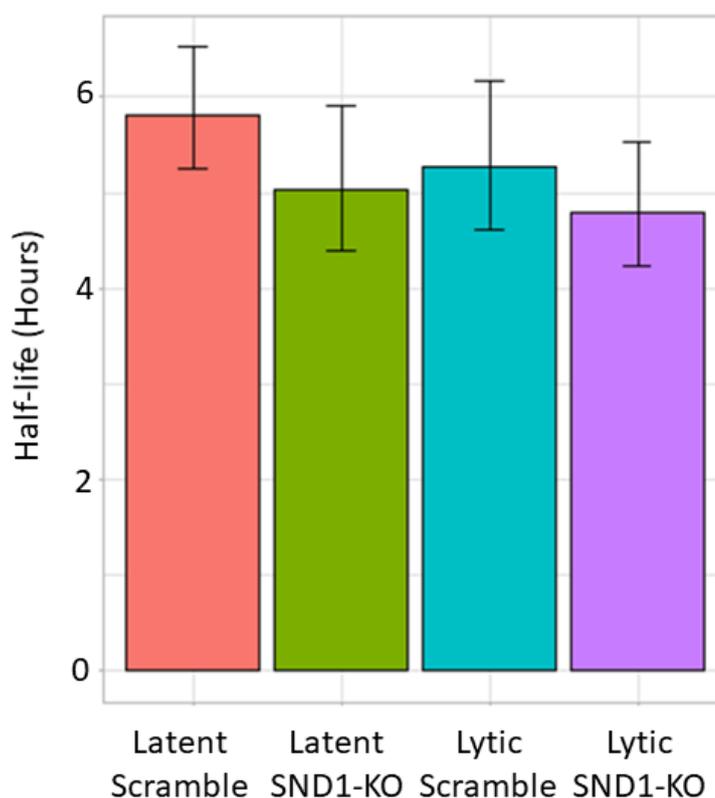


**Figure 89.** Heatmap showing the expression of KSHV transcripts in lytic BCBL-1 cells in SND1 knockdown and scramble control samples, 3 time points, 2 replicates.

In KSHV viral transcriptome, SND1 knockdown was found to severely impact the expression of the majority of KSHV viral transcripts, halting viral lytic replication in BCBL-1 cells (**Figure 89**). This effect may be a result of a direct global repression of all viral transcripts, or indirectly due to the inability to activate the RTA gene upon SND1 knockdown, which is required for the latent-lytic KSHV switch. On further investigation, however, it was found that RTA showed neither any alteration in splicing patterns (**Figure 90A**), nor any differences in the rate of RNA degradation (**Figure 90B**). No significant differences could be found for other viral transcripts. This would then suggest that SND1 may be involved in transcriptional regulation of KSHV transcriptome; or, as SND1 was found to specifically bind methylated RTA transcript, it may act as a post-transcriptional regulator further downstream. However, this effect is unlikely to be related to RNA translation mechanisms, as the effects on KSHV lytic replication were identified at RNA level.



**Figure 90A.** RTA transcript sashimi and read coverage plot showing splicing graph in lytic BCBL-1 cells in SND1 knockdown (top, red) and scramble control (bottom, blue) samples. The numbers over arches joining exons show the number of split-read alignments supporting the junction. As RTA overlaps an anti-sense transcript, ORF49, only correctly stranded reads were used to produce the above plot.



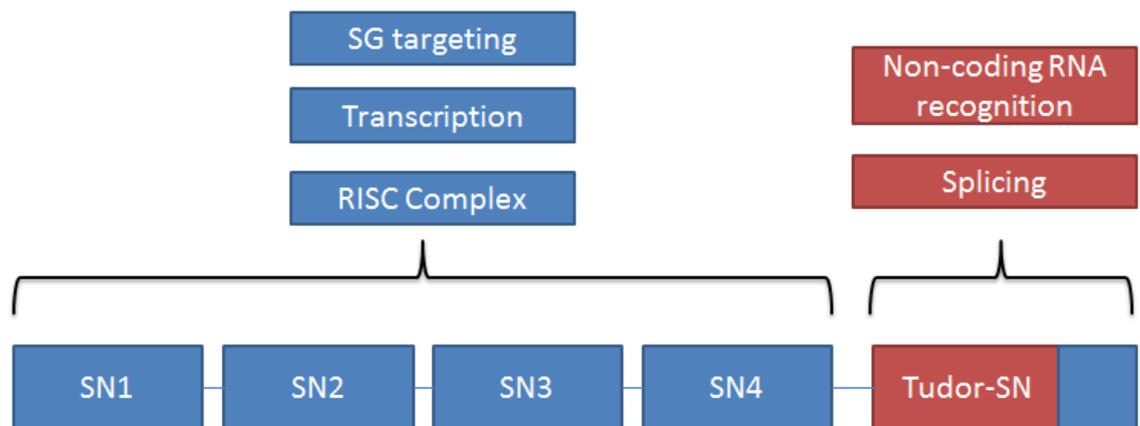
**Figure 90 B.** RTA transcript half-life in latent and lytic BCBL-1 cells in SND1 knockdown and scramble control samples. Half-lives from left to right: latent scramble, latent SND1 knockdown, lytic scramble, lytic SND1 knockdown.

### 6.4.5 Discussion

While it was previously unknown whether KSHV viral transcriptome is subject to post-transcriptional m<sup>6</sup>A modification, very recently Ye *et al* (2017) described similar findings to those presented here. Ye *et al* (2017) found the KSHV transcriptome to be heavily methylated, including the methylation of latent-lytic switch master regulator RTA. The group reported the YTHDC1 reader protein binding of m<sup>6</sup>A-RTA, which was also identified here. Ye *et al* further postulate that the lytic-latent switch is first induced by m<sup>6</sup>A methylation of the RTA transcript, which facilitates its pre-mRNA splicing and processing.

These results are in line with the hypothesis presented here, with a key difference - m<sup>6</sup>A-mediated processing of RTA is likely to be at least partially facilitated via a putative novel m<sup>6</sup>A reader protein SND1, rather than YTHDC1.

SND1 is an evolutionary conserved protein consisting of tandem-repeated staphylococcal nuclease domains and a fusion of Tudor domain, which is known to recognise methylated lysine and arginine residues (Liu et al. 2010a; Tripsianes et al. 2011; Liu et al. 2010b), with a partial SN domain at the C-terminus. SND1 is a multi-functional protein and has been shown to act in a number of different pathways as a regulator of gene expression (**Figure 91**).



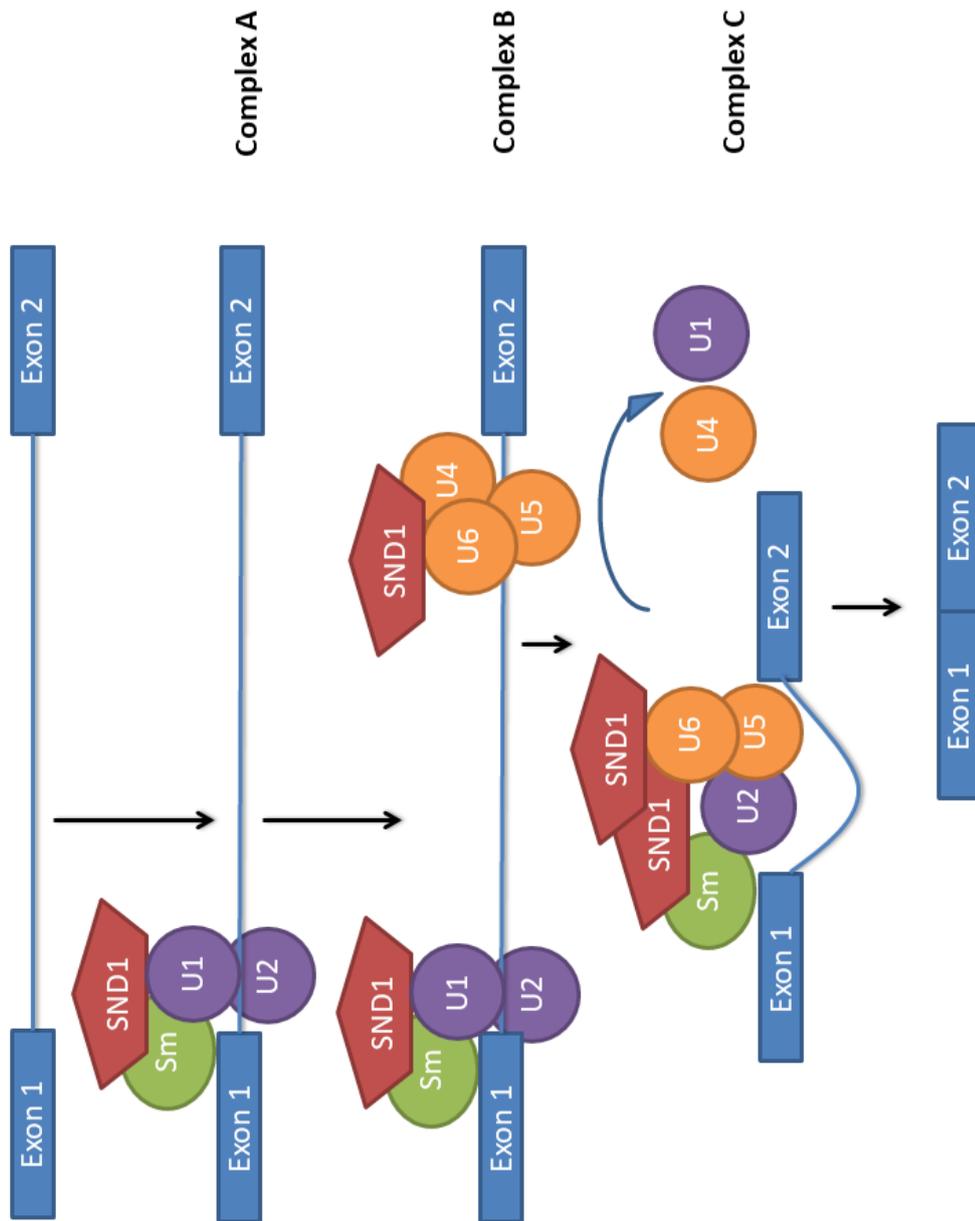
**Figure 91.** Summary of SND1 domain organisation and main functions. Different SND1 domains have been attributed differing roles in gene expression pathways, with SN domains participating in RISC complex activity, transcription factor activity and targeting to stress granules (SG), whereas Tudor-SN domain has been suggested to participate in RNA splicing.

To date, SND1 has been implicated as a regulator of many gene-expression pathways, including transcription, splicing and RNA silencing. Initially discovered as a transcriptional co-activator, SND1 has been shown to interact with a number of transcription factors, such as STAT5, PPAR $\gamma$  and NF- $\kappa$ B (Duan et al. 2014; Rawlings et al. 2004; Santhekadur et al. 2012). These interactions are mediated via concurrent binding of these transcription factors and DNA by the SN domains.

SND1 has also been demonstrated to enhance the rate of splicing through interactions with multiple spliceosome components. SND1 reportedly binds U1, U2, U4, U5 and U6 snRNPs, as well as SmB and SmD1/D3 (Gao et al. 2012; Cappellari et al. 2014; Yang et al. 2007; Will and Lührmann 2001), enabling

spliceosome complex assembly, but also directing the transition from spliceosome complex A to complex B (**Figure 92**). To date, however, there is limited evidence of SND1 role in alternative splicing, with only alternative splicing of CD44 transcript thus far reported in literature.

Furthermore, SND1 was found to be associated with RISC, a complex which mediates RNA interference (RNAi) silencing. Staphylococcal nuclease inhibitors have been shown to also inhibit RISC activity, suggesting that SND1 SN domains may be contributing to nucleolytic RISC action. There is evidence that SND1 may degrade A-to-I edited double stranded RNA (I-dsRNA), as well as hyper-edited miRNA precursors, although the precise cellular function of hyper-edited RNA remains unclear (Yang et al. 2006; Scadden 2005; Sontheimer 2005; Caudy et al. 2003).



**Figure 92.** Dual role of SND1 in splicing. SND1 is proposed to interact with snRNPs U1, U2, U4, U5 and U6, facilitating assembly of spliceosome Complex A, but is also putatively involved in 'bridging' the transition to Complex B.

Under stress conditions, cytoplasmic RNAs are either stored in stress granules, or degraded in P-bodies. SND1 has also been found to interact with proteins forming the core of stress granules, and it has been suggested that SND1 may act as a scaffold or a recruiting factor. Certain RNAs localised in stress granules may be stabilised by specific RBPs, which protects them from degradation. Indeed, SND1 has been shown to bind some RNAs during stress conditions and in doing so, increase their stability (Lokdarshi et al. 2016; Yan et al. 2014; Zhu et al. 2013; Weissbach and Scadden 2012).

Accordingly, SND1 is involved in multiple gene expression regulation processes. Under normal conditions, it appears to act as a regulator of transcription, enhancing transcription via its roles as transcriptional co-activator, stimulating splicing and regulating RNA levels via RNAi pathways. When subjected to stress, SND1 localises to the cytoplasm, where it contributes to stabilisation or degradation of specific RNAs, conferring stress resistance to the cell and promoting survival (reviewed in (Gutierrez-Beltran et al. 2016)). In addition, localisation to the cytoplasm impacts nuclear SND1 functions, thus resulting in global repression of transcription of nuclear SND1 target genes.

It is thus possible that m<sup>6</sup>A modified RNA plays a role in one or more of these pathways by directly associating with SND1. Indeed, it has been shown that the roles of m<sup>6</sup>A are diverse, and can promote splicing; increase in RNA stability; increase of degradation by targeting m<sup>6</sup>A-modified RNAs to P-bodies; or enhance translation. It is evident that there is considerable overlap between known m<sup>6</sup>A functions and those of SND1.

Here, SND1 is shown to specifically bind m<sup>6</sup>A methylated RTA; and SND1 knockdown halts KSHV viral gene expression, suggesting that m<sup>6</sup>A-mediated SND1 binding of RTA is required for KSHV lytic replication.

The mechanism of SND1 action is less clear. Upon SND1 depletion, no significant effects could be detected in the splicing of viral transcripts or in the rate of KSHV transcript degradation. Some of these effects may be difficult to detect, however, as KSHV transcriptome is tightly packaged in overlapping open reading frames, preventing accurate characterisation of individual transcripts using short read sequencing technologies. Thus, substantial

inaccuracies may be present in the data that are often impossible to address using the short-read sequencing technologies.

On the other hand, a clear effect on the stability of SND1 cellular target transcripts was observed in lytic cells on SND1 knockdown, suggesting that SND1 is involved cellular stress response. This effect may be mediated via m<sup>6</sup>A-specific recognition of RNA; however, this could not be conclusively demonstrated using sequencing data.

Many intronic SND1 binding sites were identified, often tightly co-localising with m<sup>6</sup>A methylation, suggesting that these processes are coupled in some way. Altered - or perhaps intermediate- splicing products can also be detected in these binding sites in immunoprecipitated fractions, further suggesting that SND1 plays role in splicing. However, no significant effects on splicing could be detected in SND1 knockdown samples, indicating some other role for SND1; or perhaps suggesting that there is redundancy in SND1 involvement in splicing processes and other proteins are able to compensate in the absence of SND1.

The work presented in this chapter remains on-going, thus many avenues remain unexplored. Further bioinformatics analysis of the SND1-fRIP-Seq, RNA lifetime profiling and m<sup>6</sup>A-Seq datasets may reveal insights into other potential mechanisms of SND1 action. For example, SND1 has been implicated in RISC/RNAi-mediated transcript degradation. It is thus feasible that the m<sup>6</sup>A modification in the 3'UTR of transcripts enables this function. Identification of A-to-I hyper-edited regions could shed light on this potential interaction.

## 7. Conclusion

The decreasing cost and increasing through-put of 'Next-generation' sequencing has led to rapid development of novel transcriptome and genome-wide sequencing applications. As a result, the bottle-neck in both research and clinical sequencing has shifted from data generation to analysis and interpretation. In these areas, the development of robust and scalable bioinformatics software and algorithms has lagged considerably behind data generation, and with few community gold standard tools available, researchers often use in-house scripts and *ad hoc* approaches. Thus, to bridge the gap between data generation and analysis and interpretation, more focus is needed on the development and new and improved bioinformatics tools.

This work presents three such applications: GeneTiER, OVA and m6aViewer. GeneTiER explores the use of data mining tissue-specific RNA-Seq and microarray expression data for candidate gene prioritisation. The resulting application is capable of unbiased candidate gene prioritisation and performs well in cases where disease phenotype is localised to few tissues or systems. The OVA application has been built to utilise a more diverse knowledge base, exploiting multiple ontologies to prioritise candidate disease genes and variants. While subject to 'guilt-by-association' bias towards the better studied genes, overall OVA achieves higher precision than GeneTiER, and therefore is more broadly applicable. However, in cases where knowledge-based approaches fail – for example, where the disease gene is poorly annotated – GeneTiER could identify disease genes in an unbiased way.

Shifting the focus from genomic sequencing to transcriptomics, the third application presented in this work is m6aViewer, a cross-platform GUI-driven desktop tool for the detection of visualisation of transcriptome-wide m<sup>6</sup>A methylation. m6aViewer was developed to meet the data analysis needs of an investigation into KSHV m<sup>6</sup>A methylome. The application of m6aViewer methodology and other related analyses have been presented in the last chapter, reporting previously uncharacterised RNA methylome of KSHV virus. Furthermore, this work introduced a putative novel N6-methyl-adenosine -RNA

binding protein, SND1, discussing its possible roles in the progression of viral infection and outlining future research avenues that remain unexplored.

## **8. Publications**

The work presented in sections 3, 4 and 5 of this thesis are based on the following publications respectively:

**Agne Antanaviciute**, Catherine Daly, Laura A Crinnion, Alexander F Markham, Christopher M Watson, David T Bonthron, Ian M Carr, 2015. *GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles*, *Bioinformatics* 31 (16), 2728-2735

**Agne Antanaviciute**, Christopher M Watson, Sally M Harrison, Carolina Lascelles, Laura Crinnion, Alexander F Markham, David T Bonthron, Ian M Carr, 2015 *OVA: integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization*, *Bioinformatics* 31 (23), 3822-3829

**Agne Antanaviciute**, Belinda Baquero-Perez, Christopher M Watson, Sally M Harrison, Carolina Lascelles, Laura Crinnion, Alexander F Markham, David T Bonthron, Adrian Whitehouse, Ian M Carr , 2017 *m6aViewer: software for the detection, analysis and visualization of N6-methyl-adenosine peaks from m6A-seq/ME-RIP sequencing data*, *RNA*, rna. 058206.116

## 9. Bibliography

- ADAMS JM, JEPPESEN PGN, SANGER F, BARRELL BG. 1969. Nucleotide Sequence from the Coat Protein Cistron of R17 Bacteriophage RNA. *Nature* **223**: 1009–1014.  
<http://www.nature.com/doi/10.1038/2231009a0> (Accessed May 11, 2017).
- Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. 2006. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* **22**: 773–774.  
<http://www.ncbi.nlm.nih.gov/pubmed/16423925>.
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248–249. <http://www.ncbi.nlm.nih.gov/pubmed/20354512> (Accessed August 18, 2017).
- Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, et al. 2006. Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**: 537–544. <http://www.ncbi.nlm.nih.gov/pubmed/16680138>.
- Agarwala SD, Blitzblau HG, Hochwagen A, Fink GR. 2012. RNA methylation by the MIS complex regulates a cell fate decision in yeast. *PLoS Genet* **8**: e1002732.  
<http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002732>  
(Accessed May 12, 2016).
- Alarcón CR, Goodarzi H, Lee H, Liu X, Tavazoie S, Tavazoie SF. 2015a. HNRNPA2B1 Is a Mediator of m(6)A-Dependent Nuclear RNA Processing Events. *Cell* **162**: 1299–308.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4673968&tool=pmcentrez&rendertype=abstract> (Accessed September 1, 2015).
- Alarcón CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF. 2015b. N6-methyladenosine marks primary microRNAs for processing. *Nature* **519**: 482–5.  
<http://dx.doi.org/10.1038/nature14281> (Accessed December 7, 2015).
- Ali MU, Rahman MSU, Cao J, Yuan PX. 2017. Genetic characterization and disease mechanism of retinitis pigmentosa; current scenario. *3 Biotech* **7**: 251.  
<http://www.ncbi.nlm.nih.gov/pubmed/28721681> (Accessed August 18, 2017).

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–402. <http://www.ncbi.nlm.nih.gov/pubmed/9254694> (Accessed August 17, 2017).
- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**: D789-98. <http://www.ncbi.nlm.nih.gov/pubmed/25428349> (Accessed August 17, 2017).
- Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–9. <http://www.ncbi.nlm.nih.gov/pubmed/25260700> (Accessed October 12, 2017).
- Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, et al. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290**: 457–65. <http://www.ncbi.nlm.nih.gov/pubmed/7219534> (Accessed May 11, 2017).
- Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.
- Ansorge W, Sproat BS, Stegemann J, Schwager C. 1986. A non-radioactive automated method for DNA sequence determination. *J Biochem Biophys Methods* **13**: 315–23. <http://www.ncbi.nlm.nih.gov/pubmed/3559035> (Accessed May 11, 2017).
- Arias C, Weisburd B, Stern-Ginossar N, Mercier A, Madrid AS, Bellare P, Holdorf M, Weissman JS, Ganem D. 2014. KSHV 2.0: A Comprehensive Annotation of the Kaposi's Sarcoma-Associated Herpesvirus Genome Using Next-Generation Sequencing Reveals Novel Genomic and Functional Features ed. D.P. Dittmer. *PLoS Pathog* **10**: e1003847. <http://www.ncbi.nlm.nih.gov/pubmed/24453964> (Accessed August 19, 2017).
- Armstrong RA. 2014. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* **34**: 502–8. <http://www.ncbi.nlm.nih.gov/pubmed/24697967> (Accessed May 5, 2015).
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29. <http://www.ncbi.nlm.nih.gov/pubmed/10802651>.

- Atkinson MR, Deutscher MP, Kornberg A, Russell AF, Moffatt JG. 1969. Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. *Biochemistry* **8**: 4897–904.  
<http://www.ncbi.nlm.nih.gov/pubmed/4312461> (Accessed March 31, 2017).
- Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74. <http://www.ncbi.nlm.nih.gov/pubmed/26432245> (Accessed August 18, 2017).
- Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G, Hudson GS, Satchwell SC, Séguin C, et al. 1984. DNA sequence and expression of the B95-8 Epstein—Barr virus genome. *Nature* **310**: 207–211.  
<http://www.nature.com/doi/10.1038/310207a0> (Accessed May 11, 2017).
- Bagn ris C, Ageichik A V., Cronin N, Wallace B, Collins M, Boshoff C, Waksman G, Barrett T. 2008. Crystal Structure of a vFlip-IKK  Complex: Insights into Viral Activation of the IKK Signalosome. *Mol Cell* **30**: 620–631. <http://www.ncbi.nlm.nih.gov/pubmed/18538660> (Accessed August 19, 2017).
- Bahn JH, Lee J-H, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**: 142–50.  
[http://genome.cshlp.org/content/22/1/142.abstract?ijkey=38733f14c8c55148d6e1bebce339c0133e1629c4&keytype=tf\\_ipsecsha](http://genome.cshlp.org/content/22/1/142.abstract?ijkey=38733f14c8c55148d6e1bebce339c0133e1629c4&keytype=tf_ipsecsha) (Accessed May 4, 2016).
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-8.  
[http://nar.oxfordjournals.org/content/37/suppl\\_2/W202.full](http://nar.oxfordjournals.org/content/37/suppl_2/W202.full) (Accessed July 9, 2014).
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**: D991-5.  
<http://nar.oxfordjournals.org/content/41/D1/D991.full> (Accessed April 3, 2016).
- Batista PJ, Molin  B, Wang J, Qu K, Zhang J, Li L, Bouley DM, Lujan E, Haddad B, Daneshvar K, et al. 2014. m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* **15**: 707–19.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4278749&tool=pmcentrez&rendertype=abstract> (Accessed April 22, 2016).

Baum LE, Petrie T. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Ann Math Stat* **37**: 1554–1563. <http://projecteuclid.org/euclid.aoms/1177699147> (Accessed May 23, 2016).

Bell GI, Polonsky KS. 2001. Diabetes mellitus and genetically programmed defects in  $\beta$ -cell function. *Nature* **414**: 788–791. <http://www.ncbi.nlm.nih.gov/pubmed/11742410> (Accessed August 17, 2017).

Benjamini Y, Hochberg Y. 1995. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. *J R Stat Soc Ser B Methodol* **57**: 289–300. [http://www.researchgate.net/publication/221995234\\_Controlling\\_The\\_False\\_Discovery\\_Rate\\_-\\_A\\_Practical\\_And\\_Powerful\\_Approach\\_To\\_Multiple\\_Testing](http://www.researchgate.net/publication/221995234_Controlling_The_False_Discovery_Rate_-_A_Practical_And_Powerful_Approach_To_Multiple_Testing) (Accessed October 1, 2015).

Bennett ST, Barnes C, Cox A, Davies L, Brown C. 2005. Toward the \$1000 human genome. *Pharmacogenomics* **6**: 373–382. <http://www.ncbi.nlm.nih.gov/pubmed/16004555> (Accessed April 1, 2017).

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2005. GenBank. *Nucleic Acids Res* **33**: D34-8. <http://www.ncbi.nlm.nih.gov/pubmed/15608212> (Accessed May 11, 2017).

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59. <http://www.ncbi.nlm.nih.gov/pubmed/18987734> (Accessed May 11, 2017).

Berchtold A, Raftery A. 2002. The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Stat Sci* **17**: 328–356. <http://projecteuclid.org/euclid.ss/1042727943> (Accessed March 9, 2016).

Berson A, Barbash S, Shaltiel G, Goll Y, Hanin G, Greenberg DS, Ketzeff M, Becker AJ, Friedman A, Soreq H. 2012. Cholinergic-associated loss of hnRNP-A/B in Alzheimer's disease impairs cortical splicing and cognitive function in mice. *EMBO Mol Med* **4**: 730–42. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3494073&tool=pmcentrez&>

rendertype=abstract (Accessed May 11, 2016).

Berulava T, Rahmann S, Rademacher K, Klein-Hitpass L, Horsthemke B. 2015. N6-adenosine methylation in MiRNAs. *PLoS One* **10**: e0118438.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4344304&tool=pmcentrez&rendertype=abstract> (Accessed May 5, 2016).

Blackbourn DJ, Lennette E, Klencke B, Moses A, Chandran B, Weinstein M, Glogau RG, Witte MH, Way DL, Kutzkey T, et al. 2000. The restricted cellular host range of human herpesvirus 8. *AIDS* **14**: 1123–33. <http://www.ncbi.nlm.nih.gov/pubmed/10894276> (Accessed August 19, 2017).

Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE. 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* **42**: D810–7. <http://www.ncbi.nlm.nih.gov/pubmed/24285300>.

Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ, the Mouse Genome Database Group. 2017. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res* **45**: D723–D729. <http://www.ncbi.nlm.nih.gov/pubmed/27899570> (Accessed August 18, 2017).

Boccuto L, Aoki K, Flanagan-Steet H, Chen C-F, Fan X, Bartel F, Petukh M, Pittman A, Saul R, Chaubey A, et al. 2014. A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation. *Hum Mol Genet* **23**: 418–433.  
<http://www.ncbi.nlm.nih.gov/pubmed/24026681> (Accessed August 18, 2017).

Boeva V, Lermine A, Barette C, Guillouf C, Barillot E. 2012. Nebula--a web-server for advanced ChIP-seq data analysis. *Bioinformatics* **28**: 2517–9.  
<http://bioinformatics.oxfordjournals.org/content/28/19/2517.long> (Accessed May 17, 2016).

Bokar JA, Shambaugh ME, Polayes D, Matera AG, Rottman FM. 1997. Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase. *RNA* **3**: 1233–47.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1369564&tool=pmcentrez&rendertype=abstract> (Accessed April 22, 2016).

- Bornigen D, Tranchevent LC, Bonachela-Capdevila F, Devriendt K, De Moor B, De Causmaecker P, Moreau Y. 2012. An unbiased evaluation of gene prioritization tools. *Bioinformatics* **28**: 3081–3088. <http://www.ncbi.nlm.nih.gov/pubmed/23047555>.
- Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33**: 228–237. <http://www.ncbi.nlm.nih.gov/pubmed/12610532> (Accessed October 10, 2017).
- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**: 365–371. <http://www.ncbi.nlm.nih.gov/pubmed/11726920> (Accessed August 18, 2017).
- Breiman L. Random Forests. *Mach Learn* **45**: 5–32. <http://link.springer.com/article/10.1023/A%3A1010933404324> (Accessed July 3, 2015).
- Bringmann P, Lührmann R. 1987. Antibodies specific for N 6 -methyladenosine react with intact snRNPs U2 and U4/U6. *FEBS Lett* **213**: 309–315. <http://www.sciencedirect.com/science/article/pii/0014579387815120> (Accessed May 5, 2016).
- Britto R, Sallou O, Collin O, Michaux G, Primig M, Chalmel F. 2012. GPSy: a cross-species gene prioritization system for conserved biological processes--application in male gamete development. *Nucleic Acids Res* **40**: W458-65. <http://www.ncbi.nlm.nih.gov/pubmed/22570409>.
- Broadgate S, Yu J, Downes SM, Halford S. 2017. Unravelling the genetics of inherited retinal dystrophies: Past, present and future. *Prog Retin Eye Res* **59**: 53–96. <http://www.ncbi.nlm.nih.gov/pubmed/28363849> (Accessed August 18, 2017).
- Brownlee GG, Sanger F. 1967. Nucleotide sequences from the low molecular weight ribosomal RNA of Escherichia coli. *J Mol Biol* **23**: 337–53. <http://www.ncbi.nlm.nih.gov/pubmed/4291728> (Accessed May 11, 2017).
- Bykhovskaya Y, Casas K, Mengesha E, Inbal A, Fischel-Ghodsian N. 2004. Missense mutation in pseudouridine synthase 1 (PUS1) causes mitochondrial myopathy and sideroblastic anemia (MLASA). *Am J Hum Genet* **74**: 1303–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1182096&tool=pmcentrez&>

rendertype=abstract (Accessed May 4, 2016).

Calderone A, Castagnoli L, Cesareni G. 2013. mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods* **10**: 690–1.

<http://dx.doi.org/10.1038/nmeth.2561> (Accessed May 5, 2015).

Campbell JD, Liu G, Luo L, Xiao J, Gerrein J, Juan-Guardela B, Tedrow J, Alekseyev YO, Yang I V., Correll M, et al. 2015. Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data. *RNA* **21**: 164–171.

<http://www.ncbi.nlm.nih.gov/pubmed/25519487> (Accessed August 16, 2017).

Camper SA, Albers RJ, Coward JK, Rottman FM. 1984. Effect of undermethylation on mRNA cytoplasmic appearance and half-life. *Mol Cell Biol* **4**: 538–543.

<http://mcb.asm.org/content/4/3/538> (Accessed May 5, 2016).

Cappellari M, Bielli P, Paronetto MP, Ciccocanti F, Fimia GM, Saarikettu J, Silvennoinen O, Sette C. 2014. The transcriptional co-activator SND1 is a novel regulator of alternative splicing in prostate cancer cells. *Oncogene* **33**: 3794–3802.

<http://www.ncbi.nlm.nih.gov/pubmed/23995791> (Accessed August 19, 2017).

Carlile TM, Rojas-Duran MF, Gilbert W V. 2015. Pseudo-Seq: Genome-Wide Detection of Pseudouridine Modifications in RNA. *Methods Enzymol* **560**: 219–45.

<http://www.ncbi.nlm.nih.gov/pubmed/26253973> (Accessed May 4, 2016).

Carrillo Oesterreich F, Preibisch S, Neugebauer KM. 2010. Global Analysis of Nascent RNA Reveals Transcriptional Pausing in Terminal Exons. *Mol Cell* **40**: 571–581.

<http://www.ncbi.nlm.nih.gov/pubmed/21095587> (Accessed August 16, 2017).

Carroll SM, Narayan P, Rottman FM. 1990. N6-methyladenosine residues in an intron-specific region of prolactin pre-mRNA. *Mol Cell Biol* **10**: 4456–65.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=361031&tool=pmcentrez&rendertype=abstract> (Accessed May 18, 2016).

Caselli E, Galvan M, Cassai E, Caruso A, Sighinolfi L, Di Luca D. 2005. Human herpesvirus 8 enhances human immunodeficiency virus replication in acutely infected cells and induces reactivation in latently infected cells. *Blood* **106**: 2790–2797.

<http://www.ncbi.nlm.nih.gov/pubmed/15976177> (Accessed August 19, 2017).

- Caudy AA, Ketting RF, Hammond SM, Denli AM, Bathoorn AMP, Tops BBJ, Silva JM, Myers MM, Hannon GJ, Plasterk RHA. 2003. A micrococcal nuclease homologue in RNAi effector complexes. *Nature* **425**: 411–414. <http://www.ncbi.nlm.nih.gov/pubmed/14508492> (Accessed August 19, 2017).
- Cesarman E, Chang Y, Moore PS, Said JW, Knowles DM. 1995. Kaposi's Sarcoma–Associated Herpesvirus-Like DNA Sequences in AIDS-Related Body-Cavity–Based Lymphomas. *N Engl J Med* **332**: 1186–1191. <http://www.ncbi.nlm.nih.gov/pubmed/7700311> (Accessed August 19, 2017).
- Chang C-C, Lin C-J. 2011. LIBSVM. *ACM Trans Intell Syst Technol* **2**: 1–27. <http://dl.acm.org/citation.cfm?id=1961189.1961199> (Accessed July 10, 2014).
- Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, Moore PS. 1994. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266**: 1865–9. <http://www.ncbi.nlm.nih.gov/pubmed/7997879> (Accessed August 19, 2017).
- Check Hayden E. 2014. Is the \$1,000 genome for real? *Nature*. <http://www.nature.com/doifinder/10.1038/nature.2014.14530> (Accessed May 11, 2017).
- Chen H-M, Fitcher B, Leatherwood J. 2011a. The fission yeast RNA binding protein Mmi1 regulates meiotic genes by controlling intron specific splicing and polyadenylation coupled RNA turnover. *PLoS One* **6**: e26804. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3203177&tool=pmcentrez&rendertype=abstract> (Accessed May 10, 2016).
- Chen H, Zhou H-X. 2005. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res* **33**: 3193–9. <http://www.ncbi.nlm.nih.gov/pubmed/15937195> (Accessed August 18, 2017).
- Chen J, Bardes EE, Aronow BJ, Jegga AG. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**: W305-11. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2703978&tool=pmcentrez&rendertype=abstract> (Accessed January 9, 2015).
- Chen K, Lu Z, Wang X, Fu Y, Luo G-Z, Liu N, Han D, Dominissini D, Dai Q, Pan T, et al. 2015a. High-resolution N(6) -methyladenosine (m(6) A) map using photo-crosslinking-assisted

- m(6) A sequencing. *Angew Chem Int Ed Engl* **54**: 1587–90.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4396828&tool=pmcentrez&rendertype=abstract> (Accessed May 11, 2016).
- Chen T, Hao Y-J, Zhang Y, Li M-M, Wang M, Han W, Wu Y, Lv Y, Hao J, Wang L, et al. 2015b. m6A RNA Methylation Is Regulated by MicroRNAs and Promotes Reprogramming to Pluripotency. *Cell Stem Cell* **16**: 289–301.  
<http://www.cell.com/article/S193459091500017X/fulltext> (Accessed February 13, 2015).
- Chen W, Feng P, Ding H, Lin H, Chou K-C. 2015c. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal Biochem* **490**: 26–33.  
<http://www.ncbi.nlm.nih.gov/pubmed/26314792> (Accessed April 27, 2016).
- Chen W, Tran H, Liang Z, Lin H, Zhang L. 2015d. Identification and analysis of the N(6)-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci Rep* **5**: 13859.  
<http://www.nature.com/srep/2015/150907/srep13859/full/srep13859.html> (Accessed March 8, 2016).
- Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou K-C. 2015e. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* **31**: 119–20. <http://www.ncbi.nlm.nih.gov/pubmed/25231908> (Accessed March 10, 2016).
- Chen Y-AA, Tripathi LP, Mizuguchi K. 2011b. TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery ed. V. Uversky. *PLoS One* **6**: e17844.  
<http://www.ncbi.nlm.nih.gov/pubmed/21408081> (Accessed April 14, 2015).
- Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, Liu T, Zhang Y, Kim T-K, He HH, Zieba J, et al. 2012. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat Methods* **9**: 609–14. <http://dx.doi.org/10.1038/nmeth.1985> (Accessed May 12, 2016).
- Chen Y, Wang W, Zhou Y, Shields R, Chanda SK, Elston RC, Li J. 2011c. In silico gene prioritization by integrating multiple data sources. *PLoS One* **6**: e21137.  
<http://www.ncbi.nlm.nih.gov/pubmed/21731658>.
- Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**: 3497–500.  
<http://www.ncbi.nlm.nih.gov/pubmed/12824352> (Accessed August 17, 2017).

- Chepelev I. 2012. Detection of RNA editing events in human cells using high-throughput sequencing. *Methods Mol Biol* **815**: 91–102.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4184133&tool=pmcentrez&rendertype=abstract> (Accessed May 13, 2016).
- Chou KC. 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **43**: 246–55. <http://www.ncbi.nlm.nih.gov/pubmed/11288174> (Accessed April 21, 2016).
- Choudhry Z, Sengupta SM, Grizenko N, Thakur GA, Fortier M-E, Schmitz N, Joobar R. 2013. Association between obesity-related gene FTO and ADHD. *Obesity (Silver Spring)* **21**: E738–44. <http://www.ncbi.nlm.nih.gov/pubmed/23512716> (Accessed May 12, 2016).
- Cieniková Z, Damberger FF, Hall J, Allain FH-T, Maris C. 2014. Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. *J Am Chem Soc* **136**: 14536–44. <http://dx.doi.org/10.1021/ja507690d> (Accessed May 10, 2016).
- Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puvion-Vandier V, et al. 2015. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373**: 895–907.  
<http://www.ncbi.nlm.nih.gov/pubmed/26287746> (Accessed May 11, 2016).
- Cochrane G, Cook CE, Birney E. 2012. The future of DNA sequence archiving. *Gigascience* **1**: 2. <http://gigascience.biomedcentral.com/articles/10.1186/2047-217X-1-2> (Accessed May 24, 2016).
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.  
<http://www.ncbi.nlm.nih.gov/pubmed/19304878> (Accessed August 17, 2017).
- Codd EF, 1982. Relational database: a practical foundation for productivity. *Commun ACM* **25**: 109–117. <http://portal.acm.org/citation.cfm?doid=358396.358400> (Accessed October 11, 2017).
- Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LELM, et al. 2014. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46**: 1063–1071.

- <http://www.ncbi.nlm.nih.gov/pubmed/25217958> (Accessed July 29, 2017).
- Cohen A, Brodie C, Sarid R. 2006. An essential role of ERK signalling in TPA-induced reactivation of Kaposi's sarcoma-associated herpesvirus. *J Gen Virol* **87**: 795–802.  
<http://www.ncbi.nlm.nih.gov/pubmed/16528027> (Accessed August 19, 2017).
- Cohen AM, Hersh WR. 2005. A survey of current work in biomedical text mining. *Brief Bioinform* **6**: 57–71. <http://www.ncbi.nlm.nih.gov/pubmed/15826357> (Accessed August 18, 2017).
- COHN WE, VOLKIN E. 1951. Nucleoside-5'-Phosphates from Ribonucleic Acid. *Nature* **167**: 483–484. <http://dx.doi.org/10.1038/167483a0> (Accessed May 4, 2016).
- Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, Ohler U. 2011. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol* **12**: R79. <http://www.ncbi.nlm.nih.gov/pubmed/21851591> (Accessed November 30, 2017).
- Cory S, Marcker KA, Dube SK, Clark BF. 1968. Primary structure of a methionine transfer RNA from Escherichia coli. *Nature* **220**: 1039–40.  
<http://www.ncbi.nlm.nih.gov/pubmed/4883023> (Accessed May 11, 2017).
- Couto FM, Silva MJ, Coutinho PM. 2007. Measuring semantic similarity between Gene Ontology terms. *Data Knowl Eng* **61**: 137–152.  
<http://linkinghub.elsevier.com/retrieve/pii/S0169023X06000875> (Accessed August 19, 2017).
- Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**: 123–31.  
<http://www.ncbi.nlm.nih.gov/pubmed/16344561> (Accessed August 16, 2017).
- Crick F. 1970. Central dogma of molecular biology. *Nature* **227**: 561–3.  
<http://www.ncbi.nlm.nih.gov/pubmed/4913914> (Accessed February 3, 2017).
- Csepany T, Lin A, Baldick CJ, Beemon K. 1990. Sequence specificity of mRNA N6-adenosine methyltransferase. *J Biol Chem* **265**: 20117–22.  
<http://www.ncbi.nlm.nih.gov/pubmed/2173695> (Accessed October 30, 2015).

- Cui X, Meng J, Rao MK, Chen Y, Huang Y. 2015. HEPeak: an HMM-based exome peak-finding package for RNA epigenome sequencing data. *BMC Genomics* **16 Suppl 4**: S2.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4416174&tool=pmcentrez&rendertype=abstract> (Accessed September 25, 2015).
- Cui X, Meng J, Zhang S, Chen Y, Huang Y. 2016. A novel algorithm for calling mRNA m6A peaks by modeling biological variances in MeRIP-seq data. *Bioinformatics* **32**: i378–i385.  
<http://www.ncbi.nlm.nih.gov/pubmed/27307641> (Accessed July 6, 2016).
- Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2015. *Nucleic Acids Res* **43**: D662–669.  
<http://nar.oxfordjournals.org/content/43/D1/D662> (Accessed November 25, 2014).
- Dai Q, Fong R, Saikia M, Stephenson D, Yu Y, Pan T, Piccirilli JA. 2007. Identification of recognition residues for ligation-based detection and quantitation of pseudouridine and N6-methyladenosine. *Nucleic Acids Res* **35**: 6322–9.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2094055&tool=pmcentrez&rendertype=abstract> (Accessed May 13, 2016).
- Dasgupta A, Raftery A. 1998. Detecting features in spatial point processes with clutter via model-based clustering. *J Am Stat*.  
<http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1998.10474110> (Accessed August 3, 2016).
- DAVIS FF, ALLEN FW. 1957. Ribonucleic acids from yeast which contain a fifth nucleotide. *J Biol Chem* **227**: 907–15. <http://www.ncbi.nlm.nih.gov/pubmed/13463012> (Accessed May 1, 2016).
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J R Stat Soc Ser B* **39**: 1–38. <http://www.jstor.org/stable/2984875>.
- Deng X, Chen K, Luo G-Z, Weng X, Ji Q, Zhou T, He C. 2015. Widespread occurrence of N6-methyladenosine in bacterial mRNA. *Nucleic Acids Res* **43**: 6557–67.  
<http://nar.oxfordjournals.org/content/early/2015/06/11/nar.gkv596.full> (Accessed May 5, 2016).
- Desrosiers R, Friderici K, Rottman F. 1974. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc Natl Acad Sci U S A* **71**: 3971–5.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=434308&tool=pmcentrez&rendertype=abstract> (Accessed February 16, 2016).

Dhammi IK, Kumar S. 2014. Medical subject headings (MeSH) terms. *Indian J Orthop* **48**: 443–4. <http://www.ncbi.nlm.nih.gov/pubmed/25298548> (Accessed August 18, 2017).

Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Kazan J, Seboun E, et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**: 152–154. <http://www.ncbi.nlm.nih.gov/pubmed/8600387> (Accessed August 17, 2017).

Dina C, Meyre D, Gallina S, Durand E, Körner A, Jacobson P, Carlsson LMS, Kiess W, Vatin V, Lecoer C, et al. 2007. Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet* **39**: 724–6. <http://www.ncbi.nlm.nih.gov/pubmed/17496892> (Accessed May 11, 2016).

Ding L, Rath E, Bai Y. 2017. Comparison of Alternative Splicing Junction Detection Tools Using RNASeq Data. *Curr Genomics* **18**: 268–277. <http://www.ncbi.nlm.nih.gov/pubmed/28659722> (Accessed August 16, 2017).

Do CB, Batzoglou S. 2008. What is the expectation maximization algorithm? *Nat Biotechnol* **26**: 897–9. <http://dx.doi.org/10.1038/nbt1406> (Accessed January 28, 2015).

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21. <http://bioinformatics.oxfordjournals.org/content/29/1/15> (Accessed July 13, 2014).

Dominissini D, Moshitch-Moshkovitz S, Salmon-Divon M, Amariglio N, Rechavi G. 2013. Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing. *Nat Protoc* **8**: 176–89. <http://www.readcube.com/articles/10.1038/nprot.2012.148> (Accessed March 21, 2016).

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, et al. 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**: 201–6. <http://www.ncbi.nlm.nih.gov/pubmed/22575960> (Accessed December 28, 2014).

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden

- DW, Smith DR, Lander ES. 1987. A genetic linkage map of the human genome. *Cell* **51**: 319–37. <http://www.ncbi.nlm.nih.gov/pubmed/3664638> (Accessed August 17, 2017).
- Dozmorov MG, Adrianto I, Giles CB, Glass E, Glenn SB, Montgomery C, Sivils KL, Olson LE, Iwayama T, Freeman WM, et al. 2015. Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. *BMC Bioinformatics* **16 Suppl 1**: S10. <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S13-S10> (Accessed May 18, 2016).
- Dreyfuss G, Matunis MJ, Piñol-Roma S, Burd CG. 1993. hnRNP proteins and the biogenesis of mRNA. *Annu Rev Biochem* **62**: 289–321. <http://www.scopus.com/inward/record.url?eid=2-s2.0-0027294031&partnerID=tZOtx3y1>.
- Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science (80- )* **327**: 78–81. <http://www.ncbi.nlm.nih.gov/pubmed/19892942> (Accessed May 11, 2017).
- Du T, Rao S, Wu L, Ye N, Liu Z, Hu H, Xiu J, Shen Y, Xu Q. 2015. An association study of the m6A genes with major depressive disorder in Chinese Han population. *J Affect Disord* **183**: 279–86. <http://www.sciencedirect.com/science/article/pii/S0165032715003250> (Accessed September 3, 2015).
- Duan Z, Zhao X, Fu X, Su C, Xin L, Saarikettu J, Yang X, Yao Z, Silvennoinen O, Wei M, et al. 2014. Tudor-SN, a Novel Coactivator of Peroxisome Proliferator-activated Receptor  $\gamma$  Protein, Is Essential for Adipogenesis. *J Biol Chem* **289**: 8364–8374. <http://www.ncbi.nlm.nih.gov/pubmed/24523408> (Accessed August 19, 2017).
- Dudley JT, Kim Y, Liu L, Markov GJ, Gerold K, Chen R, Butte AJ, Kumar S. 2012. Human genomic disease variants: a neutral evolutionary explanation. *Genome Res* **22**: 1383–94. <http://www.ncbi.nlm.nih.gov/pubmed/22665443> (Accessed August 18, 2017).
- Dunham I, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, Ainscough R, Almeida JP, Babbage A, et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402**: 489–495. <http://www.ncbi.nlm.nih.gov/pubmed/10591208> (Accessed April 1, 2017).
- Eddy SR. 2004. What is a hidden Markov model? *Nat Biotechnol* **22**: 1315–6.

<http://dx.doi.org/10.1038/nbt1004-1315> (Accessed January 2, 2015).

Edsgård D, Iglesias MJ, Reilly S-J, Hamsten A, Tornvall P, Odeberg J, Emanuelsson O. 2016. GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Sci Rep* **6**: 21134. <http://www.ncbi.nlm.nih.gov/pubmed/26887787> (Accessed August 16, 2017).

Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A. 2005. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* **434**: 857–863. <http://www.ncbi.nlm.nih.gov/pubmed/15829955> (Accessed October 10, 2017).

ENCODE Project Consortium T, coordination O, production leads D, analysts L, group W, project management N, investigators P, State University B, of North Carolina at Chapel Hill Proteomics groups data production U, Institute Group data production B, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **488**. <https://www.nature.com/nature/journal/v489/n7414/pdf/nature11247.pdf> (Accessed August 16, 2017).

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rättsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* **10**: 1185–91. <http://dx.doi.org/10.1038/nmeth.2722> (Accessed July 9, 2014).

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, K€orninger F, McKay S, et al. 2016. The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**: D481–D487. <http://www.ncbi.nlm.nih.gov/pubmed/26656494> (Accessed August 17, 2017).

Fawcett KA, Barroso I. 2010. The genetics of obesity: FTO leads the way. *Trends Genet* **26**: 266–74. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2906751&tool=pmcentrez&rendertype=abstract> (Accessed February 25, 2016).

Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM. 2008. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**: 1729–30.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2638869&tool=pmcentrez&rendertype=abstract> (Accessed May 19, 2016).

Firth H V., Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Vooren S Van, Moreau Y, Pettett RM, Carter NP. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* **84**: 524–533. <http://linkinghub.elsevier.com/retrieve/pii/S0002929709001074> (Accessed August 18, 2017).

Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, Brüning JC, Rütger U. 2009. Inactivation of the Fto gene protects from obesity. *Nature* **458**: 894–8. <http://www.ncbi.nlm.nih.gov/pubmed/19234441> (Accessed January 20, 2016).

Fisher RA. 1925. *Statistical Methods For Research Workers*. Cosmo Publications <https://books.google.com/books?id=4bTtAJR5kEC&pgis=1> (Accessed May 17, 2016).

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–512. <http://www.ncbi.nlm.nih.gov/pubmed/7542800> (Accessed May 11, 2017).

Forman J, Taruscio D, Llera VA, Barrera LA, Coté TR, Edfjäll C, Gavhed D, Haffner ME, Nishimura Y, Posada M, et al. 2012. The need for worldwide policy and action plans for rare diseases. *Acta Paediatr* **101**: 805–7. <http://www.ncbi.nlm.nih.gov/pubmed/22519914> (Accessed August 17, 2017).

Fraley C, Raftery AE. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JRB, Elliott KS, Lango H, Rayner NW, et al. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**: 889–94. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2646098&tool=pmcentrez&rendertype=abstract> (Accessed January 10, 2015).

Fu Y, Jia G, Pang X, Wang RN, Wang X, Li CJ, Smemo S, Dai Q, Bailey KA, Nobrega MA, et al. 2013. FTO-mediated formation of N6-hydroxymethyladenosine and N6-formyladenosine

in mammalian RNA. *Nat Commun* **4**: 1798.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3658177&tool=pmcentrez&rendertype=abstract> (Accessed April 24, 2016).

Furniss D, Lettice LA, Taylor IB, Critchley PS, Giele H, Hill RE, Wilkie AOM. 2008. A variant in the sonic hedgehog regulatory sequence (ZRS) is associated with triphalangeal thumb and deregulates expression in the developing limb. *Hum Mol Genet* **17**: 2417–2423.

<http://www.ncbi.nlm.nih.gov/pubmed/18463159> (Accessed October 10, 2017).

Fustin J-M, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka MS, Kakeya H, Manabe I, et al. 2013. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* **155**: 793–806.

<http://www.sciencedirect.com/science/article/pii/S0092867413013445> (Accessed March 5, 2016).

G Hendrickson D, Kelley DR, Tenen D, Bernstein B, Rinn JL. 2016. Widespread RNA binding by chromatin-associated proteins. *Genome Biol* **17**: 28.

<http://genomebiology.com/2016/17/1/28> (Accessed August 16, 2017).

GAMOW G. 1954. Possible Relation between Deoxyribonucleic Acid and Protein Structures.

*Nature* **173**: 318–318. <http://www.nature.com/doifinder/10.1038/173318a0> (Accessed March 31, 2017).

Gao X, Zhao X, Zhu Y, He J, Shao J, Su C, Zhang Y, Zhang W, Saarikettu J, Silvennoinen O, et al. 2012. Tudor Staphylococcal Nuclease (Tudor-SN) Participates in Small Ribonucleoprotein (snRNP) Assembly via Interacting with Symmetrically Dimethylated Sm Proteins. *J Biol Chem* **287**: 18130–18141. <http://www.ncbi.nlm.nih.gov/pubmed/22493508> (Accessed August 19, 2017).

Garcia O, Saveanu C, Cline M, Fromont-Racine M, Jacquier A, Schwikowski B, Aittokallio T.

2007. GOLORize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics* **23**: 394–6.

<http://www.ncbi.nlm.nih.gov/pubmed/17127678> (Accessed May 6, 2016).

Gartner JJ, Parker SCJ, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, Davis S, Simhadri VL, Jha S, Katagiri N, et al. 2013. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc Natl Acad Sci U S A* **110**: 13481–6.

<http://www.ncbi.nlm.nih.gov/pubmed/23901115> (Accessed August 18, 2017).

Gaspar P, Carbonell J, Oliveira JL. 2012. On the parameter optimization of Support Vector Machines for binary classification. *J Integr Bioinform* **9**: 201.

<http://www.ncbi.nlm.nih.gov/pubmed/22829572> (Accessed May 20, 2016).

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

<http://www.ncbi.nlm.nih.gov/pubmed/15461798> (Accessed August 17, 2017).

Glick M, Jenkins JL, Nettles JH, Hitchings H, Davies JW. 2006. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J Chem Inf Model* **46**: 193–200. <http://dx.doi.org/10.1021/ci050374h> (Accessed May 20, 2016).

Godden-Kent D, Talbot SJ, Boshoff C, Chang Y, Moore P, Weiss RA, Mittnacht S. 1997. The cyclin encoded by Kaposi's sarcoma-associated herpesvirus stimulates cdk6 to phosphorylate the retinoblastoma protein and histone H1. *J Virol* **71**: 4193–8.

<http://www.ncbi.nlm.nih.gov/pubmed/9151805> (Accessed August 19, 2017).

Golbreich C, Zhang S, Bodenreider O. 2006. The foundational model of anatomy in OWL: Experience and perspectives. *Web Semant* **4**: 181–195.

<http://www.ncbi.nlm.nih.gov/pubmed/18360535>.

Golovina AY, Dzama MM, Osterman IA, Sergiev P V, Serebryakova M V, Bogdanov AA, Dontsova OA. 2012. The last rRNA methyltransferase of E. coli revealed: the yhiR gene encodes adenine-N6 methyltransferase specific for modification of A2030 of 23S ribosomal RNA. *RNA* **18**: 1725–34.

[http://rnajournal.cshlp.org/content/18/9/1725.abstract?ijkey=6dbc60b950c942bba1b17e4eb83af24608cd11ea&keytype2=tf\\_ipsecsha](http://rnajournal.cshlp.org/content/18/9/1725.abstract?ijkey=6dbc60b950c942bba1b17e4eb83af24608cd11ea&keytype2=tf_ipsecsha) (Accessed May 9, 2016).

Golovina AY, Dzama MM, Petriukov KS, Zatsepin TS, Sergiev P V, Bogdanov AA, Dontsova OA. 2014. Method for site-specific detection of m6A nucleoside presence in RNA based on high-resolution melting (HRM) analysis. *Nucleic Acids Res* **42**: e27.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3936739&tool=pmcentrez&rendertype=abstract> (Accessed May 13, 2016).

- Gordon CT, Attanasio C, Bhatia S, Benko S, Ansari M, Tan TY, Munnich A, Pennacchio LA, Abadie V, Temple IK, et al. 2014. Identification of Novel Craniofacial Regulatory Domains Located far Upstream of *SOX9* and Disrupted in Pierre Robin Sequence. *Hum Mutat* **35**: 1011–1020. <http://www.ncbi.nlm.nih.gov/pubmed/24934569> (Accessed October 10, 2017).
- Greenleaf WJ, Sidow A. 2014. The future of sequencing: convergence of intelligent design and market Darwinism. *Genome Biol* **15**: 303. <http://www.ncbi.nlm.nih.gov/pubmed/25000818> (Accessed May 11, 2017).
- Guo S-H, Deng E-Z, Xu L-Q, Ding H, Lin H, Chen W, Chou K-C. 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **30**: 1522–9. <http://www.ncbi.nlm.nih.gov/pubmed/24504871> (Accessed May 20, 2016).
- Gupta U Das, Menon V, Babbar U. 2010. Detecting the Number of Clusters during Expectation-Maximization Clustering Using Information Criterion. In *2010 Second International Conference on Machine Learning and Computing*, pp. 169–173, IEEE <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5460748> (Accessed August 3, 2016).
- Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**: 234–8. <http://www.ncbi.nlm.nih.gov/pubmed/6316146> (Accessed August 17, 2017).
- Gutierrez-Beltran E, Denisenko T V, Zhivotovsky B, Bozhkov P V. 2016. Tudor staphylococcal nuclease: biochemistry and functions. *Cell Death Differ* **23**: 1739–1748. <http://www.nature.com/doifinder/10.1038/cdd.2016.93> (Accessed August 19, 2017).
- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software. *ACM SIGKDD Explor Newsl* **11**: 10. <http://dl.acm.org/citation.cfm?id=1656274.1656278> (Accessed November 25, 2014).
- Hamerly Y. 2007. PG-means: learning the number of clusters in data. *Adv neural Inf Process*. <https://books.google.co.uk/books?hl=en&lr=&id=Tbn1I9P1220C&oi=fnd&pg=PA393&dq=PG-means:+learning+the+number+of+clusters+in+data&ots=V3r6Dftr00&sig=prKtJ228Q->

x3qfKEe8YiPHovKJk (Accessed August 3, 2016).

Harcourt EM, Ehrenschwender T, Batista PJ, Chang HY, Kool ET. 2013. Identification of a selective polymerase enables detection of N(6)-methyladenosine in RNA. *J Am Chem Soc* **135**: 19079–82. <http://pubs.acs.org/doi/abs/10.1021/ja4105792> (Accessed May 16, 2016).

He Y, Smith R. 2009. Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B. *Cell Mol Life Sci* **66**: 1239–56. <http://www.ncbi.nlm.nih.gov/pubmed/19099192> (Accessed May 10, 2016).

Heiss NS, Knight SW, Vulliamy TJ, Klauck SM, Wiemann S, Mason PJ, Poustka A, Dokal I. 1998. X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. *Nat Genet* **19**: 32–8. <http://www.ncbi.nlm.nih.gov/pubmed/9590285> (Accessed May 4, 2016).

Hess ME, Hess S, Meyer KD, Verhagen LAW, Koch L, Brönneke HS, Dietrich MO, Jordan SD, Saletore Y, Elemento O, et al. 2013. The fat mass and obesity associated gene (Fto) regulates activity of the dopaminergic midbrain circuitry. *Nat Neurosci* **16**: 1042–8. <http://dx.doi.org/10.1038/nn.3449> (Accessed April 17, 2016).

Hinney A, Nguyen TT, Scherag A, Friedel S, Brönner G, Müller TD, Grallert H, Illig T, Wichmann H-E, Rief W, et al. 2007. Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLoS One* **2**: e1361. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2137937&tool=pmcentrez&rendertype=abstract> (Accessed May 11, 2016).

Hirose K, Kawano S, Konishi S, Ichikawa M. 2011. Bayesian information criterion and selection of the number of factors in factor analysis models. *J Data Sci*. [http://www.jdsruc.org/upload/JDS-682\(2011-04-01164527\).pdf](http://www.jdsruc.org/upload/JDS-682(2011-04-01164527).pdf) (Accessed April 13, 2016).

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. *Genet Med* **4**: 45–61. <http://dx.doi.org/10.1097/00125817-200203000-00002> (Accessed September 25, 2015).

Ho AJ, Stein JL, Hua X, Lee S, Hibar DP, Leow AD, Dinov ID, Toga AW, Saykin AJ, Shen L, et al.

2010. A commonly carried allele of the obesity-related FTO gene is associated with reduced brain volume in the healthy elderly. *Proc Natl Acad Sci U S A* **107**: 8404–9. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2889537&tool=pmcentrez&rendertype=abstract> (Accessed May 12, 2016).

Holley RW, Apgar J, Merrill SH, Zubkoff PL. 1961. NUCLEOTIDE AND OLIGONUCLEOTIDE COMPOSITIONS OF THE ALANINE-, VALINE-, AND TYROSINE-ACCEPTOR “SOLUBLE” RIBONUCLEIC ACIDS OF YEAST. *J Am Chem Soc* **83**: 4861–4862. <http://pubs.acs.org/doi/abs/10.1021/ja01484a040> (Accessed May 11, 2017).

HOLLEY RW, EVERETT GA, MADISON JT, ZAMIR A. 1965. NUCLEOTIDE SEQUENCES IN THE YEAST ALANINE TRANSFER RIBONUCLEIC ACID. *J Biol Chem* **240**: 2122–8. <http://www.ncbi.nlm.nih.gov/pubmed/14299636> (Accessed February 14, 2017).

Hongay CF, Orr-Weaver TL. 2011. Drosophila Inducer of MEiosis 4 (IME4) is required for Notch signaling during oogenesis. *Proc Natl Acad Sci U S A* **108**: 14855–60. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3169142&tool=pmcentrez&rendertype=abstract> (Accessed May 12, 2016).

Horiuchi K, Kawamura T, Iwanari H, Ohashi R, Naito M, Kodama T, Hamakubo T. 2013. Identification of Wilms’ tumor 1-associating protein complex and its role in alternative splicing and the cell cycle. *J Biol Chem* **288**: 33292–302. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3829175&tool=pmcentrez&rendertype=abstract> (Accessed April 28, 2016).

Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. 2011. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* **21**: 734–40. <http://genome.cshlp.org/content/21/5/734.long> (Accessed May 24, 2016).

Hsieh Y-F, Liu G-Y, Lee Y-J, Yang J-J, Sándor K, Sarang Z, Bononi A, Pinton P, Tretter L, Szondy Z, et al. 2013. Transglutaminase 2 contributes to apoptosis induction in Jurkat T cells by modulating Ca<sup>2+</sup> homeostasis via cross-linking RAP1GDS1. *PLoS One* **8**: e81516. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3859493&tool=pmcentrez&rendertype=abstract> (Accessed May 6, 2016).

Huang DW, Sherman BT, Lempicki RA. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44–57.

<http://www.ncbi.nlm.nih.gov/pubmed/19131956> (Accessed August 19, 2017).

Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN, Smith D, Albà MM, et al. 2004. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol* **5**: R47.

<http://www.ncbi.nlm.nih.gov/pubmed/15239832> (Accessed August 18, 2017).

Huang Y, Yan J, Li Q, Li J, Gong S, Zhou H, Gan J, Jiang H, Jia G-F, Luo C, et al. 2015.

Meclofenamic acid selectively inhibits FTO demethylation of m6A over ALKBH5. *Nucleic Acids Res* **43**: 373–84.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4288171&tool=pmcentrez&rendertype=abstract> (Accessed May 9, 2016).

Hutz JE, Kraja AT, McLeod, HL, Province MA. 2008. CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet Epidemiol* **32**: 779–790.

<http://www.ncbi.nlm.nih.gov/pubmed/18613097> (Accessed August 18, 2017).

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science (80- )* **324**: 218–223. <http://www.ncbi.nlm.nih.gov/pubmed/19213877> (Accessed August 16, 2017).

Ioannidis JPA, Trikalinos TA, Khoury MJ. 2006. Implications of Small Effect Sizes of Individual Genetic Variants on the Design and Interpretation of Genetic Association Studies of Complex Diseases. *Am J Epidemiol* **164**: 609–614.

<http://www.ncbi.nlm.nih.gov/pubmed/16893921> (Accessed August 17, 2017).

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, et al. 2005. Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2**: 345–350. <http://www.ncbi.nlm.nih.gov/pubmed/15846361>.

Iwanami Y, Brown GM. 1968. Methylated bases of transfer ribonucleic acid from hela and L cells. *Arch Biochem Biophys* **124**: 472–482.

<http://www.sciencedirect.com/science/article/pii/000398616890355X> (Accessed May 5, 2016).

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, et al. 2009. STRING 8--a global view on proteins and their functional

interactions in 630 organisms. *Nucleic Acids Res* **37**: D412-6.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686466&tool=pmcentrez&rendertype=abstract> (Accessed May 5, 2015).

Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, Yi C, Lindahl T, Pan T, Yang Y-G, et al. 2011. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol* **7**: 885–7.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3218240&tool=pmcentrez&rendertype=abstract> (Accessed June 1, 2015).

Jiang JJ, Conrath DW. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. <https://arxiv.org/pdf/cmp-lg/9709008.pdf> (Accessed August 19, 2017).

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* **21**: 1543–51.

<http://www.ncbi.nlm.nih.gov/pubmed/21816910> (Accessed November 30, 2017).

Jiang R. 2015. Walking on multiple disease-gene networks to prioritize candidate genes. *J Mol Cell Biol*. <http://www.ncbi.nlm.nih.gov/pubmed/25681405> (Accessed February 18, 2015).

Joshi T, Xu D. 2007. Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics* **8**: 222.

<http://www.ncbi.nlm.nih.gov/pubmed/17620139> (Accessed August 10, 2017).

Jung M, Kadam S, Xiong W, Rauch TA, Jin S-G, Pfeifer GP. 2015. MIRA-seq for DNA methylation analysis of CpG islands. *Epigenomics* **7**: 695–706.

<http://www.ncbi.nlm.nih.gov/pubmed/25881900> (Accessed August 16, 2017).

Kahles A, Ong CS, Zhong Y, Rättsch G. 2016. *SplAdder* : identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* **32**: 1840–1847.

<http://www.ncbi.nlm.nih.gov/pubmed/26873928> (Accessed August 19, 2017).

Kane SE, Beemon K. 1985. Precise localization of m6A in Rous sarcoma virus RNA reveals clustering of methylation sites: implications for RNA processing. *Mol Cell Biol* **5**: 2298–306.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=366956&tool=pmcentrez&rendertype=abstract> (Accessed April 24, 2016).

- Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27–30. <http://www.ncbi.nlm.nih.gov/pubmed/10592173> (Accessed August 17, 2017).
- Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. 2010. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res* **38**: D690-8. <http://www.ncbi.nlm.nih.gov/pubmed/19906730> (Accessed August 18, 2017).
- Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-6. [http://nar.oxfordjournals.org/content/32/suppl\\_1/D493.abstract?ijkey=06tIQcBr2VZNz&keytype=ref](http://nar.oxfordjournals.org/content/32/suppl_1/D493.abstract?ijkey=06tIQcBr2VZNz&keytype=ref) (Accessed December 24, 2014).
- Ke S, Alemu EA, Mertens C, Gantman EC, Fak JJ, Mele A, Haripal B, Zucker-Scharff I, Moore MJ, Park CY, et al. 2015. A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev* **29**: 2037–53. <http://genesdev.cshlp.org/content/29/19/2037.long> (Accessed September 28, 2015).
- Kedes DH, Operskalski E, Busch M, Kohn R, Flood J, Ganem D. 1996. The seroepidemiology of human herpesvirus 8 (Kaposi's sarcoma-associated herpesvirus): Distribution of infection in KS risk groups and evidence for sexual transmission. *Nat Med* **2**: 918–924. <http://www.nature.com/doi/10.1038/nm0896-918> (Accessed August 19, 2017).
- Keller M, Mazuch J, Abraham U, Eom GD, Herzog ED, Volk H-D, Kramer A, Maier B. 2009. A circadian clock in macrophages controls inflammatory immune responses. *Proc Natl Acad Sci U S A* **106**: 21407–12. <http://www.pnas.org/content/106/50/21407.full> (Accessed April 11, 2016).
- Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**: 1073–80. <http://www.ncbi.nlm.nih.gov/pubmed/2570460> (Accessed August 17, 2017).
- Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**: 103–7. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3847670&tool=pmcentrez&rendertype=abstract> (Accessed May 9, 2016).
- Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D,

et al. 2015. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* **43**: D1071-8.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4383880&tool=pmcentrez&rendertype=abstract> (Accessed March 10, 2015).

Kierzek E, Kierzek R. 2003. The thermodynamic stability of RNA duplexes and hairpins containing N6-alkyladenosines and 2-methylthio-N6-alkyladenosines. *Nucleic Acids Res* **31**: 4472–80.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=169893&tool=pmcentrez&rendertype=abstract> (Accessed May 10, 2016).

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053844&tool=pmcentrez&rendertype=abstract> (Accessed July 9, 2014).

Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (80- )* **308**: 385–389.

<http://www.ncbi.nlm.nih.gov/pubmed/15761122> (Accessed August 17, 2017).

Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth H V, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, et al. 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* **42**: D966-74.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965098&tool=pmcentrez&rendertype=abstract> (Accessed March 11, 2015).

Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, Gkoutos G, Schofield P, Smedley D, Lewis SE, et al. 2013. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research.

*F1000Research* **2**: 30.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3799545&tool=pmcentrez&rendertype=abstract> (Accessed March 29, 2015).

Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S,

- Robinson PN. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* **85**: 457–64.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2756558&tool=pmcentrez&rendertype=abstract> (Accessed March 29, 2015).
- Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, et al. 2015. ArrayExpress update--simplifying data submissions. *Nucleic Acids Res* **43**: D1113-6. <http://europepmc.org/articles/PMC4383899> (Accessed April 21, 2015).
- Kong H, Lin LF, Porter N, Stickel S, Byrd D, Posfai J, Roberts RJ. 2000. Functional analysis of putative restriction-modification system genes in the *Helicobacter pylori* J99 genome. *Nucleic Acids Res* **28**: 3216–23. <http://www.ncbi.nlm.nih.gov/pubmed/10954588> (Accessed July 20, 2016).
- König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909–15. <http://dx.doi.org/10.1038/nsmb.1838> (Accessed April 24, 2016).
- Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, Blake A, Chen C-K, Easty R, Di Fenza A, et al. 2014. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res* **42**: D802–D809. <http://www.ncbi.nlm.nih.gov/pubmed/24194600> (Accessed August 18, 2017).
- Kowalak JA, Pomerantz SC, Crain PF, McCloskey JA. 1993. A novel method for the determination of post-transcriptional modification in RNA by mass spectrometry. *Nucleic Acids Res* **21**: 4577–4585. <http://www.scopus.com/inward/record.url?eid=2-s2.0-0027423649&partnerID=tZOtx3y1>.
- Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A. 2012. RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics* **28**: 1184–1185. <http://www.ncbi.nlm.nih.gov/pubmed/22345621> (Accessed August 18, 2017).
- Kryukov G V., Pennacchio LA, Sunyaev SR. 2007. Most Rare Missense Alleles Are Deleterious in

- Humans: Implications for Complex Disease and Association Studies. *Am J Hum Genet* **80**: 727–739. <http://www.ncbi.nlm.nih.gov/pubmed/17357078> (Accessed August 18, 2017).
- Kumar P, Ma X, Liu X, Jia J, Bucong H, Xue Y, Li ZR, Yang SY, Wei YQ, Chen YZ. 2011. Effect of training data size and noise level on support vector machines virtual screening of genotoxic compounds from large compound libraries. *J Comput Aided Mol Des* **25**: 455–67. <http://www.ncbi.nlm.nih.gov/pubmed/21556903> (Accessed May 20, 2016).
- Kvam VM, Liu P, Si Y. 2012. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot* **99**: 248–256. <http://www.ncbi.nlm.nih.gov/pubmed/22268221>.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921. <http://www.ncbi.nlm.nih.gov/pubmed/11237011> (Accessed April 1, 2017).
- Landfors M, Nakken S, Fusser M, Dahl J-A, Klungland A, Fedorcsak P. 2016. Sequencing of FTO and ALKBH5 in men undergoing infertility work-up identifies infertility-associated variant and two missense mutations. *Fertil Steril* **105**: 1170–1179.e5. <http://www.ncbi.nlm.nih.gov/pubmed/26820768> (Accessed April 16, 2016).
- Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**: D980-5. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965032&tool=pmcentrez&rendertype=abstract> (Accessed January 14, 2015).
- Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559. <http://www.ncbi.nlm.nih.gov/pubmed/19114008> (Accessed October 11, 2017).
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359. <http://www.ncbi.nlm.nih.gov/pubmed/22388286> (Accessed August 17, 2017).
- Lavi S, Shatkin AJ. 1975. Methylated simian virus 40-specific RNA from nuclei and cytoplasm of infected BSC-1 cells. *Proc Natl Acad Sci U S A* **72**: 2012–6. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=432682&tool=pmcentrez&rendertype=abstract> (Accessed May 4, 2016).

- Leach RA, Tuck MT. 2001. Methionine depletion induces transcription of the mRNA (N6-adenosine)methyltransferase. *Int J Biochem Cell Biol* **33**: 1116–1128.  
<http://www.sciencedirect.com/science/article/pii/S1357272501000723> (Accessed May 5, 2016).
- Lee YY, Yu YB, Gunawardena HP, Xie L, Chen X. 2012. BCLAF1 is a radiation-induced H2AX-interacting partner involved in  $\gamma$ H2AX-mediated regulation of apoptosis and DNA repair. *Cell Death Dis* **3**: e359.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3406578&tool=pmcentrez&rendertype=abstract> (Accessed April 8, 2016).
- Lehmann R, Childs L, Thomas P, Abreu M, Fuhr L, Herzel H, Leser U, Relógio A. 2015. Assembly of a comprehensive regulatory network for the mammalian circadian clock: a bioinformatics approach. *PLoS One* **10**: e0126283.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4422523&tool=pmcentrez&rendertype=abstract> (Accessed May 12, 2016).
- Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. 2011a. The sequence read archive. *Nucleic Acids Res* **39**: D19–21.  
<http://www.ncbi.nlm.nih.gov/pubmed/21062823> (Accessed May 11, 2017).
- Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. 2011b. The sequence read archive. *Nucleic Acids Res* **39**: D19–21.  
<http://www.ncbi.nlm.nih.gov/pubmed/21062823> (Accessed August 16, 2017).
- Li F, Zhao D, Wu J, Shi Y. 2014a. Structure of the YTH domain of human YTHDF2 in complex with an m(6)A mononucleotide reveals an aromatic cage for m(6)A recognition. *Cell Res* **24**: 1490–2.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4260351&tool=pmcentrez&rendertype=abstract> (Accessed May 10, 2016).
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–60.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract> (Accessed July 9, 2014).
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.

2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–9. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract> (Accessed July 9, 2014).
- Li Y, Song S, Li C, Yu J. 2013. MeRIP-PF: an easy-to-use pipeline for high-resolution peak-finding in MeRIP-Seq data. *Genomics Proteomics Bioinformatics* **11**: 72–5. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4357668&tool=pmcentrez&rendertype=abstract> (Accessed May 17, 2016).
- Li Y, Wang X, Li C, Hu S, Yu J, Song S. 2014b. Transcriptome-wide N<sup>6</sup>-methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification. *RNA Biol* **11**: 1180–8. <http://www.ncbi.nlm.nih.gov/pubmed/25483034> (Accessed May 9, 2016).
- Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**: e108. <http://www.ncbi.nlm.nih.gov/pubmed/23558742> (Accessed August 19, 2017).
- Lichinchi G, Gao S, Saletore Y, Gonzalez GM, Bansal V, Wang Y, Mason CE, Rana TM. 2016. Dynamics of the human and viral m6A RNA methylomes during HIV-1 infection of T cells. *Nat Microbiol* **1**: 16011. <http://www.nature.com/articles/nmicrobiol201611> (Accessed February 23, 2016).
- Lifton RP. Genetic dissection of human blood pressure variation: common pathways from rare phenotypes. *Harvey Lect* **100**: 71–101. <http://www.ncbi.nlm.nih.gov/pubmed/16970175> (Accessed August 17, 2017).
- Lin D. 1988. An information-theoretic definition of similarity. *Proc 15th Int Mach Learn*.
- Lin S, Choe J, Du P, Triboulet R, Gregory RI. 2016. The m(6)A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. *Mol Cell*. <http://www.ncbi.nlm.nih.gov/pubmed/27117702> (Accessed April 28, 2016).
- Linder B, Grozhik A V, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. 2015. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* **12**: 767–72. <http://dx.doi.org/10.1038/nmeth.3453> (Accessed July 6, 2015).
- Liu HC, Chen CY, Liu YT, Chu CB, Liang DC, Shih LY, Lin CJ. 2008. Cross-generation and cross-

- laboratory predictions of Affymetrix microarrays by rank-based methods. *J Biomed Inf* **41**: 570–579. <http://www.ncbi.nlm.nih.gov/pubmed/18234562>.
- Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X, et al. 2014. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat Chem Biol* **10**: 93–5. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3911877&tool=pmcentrez&rendertype=abstract> (Accessed February 29, 2016).
- Liu K, Chen C, Guo Y, Lam R, Bian C, Xu C, Zhao DY, Jin J, MacKenzie F, Pawson T, et al. 2010a. Structural basis for recognition of arginine methylated Piwi proteins by the extended Tudor domain. *Proc Natl Acad Sci* **107**: 18398–18403. <http://www.ncbi.nlm.nih.gov/pubmed/20937909> (Accessed August 19, 2017).
- Liu L, Eby MT, Rathore N, Sinha SK, Kumar A, Chaudhary PM. 2002. The Human Herpes Virus 8-encoded Viral FLICE Inhibitory Protein Physically Associates with and Persistently Activates the I $\kappa$ B Kinase Complex. *J Biol Chem* **277**: 13745–13751. <http://www.ncbi.nlm.nih.gov/pubmed/11830587> (Accessed August 19, 2017).
- Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. 2015. N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* **518**: 560–564. <http://dx.doi.org/10.1038/nature14234> (Accessed February 25, 2015).
- Liu N, Parisien M, Dai Q, Zheng G, He C, Pan T. 2013. Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA* **19**: 1848–56. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3884656&tool=pmcentrez&rendertype=abstract> (Accessed March 8, 2016).
- Liu S, Jia J, Gao Y, Zhang B, Han Y. 2010b. The AtTudor2, a protein with SN-Tudor domains, is involved in control of seed germination in Arabidopsis. *Planta* **232**: 197–207. <http://www.ncbi.nlm.nih.gov/pubmed/20396901> (Accessed August 19, 2017).
- Liu Z, Xiao X, Yu D-J, Jia J, Qiu W-R, Chou K-C. 2016. pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. *Anal Biochem* **497**: 60–7. <http://www.sciencedirect.com/science/article/pii/S0003269715005795> (Accessed May 20, 2016).

- Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**: 177–182.  
<http://www.ncbi.nlm.nih.gov/pubmed/12524541> (Accessed August 17, 2017).
- Lokdarshi A, Conner WC, McClintock C, Li T, Roberts DM. 2016. Arabidopsis CML38, a Calcium Sensor That Localizes to Ribonucleoprotein Complexes under Hypoxia Stress. *Plant Physiol* **170**: 1046–1059. <http://www.ncbi.nlm.nih.gov/pubmed/26634999> (Accessed August 19, 2017).
- Lopez-Bigas N, Ouzounis CA. 2004. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* **32**: 3108–3114.  
<http://www.ncbi.nlm.nih.gov/pubmed/15181176> (Accessed August 18, 2017).
- Lord CJ, Ashworth A. 2017. PARP inhibitors: Synthetic lethality in the clinic. *Science (80- )* **355**: 1152–1158. <http://www.ncbi.nlm.nih.gov/pubmed/28302823> (Accessed August 18, 2017).
- Lord PW, Stevens RD, Brass A, Goble CA. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**: 1275–83. <http://www.ncbi.nlm.nih.gov/pubmed/12835272> (Accessed May 5, 2015).
- Lorenz R, Bernhart SH, Höner Zu Siederdisen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3319429&tool=pmcentrez&rendertype=abstract> (Accessed May 20, 2016).
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.  
<http://www.ncbi.nlm.nih.gov/pubmed/25516281> (Accessed August 19, 2017).
- Lovejoy AF, Riordan DP, Brown PO. 2014. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in *S. cerevisiae*. *PLoS One* **9**: e110799.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4212993&tool=pmcentrez&rendertype=abstract> (Accessed May 4, 2016).
- Luo G-Z, MacQueen A, Zheng G, Duan H, Dore LC, Lu Z, Liu J, Chen K, Jia G, Bergelson J, et al.

2014. Unique features of the m6A methylome in *Arabidopsis thaliana*. *Nat Commun* **5**: 5630.  
<http://www.nature.com/ncomms/2014/141128/ncomms6630/abs/ncomms6630.html>  
(Accessed May 5, 2016).
- Luo S, Tong L. 2014. Molecular basis for the recognition of methylated adenines in RNA by the eukaryotic YTH domain. *Proc Natl Acad Sci U S A* **111**: 13834–9.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4183320&tool=pmcentrez&rendertype=abstract> (Accessed May 9, 2016).
- MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**: 823–8.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3299548&tool=pmcentrez&rendertype=abstract> (Accessed March 9, 2015).
- Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, Olchowik A, Januszewski W, Kalinowski S, Dunin-Horkawicz S, Rother KM, et al. 2013. MODOMICS: a database of RNA modification pathways--2013 update. *Nucleic Acids Res* **41**: D262–7.  
[http://nar.oxfordjournals.org/content/41/D1/D262.abstract?ijkey=5a183ae25317f3c6525b57d1febdc84218e65d52&keytype=tf\\_ipsecsha](http://nar.oxfordjournals.org/content/41/D1/D262.abstract?ijkey=5a183ae25317f3c6525b57d1febdc84218e65d52&keytype=tf_ipsecsha) (Accessed February 16, 2016).
- Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**: 3448–9.  
<http://www.ncbi.nlm.nih.gov/pubmed/15972284> (Accessed July 14, 2014).
- Maes HH, Neale MC, Eaves LJ. 1997. Genetic and environmental factors in relative body weight and human adiposity. *Behav Genet* **27**: 325–51.  
<http://www.ncbi.nlm.nih.gov/pubmed/9519560> (Accessed February 25, 2016).
- Magrane M, Consortium U. 2011. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**: bar009. <http://www.ncbi.nlm.nih.gov/pubmed/21447597>.
- Maimon O, Rokach L, eds. 2010. *Data Mining and Knowledge Discovery Handbook*. Springer US, Boston, MA <http://www.springerlink.com/index/10.1007/978-0-387-09823-4>  
(Accessed April 6, 2016).
- Maitra R. 2009. Initializing partition-optimization algorithms. *IEEE/ACM Trans Comput Biol*

- Bioinform* **6**: 144–57. <http://www.ncbi.nlm.nih.gov/pubmed/19179708> (Accessed April 28, 2016).
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–80. <http://www.ncbi.nlm.nih.gov/pubmed/16056220> (Accessed February 9, 2017).
- Martin JN, Ganem DE, Osmond DH, Page-Shafer KA, Macrae D, Kedes DH. 1998. Sexual Transmission and the Natural History of Human Herpesvirus 8 Infection. *N Engl J Med* **338**: 948–954. <http://www.ncbi.nlm.nih.gov/pubmed/9521982> (Accessed August 19, 2017).
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10. <http://journal.embnet.org/index.php/embnetjournal/article/view/200> (Accessed January 4, 2017).
- Martinez-Contreras R, Cloutier P, Shkreta L, Fiset J-F, Revil T, Chabot B. 2007. hnRNP proteins and splicing control. *Adv Exp Med Biol* **623**: 123–47. <http://www.ncbi.nlm.nih.gov/pubmed/18380344> (Accessed May 10, 2016).
- Mashima J, Kodama Y, Kosuge T, Fujisawa T, Katayama T, Nagasaki H, Okuda Y, Kaminuma E, Ogasawara O, Okubo K, et al. 2016. DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Res* **44**: D51–D57. <http://www.ncbi.nlm.nih.gov/pubmed/26578571> (Accessed November 30, 2017).
- Masotti D, Nardini C, Rossi S, Bonora E, Romeo G, Volinia S, Benini L. 2008. TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders. *Bioinformatics* **24**: 428–429. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm588> (Accessed August 18, 2017).
- Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**: 560–4. <http://www.ncbi.nlm.nih.gov/pubmed/265521> (Accessed February 14, 2017).
- McCloskey A, Taniguchi I, Shinmyozu K, Ohno M. 2012. hnRNP C tetramer measures RNA length to classify RNA polymerase II transcripts for export. *Science* **335**: 1643–6. <http://science.sciencemag.org/content/335/6076/1643.abstract> (Accessed March 28,

2016).

- McGraw S, Vigneault C, Sirard M-A. 2007. Temporal expression of factors involved in chromatin remodeling and in gene regulation during early bovine in vitro embryo development. *Reproduction* **133**: 597–608. <http://www.reproduction-online.org/content/133/3/597.abstract> (Accessed May 5, 2016).
- McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**: 1527–1541. <http://www.ncbi.nlm.nih.gov/pubmed/19546169> (Accessed February 9, 2017).
- McMurray F, Church CD, Larder R, Nicholson G, Wells S, Teboul L, Tung YCL, Rimmington D, Bosch F, Jimenez V, et al. 2013. Adult onset global loss of the *fto* gene alters body composition and metabolism in the mouse. *PLoS Genet* **9**: e1003166. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3536712&tool=pmcentrez&rendertype=abstract> (Accessed May 11, 2016).
- Meehan TF, Conte N, West DB, Jacobsen JO, Mason J, Warren J, Chen C-K, Tudose I, Relac M, Matthews P, et al. 2017. Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat Genet* **49**: 1231–1238. <http://www.ncbi.nlm.nih.gov/pubmed/28650483> (Accessed August 18, 2017).
- Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis NA, Li T, Miteva Y V, Hauri S, Sardi ME, Low TY, et al. 2013. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* **10**: 730–6. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3773500&tool=pmcentrez&rendertype=abstract> (Accessed February 21, 2016).
- Melnykov V, Melnykov I. 2012. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *Comput Stat Data Anal* **56**: 1381–1395. <http://www.sciencedirect.com/science/article/pii/S0167947311003963> (Accessed March 10, 2016).
- Meng J, Cui X, Rao MK, Chen Y, Huang Y. 2013. Exome-based analysis for RNA epigenome sequencing data. *Bioinformatics* **29**: 1565–7.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3673212&tool=pmcentrez&rendertype=abstract> (Accessed September 15, 2015).

Meng J, Lu Z, Liu H, Zhang L, Zhang S, Chen Y, Rao MK, Huang Y. 2014. A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. *Methods* **69**: 274–81.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4194139&tool=pmcentrez&rendertype=abstract> (Accessed September 3, 2015).

Meyer KD, Jaffrey SR. 2014. The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nat Rev Mol Cell Biol* **15**: 313–26.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4393108&tool=pmcentrez&rendertype=abstract> (Accessed July 28, 2015).

Meyer KD, Patil DP, Zhou J, Zinoviev A, Skabkin MA, Elemento O, Pestova TV, Qian S-B, Jaffrey SR. 2015. 5' UTR m6A Promotes Cap-Independent Translation. *Cell* **163**: 999–1010.

<http://www.cell.com/article/S0092867415013252/fulltext> (Accessed October 26, 2015).

Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. 2012. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**: 1635–46.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3383396&tool=pmcentrez&rendertype=abstract> (Accessed September 16, 2014).

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* **45**: D183–D189.

<https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1138> (Accessed August 19, 2017).

Milaneschi Y, Lamers F, Mbarek H, Hottenga J-J, Boomsma DI, Penninx BWJH. 2014. The effect of FTO rs9939609 on major depression differs across MDD subtypes. *Mol Psychiatry* **19**: 960–2. <http://dx.doi.org/10.1038/mp.2014.4> (Accessed April 25, 2016).

Monini P, Colombini S, Stürzl M, Goletti D, Cafaro A, Sgadari C, Buttò S, Franco M, Leone P, Fais S, et al. 1999. Reactivation and persistence of human herpesvirus-8 infection in B cells and monocytes by Th-1 cytokines increased in Kaposi's sarcoma. *Blood* **93**: 4044–58.

<http://www.ncbi.nlm.nih.gov/pubmed/10361101> (Accessed August 19, 2017).

- Morrison JL, Breitling R, Higham DJ, Gilbert DR. 2005. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* **6**: 233. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1261158&tool=pmcentrez&rendertype=abstract> (Accessed May 5, 2015).
- Moskvina V, O'Donovan MC. 2007. Detailed Analysis of the Relative Power of Direct and Indirect Association Studies and the Implications for Their Interpretation. *Hum Hered* **64**: 63–73. <http://www.ncbi.nlm.nih.gov/pubmed/17483598> (Accessed August 17, 2017).
- Mueller WF, Larsen LSZ, Garibaldi A, Hatfield GW, Hertel KJ. 2015. The Silent Sway of Splicing by Synonymous Substitutions. *J Biol Chem* **290**: 27700–11. <http://www.ncbi.nlm.nih.gov/pubmed/26424794> (Accessed August 18, 2017).
- Mungall CJ, Torniai C, Gkoutos G V, Lewis SE, Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* **13**: R5. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3334586&tool=pmcentrez&rendertype=abstract> (Accessed February 3, 2015).
- Munkres J. 2006. Algorithms for the Assignment and Transportation Problems. <http://dx.doi.org/10.1137/0105003>.
- Munns TW, Liszewski MK, Sims HF. 1977. Characterization of antibodies specific for N6-methyladenosine and for 7-methylguanosine. *Biochemistry* **16**: 2163–8. <http://www.ncbi.nlm.nih.gov/pubmed/861202> (Accessed July 20, 2016).
- Nabel GJ, Friberg J, Kong W, Hottiger MO. p53 inhibition by the LANA protein of KSHV protects against cell death. *Nature* **402**: 889–894. <http://www.ncbi.nlm.nih.gov/pubmed/10622254> (Accessed August 19, 2017).
- Narayan P, Ludwiczak RL, Goodwin EC, Rottman FM. 1994. Context effects on N6-adenosine methylation sites in prolactin mRNA. *Nucleic Acids Res* **22**: 419–26. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=523598&tool=pmcentrez&rendertype=abstract> (Accessed March 11, 2016).
- Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453.

<http://linkinghub.elsevier.com/retrieve/pii/0022283670900574> (Accessed August 17, 2017).

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, et al. 2010a. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42**: 790–793.

<http://www.ncbi.nlm.nih.gov/pubmed/20711175> (Accessed August 17, 2017).

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al. 2010b. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**: 30–35. <http://www.ncbi.nlm.nih.gov/pubmed/19915526> (Accessed July 29, 2017).

Nitsch D, Tranchevent L-C, Gonçalves JP, Vogt JK, Madeira SC, Moreau Y. 2011. PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Res* **39**: W334-8.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125740&tool=pmcentrez&rendertype=abstract> (Accessed May 5, 2015).

Nyrén P, Lundin A. 1985. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem* **151**: 504–9.

<http://www.ncbi.nlm.nih.gov/pubmed/3006540> (Accessed May 11, 2017).

Parkin DM, Sitas F, Chirenje M, Stein L, Abratt R, Wabinga H. 2008. Part I: Cancer in Indigenous Africans—burden, distribution, and trends. *Lancet Oncol* **9**: 683–692.

<http://www.ncbi.nlm.nih.gov/pubmed/18598933> (Accessed August 19, 2017).

Pavan S, Rommel K, Mateo Marquina ME, Höhn S, Lanneau V, Rath A. 2017. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLoS One* **12**: e0170365.

<http://www.ncbi.nlm.nih.gov/pubmed/28099516> (Accessed August 18, 2017).

Pelleg D, Moore A. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *ICML*. <http://cs.uef.fi/~zhao/Courses/Clustering2012/Xmeans.pdf> (Accessed August 3, 2016).

PELLICER J, FAY MF, LEITCH IJ. 2010. The largest eukaryotic genome of them all? *Bot J Linn Soc* **164**: 10–15. <http://doi.wiley.com/10.1111/j.1095-8339.2010.01072.x> (Accessed May 26, 2016).

- Pers TH, Dworzynski P, Thomas CE, Lage K, Brunak S. 2013. MetaRanker 2.0: a web server for prioritization of genetic variation data. *Nucleic Acids Res* **41**: W104-8.  
<http://www.ncbi.nlm.nih.gov/pubmed/23703204> (Accessed August 18, 2017).
- Persson H, Søkilde R, Pirona AC, Rovira C. 2017. Preparation of highly multiplexed small RNA sequencing libraries. *Biotechniques* **63**: 57–64.  
<http://www.ncbi.nlm.nih.gov/pubmed/28803540> (Accessed August 16, 2017).
- Pesquita C, Faria D, Bastos H, Ferreira AEN, Falcão AO, Couto FM. 2008. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* **9 Suppl 5**: S4.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2367622&tool=pmcentrez&rendertype=abstract> (Accessed May 5, 2015).
- Pesquita C, Pessoa D, Faria D, Couto FM. 2009. CESSM: collaborative evaluation of semantic similarity measures. In *JB2009: Challenges in \dots*.
- Petri V, Jayaraman P, Tutaj M, Hayman GT, Smith JR, De Pons J, Laudederkind SJ, Lowry TF, Nigam R, Wang S-J, et al. 2014. The pathway ontology - updates and applications. *J Biomed Semantics* **5**: 7.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3922094&tool=pmcentrez&rendertype=abstract> (Accessed May 5, 2015).
- Petryk A, Graf D, Marcucio R. 2015. Holoprosencephaly: signaling interactions between the brain and the face, the environment and the genes, and the phenotypic variability in animal models and humans. *Wiley Interdiscip Rev Dev Biol* **4**: 17–32.  
<http://www.ncbi.nlm.nih.gov/pubmed/25339593> (Accessed August 18, 2017).
- Ping X-L, Sun B-F, Wang L, Xiao W, Yang X, Wang W-J, Adhikari S, Shi Y, Lv Y, Chen Y-S, et al. 2014. Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Res* **24**: 177–89.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3915904&tool=pmcentrez&rendertype=abstract> (Accessed January 27, 2016).
- Project Consortium G, author C, committee S, group P, College of Medicine B, Institute of MIT B, Technologies L, Planck Institute for Molecular Genetics M, Applied Science R, University in St Louis W, et al. 2011. A map of human genome variation from population-scale sequencing. *Nature* **467**.

<https://www.nature.com/nature/journal/v467/n7319/pdf/nature09534.pdf> (Accessed August 17, 2017).

Project Consortium G, Consortium Participants are arranged by project role G, by institution alphabetically then, alphabetically within institutions except for Principal Investigators finally, Leaders P, indicated as, author C, committee S, group P, College of Medicine B, et al. 2012. An integrated map of genetic variation from 1,092 human genomes.

<https://www.nature.com/nature/journal/v491/n7422/pdf/nature11632.pdf> (Accessed May 11, 2017).

Przyborowski J, Wilenski H. 1940. Homogeneity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder. *Biometrika* **31**: 313–323. <http://www.jstor.org/stable/2332612>.

Qu S, Yang X, Li X, Wang J, Gao Y, Shang R, Sun W, Dou K, Li H. 2015. Circular RNA: a new star of noncoding RNAs. *Cancer Lett* **365**: 141–8. <http://www.ncbi.nlm.nih.gov/pubmed/26052092> (Accessed June 9, 2015).

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–2. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832824&tool=pmcentrez&rendertype=abstract> (Accessed July 9, 2014).

Rabbani B, Tekin M, Mahdieh N. 2014. The promise of whole-exome sequencing in medical genetics. *J Hum Genet* **59**: 5–15. <http://www.ncbi.nlm.nih.gov/pubmed/24196381> (Accessed July 29, 2017).

Rajagopalan LE, Westmark CJ, Jarzembowski JA, Malter JS. 1998. hnRNP C increases amyloid precursor protein (APP) production by stabilizing APP mRNA. *Nucleic Acids Res* **26**: 3418–3423. <http://nar.oxfordjournals.org/content/26/14/3418> (Accessed May 10, 2016).

Ramasamy A, Mondry A, Holmes CC, Altman DG. 2008. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* **5**: e184. <http://www.ncbi.nlm.nih.gov/pubmed/18767902> (Accessed August 18, 2017).

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbelt JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3436805&tool=pmcentrez&>

rendertype=abstract (Accessed April 4, 2016).

Rawlings JS, Rosler KM, Harrison DA. 2004. The JAK/STAT signaling pathway. *J Cell Sci* **117**: 1281–1283. <http://www.ncbi.nlm.nih.gov/pubmed/15020666> (Accessed August 19, 2017).

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-Wide Location and Function of DNA Binding Proteins. *Science (80- )* **290**: 2306–2309. <http://www.ncbi.nlm.nih.gov/pubmed/11125145> (Accessed August 16, 2017).

Resnik P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. <https://arxiv.org/pdf/cmp-lg/9511007.pdf> (Accessed August 19, 2017).

Rivera M, Cohen-Woods S, Kapur K, Breen G, Ng MY, Butler AW, Craddock N, Gill M, Korszun A, Maier W, et al. 2012. Depressive disorder moderates the effect of the FTO gene on body mass index. *Mol Psychiatry* **17**: 604–11. <http://dx.doi.org/10.1038/mp.2011.45> (Accessed May 12, 2016).

Rivière J-B, Mirzaa GM, O’Roak BJ, Beddaoui M, Alcantara D, Conway RL, St-Onge J, Schwartztruber JA, Gripp KW, Nikkel SM, et al. 2012. De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nat Genet* **44**: 934–940. <http://www.ncbi.nlm.nih.gov/pubmed/22729224> (Accessed August 17, 2017).

Robey RC, Mletzko S, Gotch FM. 2010. The T-Cell Immune Response against Kaposi’s Sarcoma-Associated Herpesvirus. *Adv Virol* **2010**: 340356. <http://www.ncbi.nlm.nih.gov/pubmed/22331985> (Accessed August 19, 2017).

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–6. <http://www.ncbi.nlm.nih.gov/pubmed/21221095> (Accessed August 19, 2017).

Robinson MD, Oshlack A, Wang E, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore S, Schroth G, et al. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25. <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25> (Accessed July 28, 2016).

- Robinson PN, Köhler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D, Krawitz P, et al. 2014. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* **24**: 340–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3912424&tool=pmcentrez&rendertype=abstract> (Accessed February 18, 2015).
- Roost C, Lynch SR, Batista PJ, Qu K, Chang HY, Kool ET. 2015. Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification. *J Am Chem Soc* **137**: 2107–15. <http://dx.doi.org/10.1021/ja513080v> (Accessed February 9, 2016).
- Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348–352. <http://www.ncbi.nlm.nih.gov/pubmed/21776081> (Accessed May 11, 2017).
- Salzberg S, Delcher A, Kasif S. 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* <http://nar.oxfordjournals.org/content/26/2/544.short> (Accessed August 4, 2016).
- Salzberg S, Pertea M, Delcher A, Gardner M. 1999. Interpolated Markov models for eukaryotic gene finding. *Genomics*. <http://www.sciencedirect.com/science/article/pii/S0888754399958548> (Accessed August 4, 2016).
- Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. 2012. Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types ed. T. Preiss. *PLoS One* **7**: e30733. <http://www.ncbi.nlm.nih.gov/pubmed/22319583> (Accessed August 16, 2017).
- Saneyoshi M, Harada F, Nishimura S. 1969. Isolation and characterization of N6-methyladenosine from Escherichia coli valine transfer RNA. *Biochim Biophys Acta - Nucleic Acids Protein Synth* **190**: 264–273. <http://www.sciencedirect.com/science/article/pii/0005278769900781> (Accessed May 5, 2016).
- Sanger F. 1949. The terminal peptides of insulin. *Biochem J.* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1275055/> (Accessed February 14, 2017).

- Sanger F, Brownlee GG, Barrell BG. 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *J Mol Biol* **13**: 373–98.  
<http://www.ncbi.nlm.nih.gov/pubmed/5325727> (Accessed May 11, 2017).
- Sanger F, Coulson A, Friedmann T, Air G. 1978. The nucleotide sequence of bacteriophage  $\phi$ X174. *J Mol Biol*. <http://www.sciencedirect.com/science/article/pii/0022283678903467> (Accessed February 13, 2017).
- Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**: 441–8.  
<http://www.ncbi.nlm.nih.gov/pubmed/1100841> (Accessed February 13, 2017).
- Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. 1982. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162**: 729–73.  
<http://www.ncbi.nlm.nih.gov/pubmed/6221115> (Accessed May 11, 2017).
- Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463–7. <http://www.ncbi.nlm.nih.gov/pubmed/271968> (Accessed February 13, 2017).
- Sanger F, Tuppy H. 1951. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem J*.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1197536/> (Accessed February 14, 2017).
- Santhekadur PK, Das SK, Gredler R, Chen D, Srivastava J, Robertson C, Baldwin AS, Fisher PB, Sarkar D. 2012. Multifunction Protein Staphylococcal Nuclease Domain Containing 1 (SND1) Promotes Tumor Angiogenesis in Human Hepatocellular Carcinoma through Novel Pathway That Involves Nuclear Factor  $\kappa$ B and miR-221. *J Biol Chem* **287**: 13952–13958.  
<http://www.ncbi.nlm.nih.gov/pubmed/22396537> (Accessed August 19, 2017).
- Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, et al. 2012. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **40**: D13–D25.  
<http://www.ncbi.nlm.nih.gov/pubmed/22140104> (Accessed August 18, 2017).
- Scadden ADJ. 2005. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. *Nat Struct Mol Biol* **12**: 489–496.  
<http://www.ncbi.nlm.nih.gov/pubmed/15895094> (Accessed August 19, 2017).

- Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Hum Mol Genet* **19**: R227–R240. <http://www.ncbi.nlm.nih.gov/pubmed/20858600> (Accessed May 11, 2017).
- Schaefer M, Pollex T, Hanna K, Lyko F. 2009. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res* **37**: e12. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2632927&tool=pmcentrez&rendertype=abstract> (Accessed May 4, 2016).
- Schalling M, Ekman M, Kaaya EE, Linde A, Biberfeld P. 1995. A role for a new herpes virus (KSHV) in different forms of Kaposi's sarcoma. *Nat Med* **1**: 707–8. <http://www.ncbi.nlm.nih.gov/pubmed/7585156> (Accessed August 19, 2017).
- Schloss JA. 2008. How to get genomes at one ten-thousandth the cost. *Nat Biotechnol* **26**: 1113–1115. <http://www.nature.com/doifinder/10.1038/nbt1008-1113> (Accessed April 7, 2017).
- Schwartz S, Agarwala SD, Mumbach MR, Jovanovic M, Mertins P, Shishkin A, Tabach Y, Mikkelsen TS, Satija R, Ruvkun G, et al. 2013. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* **155**: 1409–21. <http://www.sciencedirect.com/science/article/pii/S0092867413013652> (Accessed December 31, 2015).
- Schwartz S, Mumbach MR, Jovanovic M, Wang T, Maciag K, Bushkin GG, Mertins P, Ter-Ovanesyan D, Habib N, Cacchiarelli D, et al. 2014. Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. *Cell Rep* **8**: 284–96. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4142486&tool=pmcentrez&rendertype=abstract> (Accessed September 22, 2015).
- Schwarz G. 1978. Estimating the Dimension of a Model. *Ann Stat* **6**: 461–464. <http://projecteuclid.org/euclid.aos/1176344136> (Accessed August 3, 2016).
- Scuteri A, Sanna S, Chen W-M, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orrú M, Usala G, et al. 2007. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* **3**: e115. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1934391&tool=pmcentrez&rendertype=abstract> (Accessed May 11, 2016).

- Seelow D, Schwarz JM, Schuelke M. 2008. GeneDistiller--distilling candidate genes from linkage intervals. *PLoS One* **3**: e3874.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2587712&tool=pmcentrez&rendertype=abstract> (Accessed May 5, 2015).
- Seifert M, Cortijo S, Colomé-Tatché M, Johannes F, Roudier F, Colot V. 2012. MeDIP-HMM: genome-wide identification of distinct DNA methylation states from high-density tiling arrays. *Bioinformatics* **28**: 2930–9. <http://www.ncbi.nlm.nih.gov/pubmed/22989518> (Accessed August 4, 2016).
- Sergiev P V, Serebryakova M V, Bogdanov AA, Dontsova OA. 2008. The ybiN gene of *Escherichia coli* encodes adenine-N6 methyltransferase specific for modification of A1618 of 23 S ribosomal RNA, a methylated residue located close to the ribosomal exit tunnel. *J Mol Biol* **375**: 291–300.  
<http://www.sciencedirect.com/science/article/pii/S0022283607013873> (Accessed May 9, 2016).
- Sevgi M, Rigoux L, Kühn AB, Mauer J, Schilbach L, Hess ME, Gruendler TOJ, Ullsperger M, Stephan KE, Brüning JC, et al. 2015. An Obesity-Predisposing Variant of the FTO Gene Regulates D2R-Dependent Reward Learning. *J Neurosci* **35**: 12584–92.  
<http://www.jneurosci.org/content/35/36/12584.long> (Accessed May 12, 2016).
- Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. 2014. Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis. *Biomed Res Int* **2014**: 1–16. <http://www.ncbi.nlm.nih.gov/pubmed/24779008> (Accessed August 17, 2017).
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–504.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=403769&tool=pmcentrez&rendertype=abstract> (Accessed July 9, 2014).
- Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science (80- )* **309**: 1728–1732.  
<http://www.ncbi.nlm.nih.gov/pubmed/16081699> (Accessed May 11, 2017).

- Si H, Robertson ES. 2006. Kaposi's Sarcoma-Associated Herpesvirus-Encoded Latency-Associated Nuclear Antigen Induces Chromosomal Instability through Inhibition of p53 Function. *J Virol* **80**: 697–709. <http://www.ncbi.nlm.nih.gov/pubmed/16378973> (Accessed August 19, 2017).
- Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCr: visualizing classifier performance in R. *Bioinformatics* **21**: 3940–1. <http://bioinformatics.oxfordjournals.org/content/21/20/3940.full> (Accessed June 29, 2015).
- Singh OP. 2001. Functional diversity of hnRNP proteins. *Indian J Biochem Biophys* **38**: 129–34. <http://www.ncbi.nlm.nih.gov/pubmed/11693373> (Accessed May 10, 2016).
- Skinner R, Cundliffe E, Schmidt FJ. 1983. Site of action of a ribosomal RNA methylase responsible for resistance to erythromycin and other antibiotics. *J Biol Chem* **258**: 12702–6. <http://www.ncbi.nlm.nih.gov/pubmed/6195156> (Accessed May 13, 2016).
- Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G, et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* **43**: W589–98. <http://www.ncbi.nlm.nih.gov/pubmed/25897122> (Accessed August 18, 2017).
- Smedley D, Köhler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A, Veldboer J, Zemojtel T, Robinson PN. 2014. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics* **30**: 3215–22. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4221119&tool=pmcentrez&rendertype=abstract> (Accessed April 3, 2015).
- Smedley D, Oellrich A, Köhler S, Ruef B, Westerfield M, Robinson P, Lewis S, Mungall C. 2013. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database (Oxford)* **2013**: bat025. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3649640&tool=pmcentrez&rendertype=abstract> (Accessed February 27, 2015).
- Smedley D, Robinson PN. 2015. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med* **7**: 81. <http://www.ncbi.nlm.nih.gov/pubmed/26229552> (Accessed February 3, 2017).

- Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* **25**: 1251–5. <http://dx.doi.org/10.1038/nbt1346> (Accessed November 30, 2014).
- Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**: 674–679. <http://www.ncbi.nlm.nih.gov/pubmed/3713851> (Accessed May 11, 2017).
- Smyth GK. 2005. limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 397–420, Springer-Verlag, New York [http://link.springer.com/10.1007/0-387-29362-0\\_23](http://link.springer.com/10.1007/0-387-29362-0_23) (Accessed October 12, 2017).
- Sontheimer EJ. 2005. Assembly and function of RNA silencing complexes. *Nat Rev Mol Cell Biol* **6**: 127–138. <http://www.ncbi.nlm.nih.gov/pubmed/15654322> (Accessed August 19, 2017).
- Soulier J, Grollet L, Oksenhendler E, Cacoub P, Cazals-Hatem D, Babinet P, d'Agay MF, Clauvel JP, Raphael M, Degos L. 1995. Kaposi's sarcoma-associated herpesvirus-like DNA sequences in multicentric Castleman's disease. *Blood* **86**: 1276–80. <http://www.ncbi.nlm.nih.gov/pubmed/7632932> (Accessed August 19, 2017).
- Sparmann A. 2015. m6A drives structural changes in RNA. *Nat Struct Mol Biol* **22**: 184–184. <http://dx.doi.org/10.1038/nsmb.2987> (Accessed May 10, 2016).
- Spitzer J, Hafner M, Landthaler M, Ascano M, Farazi T, Wardle G, Nusbaum J, Khorshid M, Burger L, Zavolan M, et al. 2014. PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation). In *Methods in enzymology*, Vol. 539 of, pp. 113–161 <http://www.ncbi.nlm.nih.gov/pubmed/24581442> (Accessed August 16, 2017).
- Srour M, Schwartzenruber J, Hamdan FF, Ospina LH, Patry L, Labuda D, Massicotte C, Dobrzeniecka S, Capo-Chichi J-M, Papillon-Cavanagh S, et al. 2012. Mutations in C5ORF42 cause Joubert syndrome in the French Canadian population. *Am J Hum Genet* **90**: 693–700. <http://linkinghub.elsevier.com/retrieve/pii/S0002929712000961> (Accessed August 17, 2017).
- Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I,

- Lapp H, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**: 1611–8. <http://www.ncbi.nlm.nih.gov/pubmed/12368254> (Accessed August 17, 2017).
- Stebbing J, Bourboulia D, Johnson M, Henderson S, Williams I, Wilder N, Tyrer M, Youle M, Imami N, Kobu T, et al. 2003. Kaposi's sarcoma-associated herpesvirus cytotoxic T lymphocytes recognize and target Darwinian positively selected autologous K1 epitopes. *J Virol* **77**: 4306–14. <http://www.ncbi.nlm.nih.gov/pubmed/12634388> (Accessed August 19, 2017).
- Steele R, Raftery A. 2010. Performance of Bayesian model selection criteria for Gaussian mixture models. *Front Stat Decis Mak*.  
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.436.1669&rep=rep1&type=pdf> (Accessed August 3, 2016).
- Stein LD. 2010. The case for cloud computing in genome informatics. *Genome Biol* **11**: 207. <http://www.ncbi.nlm.nih.gov/pubmed/20441614> (Accessed May 11, 2017).
- Steinman RA, Yang Q, Gasparetto M, Robinson LJ, Liu X, Lenzner DE, Hou J, Smith C, Wang Q. 2013. Deletion of the RNA-editing enzyme ADAR1 causes regression of established chronic myelogenous leukemia in mice. *Int J cancer* **132**: 1741–50. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3565035&tool=pmcentrez&rendertype=abstract> (Accessed May 4, 2016).
- Strittmatter WJ, Roses AD. 1996. Apolipoprotein E and Alzheimer's Disease. *Annu Rev Neurosci* **19**: 53–77. <http://www.ncbi.nlm.nih.gov/pubmed/8833436> (Accessed August 17, 2017).
- Sullivan PF. 2007. Spurious genetic associations. *Biol Psychiatry* **61**: 1121–6. <http://www.sciencedirect.com/science/article/pii/S0006322306014703> (Accessed May 12, 2016).
- Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M, Blurb E, To C. 2016. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia HHS Public Access. *Neuron March* **2**: 940–947. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4793939/pdf/nihms764213.pdf> (Accessed August 18, 2017).
- Theler D, Dominguez C, Blatter M, Boudet J, Allain FH-T. 2014. Solution structure of the YTH domain in complex with N6-methyladenosine RNA: a reader of methylated RNA. *Nucleic*

*Acids Res* **42**: 13911–9.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4267619&tool=pmcentrez&rendertype=abstract> (Accessed May 10, 2016).

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–11.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2672628&tool=pmcentrez&rendertype=abstract> (Accessed July 10, 2014).

Tripsianes K, Madl T, Machyna M, Fessas D, Englbrecht C, Fischer U, Neugebauer KM, Sattler M. 2011. Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins. *Nat Struct Mol Biol* **18**: 1414–1420.

<http://www.ncbi.nlm.nih.gov/pubmed/22101937> (Accessed August 19, 2017).

Trujillano D, Bertoli-Avella AM, Kumar Kandaswamy K, Weiss ME, Köster J, Marais A, Paknia O, Schröder R, Garcia-Aznar JM, Werber M, et al. 2017. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur J Hum Genet* **25**: 176–182.

<http://www.ncbi.nlm.nih.gov/pubmed/27848944> (Accessed July 29, 2017).

Tuck MT, James CBL, Kelder B, Kopchick JJ. 1996. Elevation of internal 6-methyladenine mRNA methyltransferase activity after cellular transformation. *Cancer Lett* **103**: 107–113.

<http://www.sciencedirect.com/science/article/pii/0304383596042036> (Accessed May 5, 2016).

Turcatti G, Romieu A, Fedurco M, Tairi A-P. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis †. *Nucleic Acids Res* **36**: e25–e25.

<http://www.ncbi.nlm.nih.gov/pubmed/18263613> (Accessed May 11, 2017).

Turner FS, Clutterbuck DR, Semple CAM. 2003. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* **4**: R75.

<http://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-4-11-r75> (Accessed August 18, 2017).

Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, Raja A, Baker C, Hoekzema K, Stessman HA, et al. 2016. Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am J Hum Genet* **98**: 58–74.

<http://www.ncbi.nlm.nih.gov/pubmed/26749308> (Accessed October 10, 2017).

- Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on CHIP-Seq data. *Nat Methods* **5**: 829–34. <http://dx.doi.org/10.1038/nmeth.1246> (Accessed May 19, 2016).
- van Dam S, Cordeiro R, Craig T, van Dam J, Wood SH, de Magalhães JP. 2012. GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics* **13**: 535. <http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-535> (Accessed August 18, 2017).
- van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. 2006. A text-mining analysis of the human phenome. *Eur J Hum Genet* **14**: 535–542. <http://www.ncbi.nlm.nih.gov/pubmed/16493445> (Accessed October 12, 2017).
- Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**: 508–514. <http://www.ncbi.nlm.nih.gov/pubmed/27018577> (Accessed August 16, 2017).
- Vasan SK, Karpe F, Gu HF, Brismar K, Fall CH, Ingelsson E, Fall T. 2014. FTO genetic variants and risk of obesity and type 2 diabetes: a meta-analysis of 28,394 Indians. *Obesity (Silver Spring)* **22**: 964–70. <http://www.ncbi.nlm.nih.gov/pubmed/23963770> (Accessed May 4, 2016).
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The Sequence of the Human Genome. *Science (80- )* **291**: 1304–1351. <http://www.ncbi.nlm.nih.gov/pubmed/11181995> (Accessed April 1, 2017).
- Vespa L, Vachon G, Berger F, Perazza D, Faure J-D, Herzog M. 2004. The immunophilin-interacting protein AtFIP37 from Arabidopsis is essential for plant development and is involved in trichome endoreduplication. *Plant Physiol* **134**: 1283–92. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=419804&tool=pmcentrez&endertype=abstract> (Accessed May 12, 2016).
- Villarroya-Beltri C, Gutiérrez-Vázquez C, Sánchez-Cabo F, Pérez-Hernández D, Vázquez J,

- Martin-Cofreces N, Martinez-Herrera DJ, Pascual-Montano A, Mittelbrunn M, Sánchez-Madrid F. 2013. Sumoylated hnRNPA2B1 controls the sorting of miRNAs into exosomes through binding to specific motifs. *Nat Commun* **4**: 2980.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3905700&tool=pmcentrez&rendertype=abstract> (Accessed March 12, 2016).
- Vissers LELM, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, et al. 2010. A de novo paradigm for mental retardation. *Nat Genet* **42**: 1109–1112. <http://www.ncbi.nlm.nih.gov/pubmed/21076407> (Accessed July 29, 2017).
- Vlasblom J, Wu S, Pu S, Superina M, Liu G, Orsi C, Wodak SJ. 2006. GenePro: a cytoscape plugin for advanced visualization and analysis of interaction networks. *Bioinformatics* **22**: 2178–2179. <http://www.ncbi.nlm.nih.gov/pubmed/16921162> (Accessed August 18, 2017).
- Wabinga HR, Parkin DM, Wabwire-Mangen F, Mugerwa JW. 1993. Cancer in Kampala, Uganda, in 1989-91: changes in incidence in the era of AIDS. *Int J cancer* **54**: 26–36.  
<http://www.ncbi.nlm.nih.gov/pubmed/8478145> (Accessed August 19, 2017).
- Wagner L, Agarwala R. 2013. UniGene. <https://www.ncbi.nlm.nih.gov/books/NBK169437/> (Accessed August 18, 2017).
- Wan Y, Qu K, Ouyang Z, Chang HY. 2013. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nat Protoc* **8**: 849–869.  
<http://www.ncbi.nlm.nih.gov/pubmed/23558785> (Accessed August 16, 2017).
- Wang C, Zhu Y, Bao H, Jiang Y, Xu C, Wu J, Shi Y. 2016. A novel RNA-binding mode of the YTH domain reveals the mechanism for recognition of determinant of selective removal by Mmi1. *Nucleic Acids Res* **44**: 969–82.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4737157&tool=pmcentrez&rendertype=abstract> (Accessed May 13, 2016).
- Wang L, Liu X, Luo X, Zeng M, Zuo L, Wang K-S. 2013. Genetic Variants in the Fat Mass- and Obesity-Associated (FTO) Gene are Associated with Alcohol Dependence. *J Mol Neurosci* **51**: 416–424. <http://www.ncbi.nlm.nih.gov/pubmed/23771786> (Accessed November 21, 2017).

- Wang X, Li X, Cheng Y, Sun X, Sun X, Self S, Kooperberg C, Dai JY. 2015a. Copy number alterations detected by whole-exome and whole-genome sequencing of esophageal adenocarcinoma. *Hum Genomics* **9**: 22. <http://www.ncbi.nlm.nih.gov/pubmed/26374103> (Accessed July 29, 2017).
- Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, et al. 2014a. N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**: 117–20. <http://dx.doi.org/10.1038/nature12730> (Accessed July 10, 2014).
- Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, Weng X, Chen K, Shi H, He C. 2015b. N6-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* **161**: 1388–1399. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4825696&tool=pmcentrez&rendertype=abstract> (Accessed June 4, 2015).
- Wang X, Zhu L, Chen J, Wang Y. 2015c. mRNA m<sup>6</sup>A methylation downregulates adipogenesis in porcine adipocytes. *Biochem Biophys Res Commun* **459**: 201–7. <http://www.ncbi.nlm.nih.gov/pubmed/25725156> (Accessed May 11, 2016).
- Wang Y, Li Y, Toth JI, Petroski MD, Zhang Z, Zhao JC. 2014b. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat Cell Biol* **16**: 191–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4640932&tool=pmcentrez&rendertype=abstract> (Accessed May 11, 2016).
- Wang Y, Zhang NL. Severity of Local Maxima for the EM Algorithm: Experiences with Hierarchical Latent Class Models.
- Waterman MS. 1984. Efficient sequence alignment algorithms. *J Theor Biol* **108**: 333–7. <http://www.ncbi.nlm.nih.gov/pubmed/6748696> (Accessed July 30, 2017).
- WATSON JD, CRICK FH. 1953a. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**: 964–7. <http://www.ncbi.nlm.nih.gov/pubmed/13063483> (Accessed February 14, 2017).
- WATSON JD, CRICK FH. 1953b. The structure of DNA. *Cold Spring Harb Symp Quant Biol* **18**: 123–31. <http://www.ncbi.nlm.nih.gov/pubmed/13168976> (Accessed February 14, 2017).
- Wei CM, Gershowitz A, Moss B. 1975. Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. *Cell* **4**: 379–86. <http://www.ncbi.nlm.nih.gov/pubmed/164293> (Accessed

April 23, 2016).

- Weissbach R, Scadden ADJ. 2012. Tudor-SN and ADAR1 are components of cytoplasmic stress granules. *RNA* **18**: 462–471. <http://www.ncbi.nlm.nih.gov/pubmed/22240577> (Accessed August 19, 2017).
- Welsh EA, Eschrich SA, Berglund AE, Fenstermacher DA. 2013. Iterative rank-order normalization of gene expression microarray data. *BMC Bioinformatics* **14**: 153. <http://www.ncbi.nlm.nih.gov/pubmed/23647742>.
- Wessels H-H, Hirsekorn A, Ohler U, Mukherjee N. 2016. Identifying RBP Targets with RIP-seq. pp. 141–152 [http://link.springer.com/10.1007/978-1-4939-3067-8\\_9](http://link.springer.com/10.1007/978-1-4939-3067-8_9) (Accessed August 16, 2017).
- Will CL, Lührmann R. 2001. Spliceosomal UsnRNP biogenesis, structure and function. *Curr Opin Cell Biol* **13**: 290–301. <http://www.ncbi.nlm.nih.gov/pubmed/11343899> (Accessed August 19, 2017).
- Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, Taiwo O, Beck S, Butcher LM. 2012. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* **7**. <https://www.nature.com/nprot/journal/v7/n4/pdf/nprot.2012.012.pdf> (Accessed August 16, 2017).
- Xiao W, Adhikari S, Dahal U, Chen Y-S, Hao Y-J, Sun B-F, Sun H-Y, Li A, Ping X-L, Lai W-Y, et al. 2016. Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing. *Mol Cell* **61**: 507–19. <http://www.ncbi.nlm.nih.gov/pubmed/26876937> (Accessed February 16, 2016).
- Yan C, Yan Z, Wang Y, Yan X, Han Y. 2014. Tudor-SN, a component of stress granules, regulates growth under salt stress by modulating GA20ox3 mRNA levels in Arabidopsis. *J Exp Bot* **65**: 5933–5944. <http://www.ncbi.nlm.nih.gov/pubmed/25205572> (Accessed August 19, 2017).
- Yang H, Robinson PN, Wang K. 2015. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods* **12**: 841–843. <http://www.nature.com/doifinder/10.1038/nmeth.3484> (Accessed October 12, 2017).
- Yang J, Välineva T, Hong J, Bu T, Yao Z, Jensen ON, Frilander MJ, Silvennoinen O. 2007. Transcriptional co-activator protein p100 interacts with snRNP proteins and facilitates the

assembly of the spliceosome. *Nucleic Acids Res* **35**: 4485–4494.  
<http://www.ncbi.nlm.nih.gov/pubmed/17576664> (Accessed August 19, 2017).

Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, Nishikura K. 2006. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol* **13**: 13–21.  
<http://www.ncbi.nlm.nih.gov/pubmed/16369484> (Accessed August 19, 2017).

Yang Y, Huang W, Huang J-T, Shen F, Xiong J, Yuan E-F, Qin S-S, Zhang M, Feng Y-Q, Yuan B-F, et al. 2016. Increased N(6)-methyladenosine in Human Sperm RNA as a Risk Factor for Asthenozoospermia. *Sci Rep* **6**: 24345.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4829835&tool=pmcentrez&rendertype=abstract> (Accessed April 22, 2016).

Ye F, Chen ER, Nilsen TW. 2017. Kaposi's Sarcoma-Associated Herpesvirus Utilizes and Manipulates RNA N(6)-Adenosine Methylation To Promote Lytic Replication. *J Virol* **91**: JVI.00466-17. <http://www.ncbi.nlm.nih.gov/pubmed/28592530> (Accessed August 19, 2017).

Yeap W-C, Namasivayam P, Ho C-L. 2014. HnRNP-like proteins as post-transcriptional regulators. *Plant Sci* **227**: 90–100.  
<http://www.sciencedirect.com/science/article/pii/S0168945214001654> (Accessed May 10, 2016).

Yu F, Harada JN, Brown HJ, Deng H, Song MJ, Wu T-T, Kato-Stankiewicz J, Nelson CG, Vieira J, Tamanoi F, et al. 2007. Systematic Identification of Cellular Signals Reactivating Kaposi Sarcoma-Associated Herpesvirus. *PLoS Pathog* **3**: e44.  
<http://www.ncbi.nlm.nih.gov/pubmed/17397260> (Accessed August 19, 2017).

Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**: 284–7.  
<http://www.ncbi.nlm.nih.gov/pubmed/22455463> (Accessed August 18, 2017).

Yu G, Wang L-G, Yan G-R, He Q-Y. 2015. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**: 608–609.  
<http://www.ncbi.nlm.nih.gov/pubmed/25677125> (Accessed August 18, 2017).

Zhang C, Samanta D, Lu H, Bullen JW, Zhang H, Chen I, He X, Semenza GL. 2016. Hypoxia

induces the breast cancer stem cell phenotype by HIF-dependent and ALKBH5-mediated m6A-demethylation of NANOG mRNA. *Proc Natl Acad Sci U S A*.

<http://www.ncbi.nlm.nih.gov/pubmed/27001847> (Accessed March 28, 2016).

Zhang L, Zhang J, Yang J, Ying D, Lau YL, Yang W. 2013. PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data. *Bioinformatics* **29**: 124–5.

<http://www.ncbi.nlm.nih.gov/pubmed/23104884> (Accessed May 5, 2015).

Zhang M, Zhang Y, Ma J, Guo F, Cao Q, Zhang Y, Zhou B, Chai J, Zhao W, Zhao R. 2015. The Demethylase Activity of FTO (Fat Mass and Obesity Associated Protein) Is Required for Preadipocyte Differentiation. *PLoS One* **10**: e0133788.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4517749&tool=pmcentrez&rendertype=abstract> (Accessed May 11, 2016).

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of CHIP-Seq (MACS). *Genome Biol* **9**: R137.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2592715&tool=pmcentrez&rendertype=abstract> (Accessed July 11, 2014).

Zhang Z, Theler D, Kaminska KH, Hiller M, de la Grange P, Pudimat R, Rafalska I, Heinrich B, Bujnicki JM, Allain FH-T, et al. 2010. The YTH domain is a novel RNA binding domain. *J Biol Chem* **285**: 14701–10.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2863249&tool=pmcentrez&rendertype=abstract> (Accessed March 23, 2016).

Zhao X, Yang YY-G, Sun B-F, Shi Y, Yang X, Xiao W, Hao Y-J, Ping X-L, Chen Y-S, Wang W-J, et al. 2014. FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Res* **24**: 1403–19.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4260349&tool=pmcentrez&rendertype=abstract> (Accessed July 14, 2015).

Zheng G, Dahl JA, Niu Y, Fedorcsak P, Huang C-M, Li CJ, Vågbø CB, Shi Y, Wang W-L, Song S-H, et al. 2013. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol Cell* **49**: 18–29.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3646334&tool=pmcentrez&rendertype=abstract> (Accessed October 2, 2015).

- Zhi H, Zahoor MA, Shudofsky AMD, Giam C-Z. 2015. KSHV vCyclin counters the senescence/G1 arrest response triggered by NF- $\kappa$ B hyperactivation. *Oncogene* **34**: 496–505.  
<http://www.ncbi.nlm.nih.gov/pubmed/24469036> (Accessed August 19, 2017).
- Zhong S, Li H, Bodi Z, Button J, Vespa L, Herzog M, Fray RG. 2008. MTA is an Arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell* **20**: 1278–88.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2438467&tool=pmcentrez&rendertype=abstract> (Accessed May 5, 2016).
- Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian S-B. 2015. Dynamic m(6)A mRNA methylation directs translational control of heat shock response. *Nature* **526**: 591–594.  
<http://dx.doi.org/10.1038/nature15377> (Accessed October 14, 2015).
- Zhou KI, Parisien M, Dai Q, Liu N, Diatchenko L, Sachleben JR, Pan T. 2016a. N(6)-Methyladenosine Modification in a Long Noncoding RNA Hairpin Predisposes Its Conformation to Protein Binding. *J Mol Biol* **428**: 822–33.  
<http://www.ncbi.nlm.nih.gov/pubmed/26343757> (Accessed April 15, 2016).
- Zhou W, Chen T, Zhao H, Eterovic AK, Meric-Bernstam F, Mills GB, Chen K. 2014. Bias from removing read duplication in ultra-deep sequencing experiments. *Bioinformatics*.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3982159&tool=pmcentrez&rendertype=abstract> (Accessed May 25, 2016).
- Zhou Y, Zeng P, Li Y-H, Zhang Z, Cui Q. 2016b. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res* gkw104-. <http://nar.oxfordjournals.org/content/early/2016/02/19/nar.gkw104.long> (Accessed March 8, 2016).
- Zhu L, Tatsuke T, Mon H, Li Z, Xu J, Lee JM, Kusakabe T. 2013. Characterization of Tudor-sn-containing granules in the silkworm, *Bombyx mori*. *Insect Biochem Mol Biol* **43**: 664–674.  
<http://www.ncbi.nlm.nih.gov/pubmed/23643815> (Accessed August 19, 2017).
- Zhu T, Roundtree IA, Wang P, Wang X, Wang L, Sun C, Tian Y, Li J, He C, Xu Y. 2014. Crystal structure of the YTH domain of YTHDF2 reveals mechanism for recognition of N6-methyladenosine. *Cell Res* **24**: 1493–6.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4260350&tool=pmcentrez&>

rendertype=abstract (Accessed May 10, 2016).