A systems biology approach to understanding Autosomal Dominant Polycystic Kidney Disease

A thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy

Matei Stefan Neagu

Department of Automatic Control and Systems Engineering

April 2018

Mamei mele, Daniela

Acknowledgements

First, I would like to thank my supervisors, Professor Daniel Coca and Professor Albert Ong for their constant guidance they have provided during the past four years. Next, I would like to thank Dr. Andrew Streets for turning me from someone who never looked at a cell through a microscope to designing my own experiments. I would also like to thank Dr. Paul J. Gokhale for having patience in helping me with my microscopy work.

Next, I would like to thank all the people who have helped me with my PhD work and also have become my friends during the PhD. I want to send many thanks to Dr. Adrian Alecu who has been a great mentor in my first year of the PhD when I was a disoriented student and he put me on the right path. I would also like to thank Dr. Veronica Biga for introducing me to diffusion and conveying some of her contagious enthusiasm about research. Last but not least I would like to thank (soon to be Dr.) Laura Vergoz for answering all my questions about biological work.

I would like so send some special thanks to my parents, my mother without whose constant support(which sometimes took the form of a push) I would have never gotten to where I am now and my father who have provided me with helpful tips about doing a PhD and was able to humour me in my lower moments.

Finally, during my PhD time I have made countless friends who have made it a great experience I am grateful for, a special mention in this sense being, "The Lunch Team"(Sachin, Roberto, Tonito), they have provided a refreshing break in the middle of each day.

Abstract

The purpose of the thesis was to study the pathogenesis of Autosomal Dominant Polycystic Kidney Disease, the most common genetic disease affecting the kidney, using novel bioinformatics and computational approaches combined with experimentation.

A new multi-stage framework for the analysis of time-series microarray data which identifies a set of possibly relevant genes and then builds a dynamic model for their regulatory network was produced during this work. The framework combines statistical filtering, Support Vector Machines, clustering and system identification in order to achieve these goals. As a practical application, it was employed to analyse two published microarray datasets derived from genetically modified Pkd1 mice. A defined set of genes was obtained from this analysis which provided good discrimination for the measurements coming from healthy and diseased animals. Also, it was noted that some genes previously linked to the disease and others related to cancer pathogenesis were identified. A potential model for their interactions was also derived.

In the second part of the project, time-lapse microscopy combined with mathematical modelling was used to study human normal and disease kidney tubular cells in both low-density migration and wound closure assays. It was found that disease cells migrated more slowly than normal cells due to a reduction in their velocity and diffusion coefficient. Of interest, the somatostatin analogue, octreotide, partially restored cell migration in disease cells primarily by increasing cell velocity. Disease cells also showed a reduced capacity to close a wound in a monolayer and this was associated with randomisation of the directionality of movement. Using textural analysis, it was noted that cell tightness appears lower in disease cells during cell migration after wounding suggesting a reduction in cell-cell adhesion in these cells.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Motivation 2

1.3 Thesis Aim and Objectives 2

1.4 Contributions 3

1.5 Outline of the Thesis 4

1.6 Publications 5

2 Pathogenesis of Autosomal Dominant Polycystic Kidney disease **6**

2.0.1 Disease Phenotype 6

2.0.2 Disease Genotype 7

2.0.3 Cell Proliferation and Apoptosis in ADPKD 8

2.0.4 ADPKD Effects on Cell Migration 11

2.0.5 ADPKD Effects on Cell Adhesion 12

2.0.6 Drugs Recommended for the Treatment of ADPKD 14

2.1 Summary 15

3 Microarray Data Processing and Modelling Methods for the Study of Diseases **17**

3.1 Introduction 17

3.2 Microarray Technology 18

3.2.1 Spotted Arrays 19

3.2.2 Affimetrix Microarrays 21

3.2.3 Ilumina Microarrays 23

3.3 Data Pre-Processing for Microarrays 24

3.3.1 Data Pre-Processing for Spotted Arrays 24

3.3.2 Data Pre-Processing for Affimetrix Microarrays 25

3.3.3 Data Pre-Processing for Ilumina Microarrays 27
3.4 Dimensionality Reduction 31
3.4.1 Feature Selection 31
3.5 Classification Algorithms 36
3.5.1 Support Vector Machines 39
3.5.2 Support Vector Machines with Redundant Feature Elimination 41
3.5.3 $l1$-Star 41
3.6 Cluster Analysis 42
3.7 Gene Regulatory Networks 45
3.7.1 Static Models 46
3.7.2 Dynamical models 47
3.8 Summary 52

4 Migration and Textural Analysis in Cells Studies 54
4.1 Introduction 54
4.2 Cell Migration Assays 55
4.2.1 Filter Assays 55
4.2.2 Time Lapse Assays 57
4.3 Measures and Models for Cell Motility 60
4.3.1 Measures for the Movement of Cells 60
4.3.2 Diffusivity models 64
4.4 Migration Analysis in PKD Cell Studies 66
4.4.1 Boyden Chambers Based Studies 66
4.4.2 Wound Analysis Studies 67
4.4.3 Combined Studies 68
4.5 Texture Analysis for Cell Studies 69
4.5.1 Structural Methods 70
4.5.2 Statistical Methods 70
4.5.3 Local Binary Patterns 73
4.5.4 Model-Based Methods 74
4.5.5 Transform-Based Methods 75
4.6 Summary 76

5 A Novel Framework for Time Series Gene Expression Data Analysis Combining Biomarker and GRN Identification 77
5.1 Introduction 77

5.2 Stage 1: Feature Selection Using q-Value-Based *l*1-StaR Algorithm 80
5.2.1 Welch t-test . . . 80
5.2.2 Q-value Calculation for Statistical Filtering . . . 82
5.2.3 Description of the Proposed Algorithm and its Application . 83
5.3 Stage 2: Gene Subset Augmentation Through Clustering . . . 84
5.4 Stage 3: Gene Subset Refinement Using Biological Knowledge . . . 84
5.5 Stage 4: GRN Identification . . . 86
5.5.1 4.1 Nonlinear Interpolation of Gene Expression Data . . . 86
5.5.2 GRN inference . . . 89
5.6 Experimental Dataset Description . . . 89
5.7 Data pre-Processing . . . 91
5.8 Modelling and Analysis of the PKD Datasets . . . 92
5.8.1 Genes Selection . . . 93
5.8.2 Discovery of Genes Similar to the Selected Ones . . . 96
5.8.3 Network Analysis . . . 98
5.9 Summary . . . 102

6 Analysis of Normal vs PKD1,2 Knockdown ciPTEC Cells and the Effects of Octreotide in Low-Density Free Migration Assays 104
6.1 Introduction . . . 104
6.2 Experimental Materials and Methods . . . 106
6.2.1 Materials . . . 106
6.2.2 Methods . . . 107
6.3 Analytical Methods . . . 109
6.3.1 Cell Division Time . . . 109
6.3.2 Cell Motility Analysis . . . 109
6.3.3 Cell Motility Analysis Modelling . . . 110
6.4 A Comparison Between the Phenotype in Healthy and Disease Cells 112
6.4.1 Division time . . . 112
6.4.2 Motility Analysis . . . 112
6.5 A Comparison Between DMSO and Octreotide Treated Disease Cells 115
6.5.1 Division Time . . . 115
6.5.2 Motility Analysis . . . 116
6.6 Discussion . . . 118
6.7 Summary . . . 120

7 Analysis of Normal vs PKD1,2 Knockdown ciPTEC Cells in Wound Healing Assays **122**
7.1 Introduction . . . 122
7.2 Experimental Methods . . . 123
7.2.1 Scratch Wound Healing Assay . . . 124
7.3 Analytical Methods . . . 124
7.3.1 Wound Closing Rate . . . 125
7.3.2 Individual Cell Tracking in Scratch Healing Assays . . . 127
7.3.3 Haralick Texture Features . . . 128
7.4 Comparison of Wound Healing Rates in Scratch Healing Assays . . 130
7.5 Directionality of Cell Movement During Wound Healing . . . 135
7.6 Comparative Texture Analysis . . . 139
7.7 Discussion . . . 142
7.8 Summary . . . 145

8 Conclusions and Future Work **147**

Bibliography **150**

List of Figures

2.1 Normal kidney (right) vs kidney affected by ADPKD (left). Republished with permission from PKD Charity Foundation US, 1001 E. 101st Terrace, Suite 220, Kansas City, MO 64131 7

3.1 Visual representation of the spotted cDNA preparation for a cancer vs normal cells experiment. Figure republished from https://upload .wikimedia.org/wikipedia/commons/c/c8/Microarray-schema.jpg . . 20

3.2 Visual representation of the Affymetrix protocol for sample preparation vs the sample preparation protocol for glass slide arrays. Reprinted by permission from Springer Nature Terms and Conditions for RightsLink Permissions, Springer Customer Service Centre GmbH: Springer Nature ,Leukemia,DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers, Staal et al [333], All rights reserved Springer Nature (2003) 22

3.3 Visual representation of the construction of Illumina Array using randomly located beads. Republished from http://www.ipc.nxgenomics .org/newsletter/no8.htm with permission by Dr Ken Lain 23

4.1 Schematic representation of a Boyden chamber 56

4.2 Schematic representation of a scratch assay 57

4.3 Schematic representation of an exclusion zone assay using a separator 58

4.4 Schematic representation of a fence assay 59

4.5 Schematic representation of a Dunn's chamber 59

5.1 Diagram of the application of the proposed framework. The analysis begins with the raw data which is usually stored in a set of matrices, represented as M_1 to M_n. The first step is to put the data together in one matrix and apply different pre-processing techniques to obtain the final dataset on which the framework is applied. The first step of the framework consists in the application of $l1$-StaR, leading to the selection of a set of genes. Next, the features in the dataset are organized in clusters. Once this operation take place, the genes selected by $l1$-StaR together with the genes in some of the clusters containing them undergo a selection based on the use of biological knowledge. For the final set of genes obtained after this step, models of regulatory interactions are created for the healthy and diseased conditions 79

5.2 Graph for the samples of 2 most selected genes in each dataset and the models fit through them. a. Graph for the genes selected for the first dataset b. Graph for the genes selected for the second dataset. . 88

5.3 Results for gene selection and clustering analysis. A) Frequency of apparition of genes selected for first dataset, B-Frequency of apparition of genes selected for second dataset, C) Dendrogram for the cluster for *Cphf*, D) Dendrogram for the cluster for *Dmkn*, E) Dendrogram for the cluster for *Guca2b* 94

5.4 Graph for the samples of 3 most selected genes in each dataset. A-Graph for the genes selected for the first dataset B-Graph for the genes selected for the second dataset. 95

5.5 Protein interaction map for the 2 datasets. A-Dataset 1, B-Dataset 2 97

5.6 Modelled gene network for the first dataset. A) Network for control samples B) Network for mutant samples 100

5.7 Modelled gene network for the second dataset. A) Network for control samples B) Network for mutant samples 101

6.1 Means, standard deviations and the results of a Welch t-test for the division times of healthy vs disease cells 112

6.2 MSDs and best model fits for the migration results of the first replicate of the healthy vs disease cells experiment. 113

6.3 Means, standard deviations and the results of a Welch t-test for the parameters of the models for the 2 conditions in the first replicate of the healthy vs disease cells experiment. 113
6.4 MSDs and best model fits for the migration results of the second replicate of the healthy vs disease cells experiment. 114
6.5 Means, standard deviations and the results of a Welch t-test for the parameters of the models for the 2 conditions in the second replicate of the healthy vs disease cells experiment. 114
6.6 Means, standard deviations and the results of a Welch t-test for the division times of DMSO vs octreotide treated cells 115
6.7 MSD's and the best model fit for the migration results of the first replicate of the DMSO vs Octreotide treatment experiment 116
6.8 Means, standard deviations and the results of a Welch t-test for the parameters of the models for the 2 conditions in the first replicate of the DMSO vs Octreotide cells experiment. 117
6.9 MSD's and the best model fit for the migration results of the second replicate of the DMSO vs Octreotide treatment experiment 117
6.10 Means, standard deviations and the results of a Welch t-test for the parameters of the models for the 2 conditions in the second replicate of the DMSO vs Octreotide cells experiment. 118

7.1 Example of the elimination of the wound in a picture. The white area in the middle represents the area eliminated by the algorithm . 129
7.2 Evolution in time of the ratio of the frames occupied by the cells in the wound healing assay 130
7.3 Snapshots of the wound closing and the results given by the segmentation algorithm (white - area detected to be covered by the cell layer, black-area detected as cell free) 131
7.4 Plot of the values of the two timeseries $d_{health}(t_n)$ and $d_{PKD}(t_n)$ for each replicate. 131
7.5 Means, standard deviations and the results of a Welch t-test for $d_{health}(t_n)$ and $d_{PKD}(t_n)$ for each replicate 132
7.6 Initial ratios for the four replicates 133
7.7 Plot of the values of the two timeseries $sd_{health}(t_n)$ and $sd_{PKD}(t_n)$ for each replicate. 133

7.8 Means, standard deviations and the results of a Welch t-test for $sd_{health}(t_n)$ and $sd_{PKD}(t_n)$ for each replicate 134
7.9 Trajectories of the cells in each replicate for each condition. The number of cells whose trajectories were plotted in each case is displayed 136
7.10 Linearity of cell movement (calculated using the confinement ratio) 136
7.11 Distance travelled by the cells towards the wound along the x-axis. . 137
7.12 Distance travelled towards the wound along the x-axis relative to the total length of the path of the cell 138
7.13 Proportion of frames in which a cell is closer to the wound after 1 hour of travelling 139
7.14 Trajectories in time represented through means and standard deviations across all the positions in a condition for the 5 Haralick features in the first replicate of the wound closing experiment 140
7.15 Trajectories in time represented through means and standard deviations across all the positions in a condition for the 5 Haralick features in the second replicate of the wound closing experiment . . 140
7.16 Trajectories in time represented through means and standard deviations across all the positions in a condition for the 5 Haralick features in the third replicate of the wound closing experiment . . . 141
7.17 Trajectories in time represented through means and standard deviations across all the positions in a condition for the 5 Haralick features in the fourth replicate of the wound closing experiment . . . 141
7.18 Snapshots of the cell layer in a movie. A-frame 1, B-frame 170, C-frame 380 142
7.19 Moving cell vs cell in monolayer representation. When the cell are moving their height gradually increases from margins to the nucleus. When they are in the monolayer the height is approximately constant. Taken from: Yeaman, C., Grindstaff, K. K., & Nelson, W. J. (1999). New perspectives on mechanisms involved in generating epithelial cell polarity. Physiological reviews, 79(1), 73-98. [396] . 144

List of Tables

2.1 Summary of the articles on proliferation 9
2.2 Summary of the articles on apoptosis 11
2.3 Summary of the articles on cell migration 13
2.4 Summary of articles on adhesion 14

5.1 Genes selected for the 2 datasets 96

Nomenclature

Abbreviations

ADPKD	Autosomal Dominant Polycystic Kidney Disease
ADPLD	Autosomal Dominant Polycystic Liver Disease
ARACNE	Algorithm for the Reconstruction of Accurate Cellular NEtworks
ASM	Angular Second Moment
AUC	Area Under the Curve
Avdiff	Average differences
BD	Bayesian Dirichlet evaluation score
BFGS	Broyden-Fletcher-Goldfarb-Shanno algorithm
BIC	Bayesian Information Criterion
BN	Boolean Networks
cDNA	complementary DNA
ciPTEC	conditionally immortalized PTEC
CLR	Context Likelihood of Relatedness
CLS	Concept Learning System
CMIM	Conditional Mutual Information Maximization
CNS	Central Nervous System
CO2	Carbon dioxide
cRNA	complementary RNA
CT	Computed Tomography
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DISR	Double Input Symmetrical Relevance
DMEM	Dulbecco's Modified Eagle's medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPI	Data Processing Inequality
DREAM	Dialogue for Reverse Engineering Assessments and Methods
DWD	Distance Weighted Discrimination
EB	Empirical Bayes
EGF	Epidermal Growth Factor
ESRD	End Stage Renal Disease

FARMS	Factor Analysis for Robust Microarray summarisation
FDA	United States Federal Drug and Administration Agency
FDR	False Discovery Rate
FE	Feature Extraction
FKK	Fractional Kramer Klein
FS	Feature Selection
FWER	Familywise Type I Error Rate
GCRMA	GeneChip Robust Multi-array Analysis
GLCM	Gray Level Co-occurrence Matrix
GO	Gene Ontology
GRN	Gene Regulatory Network
HEK	Human Epithelial Kidney Cells
HESC	Human Embryonic Stem cells
I/NI	Informative/Non-Informative
ID3	Iterative Dichotomizer 3
IPA	Ingenuity Pathway Analysis
IQR	Interquartile Range
LOAD	Late Onset Alzheimer's Disease
LSE	Least Squares Estimation
MAS	MicroArray Suite
MBEI	Model-Based Expression Intensities
MDCK	Madin-Darby Canine Kidney Epithelial Cells
MEF	Mouse Embryonic Fibroblasts
MI	Mutual Information
mIMCD	mouse Inner Medullary Collecting Duct Cells
MIT	Mutual Information Tests
MLE	Maximum Likelihood Estimation
MM	Mismatch probes
MRI	Magnetic Resonance Imaging
mRMR	minimum Redundancy Maximum Relevance
MRNET	Minimum Redundancy NETwork
MSD	Mean Squared Displacement
NB	Naive Bayes
PBN	Probabilistic Boolean Networks
PCNA	Proliferating Cell Nuclear Antigen

PCP	Planar Cell Polarity Pathway
PCR	Polymerase Chain Reaction
pFDR	positive False Discovery Rate
PKD	Polycystic Kidney disease
PLD	Polycystic Liver Disease
PM	Perfect Match probes
PTEC	Proximal Tubule Epithelial Kidney Cells
REVEAL	REVerse Engineering ALgorithm
RMA	Robust Multi-array Analysis
RNA	Ribonucleic acid
SAGE	Significance Analysis of microarray
SIRENE	Supervised Inference of Regulatory Networks
SLCL	Small Cell Lung Cancer
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVA	Surrogate Variables Analysis
SVM	Support Vector Machines
SVM-RFE	Support Vector Machines with Redundant Feature Elimination
VSN	Variance Stabilizing normalisation
VST	Variance Stabilizing Transformation

Operators

$\int^x$	Primitive of function f(x)
$\bar{x}$	Mean x_i
E[X]	Expected value of random variable X
Pr	Probability of the event
max/min	Maximum Minimum
inf	Infimum
.T	Transpose
sign	signum(sign) function
$min_{x_1,x_2,...x_n} f(x_1,x_2..x_n)$	$x_1,x_2,...x_n$ for which f($x_1,x_2,...x_n$) is minimum
$max_{x_1,x_2,...x_n} f(x_1,x_2..x_n)$	$x_1,x_2,...x_n$ for which f($x_1,x_2,...x_n$) is maximum
$minimize_A\, f(A)$	find A for which $f(A)$ minimum
$\|.\|_2$	Euclidean norm of vector

$\lvert . \rvert_1$	Manhattan norm
$\lVert . \rVert_2$	Frobenius norm
$\lVert . \rVert_1$	element-wise L1 norm
card	number of elements in a set
$\dot{x}$	derivative in respect to time of x
$\prec 0$	negative definite
$\in$	member of set
$\cup$	union of sets
$< X^2 >$	raw second moment for random variable X
$\circ$	Hadamard(element-wise) product

Symbols

#	number of

Chapter 1

Introduction

1.1 Background

Since the beginning of medicine, cysts have fascinated scientists as well as philosophers, the first accounts about this kind of affection being traceable to Ancient Greece [358]. They consist of fluid-filled cellular structures that can be localized in any part of the body [260].

Polycystic kidney diseases (PKDs) are a group of diseases that result in multiple cysts in the kidneys. Its various forms are responsible for a high number of cases of renal failure [164], the most common form of the disease, Autosomal Dominant Polycystic Kidney Disease (ADPKD) being the most prevalent inherited kidney disease appearing in 1 in 800 live births [383] and the underlying cause for 7-10% of all renal replacement therapy through the world [266]. An interesting feature of this disease is its extreme variability between patients with some experiencing kidney failure when they are below 40 years of age while others never reach this stage [382].

Due to its prevalence, ADPKD has been extensively studied but no cure has been developed so far. Many research approaches have been explored, such as its genetic component, its effect on cellular characteristics such as motility [59], proliferation [256], apoptosis [105], adhesion [338], planar polarity [226] and the effects of different drugs in slowing disease progression [357].

1.2 Motivation

In recent years, rapid advancements in analytical fields such as machine learning or image processing coupled with the capacity to extract vast amounts of biological data by tools such as microarrays or time-lapse microscopy have started to change the face of medical research, with laboratory work being aided by computer-based analysis in order to facilitate new discoveries in the field while being less invasive and limiting the costs of research. This can be done by either reducing the possible directions of research to only promising ones as is the case of microarray analysis where only few genes selected by analytical methods are investigated in the laboratory or by directly characterizing cells or tissue in a non-invasive manner using imaging as a tool.

Motivated by the fact that a cure for ADPKD was not developed yet, as well as the incidence of the disease, the overall aim of this thesis was to use the latest advancements in biological data analysis in order to discover new information about ADPKD while at the same time exploring novel approaches to analytically studying diseases. As ADPKD is a genetic disease that affects cellular formations, two directions of investigation were pursued. The genetic component of the disease was analysed using microarray data on which feature selection, clustering and classification as well as system identification methods were employed and the cellular behaviour in case of the disease was investigated using high content time-lapse imaging assisted by computational analysis.

1.3 Thesis Aim and Objectives

The first objective was to investigate which genes are altered by ADPKD on a genome-wide basis. As the quest to discover the major genes that affect disease severity and that could lead to novel drug discovery, the first sub-objective was to use the best performing machine learning techniques to find a small group of genes whose expression is changed by the disease which could be biologically tested to reveal their functional role in disease initiation or progression. A secondary sub-objective was to uncover potential new gene interaction networks through the latest advances in the modelling of regulatory networks with the final aim of allowing laboratory investigation. For accomplishing the two sub-objectives there is the need for creating a framework that allows a scientist to both isolate genes that are related to

a disease and model their interactions. The development of such a framework was a third objective in this thesis. By studying state of the art feature selection methods as well as network identification ones, a multi-stage framework for the analysis of time-series microarray data will be created in this work.

A second objective was to examine the effects of PKD1 and 2 knockdown in a human kidney cell line model (PTEC) in a time-lapse single cell migration assay. Two sub-objectives were simultaneously followed ie firstly to study differences in cell division since the assay allows for precise timing of this phenomenon and secondly, to computationally study changes in cell migration using advanced mathematical models of diffusion that allow for the quantification of cell speed as well as directionality. The ability of a promising drug candidate, octreotide, will be tested in this assay to see if it may correct either changes in cell division or cell migration.

A third objective was to examine the effect of knocking down PKD1 and PKD2 on cell migration in a wound healing assay. The first sub-objective was to discover the capacity of disease cells to close a wound. The second sub-objective was to investigate how loss of both genes might alter the directionality of cells migration. The third sub-objective was to explore how textural analysis could be employed to extract information about how the disease might alter cellular properties. This sub-objective has a more wide-range of benefits, by using images for computing textural features, non-invasive methods for cell studying can be developed that can be used in various diseases and cell property studies.

1.4 Contributions

The first significant contribution of this thesis, was to create a new framework for the analysis of time-course microarray gene expression data. The framework is used to select of potentially relevant genes for a specific condition and model the regulatory network governing their interactions. The proposed framework was applied to two different sets of microarray data derived from studies of Pkd1 knock-out mice which led to the identification of a set of genes previously connected to the disease as well as some new ones which seem to be promising biomarkers for the condition. The network analysis revealed some possible regulations that exist between them.

The second significant contribution of the thesis refers to the analysis of division of ciPTEC cells. It seems that the disease, simulated by knocking down Pkd1 and Pkd2 increases division time. Also, when testing a drug, octreotide, on diseased

cells it seems that it does not affect their division capacity.

The third contribution of the thesis regards migration of the ciPTEC cells. Low-density cell migration experiments were used to show that the disease reduces the speed of the cells and wound-healing experiments to prove that their directionality gets impaired.

The last contribution of the thesis is the discovery of a possible correlation between the value of the Haralick feature Contrast and the tightness of the cell layer. Considering this result, it was observed that the disease seems to reduce cells adhesion properties.

1.5 Outline of the Thesis

The remainder of the thesis contains 7 chapters and the bibliography.

Chapter 2 presents the biological part of this study by describing in detail the Autosomal Dominant Polycystic Kidney Disease as well as the current research that is done on it. Next, the technology for the extraction of microarray data is introduced and a description of 3 popular platforms used for this operation as well as the protocols for using them is done. Next, as microarray data is extremely noisy, a description of the methods used to pre-process it is presented.

Chapter 3 presents an up-to date review on the various techniques that have been employed through the years to analyse genetic data. The chapter presents methods used to extract relevant features by employing dimensionality reduction and clustering. Finally, methods to model the regulatory networks that govern the interactions between genes are presented and compared.

Chapter 4 consists of a review of methods used to analyse cell migration as well as cell images textural features. Different types of migration assays are described as well as their purpose in extracting information about the cells. Next, measures used to assess the migration capacities of the cells based on their trajectories are presented that analyse the speed, directionality as well as their diffusion properties. In the case of diffusion-based methods, a review is done on the models that have been proposed to extract meaningful parameters from the movement of the cells. The review then presents in detail studies that have been employed to analyse ADPKD's effects on cell migration. Finally techniques for textural analysis are presented.

Chapter 5 introduces a new framework to analyse genetic data and shows the results obtained when applying it on two publicly available ADPKD datasets. The

framework was designed to accomplish two objectives: to extract a small set of genes that are possibly relevant for the disease and to provide an approximation for the regulatory network that governs it. In the case of gene selection, the chapter provides a description of the main genes in both sets and connects the findings to other published studies on ADPKD

Chapter 6 presents the results obtained using low-density motility assays on PTEC cells that had genes PKD1,2 knocked-down. Two comparisons are provided, between normal PTEC cells and those lacking both genes and on cells lacking the genes with or without a drug, octreotide. The division properties of the cells are analysed by counting the number of frames between the division of a mother cell and its daughters. Also, state of the art diffusion models are applied to extract characteristics of the movement of the cells. Comparisons with the results of similar studies available in the literature are provided.

Chapter 7 presents the outcomes of a wound healing assay study on the same type of cells. Comparisons are provided between the control and disease cells. First, the rates at which the different cells close the wound are compared using segmentation to calculate the area covered by the cells layer. Next, the directionality of cell migration as examined in the previous chapter is analysed. Finally textural features are employed to discover properties of the migrating cell layer.

The last chapter present the conclusions of the thesis and possible future work.

1.6 Publications

During the course of the thesis, a poster with the title *Discovery of underlying mechanisms of genetic diseases using feature extraction and network identification methods* containing part of the work from Chapter 5 of the thesis was presented at the International Conference on Systems Biology, 2015.

Chapter 2

Pathogenesis of Autosomal Dominant Polycystic Kidney disease

Being a wide-spread genetic disease, in which the formation of cysts obstruct the correct functioning of the kidneys, ADPKD has been extensively studied by biologists. The scientists have tried to understand the disease both at a genetic as well as a cellular level. This chapter will present in more depth the current knowledge on the genetic component of the ADPKD as well as the disruptions in the behaviour of the cells affected by the disease that might be responsible for the creation of cysts and the drugs that have been proposed to deal with the affection.

The following sub-sections will further describe the disease and some of the major research findings. Subsection 2.0.1 will describe the phenotype of the disease, subsection 2.0.2 its genotype and the research taken on its genetic effects, subsection 2.0.3 on cell division and apoptosis, subsection 2.0.4 will summarise the literature on its effects on cell migration, subsection 2.0.5 on adhesion and subsection 2.0.6 on the drugs that have shown to have some effect on the disease.

2.0.1 Disease Phenotype

The disease leads to the development of large multiple fluid filled cysts that appear in the entire volume of both kidneys resulting in a significant increase in the total kidney volume and weight. In a study on a large sample of disease sufferers (214) by [134] the authors show that the average kidney volume for the patients was 1076 ml with a total cystic volume of 534 ml. By comparison the average human kidney is 196 ml Grantham et al. [134]. In terms of weight, on average normal male kidneys

have a combined weight of 153-455 grams [251] while in the case of ADPKD, Ekser and Rigotti [110] report a male patient with a combined kidney weight of 22 kg. 2.1 presents a normal kidney next to a diseased one.

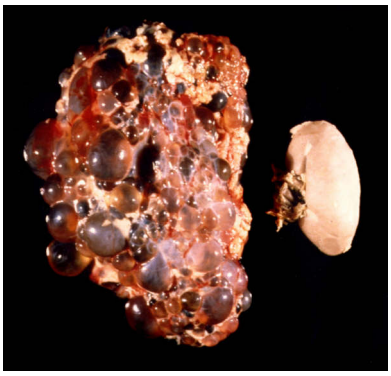

Figure 2.1: Normal kidney (right) vs kidney affected by ADPKD (left). Republished with permission from PKD Charity Foundation US, 1001 E. 101st Terrace, Suite 220, Kansas City, MO 64131

2.0.2 Disease Genotype

ADPKD results from germline mutations in one of two genes, *PKD1* [77] and *PKD2* [250]. In clinical studies, *PKD1* was shown to be responsible for 85% while *PKD2* accounts for 15% of cases [298] . In community based studies, *PKD2* mutations may account for more (29-36%) cases [348]. Individual mutations are important predictors of disease severity, with the median age of end-stage renal disease (ESRD) for *PKD1*-mutant patients being 53 years as opposed to that for *PKD2*-mutant of 69 years [164].

Although mutations in both genes are essential for the origin and severity of disease, they do not completely explain the phenotype in individual patients. In this regard, other modifying genes acting through independent or interacting pathways could play major roles in determining disease severity. Examples include genes in the EGF family, *HGF*, *IGF1* and their receptors [146], HNF family such as $Hnf4-\alpha$ [242] and $Hnf1-\beta$ [372].

When it comes to dysregulated pathways associated with the disease, a study by [289] indicates that the Wnt pathway which is involved in signalling between cells [229] could be responsible for the severity of the cystogenesis produced by the disease. The cAMP intracellular signalling pathway also seems to be involved in the disorder as it increases proliferation in disease cells but not in normal cells [394].

Mutations in both genes were also shown to initiate aberrant G-protein signalling pathways [89]. In the case of the mTOR pathway, *PKD1* mutation was shown to be associated with its activation and its inhibition led to reduced cytogenesis [318]. Another pathway that appears to be involved is the AMPK pathway which is responsible for cellular energy homeostasis [247], its activation being shown to lead to reduced cystogenesis in part through through inhibition of the mTOR pathway [346]. There continues to be active research interest in new gene discovery and novel disease pathways in ADPKD.

2.0.3 Cell Proliferation and Apoptosis in ADPKD

For the normal growth of a kidney capable of maintaining its correct structure after birth, the balance between cell division and and programmed cell death (apoptosis) must be strictly controlled [383]. ADPKD seems to alter this primary mechanism and as a result, many studies have been undertaken to investigate the role of the ADPKD polycystin proteins in cell proliferation and apoptosis.

2.0.3.1 Proliferation

The consensus in the literature is that deletion or mutation of the 2 genes results in increased cell proliferation while overexpression of the genes inhibits it. Many articles have come to backup this claim.

In a study by Nadasdy et al. [256] human ADPKD kidney was shown to have increased cell proliferation. Of interest, the proliferation rate was similar in cystic tissue and non-cystic proximal tubular and distal tubules [256]. Increased proliferation was also observed in a Han:SPRD rat model of ADPKD [292]. Similar observation were made by Tao et al [349][350] in the same rat model in two studies investigating the effect of the mTOR inhibitor rapamycin [350] and caspase inhibitor [349].

Chang et al. [65] examined normal or minimally cystic kidney tissue from two Pkd2 mouse models showing an increase in cell proliferation in cystic and non-cystic tubules compared to control wild-type animals. They confirmed these findings in human ADPKD kidney. In support of these findings, MDCK cells over-expressing Pkd1 had a halving of their proliferation rate [41]. Conversely, MDCK cells in which Pkd1 expression was knocked-down showed an increased basal proliferation rate [22]. Cystic cells also appear more responsive to ligand-activated

proliferation. In one study, lactosylceramide extracted from ADPKD cells was mitogenic while lactosylceramide extracted from healthy cells was not [67]. In another, ADPKD cells were more sensitive than normal cells to the mitogenic effects of IGF-1 [276].

Cells/Tissue	Organism	Cell proliferation quantification method	Results	Reference
Healthy kidney tissue vs tissue in cystic kidneys	Human	PCNA staining	ADPKD increases cell proliferation	Nadasdy et al. [256]
Healthy proximal tubule tissue vs cystic proximal tubule tissue	Rat	PCNA staining	ADPKD increases cell proliferation	Ramasubbu et al. [292]
MDCKZeo vs MDCKPKD1Zeo cell lines	Canine	BrdU labeling and detection kit	PKD1 overexpression reduces proliferation	Boletta et al. [41]
Healthy proximal tubule tissue vs cystic proximal tubule tissue	Rat	PCNA staining	ADPKD increases cell proliferation	Tao et al. [350]
Healthy proximal tubule tissue vs cystic proximal tubule tissue	Rat	PCNA staining	ADPKD increases cell proliferation	Tao et al. [349]
MDCK/E/siLuc vs MDCK/E/siPKD	Canine	Culture growth rate	PC-1 deletion increases proliferation	Battini et al. [22]
Wild mice vs Pkd2+/- mice tissue vs Pkd2WS25/WS25 mice tissue	Mouse	PCNA and Ki67 staining	Mutation or haploinsufficency of Pkd2 increased proliferation in mice	Chang et al. [65]
Human healthy tissue vs human ADPKD tissue	Human		ADPKD increases proliferation in humans	
HEK293cells vs HEK293cells -WT PKD2 vs HEK293cells -R742X PKD2	Human	propidium iodide cell cycle analysis	Only in the case of primary cells there is a decrease in proliferation with less cells in G0/G1 phase	Felekkis et al. [117]
NRK52-E cells -WT vs PKD2 vs NRK52-E,-R742X PKD2	Rat	PCNA		
Primary healthy rat cells vs ADPKD primary rat cells				
Tubular medula and cortex kidney tissue of Healthy mice vs PKD2 trangenic mice tissue	Mouse	PCNA	Increased proliferation in the trangenic mice	Park et al. [275]
LCLs healthy vs LCL PKD1 vs LCL PKD2	Human	Ki67 Duplication time	LCL cells from patients with PKD2 show reduced proliferation while the cells from PKD1 patients do not show changes in proliferation	Aguiari et al. [4]
Healthy mouse tissue vs tissue of mice with Pkd1 deleted	Mouse	Ki67	No significant difference was observed in the proliferative rates of healthy vs cystic mice tissue	Piontek et al. [284]
OX161/1, SKI-001, OX938 vs CL-11, RFH, UCL93	Human	Promega Cell Titer 96 Aqueous One assay	Proliferation is increased in the cystic lines. IGF-1 further increases proliferation in all cell lines	Parker et al. [276]

Table 2.1: Summary of the articles on proliferation

While the majority of studies have shown that cell proliferation was inhibited by both genes, some have presented neutral or opposite results. For example, a study overexpressing wild type and mutant PKD2 in 3 cell types (human and rat cell lines

and rat primary epithelial cells) showed conflicting results [117]. In the case of the cell lines, the overexpression of PKD2 was neutral although proliferation was reduced in the primary epithelial cells. A second study showed that transgenic PKD2 expression in mice led to increased cell proliferation and cyst formation [275]. Tam-cre mice with an inducible kidney-specific deletion of Pkd1 had slightly higher proliferation rates than controls although statistically insignificant [284]. In the case of lymphoblastoid cells, PKD2 mutation was associated with reduced proliferation [4]. Similar results were observed by Hanaoka and Guggino [143] were primary human cells coming from patients with ADPKD showed lower numbers than the controls after a period of 4 days. The authors however do not report the effects on apoptosis so it is not clear if the observed numbers are due to a reduction in proliferation or an increased in apoptosis.

It seems likely that the correct dosage of PKD1 and PKD2 is necessary to regulate cell proliferation and the threshold requirement may vary depending on tissue, cell type and stage of maturation. Table 2.1 summarises the articles on proliferation.

2.0.3.2 Apoptosis

The role of apoptosis has also been extensively studied in ADPKD. The first paper to report increased apoptosis in ADPKD was by Woo [386] who observed apoptosis in cystic ADPKD but not in control cells or tissues. Similar results have been reported by others. Ecder et al. [105] and Tao et al. [349] showed increased renal apoptosis in the rat Han:SPRD model. Shillingford et al. [319] showed that Pkd1 deletion leads to an increase in apoptosis. Apart from increased proliferation in the kidney of PKD2 transgenic mice, Park et al. [275] demonstrated increased apoptosis. Of interest, the same result was observed in a $PKD2^{ws25/-}$knockout mouse [339]. An increased apoptotic rate was reported in Pkd1 knockdown MDCK cell lines [22]. Similarly, overexpression of the gene can lead to a similar effect: [318] reported that MDCK cells overexpressing the tail terminal of PC1 showed increased apoptosis.

As in the case of cell proliferation, not all studies report an increase in the level of apoptosis in ADPKD models. In a $Pkd1^{flox(neo)}$ mouse model, Shibazaki et al. [317] show that there was almost no difference in the apoptosis level of samples coming from cystic and non-cystic kidneys. A similar lack of effect on apoptosis was reported by [284] on a Tam-cre conditional Pkd1 knockout mouse model. In contrast, mouse proximal tubule cell lines with Pkd1 deleted show a decreased level of apoptosis compared to normal proximal tubule cells [377].

Table 2.2 presents a summary of the articles on cell apoptosis.

Cells/Tissue	Organism	Apoptosis quantification method	Results	Reference
Primary cells healthy vs ADPKD	Mouse	biotin 16,21-deoxyuridine triphosphate labelling terminal transferase labelling	Apoptosis is increased in ADPKD	Woo [386]
Healthy tissue vs cystic tissue	Rat	TUNEL assay	Apoptosis is increased in ADPKD	Ecder et al. [105]
Healthy tissue vs cystic tissue	Rat	TUNEL assay	Apoptosis is increased in ADPKD	Tao et al. [349]
Healthy tissue vs tissue from Pkd1-knockout subjects	Mouse	TUNEL assay	Pkd 1 deletion increases apoptosis	Shillingford et al. [319]
Tubular medula and cortex kidney tissue of healthy mice vs PKD2 transgenic mice	Mouse	TUNEL assay	Apoptosis is increased in PKD2 trangenic mice	Park et al. [275]
Renal tissue of wild type mice vs Pkd2ws25/- mice	Mouse	ApopTag Peroxidase In Situ Apoptosis Detection Kit	Apoptosis is increased in Pkd2ws25/- mice	Stroope et al. [339]
MDCK/E/siLuc vs MDCK/E/siPKD	Canine	Vibrant Apoptosis Assay Kit #9 EnzChek Caspase-3 Assay Kit #1	Apoptosis is increased when PC1 is deleted	Battini et al. [22]
NTM-PC1 MDCK cells vs healthy MDCK cells	Canine	TUNEL assay	Overexpression of the PC1 N-tail induces apoptosis	Shillingford et al. [318]
Healthy tissue vs tissue from Pkd1-knockout subjects	Mouse	TUNEL assay	No significant difference was observed in the apoptosis rates of healthy vs cystic mice tissue	Shibazaki et al. [317]
Healthy tissue vs tissue from Pkd1-knockout subjects	Mouse	TUNEL assay	No significant difference was observed in the apoptosis rates of healthy vs cystic mice tissue	Piontek et al. [284]
PN18 PN24 PH2 PH3	Mouse	Annexin V labelling Cell counting	Pkd 1 deletion reduces apoptosis	Wei et al. [377]

Table 2.2: Summary of the articles on apoptosis

2.0.4 ADPKD Effects on Cell Migration

Normal cell migration is thought to play a major role in the homeostasis of the renal tubules, and its derangement may be involved in the transformation of normal epithelia into cysts [259]. Several groups have focussed on studying changes in cell migration in ADPKD as a possible major disease mechanism. Indeed, several studies have shown that migration is impaired in disease cells and conversely, overexpression of full-length or part of Pkd1 may stimulate cell motility.

Some studies have taken a very simple approach to characterizing cell migration by only measuring the total distance travelled. In an example of this type of study, mouse mIMCD-3 cells overexpressing the C-terminus of PKD1 had a higher migration capacity compared to controls. A similar study performed on immortalized human ADPKD cells gave the same results [384]. In a study on 3 cell lines

[60] , canine MDCK, mouse mIMCD-3 and mouse MEF cells, MDCK cells over-expressing PKD1 were more motile than control MDCK cells while the murine Pkd1 cells were less migratory than control cells. Another interesting study of this type observed kidney rudiments from mouse embryos [300]. In this report, the area occupied by migrating epithelial cells from the tissue explant was significantly decreased in ADPKD mice.

The distance travelled by cells in a period of time is a function of two variables, the speed at which the cells travel and the linearity of their movement. These two components of migration have been studied in a number of studies by time-lapse microscopy.

Some researchers have only studied the speed of cell migration. In a 2007 study, Boca et al. [40] showed that MDCK cells overexpressing PKD1 moved faster than control cells. Similar results have been observed in YPC1m-HEK cells where cells with induced expression of PC-1 demonstrated increased migration speed compared to their control counterparts [226].

In recent years, more complex analyses on cell migration have been performed in which the directionality of migration has been assessed. In a 2013 study by Castelli et al. [59], the movement of mouse MEF cells with Pkd1 mutation was significantly more random compared to control cells. Another study on disease cells (PKD1 or PKD2) in human lymphatic endothelial cells [270] found that the directionality of cells with either gene deleted was impaired compared to controls. Finally, Yao et al. [395] reported that both migration speed and its directionality were lowered in MEK cells with PKD1 mutations.

Table 2.3 presents a summary of the articles on cell migration.

2.0.5 ADPKD Effects on Cell Adhesion

Changes in cell adhesion has been another major area of research in ADPKD. In an early paper, Rocco et al. [295] reported lower expression of several epithelial cell adhesion molecules in ADPKD cells. With the discovery of the genes responsible of the disease, scientists have been able to show that the polycystin proteins are localized in several cell adhesion structures including cell-cell junctions and focal adhesions, the main cellular adhesion complex that mediates cell-matrix adhesion [99].

Most of the studies on cell adhesion in ADPKD have demonstrated that the

Cells/Tissue	Organism	Cell migration quantification method	Results	Reference
mIMCD-3- CD16.7-positive vs CD16.7.PKD1-positive	Mouse	Boyden chamber	Polycystin-1 C-terminal fragment overexpression increases migration	Nickel et al. [258]
CI normal vs ADPKD cells	Human	Boyden chamber	Cells extracted from ADPKD patients show decreased migration	Wilson et al. [384]
MDCKZeo vs MDCKPKD1Zeo	Canine	Boyden chamber	Migration is increased in MDCK cells overexpressing Pkd1 and decreased in the mouse cells with deleted PKD1	Castelli et al. [60]
mIMCD3-shCtrl vs mIMCD3-shPKD1	Mouse			
MEF PKD+/+ vs MEF PKD -/-				
Kidney rudiments from normal vs ADPKD embryos	Mouse	Time-lapse on embryonic kidney explant cultures	Cells from ADPKD mice cover a smaller area outside the tissue	Rowe and Boletta [300]
MDCKZeo vs MDCKPKD1Zeo	Canine	Boyden chamber Cells velocity in wound healing assay	PKD1 overexpression increases migration and velocity	Boca et al. [40]
YPC1m-HEK vs YPC1m-HEK with PC1 induction	Human	Cell velocity in wound healing assay	PC1 induction increases cell velocity	Luyten et al. [226]
MEF PKD+/+ vs MEF PKD -/-	Mouse	Measurement of the migration angles in wound healing assays	Loss of PKD1 makes cells to lose oriented migration	Castelli et al. [59]
Control vs PC1 vs PC2 deficient LEC	Mouse	Boyden chamber Wound closing rate Directionality and distance travelled,in time-lapse migration assay	The PC1 and PC2 defficient cells migrate less and have impaired directionality	Outeda et al. [270]
MEK DBA WT vs MEK DBA Pkd1 -/-	Mouse	Migration rate in wound healing assay Migration rate and directional persistence in low density migration assay	Migration rate and directional persistence are impaired in cells without Pkd1	Yao et al. [395]

Table 2.3: Summary of the articles on cell migration

absence of the polycystins seems to reduce the capacity of cells to adhere to surfaces or to each other. For instance, Silberberg et al. [324] found that cell-cell adhesion in primary human epithelial cells derived from ADPKD kidneys was lower compared to cells from healthy kidneys. Conversely, Castelli et al. [60] report that MDCK cells over-expressing PC-1 show increased adhesion rates to the substrate compared to normal MDCK cells. The same effects were observed in another study in which MDCK and HEK cells overexpressing PC1 showed increased adhesion to substrate, while in primary mouse tubular epithelial cells lacking Pkd1 there was a significant decrease in cell adherence [390]. A study on human immortalized cystic kidney cells showed that the use of a PKD1 antibody led to cell detachment in the case of normal cells but had no effect on cystic cells [338]. A computer modelling study in which cell mechanical interactions where simulated, showed that by lowering the parameter for cell adhesion, the cells start forming cyst-like structures [27].

Cells/Tissue	Organism	Adhesion quantification method	Results	Reference
Healthy vs ADPKD tissue	Mouse	Measurement of adhesion molecules values	Reduce level of N-CAM and E-cadherin	Rocco et al. [295]
Primary cells from normal vs ADPKD individuals	Human	Typsinizing and shaking	Cell-cell adhesion is weakened in ADPKD cells	Silberberg et al. [324]
MDCKZeo vs MDCKPKD1Zeo	Canine			
mIMCD3-shCtrl vs mIMCD3-shPKD1		Consecutive washings of freshly plated cells	Increased adhesion to surface in cells overexpressing PKD1 while decreased adhesion in cells with PKD1 deleted	Castelli et al. [60]
MEF PKD+/+ vs MEF PKD -/-	Mouse			
MDCK vs PC1 overexpressing MDCK	Canine	Cell were plated and left for 30 minutes on 96 well plates and then washed	Overexpression of PC1 increases adhesion	Wu et al. [390]
HEK293 vs PC1 overexpressing HEK293	Human			
M7 M8 OX161/1	Mouse	Washing and counting after application of IgPKD antibody	In cells expressing PKD1, the antibody leads to the cells detachment while in cells not expressing it it does not lead to cells detachment	Streets et al. [338]
-	-	Virtual simulation	Reduction of the cell-cell adhesion parameter creates cysts-like formations	Belmonte et al. [27]
Primary cells from normal vs ADPKD individuals	Human	Cell were plated and left for 60 minutes on 96 well plates and then washed	No significant change was observed in adhesion	Joly et al. [177]
MDCKZeo vs MDCKPKD1Zeo	Canine	Number of clusters obtained through mechanical dissociation divided by number of cells obtained throught trypsinization	Mechanical strength of cell-cell adhesion is lowered in cells overexpressing PKD1	Boca et al. [40]

Table 2.4: Summary of articles on adhesion

Although the overall consensus is that the expression of polycystins increases cell adhesion, there are some studies that do not show this. Joly et al. [177] report no significant difference of cell adhesion between healthy and cystic cells plated on plastic, collagen I and collagen IV. Also, Boca et al. [40] report that MDCK cells overexpressing PC1 show weakened mechanical cell-cell adhesion following trypsinization.

Table 2.4 summarises the articles on cell adhesion.

2.0.6 Drugs Recommended for the Treatment of ADPKD

Although ADPKD has been studied for a long time, finding a drug that can cure it or at least slow its progression has been an ongoing research question. A significant change occurred in 2016 when one of the drugs tested for the disease, tolvaptan in clinical trials was approved for use in patients in the European Union [2], the drug being previously approved for use in Japan in 2014 [282] but being rejected by

FDA in USA [352]. In the pivotal trial which involved over 1300 patients treated by tolvaptan or placebo [357], total kidney volume increased by around 2.8% per year as opposed to an increase of an annual increase of 5.5% in the placebo group. Since this is a modest change and the drug associated with significant side-effects, there is on-going active interest in finding more effective, safer and better tolerated compounds.

Another drug that has been tested in patients and shown promising results in early clinical studies is the somatostatin analogue octreotide. In a one-year long study on patients with severe Polycystic Liver Disease (PLD) associated with ADPKD or Autosomal Dominan polycystic liver disease (ADPLD) [154], the authors observed a slight reduction in total kidney volume in the treated patients compared to a significant increase in the placebo group. In the subsequent 4-year follow-up study, octerotide was associated with stabilisation of liver volume. However, in the case of kidney volume, there was an increase of 0.7% per year but with a very high variance (13%). In the case of the placebo group, the average was similar as the tolvaptan case (around 5.5% per year) but with a lower variance than the octreotide group (around 7% per year). These results were not statistically significant but since the number of patients was low and the variance is high, it is still possible that the drug may be effective if tested in larger numbers of patients. A study on the safety of the drug concluded that it is safe for short-term administration (7 months) [150].

2.1 Summary

The current chapter presents the state of the art research on ADPKD.

First, the phenotype of the disease is presented, followed by its genotype. The genes responsible for the disease have been highlighted as well as other genes and pathways that seem to be affected by it.

The review then moves to present the current knowledge on how the disease affects cell behaviour. The first subject to be treated is its effect on cell division and apoptosis. Although thorough the year different studies have presented various results, the dominant position that appears in most studies is that the disease seems to increase both proliferation and apoptosis. Next, articles on how migration is affected by the condition are reviewed, with the conclusion that the disease seems to reduce the capacity of the cells for directed movement. Last subject to be reviewed

as far as cell behaviour is concerned, are the effects of ADPKD on the adhesion of cells. The dominant position in literature is that of the cells adhesion seems to be weakened by ADPKD.

The last topic treated in this chapter are the effects of different drugs that were proposed to combat the disease. Two drugs that showed promising results are introduced together with the studies in which they were employed.

Chapter 3

Microarray Data Processing and Modelling Methods for the Study of Diseases

3.1 Introduction

As Oswald Avery discovered that the deoxyribonucleic acid (DNA) can transform the properties of a cell [16] a new era in understanding the functionality of organisms has started. A most important place in this area of research is occupied by the study of diseases, both common ones and genetically inherited.

Nowadays the focus in research is the identification of biomarkers which are genes whose changes in expression are signs of the development of certain diseases. Studying the expressions of single genes however is far from sufficient in understanding the complex mechanism of diseases for which the interactions of many genes forming regulatory networks are responsible [183]. This means that the extraction of the topology of these networks is needed to unveil the effects that various conditions have on the organisms they affect.

The technology that makes such studies possible are the microarrays [51] that can be used for the measurement of a large number of genes. As microarray data is extremely noisy [361], various methods have been proposed to clean them. Once the noise in the data has been reduced, the experimenter is confronted with a complex data mining problem when wanting to find biomarkers [330], as a few features have to be selected from tens of thousands of genes. In order to tackle this, methods

from statistics [336], information theory [278] and classification [140] have been proposed. Another direction for the analysis of data can the use of clustering [174] for identifying which genes have a similar behaviour as this could be a predictor for having similar functionality. When a researcher has a set of genes that appear to be an interest for a specific behaviour of the cell or a condition of the organism, methods for network identification [183] can be employed to extract their regulatory network. This chapter provides a review of the microarray technology as well as the different methods employed for the steps in analysis described above.

The remainder of this chapter is organised as follows: Section 3.2 introduces the microarray technology while Section 3.3 presents pre-processing techniques that allow scientists to eliminate the noise that comes from measurement of gene expression as well as to fuse data coming from similar but different platforms. Sections 3.4 presents methods for discovering the most interesting features while section 3.5 presents classification techniques used to predict phenotypes based on the level of gene expression. Section 3.6 introduces clustering techniques for discovering similarities between genes. Finally, Section 3.7 describes reverse engineering techniques used to identify the regulations that exist between genes.

3.2 Microarray Technology

In the beginning of research on gene expression, scientist could analyse only a small number of genes that made them unable to get the whole picture on the functioning of the genome[114]. This changed with the introduction of microarrays which revolutionised the field by allowing biologists to measure the expressions of tens of thousands of genes. As a consequence of this novel technology, genome-wide changes that appear both in healthy and disease states can be tracked [351].

Microarrays make use of nucleic acid hybridization which is the capacity of a single stranded DNA or ribonucleic acid (RNA) molecule to attach to a complementary DNA or RNA molecule [118]. For each gene of interest, a probe that contains copies of a synthetic oligonucleotide[343] or a product of polymerase chain reaction (PCR) generated from complementary DNA templates [392] which encode a sequence complementary to a unique sequence of the gene is created. The probes are arranged on a hard surface at known positions on an array [114]. Fluorescent labelled RNA products from a sample coming from the studied subjects are then allowed to hybridyze on the arrays and the light intensity exhibited at each probe

permits the quantification of the abundance of transcripts for the genes corresponding to each probe.

While many methods are used for the creation of new microarrays that try to increase the density of probes while improving their capacity to the accurately reveal the expressions of different genes, 3 methods are very popular nowadays: Spotted complimentary DNA (cDNA) arrays, Affimetrix arrays and Illumina arrays [52]. A presentation of the way they are produced as well as the protocols used to measure gene expressions with them are presented below.

3.2.1 Spotted Arrays

3.2.1.1 Production

In the glass cDNA technology, the production process starts with the selection of templates for the genes of interest which are taken from a library of genes. The templates are then cloned and PCR is used to amplify them. In the next step, purification is used to remove impurities from the obtained products. A robotic arm then prints 5 nl aliquots of the PCR products on a glass slide in a matrix pattern. Finally the board is dried and ready to be used for experiments[392].

The advantages of the glass slide cDNA microarrays consist in the reduced costs for their fabrication. Also they are very versatile, the designer of an experiment being able to customize them for the genes they are interested in. As a result they are popular in small research laboratories. The disadvantages they come with are the relatively wide areas between spots which limits the number of genes that can be measured and the reduced reproducibility between samples [52].

3.2.1.2 Protocol for Gene Expression Measurement

The first step for gene expression measurement is to extract RNA from the test and control samples, transform it into cDNA using reverse transcriptase and label it with either Cye3- or Cye5-dUTP. Next, the targets are pooled and they are let to hybridize to the clones on the glass slide. The intensities for the 2 dyes are measured with a confocal microscope through laser excitation. Because the 2 dyes have different wavelengths they can be excited separately producing 2 monochrome images. Next, the intensity of the 2 dyes for each spot are calculated with a software and the final results appear as a ratio, Cye3/Cye5.

A visual representation of the spotted cDNA labelling protocol is available in Figure 3.1

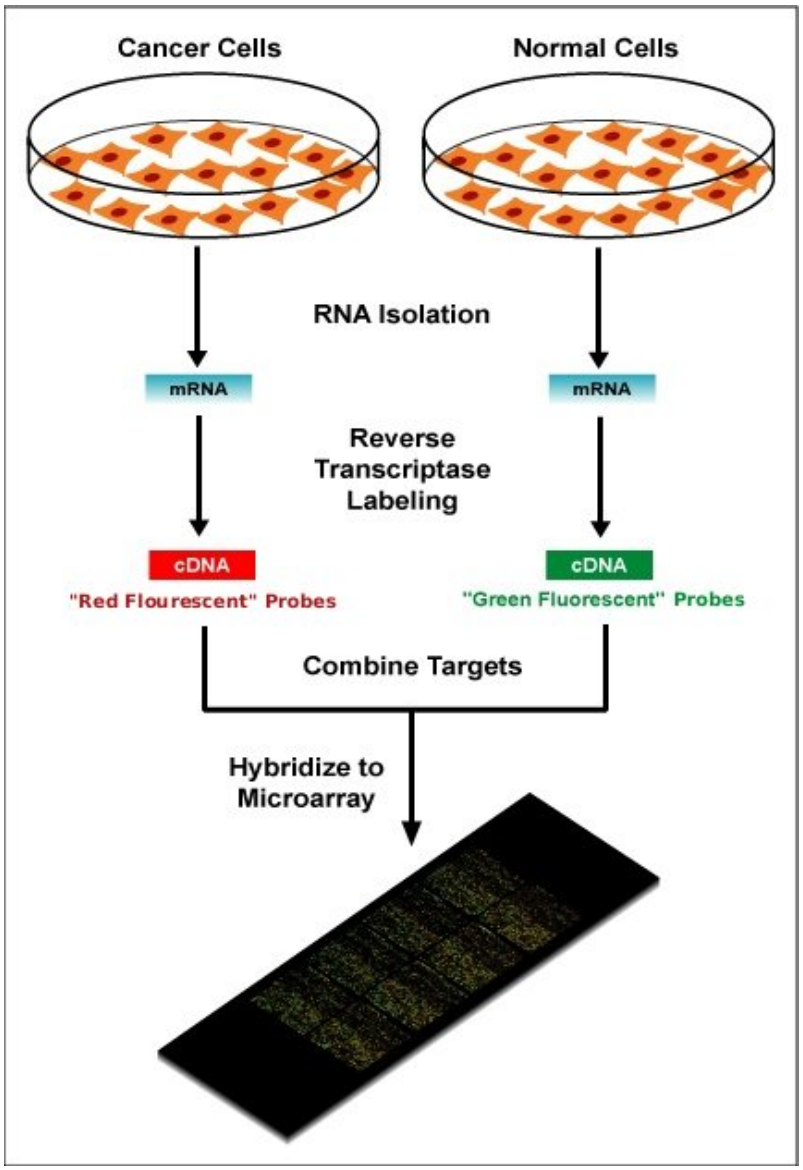

Figure 3.1: Visual representation of the spotted cDNA preparation for a cancer vs normal cells experiment. Figure republished from https://upload .wikimedia.org/wikipedia/commons/c/c8/Microarray-schema.jpg

A full description of the protocol can be found in Shalon et al. [316].

3.2.2 Affimetrix Microarrays

3.2.2.1 Production

Affimetrix microarrays use photolithographic techniques for artificially creating desired sequences [82]. Hydroxyalkyl groups are placed on a quartz wafer to produce a surface to which linker molecules with photolabile protecting groups [82] get attached. By shining near-ultraviolet light through a photolithographic mask, precise areas on the quartz are deprotected. Next, a coupling step takes place in which the wafer is washed with either A-,C-,G- or T- modified nucleotides. The 2 steps are repeated alternating the nucleotides until a 25-mer probe is produced. The obtained wafers can be diced and loaded into cartridges. The spacing between probes is around 5um. Gene expression is measured using 11-20 probe pairs. Each pair has 2 types of probes, PM (or perfect match) which are the exact probes and MM (or mismatch probes) which have the exact configuration of a perfect probe except the nucleotide in the middle. The purpose of the mismatch probes is to quantify background and non-specific hybridization which allows measurement quality assessment.

3.2.2.2 Protocol for Gene Expression Measurement

The target RNA is extracted from the sample of interest and biotin-labelled complementary RNA (cRNA) is obtained from it. The obtained solution is spread on the array and let to hybridize overnight in a hybridization oven.

Once hybridization has finished, the samples are washed in ordered to eliminate impurities and stained with streptavidin-phycoerythrin conjugate in a fluidics station. Once this operations are done, the array is scanned with a laser that excites the dye and allows the experimenter to measure the light intensity for each probe which is proportional to the quantity of RNA attached to it. The values obtained this way are stored into a computer and analysis can be performed on them.

Figure 3.2 contains a visual representation of the sample preparation for Affimetrix array together with a comparison to the glass spotted microarray preparation process. A complete description of the procedure is available in Auer et al. [15].

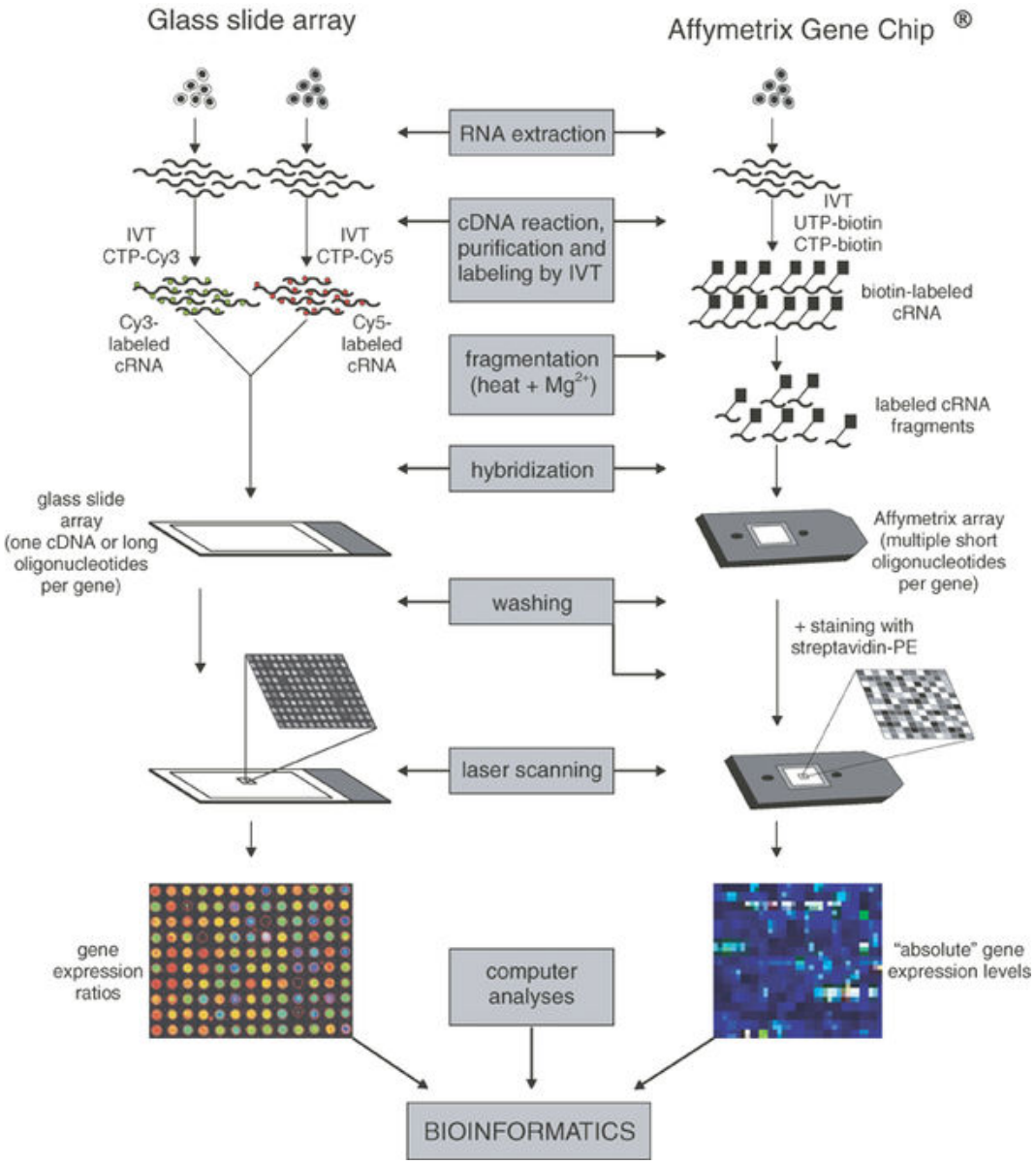

Figure 3.2: Visual representation of the Affymetrix protocol for sample preparation vs the sample preparation protocol for glass slide arrays. Reprinted by permission from Springer Nature Terms and Conditions for RightsLink Permissions, Springer Customer Service Centre GmbH: Springer Nature ,Leukemia,DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers, Staal et al [333], All rights reserved Springer Nature (2003)

3.2.3 Ilumina Microarrays

3.2.3.1 Production

In the case of Ilumina microarrays [201], plate production starts by attaching 79-mer oligonucleotides to 3um silica beads. Of the 79 monomers, 29 represent addresses which will be used to find the bead on the final slide and 50 monomers represent probes for a gene of interest. At the same time lithography is used on a glass slide to create a matrix of 96 arrays, each of them containing 50000 wells arranged in a honeycomb pattern. Next, for each array a high number of different types of beads are mixed in equal quantities and applied on it. Each bed will randomly attach to a single well on the glass plate.

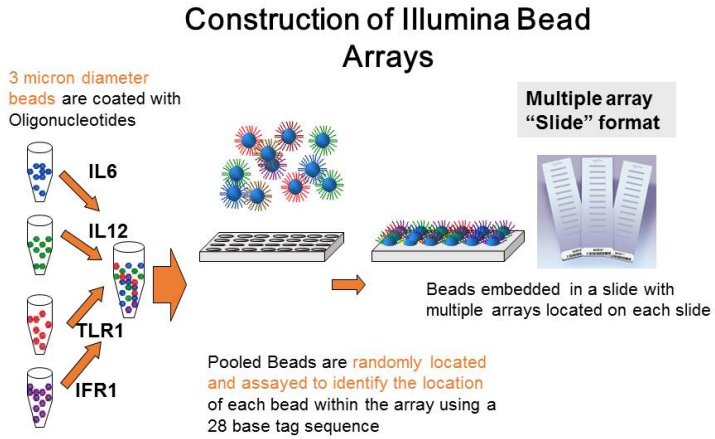

Figure 3.3: Visual representation of the construction of Illumina Array using randomly located beads. Republished from http://www.ipc.nxgenomics .org/newsletter/no8.htm with permission by Dr Ken Lain

Figure 3.3 contains a visual representation of how the beads are randomly spread on the surface of the array. With beds attached, the last part in the production is the identification of each bead on the slide. For this, a sequential method proposed by Gunderson et al. [137] is used that consists in labelling the 29-mer address section of the oligonucleotides on a bead with a different color at each step. The procedure is done until all the beads with the same oligonucleotide get a unique identifier consisting of the string of colours applied during identification.

The advantage of the Ilumina method consists in the very high number of measurements that can be taken. This allows for sequencing of whole genomes with around 30 beads per probe. The high number of replicates as well as the randomness of beads localization which eliminates systematic errors due to local defects of the plates gives measurements of Illumina microarrays high replicability. The main disadvantage of the method is the high costs of the product.

3.2.3.2 Protocol for Gene Expression Measurement

The measuring of gene expression using Ilumina microarrays starts with the extraction of RNA from the samples of interest. Next, cRNA should be obtained from the RNA of the sample. An optional step at this point is to create biotin labelled cRNA which will enhance the readings. Hybridization of the assay is done with cRNA which is added to hybridization chambers provided by ilumina that contain the BeadChips, each with a specific number of 96-arrays matrices. Once hybridization is done, the BeadChips are washed in order to eliminate impurities. The next step is to add Cye3 dye which will lead to different light intensities for the beads proportional to the quantity of cRNA they have hybridized. Last a laser scanner excites the dye and reads the intensities for each probe. A software provided by Ilumina localizes all beads corresponding to the same probe and stores their corresponding intensities. The full protocol can be found in the Ilumina datasheet[165].

3.3 Data Pre-Processing for Microarrays

In order for raw data obtained using microarrays to be analysed so that relevant information related to diseases can be extracted, a series of pre-processing steps have to be taken first. These methods are platform dependent and the remainder of this subchapter presents them as well as the objectives they try to accomplish for the 3 types of microarrays presented.

3.3.1 Data Pre-Processing for Spotted Arrays

The main objective when working with cDNA microarrays is to correct for the various sources of noises that appear with the measurements. In order to do this, 2 types of methods are employed, background correction and normalisation.

The luminous intensity of a probe on a microarray comes from 2 sources, the intensity due to hybridization and the intensity of the background in its neighbourhood [397]. The objective of background correction is to eliminate the luminous intensity of the probe that is due to the location on the microarray which it resides on. A simple way to do this is to subtract the mean or median intensity of neighbouring pixels from the intensity of the probe [294]. The problem with this method is that it can lead to negative intensities so a few alternative methods have been proposed for the problem. These include using a threshold to decide if subtraction should be done[107], use of empirical Bayes models[194] or more recently the use of background smoothing before background correction is applied [310]. More details on the background correction methods used with cDNA microarrays as well as a comparison between them can be found in the review by Ritchie et al. [294].

The second stage in the pre-processing of data coming from cDNA microarrays is represented by normalisation. In microarray data analysis, normalisation is used in order to reduce the biologically unrelated variance of measurements between different arrays. Multiple normalisation methods have been proposed, however they can be classified in 2 major categories [42]: Complete data methods and methods using a baseline array. Approaches in the first category use all the information from the studied arrays in order to obtain normalized values. By contrast, techniques in the second category choose a baseline array in respect to which the other the arrays are normalized. A problem introduced by this strategies is the choice of the baseline.Bolstad et al. [42] provide a comprehensive review of normalisation methods, in which they prove that complete data methods give better performance. The conclusion of the review is that quantile normalisation is the best method to be used for its simplicity and performance.

3.3.2 Data Pre-Processing for Affimetrix Microarrays

Data pre-processing of Affimetrix microarrays has 2 goals: noise filtering similar to cDNA microarrays and, in addition, summarisation. Summarisation is the operation through which all the measurements for one probe are brought to a single value using different averaging methods. Because of the very high density of probes, background correction methods that are used in cDNA microarrays cannot be used on Affimetrix slides, the researcher however can make use of the difference between PMs and MMs.

One of the first methods that was used to pre-process Affimetrix microarrays is called AvDiff [1] or MAS4 and was developed by Affimetrix. It is a simple method of noise filtering and summarizing Affimetrix microarray measurements, by averaging the PM-MM difference for each probe. Li and Wong [214] show that, by using MAS4 the variations due to probe effects can be up to 5 times higher than the variations in-between arrays. This suggests that AvgDiff gives poor performance in reducing the measurement related noise. In response to this problem they introduce a new algorithm called Model-Based Expression Intensities (MBEI) [214] that solves it by fitting a model through a probe set across all arrays.

Affimetrix came with a new algorithm that corrects most of the problems of MAS4 called MAS5 [1]. It uses Turkey biweight averaging [142] which leads to a summarisation method that does better noise filtering than its predecessor and comes with a new a method for background noise correction. Unlike MAS4 and MBEI whose final results are on a linear scale, MAS5 uses a log 2 scale which adds better variance stabilization and easier interpretability of the results. Also MAS5 is the first method to introduce normalisation. Another novelty in this method is the presence of a signal detection call that can be used to eliminate non-informative genes when looking for differential expressed ones.

A new method for summarizing Affymetrix microarrays was introduced by Irrizary et al [169] called robust multi-array analysis (RMA). One important difference between RMA and the previous methods is that it uses only the PM values. The reason is that MM probes could also contain signal information which will be lost by subtraction.

When it comes to comparing RMA with MAS 4, MAS5 and MBEI in terms of detecting changes in differentially expressed genes, RMA outperforms the other methods in terms of the area under the curve (AUC) [169]. This result appears, however, because RMA has a significant increase in precision but a small decrease in accuracy compared to MAS5. A new method proposed to solve the problem is GeneChip robust multi-array analysis (GCRMA) [391] which is very similar to RMA but uses information on the specific biding of each nucleotide when background correction is applied. GCRMA provides accuracy comparable to MAS5 and precision as good as RMA [391]. Its only drawback is that it does not provide a call detection value such as MAS5.

The high number of summarisation methods for Affymetrix arrays, led to a competition called Affycomp II [78]. A new method called FARMS [153] has outper-

formed all the others in terms of AUC for detecting differentially expressed genes. FARMS also uses just PM values but unlike MAS5 and RMA it does not do background correction, but just clears the data of noise on the same probesets across different arrays. The authors of FARMS also came with a method for detecting if a probe is absent, which was named I/NI calls [347].

3.3.3 Data Pre-Processing for Ilumina Microarrays

Illumina technology provides an original approach in the way the results are outputted by the machine. Instead of the PM-MM pairs, the users receive the average and standard deviation of the measurements for each probe on an array as well as a presence call for the probe so no detection call method has to be implemented. Because of these factors, the pre-processing for Illumina arrays focuses on variance stabilization, normalisation and batch effect removal.

The objective of variance stabilization is to correct unwanted dependency between the variances and the means of the measurements [217]. In the case of Affimetrix arrays the application of more complex variance stabilization methods can be quite difficult as the number of replicates is quite limited [217]. As with Ilumina technology more measurements are taken for each probe, advanced variance stabilization techniques become feasible. The simplest method for stabilising variance is to apply base 2 logarithm on the data, an approach also used in Affimetrix methods such as MAS5 and RMA. This method though has some significant short-comings [217]. First it a global solution that ignores the measurement noise characteristics of different machines and experiments. Also negative measurements coming from background corrections on low intensity signals have to be changed to 0 before logarithm can be applied. Finally, for values close to 0, the algorithm increases variance rather than reducing it. In order to solve this problems, a new variance stabilization method was proposed by Huber et al. [159] based on the model:

$$Y = \alpha + \mu e^{\eta} + \varepsilon \tag{3.1}$$

proposed by Durbin et al. [104], where Y is the raw expression measurement, α is the background noise, μ is the true expression of the gene and η and ε are normally distributed error terms with mean 0. The name of the method is variance stabilizing normalization (VSN) and it solves the problems of log2. Originally, VSN was developed for Affimetrix arrays so Lin et al. [217], proposed a new method for

doing variance stabilization on Ilumina chips based on the model 2.1 called VST. The difference between the 2 algorithms resides in the way in which the parameters are approximated. Variance stabilizing transformation (VST) relies on the much higher number of beads per probe in Illumina and uses only intra-array measurements to estimate the parameters. Lin et al. [217] proved that when using Ilumina data, VST outperforms VSN by improving the signal to noise ratio and reduces the number of false positives of differentially expressed genes. A presentation of the VST is provided below. For measurements described with the model 3.1 the mean is:

$$E(Y) = u = \alpha + m_\eta \mu \tag{3.2}$$

And the variance:

$$Var(Y) = v = s_\eta^2 \mu^2 + \sigma_\varepsilon^2 \tag{3.3}$$

where m_η and s_η^2 are the mean and the variance of e^η, and σ_ε is the standard deviation of ε . Now if μ is substituted from the first to the second equation it is possible to write the variance of measurements as a function of mean which also shows why the 2 are interdependent.

$$v(u) = (s_\eta / m_\eta)^2 (u - \alpha)^2 + \sigma_\varepsilon^2 = (c_1 u + c_2)^2 + c_3 \tag{3.4}$$

The authors use a transformation on Y so that u and v become independent proposed in [355] which is:

$$h(y) = \int^y 1/\sqrt{v(u)} du \tag{3.5}$$

by substituting c_1, c_2, c_3 in this function the resulting function is:

$$h(y) = \begin{cases} 1/c_1 \, arcsinh(c_2/\sqrt{c_3} + c_1 y/\sqrt{c_3}) & \text{when } c_3 > 0 \\ 1/c_1 \, \ln(c_2 + c_1 y) & \text{when } c_3 = 0 \end{cases} \tag{3.6}$$

Of the 3 parameters, c_3 is the average of the background probes which are defined as having a detection p-value of higher than a predefined threshold, in general 0.05. The other 2 parameters, c_1 and c_2 are estimated using linear fitting from the equation:

$$\sqrt{v(u) - c_3} = c_1 u + c_2 \tag{3.7}$$

Once the transformation has been applied, the data is summarised by calculating the mean of the transformed h(y) bead values. As the technologies are similar, using an array to measure one sample, normalisation methods for Ilumina are the same as in the case of cDNA or Affimetrix. This means that again quantile normalisation is the usual choice. In microarray technology, a batch refers to plates prepared in the same place using the same platform during short periods [68]. When the data is obtained over multiple batches which means that it is measured in different days or the biological samples are measured using different instruments, differences can appear that are unrelated to the real biological changes [213]. Normalisation methods are not enough to remove them [213] so new methods were developed. Different batch removal procedures use a wide range of techniques. Examples come from algebra such as singular value decomposition used in Surrogate Variable Analysis (SVA) [212], statistics with the Empirical Bayes(EB) estimates used in Combat [176], or artificial intelligence, support vector machines (SVMs) being used in Distance Weighted Discrimination (DWD) [31]. SVA and DWD have some shortcomings when it comes to practical implementation [176]. In the case of SVA the method requires proper selection of first several eigenvector while DWD permits comparison only between 2 batches. The only drawback of ComBat is the fact that the batches from which the data comes have to be known but as long as the information is available, the method can be applied. In a thorough study, Chen et al. [68] compare these methods on artificial and experimental data. In their analysis ComBat has shown superior performance and also it appears that its performance relative to the other approaches increases as the size of batches decreases. This is an advantage in practice as the number of samples is limited. In the Combat method it is assumed that the data has been normalized, summarised and low detection probes have been eliminated. Once these operations have been finished the measurements for each probe are characterized using the model:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg} \tag{3.8}$$

where Y_{ijg} represents the measurement for gene g in sample j from batch i, α_g represents the real value of the gene, X represents a design matrix for sample conditions, and β_g is the vector of regression coefficients corresponding to X. The error terms, ε_{ijg}, come from a zero-mean normal distribution with variance δ_g^2. γ_{ig} and δ_{ig} represent the batch effects of batch i for gene g, with γ_{ig} the additive effect and δ_{ig} the

multiplicative one.

The method proceeds to eliminate the batch effects in 3 steps:

1. Data Standardisation

   It is expected that the magnitude of expression will vastly vary among different genes and this will affect the bayesian estimates of the prior distribution of batch effect. In order to do this, least squares are used on the model to estimate α_g, β_g, γ_{ig} as $\widehat{\alpha_g}$, $\widehat{\beta_g}$, $\widehat{\gamma_{ig}}$ with the constraint:

   $$\sum_i n_i \widehat{\gamma_{ig}} = 0 \tag{3.9}$$

   for all the genes, where n_i is the number of samples for batch i. With the model obtained this way, the distribution of error terms can be approximated as having a variance:

   $$\widehat{\sigma_g^2} = \frac{1}{N}\sum_{ij}(Y_{ijg} - \widehat{\alpha_g} - X\widehat{\beta_g} - \widehat{\gamma_{ig}})^2 \tag{3.10}$$

   The standardized data are calculated as:

   $$Z_{ijg} = \frac{Y_{ijg} - \widehat{\alpha_g} - X\widehat{\beta_g}}{\widehat{\sigma_g}} \tag{3.11}$$

2. Estimation of the batch effect parameters

   The standardized data is considered to come from the distribution $N(\gamma_{ig}, \delta_{ig}^2)$. Another assumptions made for the parameters of the distribution are:

   $$\gamma_{ig} \sim N(\gamma_i, \tau_i^2) \; and \; \delta_{ig}^2 \sim InverseGamma(\lambda_i, \theta_i) \tag{3.12}$$

   The hyperparameters $\gamma_i, \tau_i^2, \lambda_i, \theta_i$ are estimated using the method of moments.Based on this assumptions the 2 parameters are approximated using the conditional posterior means:

   $$\gamma_{ig}^* = \frac{n_i\overline{\tau_i}^2\widehat{\gamma_{ig}} + \delta_{ig}^{2*}\overline{\gamma_i}}{n_i\overline{\tau_i}^2 + \delta_{ig}^{2*}} \; and \; \delta_{ig}^{2*} = \frac{\overline{\theta_i} + \frac{1}{2}\sum_j (Z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_j}{2} + \overline{\lambda_i} - 1} \tag{3.13}$$

3. Adjust the data for batch effects

   The final measurements, γ^*_{ijg} are approximated using:

$$\gamma^*_{ijg} = \frac{\widehat{\sigma_g}}{\widehat{\delta^2_{ig}}}(Z_{ijg} - \widehat{\gamma_{ig}}^*) + \widehat{\alpha_g} + X\widehat{\beta_g} \tag{3.14}$$

3.4 Dimensionality Reduction

In the era of big data, fields such as biology, engineering and economics face high dimensional measurements that create the need for new strategies in order to select the most important features[121]. Different approaches from statistics and machine learning have been proposed in order to transform the data from a high dimensional space to a lower dimensional one while preserving the important information contained by them. They represent an important pre-processing step in pattern recognition as they are a viable cure to the curse of dimensionality.

The dimensional reduction techniques come in 2 main categories, feature selection (FS) and feature extraction (FE). Feature selection procedures return a subset of the original dimensions or a ranking for them according to a criterion [39] while feature extraction methods create a novel set of attributes based on transformations applied to the original ones.

This thesis will concentrate on FS procedures for supervised classification. Feature extraction and unsupervised feature selection methods have little relevance for this study, as the classes for the data are known and the final goal is to select genes not classify data, so they will be left out of this literature review.

3.4.1 Feature Selection

Feature selection represents an efficient strategy to tackle some of the problems created by high dimensional data. Its main benefits are improvement of classifiers accuracy, reduction of the memory requirements for the data and increase in the interpretability of models.

Algorithms implementing feature selection methods have 3 main components: the generation procedure, the evaluation function and the stopping criteria [219].

The generation procedure creates subsets of features that are going to be evaluated. There are 3 primary strategies for it: exhaustive search, sequential search and

random search. Exhaustive selects the best features by checking all the possible combinations. Its advantage is the fact that it always provides the optimal solution but for a high number of dimensions this method might be impossible to use in practice because of time limitations. Sequential search methods produce subsets by iteratively adding or removing sets of features until no improvement is observed in the evaluation function. They do not do a complete search, unlike exhaustive methods but are more practical for implementation on real systems. In the case of random search, a randomly selected subset is considered initially and then the algorithm continues either by using sequential search in which some randomness has been added or just continues to generate random subsets.

The evaluation function represents the technique used to assess the goodness of the candidate feature subsets. Two classes of evaluation criteria exist, independent ones such as measure of distance, dependency measures, information and consistency measures and dependent ones that asses the goodness of results based on the performance of a data mining algorithm [83]. The stopping criteria represents the condition for which the algorithm stops. Some common examples are:

- All subsets have been evaluated
- A certain number of iterations have passed or a certain number of features have been selected
- Adding or removing features does not bring improvements in the value evaluation function
- A certain value has been obtained by the evaluation function

Many feature selection algorithms have been developed and according to their evaluation function they can be split into 2 main categories: filters and wrappers. The filters use independent evaluation functions while the wrappers use the dependent ones.

3.4.1.1 Filters

Algorithms in the filter category asses the performance of each individual features and select a subset either by putting a threshold on the number of features selected or on their evaluation score. A popular filter approach used in microarray data [209] is the imposition of a threshold on a measure of statistical significance.

The simplest approach is the application of single hypothesis testing on each of the features in a dataset. This tests the null hypothesis (which is that there is no real difference between samples of the feature coming from different classes) against the alternative hypothesis (which is that real differences exist). A measure of significance, the p-value which is the probability that the measurements came from a set where the null hypothesis is true is assigned to each feature. The last step in this case is to put a threshold on the p-value, usually 0.05 or 0.01, and pick all the features with a p-value below it. Two main types of statistical methods for testing exist, parametric ones in which the distribution of the samples of the classes is considered known, the most wide-spread of which is the Student t-test [340] and non-parametric methods such as the Wilcoxon rank sum test, where no knowledge of the distributions is needed.

In practice in a gene expression dataset multiple features are evaluated which corresponds to the case of multiple hypothesis testing. In this situation, applying a constant threshold on the probability that the null hypothesis is true is not enough. This happens because a number of features for which the null hypothesis is true will give a p-value below the threshold purely by chance [102]. The percentage of features for which this happens out of all the features considered significant is called the false discovery rate(FDR). Early methods for dealing with this problem have been based on controlling the family-wise type I error rate (FWER) [315] which is the probability that one real null hypothesis was misclassified. The problem with this approach is that it is too strict [335], leading to the elimination of too many genes in the case of the tens of thousands of features that appear in a gene dataset. A relaxation to this approach is to put a threshold on the false discovery rate (FDR) as proposed by Benjamini and Hochberg [32] . The measure they control is:

$$FDR = E[\frac{V}{R}|R > 0]Pr(R > 0) \tag{3.15}$$

where V-the number of features falsely called significant, and R the number of features that are called significant. The way they do this is:

$$reject\ H_j\ for j = 1,...,max\ \{i|p_i \leq \frac{i}{m}\alpha\} \tag{3.16}$$

where H_j is the null hypothesis for feature j, p_i is the p-value for feature i with the features arranged in the increasing order according to their p-value, m the total number of features and α the threshold for the p-value assigned by the scientist

when doing single hypothesis testing on the feature. As pointed out in Storey [335] there are 2 problems with the initial method for bounding the FDR. The first one is that there is no guarantee that the threshold chosen imposes a hard upper-bound on the percentage of misclassified features. The actual threshold to which it limits the FDR is $\alpha/P(R>0)$ once one feature has been classified as insignificant. The other one is that no information from the data about the number of features in which the null hypothesis is true is actually used to approximate the FDR. A new method, the q-value which measures the positive false discovery rate (pFDR) was proposed to solve these weaknesses. The positive false discovery rate is defined as:

$$pFDR = E[\frac{V}{R}|R>0] \tag{3.17}$$

and it represents the false discovery rate when at least one feature was called significant. In their approach, instead of thresholding the pFDR with a fixed α, the authors use the features arranged in increasing order of their p-values to define nested rejection regions [0,γ]. Because the values are nested, a region γ contains the first k features, k<γ*m with the smallest k p-values. Next for each of these regions the pFDR (γ) is:

$$pFDR(\gamma) = \frac{\pi_0\gamma}{Pr(P\leq\gamma)} \tag{3.18}$$

with π_0 the probability that a feature is non-significant and P the random variable which contains all the p-values of the features. The pFDR is estimated for all regions of interest (in practice this means the regions that have different numbers of elements) using:

$$\widehat{pFDR_\lambda}(\gamma) = \frac{\widehat{\pi_0}(\lambda)\gamma}{\widehat{Pr}(P\leq\gamma)\{1-(1-\gamma)^m\}} \tag{3.19}$$

with

$$\widehat{\pi_0}(\lambda)\gamma = \frac{W(\lambda)}{(1-\lambda)m} \tag{3.20}$$

and

$$\widehat{Pr}(P\leq\gamma) = \frac{R(\gamma)\vee 1}{m} \tag{3.21}$$

where $R(\gamma) = \#p_i \leq \gamma$ and $W(\lambda) = \#p_i > \gamma$.

λ can be optimally chosen using an automatic approach proposed by the authors.

The last step is to assign q-values to each of the features in the set. The q-value of the feature i in a set of nested regions A that contain it is:

$$q(i) = inf_{\gamma \geq p_i} \widehat{pFDR}_{\lambda}(\gamma)$$

This approach however does not take in consideration the correlation between the features. As shown in [139], not considering the correlation between features during selection might lead to decreased performance in later classification. To tackle this problem, new multivariate filters emerged. Popular representatives for this category are minimum redundancy maximum relevance (mRMR) [278], conditional mutual information maximization (CMIM) [120] and the double input symmetrical relevance (DISR) [246] .In the mRMR criterion mutual information(MI) is used to combine 2 previous principles which are to obtain maximum correlation between a selected feature and the vector of classes for samples and to have as little correlation as possible between the selected features.The CMIM criterion selects features whose capability of predicting which class the sample came from was not caught by other attributes.The DISR criterion employs feature complementarity which means that for a set of features to be selected, their combined classification capacity is higher than the sum of individual classification performance. Meyer et al. [246] have proven that mRMR returns a set of features as relevant as those selected by the newer DISR-based method but at a significantly reduced computational cost.

3.4.1.2 Wrappers

The wrapper approach is classifier dependent so the subset selected with this method can differ a lot if the predictor changes. Because a machine learning method is used for evaluating each selected subset, wrappers are much more computationally expensive than filters [304]. However, features selected using wrappers are expected to obtain better classification accuracy than filters [151] as they directly optimize features that can discriminate between classes.

The combination of the 2 methods lead to new feature selection algorithms represented by embedded methods [139] and hybrid algorithms [219]. In the embedded methods the whole feature set is used as input for a classification method which will output a classifier using just a reduced set of features.

Hybrid methods come with a different approach which works in 2 stages. Filters are used in the first instance to create subsets with different sizes. Next, one of the subsets is selected as the best one by using a wrapper. The disadvantage of this method is that it need a predefined stopping criterion.

Section 3.5 presents a review on some classification methods widely used for microarray studies as well as history of the wrappers that have been used with them.

3.5 Classification Algorithms

Classification is a main task in data mining and machine learning. Its objective is to produce models that can identify to which class a specific instance belongs to. Methods used for classification have 2 phases, training and testing. The training phase consists of building the model using samples from known classes. In the testing stage the class of different instances is predicted using the model obtained in the training stage and the resulted labels are compared with the real ones.

Classifiers are a vast area of research, many of them being developed over the years. A thorough review on the subject is provided by Kotsiantis et al. [196]. Although there are many classification methods that have been created over the years, many of them combining existing methods there are some algorithms that are fundamental to the field. They include Naive Bayes (NB) classifiers, decision trees and support vector machines. The remainder of this section will provide a description on them with a focus on Support vector machines as well as their application on feature selection using either wrapper, hybrid or embedded methods.

NB methods represent a class of classification methods in which Bayes theory is used to predict the class from which a sample came. The ancestor of all this methods was proposed as a text-classification algorithm [101], a field in which they are still very popular [389].

The idea behind the NB is to approximate the distribution of each class from which the data come using the training data, and new samples are assigned the class they most probably come from. In the use of Bayesian method there is one assumption made about the data, which is that all the features of a sample are independent random variables. This assumption is unrealistic for most data but the method showed to work surprisingly well in practice, even when compared to more complex methods [125].

The NB methods have a history of being applied in wrapper approaches, the first paper in which they have been used for feature selection[207], appearing in the same year as the introduction of wrapper methods [175]. In this approach the authors try to eliminate correlated features in order to improve the classification performance of NB. Their method was successful, providing improvements of per-

formance to the standard NB that were close to those of other methods even in the fields where NB was showed to perform poorly. Further improvements in wrappers based on Bayesian methods have been introduced in the works of Kohavi and John in which the NB are used on wrappers for which the candidate features are selected using techniques such as best-fit search and hill-climbing search. Closer to our time NBs have been used to improve candidate subset feature selection for wrapper methods[34]. In the field of microarray studies, NB based algorithms have been to select biomarkers for diseases such as cancer [167], Chronic Fatigue Syndrome [157] or dermatological diseases [12].

Decision trees, similar to the Naive Bayes classifiers are one of the oldest methods in machine learning. The first decision tree technique ever implemented was the Concept Learning System framework(CLS) [161]. Based on it, new methods were proposed such as Iterative Dichotomiser 3(ID3) [290] and its successor C4.5 [291] which are still used nowadays.

The idea behind decision trees is to create a set of rules based on the values of the attributes for the measurements so that each resulting subset contains samples from only one class. In the case of continuous attributes, this means creating thresholds on the values of each attribute so that the samples in the resulting partitions are as homogeneous as possible from the perspective of their corresponding class. New samples are classified automatically considering the partition they are in. The main reason for the popularity of the decision trees is that they give intuitive explanations of the classification process and can be used to easily combine discrete and continuous features.

Wrappers based on decision trees are as old as the idea of wrapper itself, the first wrapper method ever proposed [175] using the ID3 and C4.5 algorithms as classifiers. An improvement has been proposed to the decision tree based wrappers in the work of Cherkauer and Shavlik [81] by creating candidate subsets using genetic algorithms. The approach found popularity in the field of intrusion detection [334], an upgrade being proposed by incorporating neural networks as well [326]. Another direction for the decision tree based wrappers is the use of random forests [152] as classifiers as in the Boruta algorithm [202], an approach where several decision trees are combined to provide improved classification. In the field of microarray data analysis decision trees have been used to detect biomarkers for cancer [400], [69].

In the recent years, a classification method that is becoming extremely popular

is the artificial multi-layer neural network [307], which is a classifier that works by trying to mimic the functioning of the biological neural networks[173]. The classifier is organized as many layers of functions, the output of a layer being the input of the following one. This allows the network to discover increasingly complex patterns after each series of layers, making it fit for applications such as object recognition to the level of dog breeds from pictures [302], understanding human language [30] and the creation of self-driving cars [163].

While not as popular as the other methods for the creation of wrappers, neural networks have been used as a classifier for a wrapper. One example in this sense is the NNFS [314], which use a penalty on the error term to eliminate features with low weights in the final network. Another example was created by De Rajat et al. [87] in which they combine a multi-layer neural network with fuzzy logic to select the most relevant features. In biology neural networks-base wrappers have been used to predict the outcome of osteoporosis by using genetic factors [64]. The reason why neural networks are not used too frequently in wrappers although they produce accurate classifiers [64] is their complexity which makes them impractical [375]. Also they are not popular in medicine as they produce complex models which are difficult to interpret [61].

One of the classifiers with the widest range of applicability in data mining is the support vector machine SVM [45]. The method produces a separation hyperplane which maximizes the smallest distance between the decision boundary and any data points(called the margin) in order to obtain a model capable of a high degree of generalization. The new data points are classified considering the side of the hyperplane they appear in. Since most of the real datasets are not linearly separable, the introduction of kernels have appeared, which map the original data in a higher dimensional space where it becomes linearly separable.

As linear support vector machines produce an array of weights for the features used, wrapper approaches come naturally to the method. The algorithm has been used in all type of classification-based feature selection approaches. A wrapper using SVMs came from the creators of the strategy and it is called support vector machines with redundant features elimination (SVM-RFE) [140]. In this approach the authors use linear support vector machines to sort features in order of their weight and eliminate half of them at each step.

An embedded method for feature selection using SVMs was proposed by Weston et al. [379] where gradient descent is used to minimize the number the features

until a specific number is reached. The problem with this approach is the number of selected features is predefined by the user without taking the data into consideration. Zhu et al. [406] solved this problem by using the l1-norm to minimize the weights of the features. This naturally eliminates them leaving a optimal number for classification thus leading to an embedded SVM-based feature selection method. Newer approaches use embedded methods to penalise the kernel into selecting the most informative features [232].

SMV-based hybrid methods have been proposed in which filters are the average difference in a measurement closely related to the Fisher criterion score [37] of the samples in different classes Furey et al. [129]. A method that combines most of the presented strategies can be found in the work of Ahsen et al. [5], an l1-norm SVM with a recursive feature elimination being combined with a statistical testing based filter.

Microarray studies are one of the first fields in which support vector machine wrappers have been used [140]. Their capacity of dealing with high dimensional data and dealing with irrelevant and redundant attributes [197], makes them extremely popular in molecular biology where the number of features greatly exceed the number of samples. Support vector machines have been applied to find biomarkers for diseases such as cancer [140], Alzheimer [96], cerebral accidents [287], multiplesclerosis [405]. Various studies have shown that the SVM classifiers provide best results for microarray data [211][285][195] , so the following sections describe more in depth the SVM methods and feature selection methods using them.

3.5.1 Support Vector Machines

The simplest case in which support vector machines can be used are the ones in which the data is linearly separable. This means that if $x_i \in \mathbb{R}^m$, where m represents the number of features is a vector of measurements for a specific sample, there can be a weight vector w and a bias b for which:

$$\begin{aligned} w^T x_i + b \geq 1, x_i \in C_1 \\ w^T x_i + b \leq -1, x_i \in C_2 \end{aligned} \tag{3.22}$$

with C_1 and C_2 the classes for the samples [45]. In this case, the function sign() produces a decision rule for the classifier

$$f_{w,b}(x_i) = sign(w^T x_i + b) \tag{3.23}$$

Let $y_i = f_{w,b}(x_i)$. In this case in order to create a decision boundary maximizing the margin, the convex programming problem

$$min_{w,b}\frac{1}{2}|w|_2^2 \tag{3.24}$$

subject to

$$y_i(w^T x_i + b) \geq 1 \tag{3.25}$$

has to be solved, where x_i, $i = 1...n$ are samples from the training group and $|w|_2$ the euclidean norm of w. Using Lagrangian multipliers the problem reduces to:

$$min_{w,b}\frac{1}{2}|w|_2^2 - \sum_{i=1}^{n}\alpha_i y_i(x_i \cdot w + b) + \sum_{i=1}^{n}\alpha_i \tag{3.26}$$

$\alpha_i \geq 0, \forall i$

Alternatively the dual form of the problem can be solved

$$\begin{aligned} max_\alpha \frac{1}{2}\sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i, x_j) \\ \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{3.27}$$

$\alpha_i \geq 0, \forall i$, where $k(x_i, x_j)$ is the linear kernel $x_i^T x_j$ and

$$w = \sum_{i=1}^{n}\alpha_i y_i x_i \tag{3.28}$$

The dual from problem is useful when the data is not linearly separable. In this case, using a kernel [309] it can be mapped in a space with a higher number of dimensions where it becomes linearly separable. The number of Lagrange multipliers is the same as the number of training samples, but most of them will have a value of 0, so their corresponding data points will not contribute to the classification of new samples. The rest of the data points will satisfy $y_i(w^T x_i + b) = 1$ and are called support vectors, giving the name of the method.

3.5.2 Support Vector Machines with Redundant Feature Elimination

As support vector machines proved to be a successful tool in classification done using gene expression [50], improvements have been proposed by the creators of the method to optimize the selection of biomarkers [140]. The new algorithm was called SVM-RFE and its purpose was to reduce as much as possible the number of genes used in classification while at the same time keep the classification capability of the obtained model.

The algorithm works by recursively following the following steps:

1. Train on the training set to obtain a vector of weights w
2. Assign a ranking for each feature.

   The ranking scoring methods proposed were w_i^2 and DJ(i)[192] where DJ(i) is the change in the objective function(for example the classification rate of the algorithm) when the weight for the respective feature is set to 0.
3. Remove the feature with lowest ranking.

In order to speed up the algorithm more than one feature could be eliminated.

3.5.3 *l*1-Star

Based on the ideas presented in the Guyon et al. [140] paper, Ahsen et al. [5] proposed a new hybrid algorithm that uses a statistical test followed by SVM-RFE called *l*1-Star. One significant difference between *l*1-Star and its predecessor, is the use of an l1-norm when minimizing the weights of the features for the classifier as in the work of Zhu et al. [406]. The problem to be solved now for the classifier is:

$$min_{w,b}|w|_1 \tag{3.29}$$

subject to $y_i(w^T x_i + b) \geq 1$ with $|w|_1$ the Manhattan norm for w. Below there are the steps of the algorithm:

1. Create a random training set with similar number of elements which is less or equal to half the number of elements from both classes and apply the l1-norm support vector machine on it.

2. Repeat step 1 several times. The authors found that for 80 randomized samples or 1000 randomized samples the results were comparable.

3. Average the weight vectors across all the classifiers. For k the average number of nonzero elements in the weight vectors of each classifier, pick the k features with highest weights.

4. Repeat steps 1-3 until no reduction is possible

5. For the final feature selection repeat step 1 and build a classifier from the weights of the top 20 classifiers

3.6 Cluster Analysis

Cluster analysis represents a set of statistical methods in which objects are assigned classes based on similarity Sadesky [303]. A high number of clustering methods exist a review being available in Berkhin et al. [33], however 2 of them are widely used, k-means clustering and hierarchical clustering.

Two main categories exist for hierarchical clustering analysis [38], agglomerative and divisive. In the agglomerative methods, every element starts as a cluster and the algorithm joins the 2 closest clusters as measured by a distance until only one cluster exists. Divisive methods use a reverse strategy, all features are considered to be in one cluster which is split in its most dissimilar parts until each feature becomes its own cluster. The rules of pairing of features representing the outcome of hierarchical clustering form what is called a tree which is visually represented by a dendrogram.

K-means clustering [230], is a method based on centroids which represent reference points situated in the middle of the desired clusters. Each new individual observation's distance to all centroids is calculated and it joins the cluster of the closest centroid. The affected centroid's position gets updated considering the new value. The method relies heavily on a good choice of initial centroids, a problem that is difficult to solve when it comes to gene expression analysis. As a result hierarchical methods are preferred for this applications so the rest of the review will focus on them. Also agglomerative methods consume less resources from a computational point of view, which makes them better fit when working with the high number of features analysis of gene expressions involves. Two elements are

of importance when using agglomerative hierarchical clustering, the metric and the linkage rule. The metric represents the measure by which the distance between the features is computed. For 2 vectors, x and y, standard metrics are:

- Euclidean distance
$$dE = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.30}$$

- Normalized Euclidean distance
$$dNE = \frac{dE}{\sqrt{n}} \tag{3.31}$$
where dE is the Euclidean distance and n the number of elements in the vectors

- Manhattan distance
$$dM = \sum_{i=1}^{n}|x_i - y_i| \tag{3.32}$$

- Pearson's correlation coefficient dP=1-r,
$$r = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \overline{x}}{\sigma_x})(\frac{y_i - \overline{y}}{\sigma_y}) \tag{3.33}$$
where $\overline{x}$ and $\overline{y}$ are the means of the element in the vectors x and y and σ_x and σ_y are their standard deviations.

- Absolute Pearson's correlation coefficient dAP = 1-|r| with r defined above

The linkage rule determines what is considered the distance between 2 different clusters with multiple features. For 2 clusters, A and B possible linkage rules between them can be:

- Maximum distance
$$D = max\{d(a,b) : a \in A, b \in B\} \tag{3.34}$$

- Minimum distance
$$D = min\{d(a,b) : a \in A, b \in B\} \tag{3.35}$$

- Average distance

$$\frac{1}{card(A) * card(B)} \sum_{a \in A} \sum_{b \in B} d(a,b) \tag{3.36}$$

In the case of agglomerative clustering a cutting method is needed in order to split the obtained tree into different entities. One simple way to do this is to cut the tree at a certain height, which would represent the maximum distance of the elements in the obtained clusters. Another method is to cut the tree based on an inconsistency value which compares how different the 2 sub clusters that form a cluster are.

For both of these methods a threshold needs to be set which can represent a difficult task. Langfelder et al. [206] proposed an automated method that permits the split of a tree in clusters without requiring a predefined threshold.

For a tree obtained through agglomerative clustering, the leafs represent the features which can be assigned a height that represents the distance between it and the closest cluster. Thus a tree can be written as a vector of heights. Dynamic tree cut is based on the function TreecutCore(). This starts by subtracting from each element in the vector a reference height l. Some heights will now be negative while other will be positive. The function defines as transition points all the elements representing a positive height with the next element a negative height. Next for each transition point it looks at its precursors and defines the first negative one as a breaking point. The new clusters are represented by the features between a transition point and a breaking point. In practice the algorithm uses 3 values for the reference height l:

$$\begin{aligned} l_m &= \frac{1}{n} \sum_{i=1}^{n} h_i \\ l_u &= \frac{1}{2}(l_m + max\{h_1, h_2, \ldots, h_n\}) \\ l_d &= \frac{1}{2}(l_m + min\{h_1, h_2, \ldots, h_n\}) \end{aligned} \tag{3.37}$$

where $\{h_1, h_2, ..., h_n\}$ the vector of heights for a cluster They are utilized in a procedure called AdaptiveTreecutCore() with the following steps:

1. For a cluster calculate l_m, l_u, l_d
2. Apply TreecutCore() with l_m as reference height

3. If at Step 2 no new cluster appears, apply TreecutCore() with l_d as reference height

4. If no new clusters appear at Step 2 or 3, apply TreecutCore() with l_u as reference height

The complete algorithm works following the next steps:

1. Cut the initial tree at a very high height to cut it in 2-3 clusters

2. For each cluster apply AdaptiveTreecutCore()

3. If new clusters are created update the list of clusters and repeat step 2

4. If no more clusters are created return the obtained clusters

3.7 Gene Regulatory Networks

Thousands of genes are encoded by the genome, their products enabling cell survival and numerous cellular functions [183]. The quantity and the moments in which these products appear in the cell are crucial to the functionality of the organism [183]. In order for this synchronization to happen, the gene need to interact with each other. The accumulation of these interactions form a gene regulatory network (GRN) [220]. By discovering the underlying structure of a GRN, new information can be obtained on the functionality of the cell and mechanisms of diseases can be studied. This in turn can lead to the development of new therapeutic strategies, so there is lots of interest in researching them.

Two main approaches exist at this moment for the identification of gene regulatory networks. The first approach consists in creating biochemical models for the genes interaction in which the underlying chemistry of gene interaction is taken into consideration. The advantage of this method consist in the reliability of the networks identified, once the scientists discover them, they get to be modified in time just by addition of newly discovered interactions [288]. Their disadvantage consists in the significant prior knowledge of the system the researchers needed when applying them [288] which makes them unfit for exploring under-examined problems. A more through presentation on them is outside the scope of this literature review but the interested reader can study the review by van Riel and Sontag [367]. The second approach is to infer the structure of the network of a set of genes using

microarray measurements. Its advantage consists in the fact that no knowledge of the network topology is needed [288] but biological knowledge of known regulators can be included when identifying it.

Identification of gene regulatory networks using microarray measurements can be done in 2 ways, using static models that just account for possible interactions between genes or dynamical models which more than just identifying the structure of the network allow the scientists to estimate future behaviour of the system represented by the genes. The following sections will provide a brief overview of the static models and a more in depth description of the dynamical ones.

3.7.1 Static Models

The simplest static method for identifying the underlying network for a set of genes based on microarray data measurements is represented by calculating the correlations between different features based on their measurements as used by Eisen et al. [109] followed by a thresholding on the correlation values. An improvement to the method is represented by the use of weighted gene co-expression networks [403] in which rather than thresholding the genes they are assigned a weight on their relation based on the value of their correlation coefficient raised to a specific power.

The problem with using correlation as a measure of genes similarity is that it cannot quantify non-linear relations between them. In order to solve this problem, mutual information was proposed as a measure of similarity between genes. Similar to the case of correlation networks the simplest way in which mutual information can be applied to retrieve the topology of a GRN is to calculate it for all pairs of genes and then threshold it with a fixed value [54]. Improvements to this method have led to various popular algorithms such as Minimum redundancy network (MRNET) [299] in which feature selection is done based on the mRMR criterion, the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Margolin 2006) in which connections due to indirect interactions between genes are eliminated using the Data Processing Inequality (DPI) [80] criterion or the Context Likelihood of Relatedness(CLR) algorithm in which the score of a pair of genes is calculated based on their MI score and the complete distribution of MI scores in the network.

A different approach for the identification of static models in gene regulatory networks is the use of machine learning methods. A good example in this sense

is represented by Supervised Inference of Regulatory Networks (SIRENE) [253] in which support vector machines are trained using known connections between genes to identify new connections based on the microarray measurements. The algorithm showed to be reliable, outperforming CLR [253] but its disadvantage consists in the need of knowing previously biologically validated connections between genes which makes it unfit for dealing with novel biological issues. A solution to this problem is represented by GENIE3 [170], an unsupervised algorithm in which Random forests are used to identify genes with a good power for predicting the values of other features. GENIE3 has shown great performance in identifying biologically validated connections, winning the Dialogue for Reverse Engineering Assessments and Methods (DREAM) competition [135] in 2009.

In order to identify the regulations that exist between genes while taking into consideration the stochastic nature of biological systems, static Bayesian networks were proposed to infer regulations between genes. In this approach to the problem, each gene is considered a node in a Bayesian network which describes the probability of a gene to take a specific value considering the values of other related genes at the same moment. If edges exist in the Bayesian network between 2 genes it is considered that regulation exist between them in the real regulatory network. The first proposal for using this approach to identify regulatory networks appeared in a paper by Friedman et al. [125]. As the search space for Bayesian network models of gene regulatory networks for a large number of genes is extremely vast various methods have been employed to narrow the search and pick a good model. In their study Friedman et al. [126] used bootstrapping for evaluating the models but other examples in this sense are represented by the use biological data to reduce the structures considered for the network [278] or using the BIC criterion [311] to penalize a large number of connections [296].

3.7.2 Dynamical models

Various models have been proposed to capture the dynamical behaviour of gene regulatory networks, thorough reviews that present them being available [183] [373]. As it is revealed in these studies, there are 4 main categories of dynamical models used in GRN identification: logical models, continuous models, single molecule level models and hybrid models that are a combination of the above. This review will focus on logical and continuous models sing linear differential equations. Com-

plete reviews on the problem are presented in the work of Karlebach and Shamir [183], Hecker et al. [148], Vijesh et al. [373], Le Novere [210].

Logical models only provide a qualitative view on the underlying network, the only available information being of the kind gene A influences gene B but no details are given on the behaviour that is influenced by the exact molecular concentrations as well as the type of regulation taking place (upregulated or downregulated).

The most used type of logical models is represented by Boolean networks (BN) [354] and Probabilistic Boolean Networks (PBN) [320]. Boolean networks consist of a set of rules that model the logical interactions between genes. Each gene is modelled as a logical switch that can be true or false. This introduces the first problem with Boolean networks which is to find a good threshold for the values of the genes. The state of the system at some point is represented by the vector of logical states for each gene. The set of rules that govern the network is inferred by checking which set of logical operations that transfer the systems from the state at time t to the state at time t+1 works on all the data. This introduces the second problem with the model, which is the need to treat time as a discrete variable and that the system evolution is considered a markovian process. Two algorithms have been proposed for dealing with the finding of the logical operations, REVEAL [216] in which the search-space is reduced by using the mutual information between the initial state of each single gene (t = 1) and the output state of the target gene. Another approach to the problem have been proposed by Akutsu et al. [6] who proves that $O(log_n)$ random measurements are enough for the identification of a network of N genes and proposes an exhaustive search method. The advantage in this case is that all the networks that can be associated with the data are identified. More advanced techniques for identification of Boolean networks by turning their logical rules into their algebraic form have been proposed in Cheng et al. [72] . An important aspect in the analysis of Boolean networks are the attractors. These are represented by states or chains of states that once are entered repeat infinitely, called limited cycle attractors, which represent the steady states of the dynamic system being used to understand the long-term behaviour of the Boolean models [393]. Also researchers have proved that the attractors can be associated to the cellular phenotype [158], meaning that attractors should be studied when we are interested in changing the long term behaviour of cells [393] as for example when trying to cure a disease.

A strong criticism of Boolean Networks is their deterministic nature. It seems unrealistic that the expression of genes is governed by constant logical rules rather

than a self- organizing stability of the dynamical system represented by the regulatory networks [320]. To address this problems, probabilistic Boolean networks have been introduced in which at every moment in time a boolean network from a predetermined set is selected with a specific probability to predict the new values for the genes. For a PBN , the attractors are represented by the set of attractors of all the Boolean networks that can be selected at a specific time [393]. The dynamic behaviour of PBNs can be represented as Markov chains [320]. The concept of attractors can be represented by an irreducible set of states in the Markov chain. A new issue compared to the deterministic Boolean networks is the study of steady state distributions which regard the probability of the system of getting into an attractor, independent of the initial state. As in the case of Boolean networks attractors represent phenotypes of the cells so the steady state distributions show the chances of cells to be of a specific phenotype. A question of interest in this case is how to push the system towards a desired behaviour. As shown in Pal et al. [272], answers to it have been proposed in Shmulevich et al. [322] by keeping the distributions constant but resetting the initial state to a more desirable one, by making minimal modifications to the steady state distribution [321] or changing external (control) variables that influence the transition probabilities of an instantaneously random PBN to steer its evolution towards a desired goal for a finite amount of time [84][85].

An improvement to the probabilistic Boolean models is represented by Bayesian network models [125] . These are similar to the static Bayesian models but the difference consists in the fact that current values of the gene expressions are used to predict the expression of a gene at the next time-step rather than the expression at the same time-point. The advantage of this approach is that the measurements do not have to be turned into binary values and weights are obtained to quantify the influence of genes on each other.

In order to approximate the parameters for dynamical Bayesian networks, various methods have been proposed for both continuous and discrete systems presented in Friedman et al. [127], Murphy and Russell [255], and Perrin et al. [281]. In order to find a good model for a set of available data, a search strategy is needed to generate new models that might describe the data and an evaluation function to score the proposed models should be used. Over the years various search strategies such as hill climbing [86], simulated annealing [186] or the junction tree algorithm [208] have been used to identify the parameters of Bayesian network models for gene regulatory networks. Evaluation methods that have been used over the years are

Bayesian scoring functions such as the Bayesian Dirichlet evaluation score (BD) [149], and its variants BDe [149] and BDeu[149] or scoring function coming from information theory such as the Bayes information criterion [311] or the newer Mutual information tests score (MIT)[57]. A presentation of the different scoring functions is available in Carvalho [58].

State space models have been developed for the description of gene regulatory networks in which linear differential equations are used to model the interactions between genes. Two main categories of state-space models exist, linear and nonlinear. Linear systems can characterize the linear relations in the systems but they cannot capture the nonlinear dynamic aspects of gene regularity which means that for higher sensitivity approximating the network there is a need for more complex models. As the complexity of nonlinear differential equations is high and reverse engineering such a system containing many variable is an extremely complex task, a good compromise solution should be found. This would be to approximate a nonlinear dynamical system around its equilibrium points using a linear state-space model based on the theory of Jacobian linearisation [271]. In general, a linear state-space model with the input $u(t) \in \mathbb{R}^m$ and output $y(t) \in \mathbb{R}^n$ can be represented as:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{3.38}$$

where $t \in \mathbb{R}$ is the time, $x(t) \in \mathbb{R}^p$ is the state vector, p is the model order, A is the system matrix, B is the input matrix, C is the output matrix and D is the feed-through matrix. A simplification of this model to be used in gene regulatory network analysis was proposed by D'haeseleer et al. [91] as the autoregressive model:

$$\dot{x}(t) = Ax(t) \tag{3.39}$$

where x(t) represents the gene expression measurements at a certain time point and A represents a matrix of constant parameters. This approach was used in the study of various mechanisms such as the Central Nervous System (CNS) development [91], yeast cell cycle [155], or with the addition of a perturbance matrix in the study of DNA-damage response pathway (SOS pathway) in the bacteria E. coli [19] [401]. Early methods have been proposed for approximating the parameters of the linear system of equations describing the networks. At first they consisted of least squares [91], simulated annealing if the measurements are not equally spaced in

time [155], and the Broyden-Fletcher-Goldfarb-Shanno algorithm [119] in solving linear models with time delays [189]. The problem with this approaches is that when they approximate the parameters they do not impose constraints specific to the gene regulatory networks. GRNs show loose connectivity [70], so that a sparse system matrix should be used and since a biologically realistic system is expected to be stable [373] a stability constraint is needed for the identification of the system. Sparsity is easy to impose, one solution being to minimize the l-1 norm of the matrix A [97]. Stability however cannot be imposed using linear programming techniques, which makes it a problem that is harder to solve. One solution for turning it into a linear programming problem is the use of Gershgorin theory [369]. This theory states that the eigenvalues of a matrix $A \in \mathbb{R}^{nxn}$ can be found in the circles having the centres C_i and radii R_i, i = 1...n with:

$$\begin{aligned} C_i &= a_{ii}, a_{ii} \in diag(A) \\ R_i &= \sum_{j=1, j\neq i}^{n} |a_{ij}|, a_{ij} \in A \end{aligned} \tag{3.40}$$

The direct consequence is that if $C_i < -R_i$ for all i = 1...n, all the eigenvalues will take negative values. The theory was used in practice [220][387], but its disadvantage is that it is just a sufficient condition for stability which means that possible good solutions will be overlooked. Another approach to this problem is the imposing of lyapunov stability criteria [227] which establishes that a necessary and sufficient condition for negative values of the real part of the eigenvalues of a matrix A is the existence of a symmetrical positive definite matrix P so that

$$A^T P + PA \prec 0 \tag{3.41}$$

Semi definite programming [368] comes with the solution for imposing the condition above. A method to infer GRNs based on Lyapunov stability criterion was proposed in [401] and the authors test it against a Gershgorin based method which it surpasses. The algorithm is described below:

1. Find a sparse matrix A that approximates well the data using adaptive lasso

   For X and $\dot{X}$ obtained by sampling as well as t, $0 \leq t \leq 1$, initialize weights $w_{ij} = 1$ for all i,j =1...n

   for it=1:10

Solve the convex problem:

$$\underset{A}{minimize}\, t\sum_{i,j=1}^{n} w_{ij}|a_{ij}| + (1-t)\varepsilon \tag{3.42}$$

subject to $\|\dot{X} - AX\|_1 \leq \varepsilon, \varepsilon > 0$

Update weights $w_{ij} = 0.01/(0.01 + |a_{ij}|)$

end for

2. Find a matrix $A' = A + D$ where D is a disturbance matrix and a Lyapunov matrix P, which satisfies Lyapunov criterion $A'^T P + PA' \prec 0$

   Solve the semidefinite problem:

$$\underset{P}{minimize}\, \|LX\|_2 \tag{3.43}$$

   subject to $A^T P + L^T + PA + L \prec 0, P \succeq 0$

   where L = PD and P symmetrical

3. Find a sparse stable matrix A_f using P found at the previous step For X and $\dot{X}$ obtained by sampling as well as P found at the previous step and t, $0 \leq t \leq 1$, initialize weights $w_{ij} = 1$ for all $i, j = 1...n$

   for it=1:10

   Solve the convex problem:

$$\underset{A}{minimize}\, t\sum_{i,j=1}^{n} w_{ij}|a_{ij}| + (1-t)\varepsilon \tag{3.44}$$

   subject to $\|\dot{X} - AX\|_1 \leq \varepsilon, \varepsilon > 0$

   $A^T P + PA \preceq 0$

   Update weights $w_i j = 0.01/(0.01 + |a_i j|)$ end for

3.8 Summary

This chapter provided an overview on the microarray technology and its use in the study of diseases. It includes a description of the microarray technology, pre-

processing methods for cleaning the data of noise, feature selection and clustering for finding biomarkers for the diseases and finally a description of gene regulatory networks and methods to approximate them from microarray data.

The chapter started by presenting the available technology at the moment from the simple but flexible spotted arrays which can be produce in small research labs to the more technologically advanced platforms provided by Affimetrix and the extremely high density arrays produced by Illumina. The production methods and the protocol for extracting data from biological samples was described for each of the platforms.

The next section reviewed pre-processing methods for extracting the real signal from the measured data. Techniques for background correction, normalization, summarisation and batch corrections have been presented and compared, in respect to the microarray technology they serve best.

Once the data is as noise-free as possible, an analysis to select the biomarkers, the genes that are affected by a disease is an important step for understanding its behaviour. In this chapter techniques that do this, that can be separated into 2 main groups: feature selection and clustering methods have been reviewed. In the case of feature selection, strategies from statistics and information theory have been presented together with the more advanced machine learning methods. For statistical methods, the review highlights the importance of introducing correction for multiple testing. The information theory based methods are given a brief review in which the most significant ones have their strengths and weaknesses presented. As machine learning methods for feature selection are more complex an entire section has been dedicated to their description together with the feature selection methods they are employed on. The 2 most popular clustering techniques, hierarchical clustering followed by a cutting method and the k-means algorithm are presented together with recommendations on their use with genetic data.

Finally, the importance of understanding disease behaviour not as a function of a few isolated genes but as a result of complex interactions in a gene regulatory network is highlighted. The biological constraints for these networks are outlined and different techniques through which they can be imposed when identifying the networks are compared based on their performance and the challenges they try to address.

Chapter 4

Migration and Textural Analysis in Cells Studies

4.1 Introduction

Study of biology by microscopy is almost as old as the invention of microscopes, the first person to create a real microscope with a magnification power of over 200x, Anton van Leeuwenhoek, used it to study yeast and bacteria [111]. His work was continued by Robert Hooke [156] who analysed corks, discovering for the first time the cell of a plant and also coined the term.

In modern times, microscopy is extensively used in biology, for cell counting [265], analysis of cell migration and cell invasion [181], studying the positioning of organelles in the cells [44], localization of gene expression [234], studying structures created by cells in 3D cultures [147], cell adhesion [187] or for analysing the texture of the cells for characterization and classification [261]. The different purposes of microscopy studies has created various methods for extracting information from cell images, either by studying cells orientation and cell trajectories [29], or using the latest developments in image processing and computer vision to extract different features that characterizes the texture of the cells [93].

Cell microscopy analysis has been extensively used in the study of ADPKD [66] as the genes responsible for the disease have been shown to affect cell migration [259], cell adhesion [324], orientation of the Golgi apparatus [60] and the cilia structure [399].

This review will focus on the subjects of cell migration and texture analysis.

First, a description of the present technology for the analysis of cell migration using different assays will be provided in Section 4.2, Cell migration assays. Next, section 4.3 Measures and models for cell motility introduces the measures used to characterize the movement of the cells. In section 4.4, studies involving ADPKD where the different technologies and measures for cell motility have been applied are provided a short summary with a focus on the methods they use. Section 4.5 Texture analysis for cell studies, different texture analysis methods are presented and their application to the analysis of cells are listed, while their advantages and disadvantages are highlighted. Last, section 4.6 Conclusions summarises the main points of this review.

4.2 Cell Migration Assays

As the study of cell migration has become more widespread, revealing different aspects of disease mechanisms and normal cell functions, different assays have been proposed for various study goals. In this review, 2 types of assays will be presented, (1) filter-based assays in which cells migrate across a cell-permeable barrier towards a defined chemo-attractant and their migratory capacity quantified by the number of migrating cells over a period of time and (2) time-lapse microscopy where cell behaviour is continuously monitored leading to detailed descriptions of their migration patterns. Other types of assays used in research on cell movement can be found in papers by Entschladen et al. [112] and Kramer et al. [200].

The following subsections present the 2 approaches as well as applications for which they were used, subsection 4.2.1 presents the transwell filter assays and subsection 4.2.2, assays compatible with time-lapse microscopy.

4.2.1 Filter Assays

The oldest and most used filter assay, based on which newer technologies have been developed is the transwell migration assay or the Boyden chamber [49](Figure 4.1). It consists of 2 chambers, one of them enclosing the other, which are separated by a porous membrane. The 2 chambers are filled with different types of media so that the substance in the external well is more attractive to cells than the one in the internal well. Examples in this sense include the use of serum free media in the inner chamber and normal media in the external chamber or normal media in the

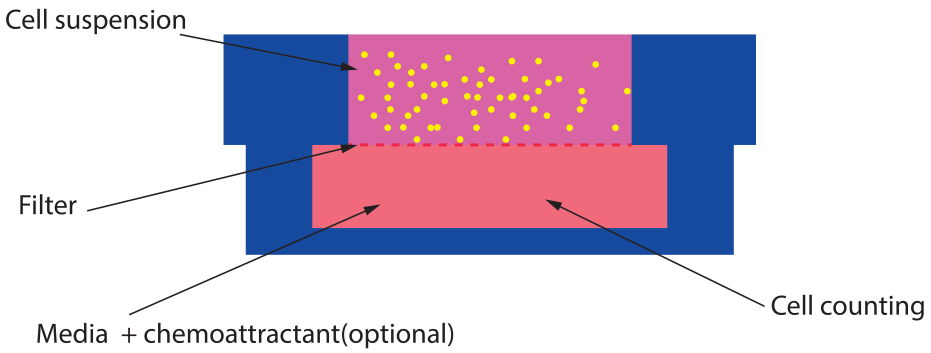

Figure 4.1: Schematic representation of a Boyden chamber

inner chamber and media combined with a chemoattractant in the external chamber. At the beginning of the experiment, cells are placed in the inner chamber and are left for a specific amount of time to pass through the porous membrane. After the time has passed, the cells that migrate to the other side of the membrane are then quantified. Two possibilities exist for this operation. The first one uses a transparent membrane and consists in removing the cells remaining in the side of the membrane corresponding to the inner chamber, fix the cells that have migrated, stain them with cytological dyes and count them.

The other method is to use a dark coloured membrane which does not let light pass through it. In this case, there is no need to remove the cells which did not migrate, just stain the side of the membrane corresponding to the outer chamber and count the cells on it. For different types of cells, membranes with corresponding pore sizes should be used, the ones that are generally employed having pore diameters ranging between 3 and 12 μm. Through the years, Boyden chambers have been used to test how cells respond to a specific substances such as antibodies [49], N-formylmethionyl peptides [308] or platelet- derived growth factor [313]. Other applications consist in studying the migration capabilities of cells affected by a disease [252][371][384] or test the capacity of various substances to enhance cell mobility [364][370][166].

The advantage of this filtering approach consists in the fact that it is cheap and easy to use, few resources being involved in the experimental setup as well as in the analysis of data. The disadvantage of the method is that no information is produced about the intermediary steps of cell migration and no details about the directionality

of the cells can be extracted.

4.2.2 Time Lapse Assays

The simplest assay for time lapse analysis is to grow sparsely distributed cells on a glass or plastic surface or to use a matrix-coated surface that recreates conditions closer to the in-vivo ones [112]. Then a microscope can be used to take photos of them at certain intervals (10-20 minutes) generally at different positions on the plate. This concept allows for a detailed analysis of their individual trajectories. As the free movement of cells presents Brownian motion patterns [94], different models have been proposed to extract information from them and will be detailed in the following section. This approach has been used to investigate T cell migration considering regulation of myosin [327], adapted stem cells migration [20] or for investigation of the role of the Arp2/3 complex in cell migration [344].

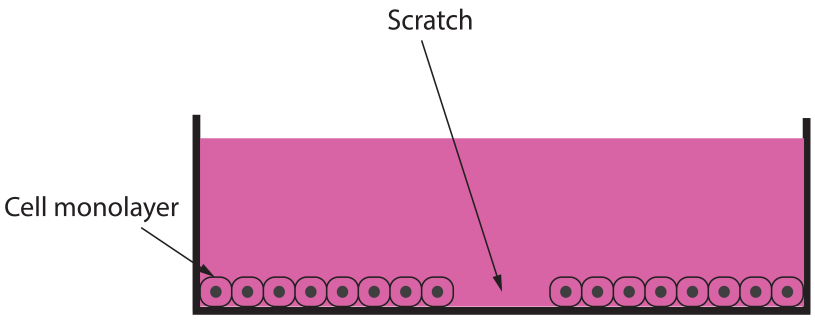

Figure 4.2: Schematic representation of a scratch assay

A modification of the previous assay is represented by a wound closure or scratch assay (Figure 4.2). In it, cells, usually epithelial or endothelial [112] are grown in a Petri dish or a multi-well plate until they reach 100% confluence. Next the cell layer formed this way is scratched with a pipette tip although more advanced setups allow for the wound to be made using an electrical signal [185]. Once the scratch has been made, a microscope can be used to film closure of the wound. This set-up allows for researchers to study the rate at which different cells are capable to close the wound, and allows for the quantification of directionality in cell migration, using the wound as a reference. One problem with the pipette tip scratching approach is that the wound may not be symmetrical and equal in diameter throughout

the wound. An improved version of the cell scratch assay that solves the problem is represented by the cell exclusion zone assay (Figure 4.3). In it after the scratch has been made, a barrier consisting of microstencils [286] is placed in the wound. This allows the cells to grow until the walls of the wound are perfectly symmetrical. In practice, wound healing assays have been used for the study of skin wound healing by mesenchymal stem cells [374], effects of Calendula extracts in wound healing [128], or the effect of HEF 1 on the migration capabilities of glioblastoma cells.

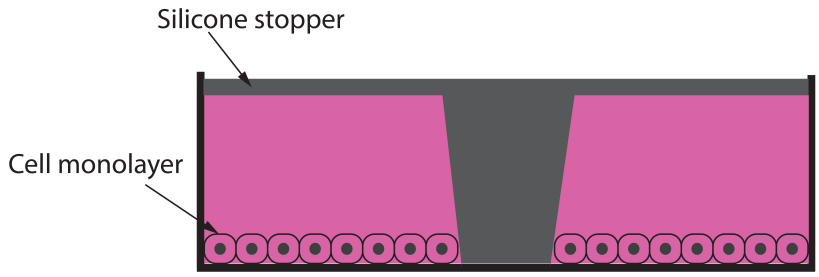

Figure 4.3: Schematic representation of an exclusion zone assay using a separator

Another assay which can be used for time lapse studies on cell migration is the fence assay or the ring assay (Figure 4.4). Its concept can be considered in opposition to the cell exclusion zone assay as a delimiter such as a Teflon glass or metal ring is used to encompass an area in which cells are grown. Once they have reached full confluence, the barrier is removed and the cells are filmed spreading away from the central point of culture. Their spread is quantified by measuring the area that they occupy at different points in time. This approach has been taken to quantify the migration properties of human endothelial cells [56] , or to study the importance of the Angiotensin system on endothelial and smooth bovine aortic muscle cell migration [25].

Techniques using the same concept as Boyden chambers but allowing time-lapse analysis are represented by capillary chamber migration assays. In this approach, 2 chambers are connected by a narrow bridge. The cells are then placed in one of the chambers and a chemoattractant is placed in the other chamber. After a period of time, the cells that migrate are counted on the separation bridge. Two popular implementations of the concept are represented by Zigmond chambers [408] in which

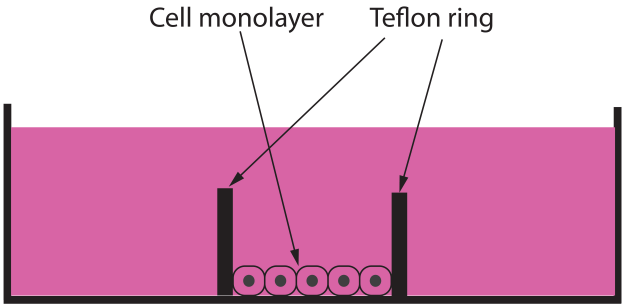

Figure 4.4: Schematic representation of a fence assay

the chambers are arranged horizontally side by side and Dunn chambers[407] (Figure 4.5) in which they are concentric. The main reason for the development of these assays is represented by the fact that very low volumes are needed for the experiment, making them suitable for working with rare and expensive substances and cell types [200]. These assays are frequently used for studying leukocytes [8].

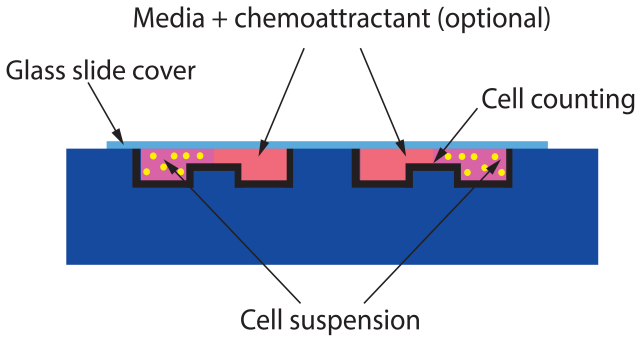

Figure 4.5: Schematic representation of a Dunn's chamber

The advantages of using assays that are compatible with time-lapse analysis consists in the fact that details about the migration of cells can be extracted, leading to accurate characterization of their behaviour. Their disadvantages consist in the availability of resources (eg. microscope time) and in the complexity of their analysis, having to deal with cell tracking which is extremely time consuming in the case of manual methods and a difficult problem without a perfect solution in the case of the automatic methods.

4.3 Measures and Models for Cell Motility

With the increase in the popularity of time lapse studies, in which the trajectories of the cells can be reliably extracted, different methods have been proposed in order to extract as much information as possible about the movements of the cells. These can be measures in which the information is directly computed from the cell trajectories or modelling methods in which features are extracted indirectly from the measurements by fitting specific models. The following 2 sections present methods for the 2 approaches. Section 4.3.1 introduces the measures used to characterize cell movements while section 4.3.2 presents the models used to achieve this goal.

4.3.1 Measures for the Movement of Cells

In their paper, Meijering et al. [240] propose a taxonomy which separates methods for quantifying movement of cells. The authors suggest three different categories: motility measures in which just the positions of cells are used for computing a feature of interest, velocity measures in which time is used explicitly in characterizing the movement of cells and diffusivity measures which use methods from the theory of diffusion and Brownian motion to characterise the movement of cells. The following subsections present the popular measures used in the study of cell movement separated in the 3 categories presented. Section 4.3.1.1 discusses the motility methods, section 4.3.1.2 the velocity measures and section 4.3.1.3 the diffusivity measures.

4.3.1.1 Motility Measures

The simplest way to analyse the motility of cells is to just plot the tracks for the cells using their centres of mass for quantifying their position at a specific timepoint. Two approaches to this method have been used in literature. The first one is to plot the cell trajectories overlaid on the original images of the cells. This method is useful in studies where scientists try to discover if cells are moving towards a certain region of importance. In practice it has been used to study the chemotaxis of breast cancer cell towards different gradients of EGF(Epidermal Growth Factor) [375] and to analyse the effect of Laminin-1 on motility of rat and mouse Muller glial cells [239]. The problem with this approach is that in general it is difficult for a scientist to visualize the spread of the directions in which the cells move when they

have random starting points in an image. In order to tackle this problem, a second approach to trajectory visualization has been proposed in which all the trajectories are shifted to have the same starting point. This enables the scientist to quickly visualize if there is a general trend in the direction in which the cells move and also to assess if there is a significant difference in how far they travel between different experimental conditions. This approach is extremely popular and has been used in various studies, examples being the analysis of stress-induced endothelial cell polarization [385], effects of IGF-IR inhibition and ROS accumulation on glioma cells motility [98] and antioxidant and anticancer effect of extracts obtained from Chenopodium quinoa leaves [132].

The visual methods only provide the means of a qualitative analysis on the cell movements [29] and can be employed when differences between conditions are very noticeable. In order to provide a quantitative analysis, capable of capturing more subtle changes in the movement of cells various methods have been proposed.

For quantifying the distances travelled by cells, 3 methods have been commonly used in the literature:

Total distance travelled defined as:

$$d_{tot} = \sum_{i=1}^{N-1} d(p_i, p_{i+1}) \tag{4.1}$$

Net distance travelled:

$$d_{net} = d(p_1, p_N) \tag{4.2}$$

Maximum distance travelled:

$$d_{max} = max_i(d(p_1, p_i)) \tag{4.3}$$

where p_i is the position of the cell centre of mass in frame i of the movie of the cells with N frames and d is a measure of distance, generally Euclidean distance being used.

A popular measure for the characterization of the linearity of cell movement that is derived directly from distance measures is the confinement ratio also known as the chemotactic index or straightness index which is defined as:

$$r_{con} = d_{net}/d_{tot} \tag{4.4}$$

As $d_{net} \leq d_{tot}$ the ratio will be between 0 and 1, a value of 0 meaning the cell returns at exactly the same position while a value of 1 that it has a perfectly straight trajectory. This measures have been used in various studies, examples in this sense being comparisons of interactions of olfactory ensheating cells and Schwann cells with astrocytes[203], identification of genes that regulate epithelial cell migration[325] and the roles of MLCK and ROCK in cell migration[359].

A different approach to characterize motility is the study of migration angles. This can be done by using the angles of cell positioning relative to previous positions or the angle of their position relative to a point of interest.

For the first method, an instantaneous angle has to be calculated which is defined as:

$$\alpha_i = arctan(y_{i+1} - y_i)/(x_{i+1} - x_i) \tag{4.5}$$

here x_i, y_i are the positions of the cell in frame i relative to the x and y axis. Once the instantaneous angles are obtained, in order to characterize the movement the researchers can plot their distribution. In the case of a 2D cell culture if the movement of cells is random the distribution is expected to be uniform ([29]) while directed movement will show higher rates of presence for a specific range of angles. For a quantitative estimation of the cells preference for a certain direction, the average angle can be calculated. In the case of random walk this would be 90 degrees. This approach was taken in Beltman et al. [28] to characterize movements of T cells in lymph nodes. Another application was the analysis of the effects of CXCL8/Interleukin-8 on cell migration [124]. A secondary measure that can be extracted from instantaneous angles is the change in direction which can be calculated using:

$$\gamma_i = \alpha_i - \alpha_{i-1} \tag{4.6}$$

For the relative position to a point of interest a similar approach can be taken but the change in direction is calculated as:

$$\theta_i = \alpha_i - \beta_i \tag{4.7}$$

where

$$\beta_i = arctan(y_i - y_r)/(x_i - x_r) \tag{4.8}$$

with x_r, y_r the coordinates of the reference point.

This measure is useful when combined with the average cell speed in order to quantify how far cells can travel within a limited time interval [29].

4.3.1.2 Velocity Measures

The most common used velocity measure in time-lapse cell motility study is the instantaneous velocity calculated as:

$$v_i = \frac{d(p_i, p_{i+1})}{\Delta t} \tag{4.9}$$

where Δt is the period of time between 2 frames. A measure that can be easily derived from the instant velocity is the arrest coefficient[29] which is the period of time for which cells move below a certain speed. Another measurement is the average instantaneous velocity. Other velocity that can be used are the mean straight-line speed [240] which is defined as:

$$v_{lin} = \frac{d_n et}{T} \tag{4.10}$$

where T is the total duration of the tracking of the cell. Velocity measures have been used to quantify the effects of the Shc and Fak on cell motility [136] or to analyse the effects of myosin II on cell migration[113].

4.3.1.3 Diffusivity Measures

A different type of measurement that is used in the characterization of cell motility is represented by the mean squared displacement (MSD) [108] which is calculated for a population of cells as:

$$MSD(t) = \frac{1}{N} \sum_{n=1}^{N} (x_n(t) - x_n(0))^2 \tag{4.11}$$

with $x_n(t)$ the position of cell number n at time t , N the total number of cells. Although this is a more complex measure than the ones discussed above, it has the advantage that different models can be fit to the resulting MSD curve so various features can be used to characterize the motion of the cells. These are discussed in more detail in the next section.

4.3.2 Diffusivity models

The mean squared displacement is an important measure in science being used in fields such as geophysics for tracer diffusion in subsurface hydrology [341], diffusion of substances inside the cells [238], cell movement [94], or movement of animals [13]. In order to characterize movement in all of these situations, various models have been proposed. A description of all models and their applications is outside the scope of this chapter but thorough reviews such those done by Metzler and Klafter [244][245][191] or more applied to the field of biology [75] are available in the literature. The current section treats just models that have been used to characterize movement of cells.

Random movement of particles can be split in 3 different categories: Brownian motion in which the particles movement is perfectly random, supra-diffusive motion in which either an internal or external force imposes some directionality to the particles or sub-diffusive motion in which the movement of particles is hampered. The simplest model to be fit on an MSD curve obtained for 2-dimensional movement of particles is:

$$MSD(t) = 4D_{if}t^{\beta}$$

where t is the time, D_{if} the diffusion coefficient[254] and β the exponent characterizing the type of movement. If $\beta > 1$, the movement is supra-diffusive, for β=1 the motion is Brownian and $\beta < 1$ corresponds to sub-diffusive movement.

While it is capable to take into consideration the type of diffusion that takes place in the movement of the cells, this model is an oversimplification of the actual movement as it characterizes the motion when $t \to \infty$. In the field of cell movement analysis, this model has been applied to characterize the movement of endodermal hydra cells [365] or the collective movement of epithelial cells [141].

The most popular model for characterizing cell movement using the MSD was proposed by Dunn [103] who found it through measurements and confirmed by theoretical analysis by Othmer et al. [268] and Alt [9]. Its equation for 2 dimensions is:

$$MSD(t) = 2S^2P(t - P(1 - e^{-t/P})) \tag{4.12}$$

The 2 parameters used are the speed (S) and the persistence time (P) which is the period of time for which a cell moves linearly in one direction. Its popularity with biologists comes from the fact that it uses just 2 parameters that are easy to inter-

pret and it does not abide to the t$\rightarrow \infty$ condition. The connection to the diffusion coefficient as described in Othmer et al. [268] is that:

$$D_{if} = (S^2 P)/n_d \tag{4.13}$$

with n_d the number of dimensions in which the movement is studied. This model has been employed to study migration of human vascular smooth cells [95], the effects of EGF on fibroblast migration [376] and endothelial cell migration on surfaces modified with immobilized adhesive peptides [199]. Its disadvantage comes from the fact that it assumes that the movement is purely random but in the case of cell movement, some cell types display supradiffusive [277] or subdiffusive movement [20]. In order to combine the advantages of the 2 models, Dieterich et al. [94] proposes a generalization of the Othmer model [268] by using fractional Kramer equations. The resulting model for a 2-D movement is:

$$MSD(t) = 4v_{th}^2 t^2 E_{\alpha,3}(-\gamma_\alpha t^\alpha) + (2\eta)^2 \tag{4.14}$$

where v_{th}^2 is $S^2/2$ with S from the previous model, and α is 2- β with β the exponent from the first model. $E_{\alpha,3}$ is the Mittag-Leffler function [249] and η^2 is a noise term.

In the particular case where α=1 the equation becomes:

$$MSD(t) = \frac{4v_{th}^2}{\gamma_1^2}(\gamma_1 t - 1 + e^{-\gamma_1 t}) + (2\eta)^2 \tag{4.15}$$

with $\gamma_1 = 1/P$.

The problem with the fractional Kramer equation is that it is a complex function for which it is difficult to approximate its parameters. In order to solve this problem, Dieterich et al. [94] approximate the parameters using Markov chain Monte Carlo sampling. A different approach to the problem is proposed in Barbaric et al. [20]. The Mitag Lefler function is transformed in a Fox function as shown in Metzler and Klafter [244]:

$$E_{\alpha,\beta}(z) = H_{1,2}^{1,1}\left[-z\middle| \begin{matrix} (0,1) \\ (0,1),(1-\alpha,\beta) \end{matrix}\right] \tag{4.16}$$

The authors show then that for β=3 this function is in the particular case where it can be reduced to a Meijer G function [21] and the result is:

$$E_{\alpha,\beta}(z) = \frac{2\pi}{3^{\alpha-0.5}} G^{1,1}_{1,4}\left[\frac{-z}{27}\middle| \begin{matrix} 0 \\ 0, 1-\frac{\alpha}{3}, 1-\frac{\alpha+1}{3}, 1-\frac{\alpha+2}{3} \end{matrix}\right]$$

which can be computed in MATLAB using the MuPAD toolbox so that the parameters can be approximated using curve fitting algorithms.

4.4 Migration Analysis in PKD Cell Studies

Various studies have shown that migration of cells is altered in ADPKD[259] as well as by different drugs that seem to affect the evolution of the disease [384]. The remainder of this section provides a review on previous studies, highlighting the type of assays and the measurements they have used. Most of the cell migration studies for ADPKD use 2 technologies to quantify cell motility, Boyden chamber type of assays or wound healing assays. In the following sections the studies were organised based on the methods they use, subsection 4.4.1 presenting those in which the Boyden chamber approach was applied, subsection 4.4.2 describes the studies where wound healing assays were employed and section 4.4.3 discusses studies where both approaches were taken.

4.4.1 Boyden Chambers Based Studies

In a study by Nickel et al. [258], the C-terminal fragment of polycystin-1 (PC1-CTF) was shown to increase migration in mouse inner medullary tract cells (mIMCD-3). In order to test this, the authors used modified Boyden chambers [90]. During the experiment, different conditions were used for the Boyden chambers, a basal condition with serum free media in the bottom chamber, addition of HGF into the bottom well and a third condition in which both HGF and a MEK inhibitor were used simultaneously. The results showed that a greater number of cells expressing the PC1-CTF migrated through the filter under each of the 3 conditions and also that the MEK inhibitor reduced their migration capacity with the greatest effect in cells that did not express the PC1-CTF.

Modified Boyden chambers were also used to quantify the effect of different drugs on ADPKD epithelial kidney cells in a study by Wilson et al. [384]. In their experimental setup the cells came from mice which have received different medical treatments from when they were 6 weeks of age by adding different test agents to their drinking water. The mice were allowed to develop until they were up to 4

months of age, when they were sacrificed and had their kidney extracted. In the cell analysis setup, in order to stimulate migration, a serum gradient was used, media in the upper chambers where the cells were plated having 1% serum and in the lower chambers in which the cell migration was quantified, the serum concentration in media was 5%. As the cells had been fluorescent labelled, quantification was performed by measuring the fluorescence in the lower chambers. The study has shown that inhibition of HER-2 (neu/ErbB2) slows the formation of cysts in PKD1 null mice.

4.4.2 Wound Analysis Studies

A different type of study on the motility of cells affected by ADPKD was done by analysing cell migration in a wound assay. This approach was taken by Luyten et al. [226] on a study on the role of the planar cell polarity (PCP) pathway in polycystic kidney disease. In their experiments, HEK 293T kidney cells were used to assess the effects of PC1 induction and expression of Fz3, a regulator of PCP that was shown to be upregulated in ADPKD patients. The measurement used to achieve this goal was the average speed travelled by cells in a wound healing assay. Images of cells in the assay were taken every 6 minutes at 3 different positions for a period of 20 hours. The conclusions of the study were that overexpression of PC1 improves cell migration while expression of Fz3 reduces their migratory capacity.

Castelli et al. [59] have also used wound healing assays to study how PC1 affects directionality of MEF cells migration. In their experiments, they used visual representation of cell trajectories that are shifted to have the same starting point. Also for a better quantification of the change in cells directionality, the distribution of angles between the initial and final position of the cells was analysed. The results of the study which can be observed both in the visual representation of the trajectories as well as the angle distribution is that cells expressing PC1 have a more linear migration in the direction of the wound.

In a study by Yao et al. [395], wound analysis was used to assess the speed of MEK cell migration in the case of PC1 deletion. In this study, images for the analysis were taken at only 3 time-points corresponding to 0, 3 and 6 hours from the moment the cell monolayer was scratched at 6 positions corresponding to each well of the 6-well plate used for the experiment. This analysis was used to assess what percentage of wound was closed at the specific time-points. The study however

also used wells with sparsely distributed cells to quantify the directionality of the migration. Cells in this experiment were imaged every 15 minutes for a period of 24 hours and just those which travelled at least 50 μm from the starting point after 12 hours were considered for analysis. The directionality was measured using the ratio between the net distance and the total distance travelled by the cells. The conclusions of the study were that PKD-knockout cells are both slower in closing the wounds and have a less persistent direction.

4.4.3 Combined Studies

The two methods to characterize cell migration have been combined in various studies in order to extract more information about the behaviour of the cells. An example of this approach is a study by Joly et al (2003) in which the migration of human renal tubular epithelial cells from cystic and non-cystic kidneys attraction to a number of different chemoattractants were investigated. For the Boyden chamber analysis, Transwell 24-well plates (Transwell) were used by filling the upper chamber with starved cells and the lower chambers with media. Two experiments were done, one where a Ln-5 coated filter was used to attract cells and the ones that migrated were counted after 18 hours and one with Epidermal growth factor (EGF) stimulated cells where the counting was done after 12 hours. For the wound analysis, 10 fields on the wound were taken at 9 and 20 hours from plating and the relative size of the wound was measured for different conditions. The area of the wound was assed using a 100 unit eyepiece optic grid which allows researchers to quantify an area by the number of cells in the grid needed to cover it. This study have found that Ln-5 stimulates cell migration more in PKD cells than in normal ones and also that EGF induced migration is much stronger in the cystic cells.

Another study that uses the combined approach was done by Boca et al. [40]. In it Boyden chambers were used with serum-free Dulbecco's Modified Eagle's medium (DMEM) placed in the lower chambers and $MDCK^{Pkd1Zeo}$and $MDCK^{Zeo}$ cells grown in different conditions were suspended in serum free media in the upper chamber and left to migrate overnight. Wound analysis studies were done by imaging the wound every 2 or 3 minutes for a period of 8 hours. The wound closure rate was calculated by measuring the area covered by the monolayer. Cells expressing *Pkd1* showed faster closing rates and the Boyden chamber migration assays were used to confirm this finding.

Outeda et al. [270] did a study on both knockout *Pkd1* and 2 mouse dermal lymphatic endothelial cells in which they use both Boyden chambers and wound healing assays to quantify their migratory capacities. Vascular endothelial growth factor C was the attractant used in the Boyden chambers study. Wound closing assays where done for 20 hour with images of the confluent layer taken at 0 and 20 hours. Also time lapse analysis was done for the movement of individual cells detached from the monolayer by taking pictures at minimum 7 positions per condition every 5 minutes for a period of a minimum of 200 minutes. To quantify cell migration the area covered by the monolayer at the 2 time points was calculated and the initial area was subtracted from the final one and the percentage of wound that was closed was presented. For visual inspection, the cell trajectories were plotted without shifting the initial position to the same point and cells which travelled more than 20 μm in the direction of the wound had their trajectories highlighted. For quantification of their movement, the total distance and the confinement ratio were calculated. The researchers found that the closing rates were higher for control cells. This happened because although the total distance travelled is roughly the same for control cells vs *Pkd1* or 2 knockdown, the directionality is significantly raised in the control cells.

In a follow-up study, Castelli et al. [60] investigated the effects of PC1 on MEF (mouse embryonic fibroblasts), mIMCD and MDCK cells cytoskeleton. To assess how this affects migration, both a Boyden chamber and a wound healing assay were used. In the case of the filtering assays the lower chambers were filled with DMEM and DMEM with cells was placed in the upper ones. Fibroblasts where let to migrate for 3 hours while the other cells were let to migrate overnight. The cells were stained and counted in at least 10 fields per chamber. While healing assays where used, no information about the migration of cells was provided, the authors used the orientation of the Golgi apparatus as an index of cell polarisation. Again, cells expressing PKD1 were shown to migrate more than control cells.

4.5 Texture Analysis for Cell Studies

Although they do not have a formal definition, textures can be described as visual patterns composed of entities or sub-patterns that have specific brightness, colour, slope or size, that are easily perceived by humans and allow us to distinguish or characterize entities in an image [237]. With the development of computers and imaging devices, scientists have tried to model the texture processing by human eyes

which led to the development of a new field called texture analysis. Texture analysis of images has many applications in practice, being used for automatic inspection, for example to detect defects in carpets [323], identify terrain from satellite images [145], analyse cell images [184] or for facial recognition [130]. This review will concentrate on texture analysis applied to cell images.

In general, two steps are taken when analysing the texture in an image [237]. The first step is feature extraction in which different characteristics of an image are computed to numerically describe its textures. The second step is classification of regions in an image that correspond to a specific texture are identified using an automatic method generally coming from artificial intelligence. In this chapter, only the subject of feature extraction will be treated but thorough reviews done by Materka et al. [237] and Di Cataldo and Ficarra [93] discuss both steps.

Feature extraction techniques for image analysis can be classified in 5 categories: geometrical or structural methods, statistical methods, local binary patterns, model based methods and transform-based methods [93].

4.5.1 Structural Methods

The underlying assumption of structural methods is that textures appear as a repetitive collection of sub-patterns (such as different shapes or lines). The problem with this approach is that it is very limiting for the application of the methods[35], needing a strict arrangement of the identified sub-patterns. For artificial textures, this approach is acceptable but in the case of cell images where there is little to no regularity, the approach falls short [93]. However there has been a paper employing this method to analyse cell images of fetal liver cells acquired by confocal laser microscopy [23].

4.5.2 Statistical Methods

Another way a texture can be viewed is as a random distribution of pixel intensities in space. This approach has led to the creation of the statistical methods for texture analysis. Depending on the number of pixels whose joint probability distributions are combined to create statistics for the image, first order statistics can be used which means that the probabilities of each pixel to have a certain intensity are calculated, or second order statistics where the probability of 2 pixels at a specific relative position to have certain values is extracted. Higher order statistics can

be used but as shown in a study by Julész et al. [180] humans seem to perceive 2 pictures with identical second order statistics as being the same even if they have different higher order statistics. Also the computational cost increases with the order so in practice higher order statistics are rarely used. While they are easy to calculate, first order statistics do not provide information on the spatial distribution of pixels which is important if pixels with different values are grouped together or intercalated in order to characterize the smoothness of the texture [3]. As a result, second order statistics are some of the most used features in image analysis and they will be described in more detail in this section. In practice, for the calculation of the second order statistics, Haralick et al. [145] proposed the gray level co-occurrence matrix (GLCM). This is created based on the gray-levels that exist in an image.

For an image, lets suppose the values of the pixels are in the interval $I = [g_{min}, g_{max}]$. For the creation of the gray occurrence matrix the interval is split into N subintervals of equal size $I_k = [g_{min} + k * \frac{g_{max} - g_{min}}{N}, g_{min} + (k+1) * \frac{g_{max} - g_{min}}{N})$ where $k \in \{0, N-2\}$ and $I_{N-1} = [g_{min} + (N-1) * \frac{g_{max} - g_{min}}{N}, g_{max}]$. Next, a relative position has to be defined for 2 co-occuring pixels. This is done using a set of 2 parameters which can be either (Δ_x, Δ_y) which are the relative positions of the pixels on the x and y axis or (d, θ) where d is the distance in squares and θ is the angle between 2 pixels defined as co-occuring. For the following part the (Δ_x, Δ_y) displacement is used as it is easier to follow. Once the intensity intervals and the relative positions are established, the gray co-occurrence matrix can be defined as $G_{(\Delta_x, \Delta_y)} \in \mathbb{N}^{NxN}$ for which $G_{(\Delta_x, \Delta_y)}(i, j)$ represents the number of times a pixel with a value in the interval I_j occurs at a relative position (Δ_x, Δ_y) from a pixel with a value in the interval I_i in the analysed image. Two approaches can be taken for quantifying the number of apparitions, either

$$G_{i,j} = card\{P_{x,y} \in I_i | P_{x+\Delta_x, y+\Delta_y} \in I_j\} \tag{4.17}$$

or

$$G_{i,j} = card\{P_{x,y} \in I_i | P_{x+\Delta_x, y+\Delta_y} \in I_j\} + card\{P_{x,y} \in I_j | P_{x+\Delta_x, y+\Delta_y} \in I_i\} \tag{4.18}$$

where $P_{x,y}$ is the pixel intensity at position (x,y) in the image.

The matrix is then normalized by dividing all its values by the total number of pairs. The significance of the gray co-occurrence matrix in texture analysis becomes apparent when Haralick features [145] are introduced. They extract different infor-

mation from a gray co-occurrence matrix and the user obtains only one value for a feature for an image. In their original paper, Haralick introduces 14 features, some examples being:

- Contrast

  Contrast quantifies the local change in an image. If the difference in intensity between pixels occurs continually, the contrast becomes large. Its formula is:

$$Contrast = \sum_{n=0}^{N-1} n^2 \cdot \sum_{|i-j|=n} G_{i,j} \tag{4.19}$$

- Homogeneity

  Homogeneity is a measure of the similarity of pixels a homogeneity of 1 meaning an image in which all the pixel intensities are the same.

$$Homogeneity = \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \frac{G_{i,j}}{1+|i-j|} \tag{4.20}$$

- Entropy Entropy measures the randomness in an image.

$$Entropy = -\sum_{i=1}^{N-1} \sum_{j=1}^{N-1} G_{i,j} log(G_{i,j}) \tag{4.21}$$

- Correlation

  Correlation shows the linear dependency of pixel values on the neighbouring pixels [342]. This means that it will have high values in textures with a highly linear structure.

$$Correlation = \sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x \sigma_y} \tag{4.22}$$

  where μ_x,μ_y,σ_x, σ_y are the means and standard deviations across the 2 axis

and are calculated with the formulas:

$$\begin{aligned}\mu_x &= \sum_i \sum_j i p(i,j) \\ \mu_y &= \sum_i \sum_j j p(i,j) \\ \sigma_x &= \sqrt{\sum_i \sum_j (i-\mu_x)^2 p(i,j)} \\ \sigma_y &= \sqrt{\sum_i \sum_j (i-\mu_y)^2 p(i,j)}\end{aligned} \tag{4.23}$$

- Energy

  Energy is calculated based on the Angular Second Moment(ASM) which increases with the local uniformity in a picture.

$$\begin{aligned}Energy &= \sqrt{ASM} \\ ASM &= \sum_i \sum_j p^2(i,j)\end{aligned} \tag{4.24}$$

A thorough description for the rest of the features and the way they relate to an actual image can be found in Malathi and Shanthi [231]. The shortcomings of texture feature extraction using second order statistics consist in the fact that the results are dependent on the displacement parameters. In order to solve this problem, Haralick proposed calculating the GLCM matrix for different angles of rotation, computing the features for each rotation and then using the means and variances of the features obtained over all the rotation angles rather than their actual values[145].

In the field of cell imaging, Haralick features have been applied for the detection of colon cancer cells [62], quantify drug response of cancer cells [223], or to study the differentiation of stem cells [222].

4.5.3 Local Binary Patterns

Local binary patterns[262][263] are a type of feature extraction method which combines structural methods with statistical ones. It works by first dividing images into cells, each cell being a square containing a certain number of pixels. Then, the value of a pixel is used as threshold (generally the central pixel), with the remaining pixels in the cell being given a value of 0 if their intensity is smaller or a value of

1 if it is higher. Next, the values obtained this way are arranged in a vector and the cell is assigned a number that written in binary form produces that array. Next, a histogram is built for all the values a cell can take and an array of features is obtained where every feature corresponds to a bin in the histogram and its value is the number of cells which had that specific value assigned. Improvements of the method have been proposed that are rotationally invariant [264] or which also use information of the magnitude of the differences in a cell [138].

The method is popular in cell texture classification and has been used in Hep2 cells classification [261], cell phenotype classification based on protein sub-cellular localization images [257] and apoptosis detection in adherent cell populations [160].

4.5.4 Model-Based Methods

A different approach to texture characterization is the use of modelling. In this approach it is considered that the pixel intensities in a specific texture are generated using a function. Three types of models that have been used to characterize texture are presented in this chapter: autoregressive models, Markov chain models and fractal-based models.

The autoregressive models for 2D texture analysis were introduced in 1986 by Deguchi et al. [88]. The idea behind it is to define a square window of a certain size and to express the intensity of the pixel in the middle as a weighted sum of the remaining pixels, with a certain weight for each position in the window. As it is expected more than one texture exists in an image, the image should be split into many small areas of equal size and a linear model fit for the pixels in each of them. Then the areas with similar models are considered to come from the same texture. For the approximation of the model parameters 2 approaches are generally used, least square estimation (LSE) or maximum likelihood estimation (MLE) [179]. In the field of cell texture analysis, autoregressive models have been use to discriminate between leukaemia and lymphoma cells [122].

Another way to characterize a texture using a model for the intensities of the pixels in an image is to consider them generated from a random process which can be described using a Markov model. As the images are bidimensional, Markov random fields are used as they represent a multidimensional generalization of the Markov chains. Again, the first step is to split the image into small regions and calculate and fit a model for each region. In cell images this approach was used for

segmentation of leukocytes in bone marrow cell images [293], and classification of cervical cells in combination with GLCM features [225]. The main disadvantage of using this approach is that it is computationally expensive [93].

The last type of model for texture analysis that will be discussed are the fractal based models. Fractals are a class of mathematical functions that have been proposed by Mandelbrot to describe visual patterns in nature [233]. Their main properties consist in the fact that an image generated using fractals, will be similar in a statistical sense at all levels of magnification or scale [328]. They were first used for describing texture in 2D images by Pentland in 1984 [279]. In the field of cell image analysis, they have been used for analysis of apoptosis of breast cancer cells [224] and detection of breast cancer cells [283].

4.5.5 Transform-Based Methods

Transform methods work by representing the image in a space where the dimensions are better related to characteristics specific to textures such as frequency or size [237]. Two types of transform methods will be presented here: frequency based methods and wavelet based methods.

Frequency based methods work by transforming an image in the frequency domain where features such as coarseness, graininess, or repeating patterns are easy to identify [93]. Two methods are usually used for this operation, the 2D discrete Fourier transform (DFT) [11] and the discrete cosine transform (DCT) [398]. In cell imaging texture analysis the approach was used to study cells from the lenses in human and animal eyes in the case of cataract [123]. Their disadvantages come from their lack of spatial localization [237].

The lack of spatial localization for Fourier based methods is solved by the use of wavelets in which window functions are combined with frequency transformation ones in order to allow for both frequency and space representation in images [221]. A particular case of wavelets very popular in texture analysis is represented by Gabor filters [172]) in which the window function is a Gaussian. Examples of the use of wavelets in texture analysis for cells images are a study by Weyn et al. [380] in which breast cell cancers were classified and a study by Kim et al. [190] in which images of renal cancerous cells obtained using a confocal microscope were used to build a classifier for the progress of the disease using wavelet-based features.

4.6 Summary

This review has presented the current techniques employed for the analysis of cell migration and texture of cell images. First, the assays used for the study of cells have been introduced, with their respective categories, filter assays in which cells are allowed to migrate past a filter and those which do so are counted and time-lapse assays which permit the videoing of live cell movements for a detailed description of their trajectories.

Next, the measurements used to characterize cell movements whose trajectories are known are introduced. Separate sub-sections are provided for motility measures in which just the trajectory is taken into consideration, velocity measures in which time appears explicitly and diffusivity measures inspired from particle physics. For the diffusivity measures, a section is dedicated to the models that have been proposed to extract relevant information from them.

Once the measures and technology has been introduced, the review moves to focusing on cell migration studies that have been applied for elucidating the mechanisms of polycystic kidney disease. Different studies are summarised with a focus on the methods used to analyse motility, the type of cells studied and their conclusions on how ADPKD affects cell migration. The last part of the review presents current methods used to analyse image textures that have been employed in cell studies. The review allocates different sections to structural methods in which regular shapes are looked for to identify the patterns in images, statistical methods in which various statistics of pixel intensities are extracted to reveal information perceived by the human eye, local binary patterns that combine the 2 approaches, model based features in which various mathematical models are employed to characterize the way in which pixel intensities in the image were generated and finally transform based methods in which the image is first moved to a different space where dimensions are more relevant for texture characterization.

Chapter 5

A Novel Framework for Time Series Gene Expression Data Analysis Combining Biomarker and GRN Identification

5.1 Introduction

Historically, researchers chose to focus on one type of microarray data study, either to identify new biomarkers [92][168][331], or to analyse a set of known genes of interest to identify their regulatory network [115][312]. In reality, both types of analysis are complementary. Biomarker identification does not provide any information as to how the genes interact to specify a function. Network inference on the other hand needs a reduction in the number of genes being analysed, a genome-wide network identification being computationally impossible. As a result, studies that seek to combine both approaches have started to be published in recent years.

In a study from 2013, Zhang et al [404] used a combined approach for studying microarray data in the case of late-onset Alzheimer's disease (LOAD). This study included samples from different tissues obtained from both healthy patients and patients suffering from LOAD. The measurements were taken with a dual channel Agilent microarray machine. The differentially expressed genes were identified as the 33% of genes with the highest variance across the same tissue of the p-value between normal and affected channels. Next, weighted correlation network analysis

(WGCNA) was employed to create modules of well-clustered co-expressed genes. The last step was to build static Bayesian networks for genes in each of the resulting modules. The authors created 1000 models of the Bayesian network structure for each module and then selected as real connections just edges that appeared in at least 30% of the models.

Gordon et al. [133] have used a combination of the two types of analysis to study the effect of abiotic factors in the gene expression of a plant, Brachypodium distachyon. In the first stage, control samples were compared with samples coming from plants put under different stresses. The differentially expressed genes were isolated by the use of significance analysis of microarrays method (SAGE) [363] in which the percentage of genes selected as differentially expressed by chance is estimated using permutations of repeated measurements. Next, WGCNA combined with a threshold on the significance of edges was used to identify the structure of their underlying network.

Wu et al. [388] use a combined method to identify a dynamic model for gene regulatory networks in a study on viral infection in mice. In their approach, differentially expressed genes were identified as genes that showed a large difference of expression between day 0 when the animals were infected to later time-points when the animals were suffering from the disease. The set of genes detected this way were clustered in modules using k-mean clustering. A curve was fitted through the expression of genes in each module over all the time-points in order to create a model for a "supergene" which represents each cluster. The obtained models were then sampled to obtain measurements for the "supergenes". Next, a system of linear ordinary differential equations was used to model the response of the network having the obtained "supergenes" as nodes. The inference method imposes sparsity but not stability in the system matrix.

Rodius et al. [297] have also applied a 2-stage method to identify the network of genes that lead to cardiac repair in zebrafish. In the first step analysis, differentially expressed genes were selected as those whose expression displayed an q-value of <0.05 between healthy fish and those with significant heart injury. In the second step of the analysis, WGCNA with a threshold on the weight of the edges was used to identify genes that influenced each other.

This chapter proposes a new multi-stage framework for analysis of time-series microarray data to identify a set of possibly relevant genes followed by building a dynamic model for their regulatory network. The framework consists of 4 stages,

the first 3 of which are used to identify a set of genes that are as relevant as possible for the studied condition, while the last stage builds the GRN model. Using the proposed framework two ADPKD microarray datasets available in the literature have been analysed [242][243]. The two sets come from mice which had gene Pkd1 knocked-out. In the case of the first dataset the knockout was done when mice were less than 10 days of age while in the case of the second dataset when they were 40 days of age corresponding to different severities of the disease described in a previous study[284].

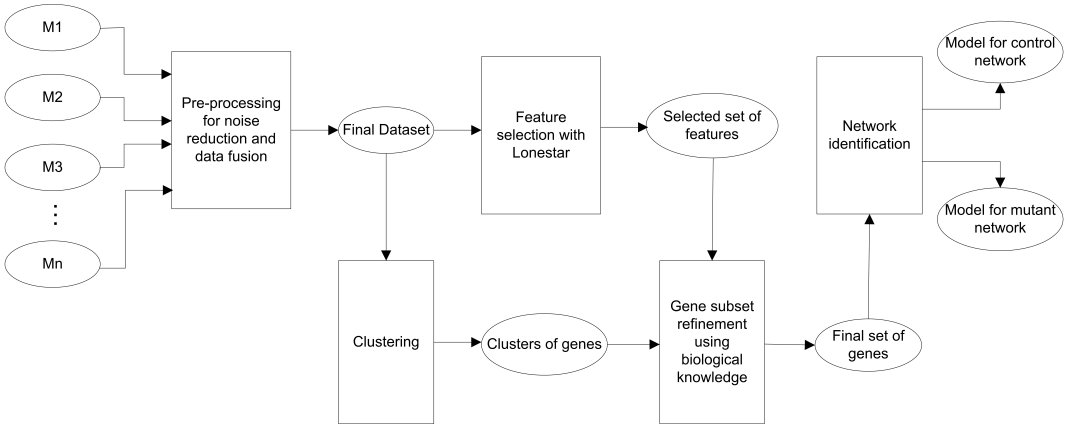

Figure 5.1: Diagram of the application of the proposed framework. The analysis begins with the raw data which is usually stored in a set of matrices, represented as M_1 to M_n. The first step is to put the data together in one matrix and apply different pre-processing techniques to obtain the final dataset on which the framework is applied. The first step of the framework consists in the application of $l1$-StaR, leading to the selection of a set of genes. Next, the features in the dataset are organized in clusters. Once this operation take place, the genes selected by $l1$-StaR together with the genes in some of the clusters containing them undergo a selection based on the use of biological knowledge. For the final set of genes obtained after this step, models of regulatory interactions are created for the healthy and diseased conditions

The next sections are organised as follows: Section 5.2 presents the first stage of the proposed framework in which an improved version of an SVM-RFE algorithm [5] is introduced and used to derive a minimal set of informative genes for classification. Section 5.3 describes the second stage of the framework in which the minimal set of genes is expanded to include other genes whose expression correlates with that of genes in the minimal sets, which provides a larger relevant set that can be analysed using expert knowledge. The third stage of the framework, detailed in section 5.4, exploits additional biological knowledge to select a final set of relevant genes that will be used to derive the dynamical model. Section 5.5 introduces the

approach to build a linear differential equations model of the regulatory network for the selected genes. The ADPKD dataset used in the study is introduced in Section 5.6 whilst Section 5.7 details data pre-processing steps. The results of the analysis and modelling study obtained by applying the proposed framework are presented in Section 5.8. The summary of the chapter are presented in Section 5.9.

Figure 5.1 presents a schematic representation of the deployment of the framework on raw data.

5.2 Stage 1: Feature Selection Using q-Value-Based *l*1-StaR Algorithm

For the first stage of the microarray data analysis framework, a hybrid feature selection method based on the *l*1-StaR algorithm is used as it provides an automatic method to reduce the number of genes to a set that allows to discriminate well between 2 conditions without the need of an user-defined threshold. In the original implementation of the algorithm, the statistical filtering is performed using a Student t-test. This approach has a number of limitations. Firstly, in the case of the Student t-test, the two tested populations are assumed to be normally distributed with equal variance. This is unrealistic in the case of gene expression data where there is reason to expect that the variance between control and mutant classes differs [273]. Secondly, the Student t-test is a single hypothesis testing algorithm and, as shown in Chapter 3 section 3.4.1.1, its sole application is likely to introduce a high number of false positives. In order to address these problems, here a hybrid feature selection method is proposed, which involves using a Welch t-test with pFDR corrected p-values as a filter followed by the SVM-RFE wrapper from the *l*1-StaR algorithm [5] described in 3 section 3.5.3. In order to obtain a ranking for genes, the proposed method uses k-fold cross-validation and a frequency-based ranking scheme.

The remainder of this section presents a description of the Welch t-test and the algorithm used to implement the pFDR based correction.

5.2.1 Welch t-test

The Welch t-test [378] is a parametric statistical test based on the original Student t-test. The difference between the two methods appear in the assumption they make

about the variances of the distributions from which the 2 sets of measurements come. In the case of the Student t-test, the assumption is that the variances are equal while the Welch t-test considers unequal variances. The following part will present how the p-value for the Welch t-test is calculated.

For 2 sets of data to be compared, A and B, their means are calculated as:

$$m_A = \frac{\sum_{i=1}^{N_A} a_i}{N_A}, a_i \in A, N_A = card(A) \tag{5.1}$$

$$m_B = \frac{\sum_{i=1}^{N_B} b_i}{N_B}, b_i \in B, N_B = card(B) \tag{5.2}$$

and variances as:

$$\sigma_A^2 = \frac{\sum_{i=1}^{N_A} (a_i - m_A)}{N_A - 1}, a_i \in A, N_A = card(A) \tag{5.3}$$

$$\sigma_B^2 = \frac{\sum_{i=1}^{N_B} (b_i - m_B)}{N_B - 1}, b_i \in B, N_B = card(B) \tag{5.4}$$

Next, their Welsch t-test statistic is:

$$T = \frac{m_A - m_B}{\sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B}}} \tag{5.5}$$

The sampling distribution can be approximated with a Student's t distribution with d degrees of freedom where

$$d = \frac{(\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B})^2}{\frac{\sigma_A^4}{(N_A-1)N_A^2} + \frac{\sigma_B^4}{(N_B-1)N_B^2}} \tag{5.6}$$

The obtained T and d are used to locate the corresponding p-value for the current test in a t-table, which contains pre-calculated values of T for different degrees of freedom and p-values.

In the case of the present analysis, the measurements are organised in the gene data matrix $G \in R^{mxn}$, where m- number of genes and n number of samples. As a result, for a gene G_i on whose measurements the Welch t-test is performed the 2 sets A and B are:

$$A = \{G_{ij} | j \in C_1\} \tag{5.7}$$

$$B = \{G_{ij} | j \in C_2\} \tag{5.8}$$

where C_1 is the set of the indexes of columns in matrix G of measurements coming from control samples and C_2 the set of the indexes of columns in matrix G of measurements coming from mutant samples.

5.2.2 Q-value Calculation for Statistical Filtering

The q-value for a statistical test was introduced in Chapter 3. In the present analysis a threshold on the q-value was used in order to select a subset of genes that are considered significant.

Storey and Tibshirani [336] proposed an algorithm for calculating these values that is presented below.

1. Let $p(1) \leq p(2) \leq ... \leq p(m)$ be the ordered p values for the genes 1...m
2. For a range of λ, in the interval [0,1], calculate:

$$\hat{\pi}_0(\lambda) = \frac{card\{p_j > \lambda\}}{m(1-\lambda)} \tag{5.9}$$

3. Fit a natural cubic spline $\hat{f}$ with 3 degrees of freedom to $\hat{\pi}_0(\lambda)$ on λ.
4. Set the estimate of π_0 to be:

$$\hat{\pi}_0 = \hat{f}(1) \tag{5.10}$$

5. Calculate

$$\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\hat{\pi}_0 m \cdot t}{card\{p_j \leq t\}} = \hat{\pi}_0 \cdot p_{(m)} \tag{5.11}$$

6. Next calculate

$$\hat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \frac{\hat{\pi}_0 m \cdot t}{card\{p_j \leq t\}} = \min(\frac{\hat{\pi}_0 m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)})) \tag{5.12}$$

for i= m-1, m-2 ,..., 1

7. For the ith most significant feature the estimated q-value is $\hat{q}(p_{(i)})$.

As shown in Storey et al. [337], the resulting estimates are guaranteed to be smaller or equal to the real q-values.

In the case of the present analysis, the threshold used on the q-value to select the statistically significant genes was 0.05.

5.2.3 Description of the Proposed Algorithm and its Application

The proposed algorithm has 3 steps:

1. For a gene dataset a p-value is computed for each feature by using a Welch t-test between the measurements coming from the control and the affected group.

2. The p-values for the features are used to calculate their corresponding q-values

3. A threshold is imposed on the q-values so a reduced set of features is obtained

4. The measurements for the remaining features are used as inputs for the SVM-RFE wrapper from the $l1$-StaR algorithm The proposed algorithm will be referred to as pFDR-corrected $l1$-StaR

For its use in practice, the following scheme is employed:

1. A gene dataset is split 10 times using a k-fold validation scheme. As a result 10 x k training sets and 10 x k testing sets are obtained

2. pFDR-corrected $l1$-StaR is trained on each training set and as a result a significantly reduced gene set is obtained as well as a SVM-classifier based on them. The trained classifier can then be used on the corresponding testing set to assess its accuracy

3. Once training was done for all training sets and a gene set was obtained in each case, all the genes that appear in at least a set are put together and ordered according to the number of gene sets obtained by the classifier in which they appear. This final set is passed to the next stage of the framework.

5.3 Stage 2: Gene Subset Augmentation Through Clustering

The aim of this step is to use a gene clustering approach to enrich the initial, minimal set of genes that are used for classification with additional potentially relevant genes that could offer insight into the disease and offer possible drug targets. The approach consists of two steps, namely statistical filtering and gene clustering and selection.

In the first instance, statistically insignificant genes with a p-value greater than 0.05 according to a Welch's t-test are eliminated to reduce the number of non-informative features.

The second step involves clustering the genes and selecting the clusters that contain genes identified at Stage 1. A hierarchical tree is built using the absolute Pearson's correlation described in Chapter 3 section 3.6 as a measure of distance and average distance as a linkage rule. The reason for using Pearson absolute correlation as a measure of distance is that it is expected that genes that regulate each other will be linearly correlated and both upregulation and downregulations are as relevant for the problem. The resulting tree is then cut using the Dynamic TreeCut algorithm presented in Chapter 3 section 3.6. For this step, only the clusters which contained genes identified at Stage 1 are selected, as they are expected to contain genes which might play a role in the disease. All the genes in clusters that contain at least one gene from the final set obtained at Stage 1 are passed to Stage 3.

5.4 Stage 3: Gene Subset Refinement Using Biological Knowledge

The aim of this step is to exploit existing biological knowledge available to identify a reduced set of genes that may offer insight into the cause and mechanism of the disease and which should be subject to more detailed analysis and modelling. At this point, the set of genes derived in the previous stage, which typically will be quite large, is the result of applying data analysis techniques alone and does not reflect any *a priori* biological knowledge.

The aim of this analysis stage is to refine the final set of genes in order to maximise the number of biologically relevant genes that are analysed and modelled further using GRN inference methods. Ideally, this set of genes should include the

genes already known to be related to the condition of interest as well as novel genes whose analysis would provide new insights into disease mechanisms. The proposed strategy uses the most complete protein interaction database currently available [345] of known and predicted protein-protein interactions in conjunction with expert medical knowledge to select a biologically informative gene set.

Four sets of genes are defined before applying the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) analysis :

- Set A: All the genes that are part of the clusters corresponding to the first 3 most frequently selected genes at Stage 1. The number of genes selected can vary depending on the frequency with which the genes were selected, 3 is used only as a rule of thumb
- Set B: All the genes that have been passed Stage 1
- Set C: Additional genes in clusters selected by the medical expert, which contain at least one gene that is known to be biologically relevant and was clustered with at least one of the genes selected at Stage 1
- Set D: Critical genes for the condition, for example *PKD1* and *PKD2* in the case of ADPKD without the need of appearing in the list of genes that have passed Stage 2

The list of genes that is used as an input for STRING, which will be named S is defined as:

$$S = A \cup B \cup C \cup D \tag{5.13}$$

The resulting network of predicted associations which can be seen as an undirected graph is used to select the genes in the final set as follows:

- genes that appear in any connected subgraph containing one of the genes used to select the clusters for set A
- genes that appear in the same connected subgraph as the genes deemed interesting by the medical expert that were used to select the clusters in set C
- genes that appear in the same connected subgraph as the genes in set D but without the genes in set D if they did not appear in the clusters selected at Stage 2

All these genes are then moved to Stage 4. Alternatively, if few genes are selected after this step, all the genes in set B can also be passed to Stage 4.

The benefits of using this approach is that genes that are of interest to the medical experts and are expected to play a role in a disease, can be put together with novel genes that are found to be differently expressed through analytical methods and have their interactions modelled, connecting the old knowledge in the field with the new one.

5.5 Stage 4: GRN Identification

The final stage of the proposed microarray data analysis framework is the identification of a dynamical model of the regulatory network based on the time-course data corresponding to the set of genes derived in the previous stage.

The aim is to derive a linear system of differential equations

$$\dot{x}(t) = Ax(t) \tag{5.14}$$

where x the vector of genes measurements and A the system matrix.

One of the biggest problems with identifying A from microarray measurements is the so called curse of dimensionality [26]. This means that as A has n^2 elements, in order to avoid overfitting, measurements at least n^2 time-points should be used, where n is the number of genes. In practice, this is impossible unless n is extremely small.

The solution used here involves interpolating the data to reconstruct the continuous-time representation of each gene expression signals and used the resulting functions to generate uniformly sampled data used to estimated the system matrix A.

5.5.1 4.1 Nonlinear Interpolation of Gene Expression Data

Unlike in D'haeseleer et al. [91], who used a cubic spline interpolation on the log values of the genes expression levels for example, here we used the exponential interpolation approach proposed by Alecu [7]. The main reason for adopting this particular type of interpolating functions, is that when A has n distinct eigenvalues, the general solution of the linear dynamical system (5.14) can be written as a

superposition:

$$x_i(t) = w_{1,i}e^{-\lambda_{1,i}t} + w_{2,i}e^{-\lambda_{2,i}t} + \cdots + w_{n,i}e^{-\lambda_{n,i}t} \tag{5.15}$$

as shown in Boyd [48] where for i=1,...,n [280] the eigenvalues λ_i have to be positive since it is expected that a GRN represents a stable system.

In the case of the present application, the system is considered to reach an equilibrium point so the model:

$$z_i(t) = w_{0,i} + w_{1,i}e^{-\lambda_{1,i}t} + w_{2,i}e^{-\lambda_{2,i}t} + \cdots + w_{n,i}e^{-\lambda_{p,i}t} \tag{5.16}$$

is interpolated, where p is chosen to be smaller than the number of data points , the parameters of the interpolation models (5.16) for each gene are estimated using the trust reflective region algorithm [55] involving a regularised cost function [356] and $w_{0,i}$ is the value of the expression for gene i when the system reaches the steady state.

$$Q_i = |y_i - Z_i|_2^2 + \alpha|w_{0,i}...w_{p,i}\lambda_{1,i}...\lambda_{p,i}|_2^2 \tag{5.17}$$

where α is the regularization parameter, Z_i is the vector of model values for each time point $[z_i(t_0), z_i(t_1), ..., z_i(t_N)]$ and y_i is the vector of real measurements for gene i.

In order to approximate a model that best describes the data, a good value for the regularization parameter α is needed.

For this purpose, α is assigned values between 10^{-6} and 10 on a logarithmic scale with 10 steps per decade and $F_m$is minimized for each α. This leads to 70 models from which one was chosen to both minimize the error but also to have small enough parameters. The operation is repeated 5 times with the parameters of the model randomly initialized in the range [0, 10) in order to avoid a local minima.

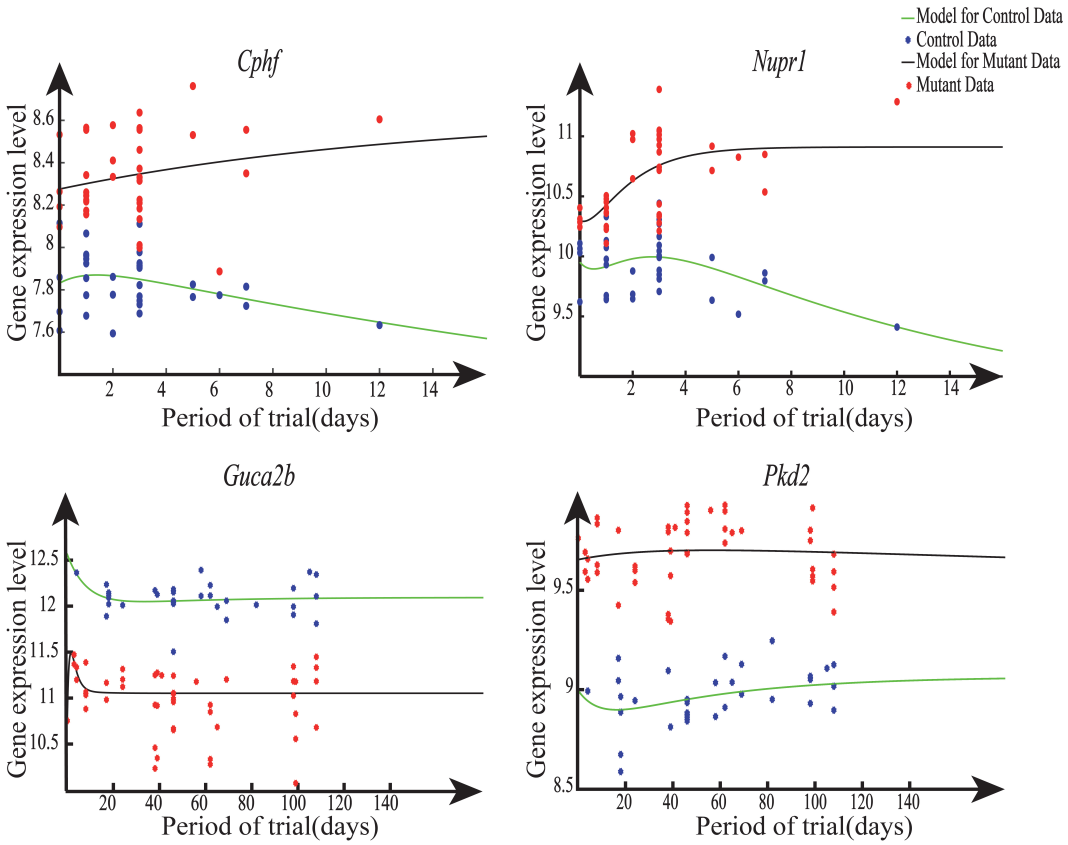

Figure 5.2: Graph for the samples of 2 most selected genes in each dataset and the models fit through them. a. Graph for the genes selected for the first dataset b. Graph for the genes selected for the second dataset.

To select a good model from those obtained for different values of α, the l-curve criterion [144] was used. For each model i, the values $a_i = log_{10}|y - Z_i|_2^2$ and $b_i = log_{10}|w_0 ... w_p \lambda_1 ... \lambda_p|_2^2$ were assigned, where Fi represents the approximated model, y the real measurements and p the index for the last parameter of the model. Then, in a manner similar to Belge et al. [24] the point $O = \binom{a_1}{b_N}$ where N is the number of models has been considered the origin and the model i for which $O - \binom{a_i}{b_i}$ is minimal was selected. This operation was done for all the 5 random initializations and the final model was chosen as the one that had the smallest fitting error out of those selected using l-curve criterion [144]. Figure 5.2 presents some examples of gene data fit using the individual gene models.

5.5.2 GRN inference

The complete matrix A is estimated based on gene trajectory data obtained by re-sampling the individual interpolation models. The sampling is performed over the original time interval with a step size τ such that:

$$\tau < \frac{t_N - t_0}{n^2} \tag{5.18}$$

to prevent overfitting.

The steady state of the system is $\begin{bmatrix} w_{0,1} \\ \vdots \\ w_{0,n} \end{bmatrix}$ with $w_{0,i}$ the parameter w_0 of the model for gene i. Its perturbation response around its equilibrium points for a time t_j is calculated as:

$$e_j = \begin{bmatrix} z_1(t_j) \\ \vdots \\ z_n(t_j) \end{bmatrix} - \begin{bmatrix} w_{0,1} \\ \vdots \\ w_{0,n} \end{bmatrix} \tag{5.19}$$

Where $z_i(t_j)$ is the value of the model for gene i at time t_j and $w_{0,i}$ defined above.

The algorithm proposed by Zavlanos et al. [401] which is described in chapter 3 section 3.7.2 was applied to estimate A for the system represented by the gene measurements based on

$$\dot{Z} = AE \tag{5.20}$$

where E=$\begin{bmatrix} z_1(t_0)-w_{0,1} & \dots & z_1(t_N)-w_{0,1} \\ \vdots & \ddots & \vdots \\ z_n(t_0)-w_{0,n} & \dots & z_n(t_N)-w_{0,n} \end{bmatrix}$ is the matrix containing the system responses for a series of time-points $t_0 ... t_N$ and $\dot{Z}$ is $\begin{bmatrix} \dot{z_1}(t_0) & \dots & \dot{z_1}(t_N) \\ \vdots & \ddots & \vdots \\ \dot{z_n}(t_0) & \dots & \dot{z_n}(t_N) \end{bmatrix}$ with $\dot{z}_i(t_j)$ the value of the derivative of the model for gene i at timepoint t_j.

The derivative samples for $z_i(t)$ are obtained by calculating the analytical derivative of the model of the gene i and sampling it at $t_0 \dots t_N$.

5.6 Experimental Dataset Description

In order to investigate new biomarkers as well as interactions between genes that might reveal mechanisms involved in the disease, the proposed framework was ap-

plied on 2 publicly available datasets coming from ADPKD studies.

A couple of datasets coming from studies on the disease are available online, but most of them have a small number of samples (maximum 20 biological replicates) [182] [305] [274] [198] [71] [284], [318] or do not provide the time points for the moments measurements were taken [331] [267]. Only two of the available datasets were suitable to be analyzed with the proposed framework, both of them providing a large number of samples and clear time labelling of the measurements and coming from the Menezes and Piontek research group [242][243] .

As discovered in a study by Piontek et al [284], a time-dependent switch exists that influences disease severity in Pkd1 knock-out mice. Pkd1 deletion when the mice were less than 13 days old leads to grossly cystic kidneys within 3 weeks, while deletion of the gene when the mice were more than 13 days old led to the onset of disease around 6 months of age. The two datasets analysed correspond to these 2 cases, the first one being a study on rapid progression of the disease while the second one was obtained during a study characterised by slower disease progression.

In the first dataset, the measurements came from 70 mice. From the total, 36 were Pkd1 conditional tamoxifen-Cre inducible mice [116] and had Pkd1 deleted between 5 and 9 days old by having their nursing mothers injected with tamoxifen. The other 34 were control mice used to provide a reference for gene expression. In order to ensure that conditions were as similar as possible to compare the effects of disease, the authors paired control and mutant mice from the same nursing mothers. When the animals were between 11 and 24 days old, their kidneys were extracted and gene expression values were measured. Microarray data was obtained using 2 versions of the Illumina sequencing systems, v1.1 and v2.0. Ilumina v1.1 was used to measure 14 samples while Ilumina v2.0 was used to measure 56 samples. The gene expression values were stored in 4 matrices, LM4, LM5, LM6 and LM8. LM4 contains the data obtained using Illumina v1.1 and in matrices LM5, LM6 and,LM8, data was obtained using Illumina v2.0.

In the second dataset, 80 mice where studied. 33 of them where control mice and 47 were Pkd1 conditional tamoxifen-Cre inducible mice which had gene Pkd1 deleted when they were 40 days of age by being injected with tamoxifen. The animals' kidneys were harvested when they were between 102 and 210 days of age and their gene expression measured. The raw results were stored on Gene Expression Omnibus in 2 tables, LM1LM48 and X1X32.

5.7 Data pre-Processing

In the present study, on the first dataset the goal of pre-processing was both to put together the measurements from the 2 machines and to eliminate the noise from the data. In the case of the second dataset the goal was just to eliminate the noise from the data. As a result the pre-processing of the second dataset had fewer steps. They are presented below with an indication of the operations that apply only to the first dataset:

1. Unify the data measured with Ilumina v2.0

   The data from dataset 1 obtained by using Illumina v2.0 was combined in one matrix by simply concatenating LM5, LM6, LM8 along their samples. This led to the data being stored in 2 matrices, one obtained with Illumina v1.1 and the other obtained using Illumina v2.0. The same was done in the case of dataset 2 the obtained matrix being arranged as X1X32LM1LM48.

2. Variance stabilisation using VST

   For its high performance as is detailed in Chapter 3, VST was chosen to provide variance stabilization on the raw data. As the method is applied individually to each sample, it was performed separately on the 2 matrices obtained after step 1 for the first dataset. In the case of dataset 2, it was just applied to the X1X32LM1LM48 matrix. After the VST step, just the means of the beads for each probe in a sample were used for further processing.

3. Probes relabelling (Applied only to the first dataset)

   Since the newer versions of Ilumina sequencing provide more probes and sometimes use different labels for measuring the same sequence of nucleotides, merging measurements of the same probe from 2 different machines becomes an issue. Du et al. [100] came up with a solution to this problem by assigning a label to each probe uniquely generated based on its sequence of oligonucleotides called NuID. In the case of the first dataset, this approach was used to generate labels for the merging of the 2 data matrices. Additionally, probes in the second matrix received a unique label based on the name of the gene they tested and a number showing their order of apparition out of the probes measuring the same gene. These labels will be used as final labels to display the results of the selection.

4. Detection p-value filtering

   In the case of the first dataset, once the probes were uniquely identified, all those that did not have a detection p-value of less than 0.05 in at least one sample in either data matrices were eliminated. For the second dataset, as there is no merging step in which features will be eliminated as well as more samples being available, the authors of the study impose a harsher condition which is that at least 7 samples must have a detection p-value below 0.05 in order to be kept. In this study the same condition is kept.

5. Merging across different platforms (Applied only to the first dataset)

   Probes identified to appear in both data matrices based on their NuID had their measurements concatenated and placed in a final matrix. The rest were eliminated.

6. Normalization

   For both datasets, quantile normalization was used to bring the measurements in each sample to the same distribution. In the case of the first dataset, doing this operation after merging is especially important as measurements done with both machines should be brought in the same range.

7. Batch effect removal

   As the authors provided information about the batches the data came from, Combat was selected to eliminate batch effects in both datasets based on its performance as described in Chapter 3.

5.8 Modelling and Analysis of the PKD Datasets

After pre-processing was done on the 2 mouse ADPKD datasets, the proposed framework was applied in order to identify possible biomarkers for the disease as well as explore possible regulation between genes. In the remainder of this section, the results obtained for the 2 datasets are presented. Since in the case of the first dataset, the authors conducted a similar analysis to the present one, a comparison between methods is provided. In the case of the second dataset, the original study was more focused on comparing differences in disease manifestation between males and females as opposed to normal vs disease and the analysis diverges too much for

a comparison to be possible. The rest of the subsections are as follows: subsection 5.8.1 Gene selection presents the potential biomarkers obtained by applying $l1$-Star on the 2 datasets, subsection 5.8.2. Discovery of genes similar to the selected ones treats the new features obtained through clustering which correlate to the genes selected by $l1$-Star as well as the results of pathway analysis and finally subsection 5.8.3 Network analysis presents the regulatory networks obtained for a final set of genes of interest.

5.8.1 Genes Selection

In the study reported by Menezes et al.[242] for the first dataset, a subset of 32 samples taken when mice were between 12 and 14 days old was used for the analysis and a subset of 38 different samples was used to validate that the selected genes show changes in value between mutant and control subjects. The criteria for genes to be selected as differentially expressed was to exhibit a fold change of > 1.2 or a fdr-adjusted p-value of < 0.05 in the group of samples taken at day 12 and a fold change > 1.2 and a fdr-adjusted p-value < 0.05 at day 14 between mutant and control subjects, resulting in a final set of 87 genes out of the around 20000 initially measured.

In the present study, the pFDR-corrected $l1$-Star was run 10 times using a 4-fold validation scheme which resulted in 40 classifiers using all 70 samples. The genes were ranked based on the number of classifiers in which they appear as a final selected feature. In the first dataset, 23 genes were identified Table 5.1 (Figure 5.3(A)), 16 of which appeared as differentially expressed in the original study. For the second dataset, a set of 13 genes were identified Table 5.1 (Figure 5.3B). The average accuracy of the classifiers was $95.85 \pm 5.89\%$ on the testing set for the first dataset and $99.25 \pm 1.81\%$ for the second one. In order to evaluate the performance of the proposed method, a NaiveBayes classifier from a standard MATLAB package was used on the same 40 training/testing partitions for the data in the 2 sets. The NaiveBayes produced a classification accuracy of $88.92 \pm 6.61\%$ on the first dataset and $92.37 \pm 5.55\%$ on the second one. A Welch t-test revealed that the difference in performance is statistically different for the 2 algorithms on both data set with a p-value of 4.02×10^{-6} for the first dataset and 1.06×10^{-10} for the second dataset.

Out of the 3 most selected genes for the first dataset, *Chpf* and *Nupr1* human homologues, *CHPF* and *NUPR1* were associated with tumours, described in the

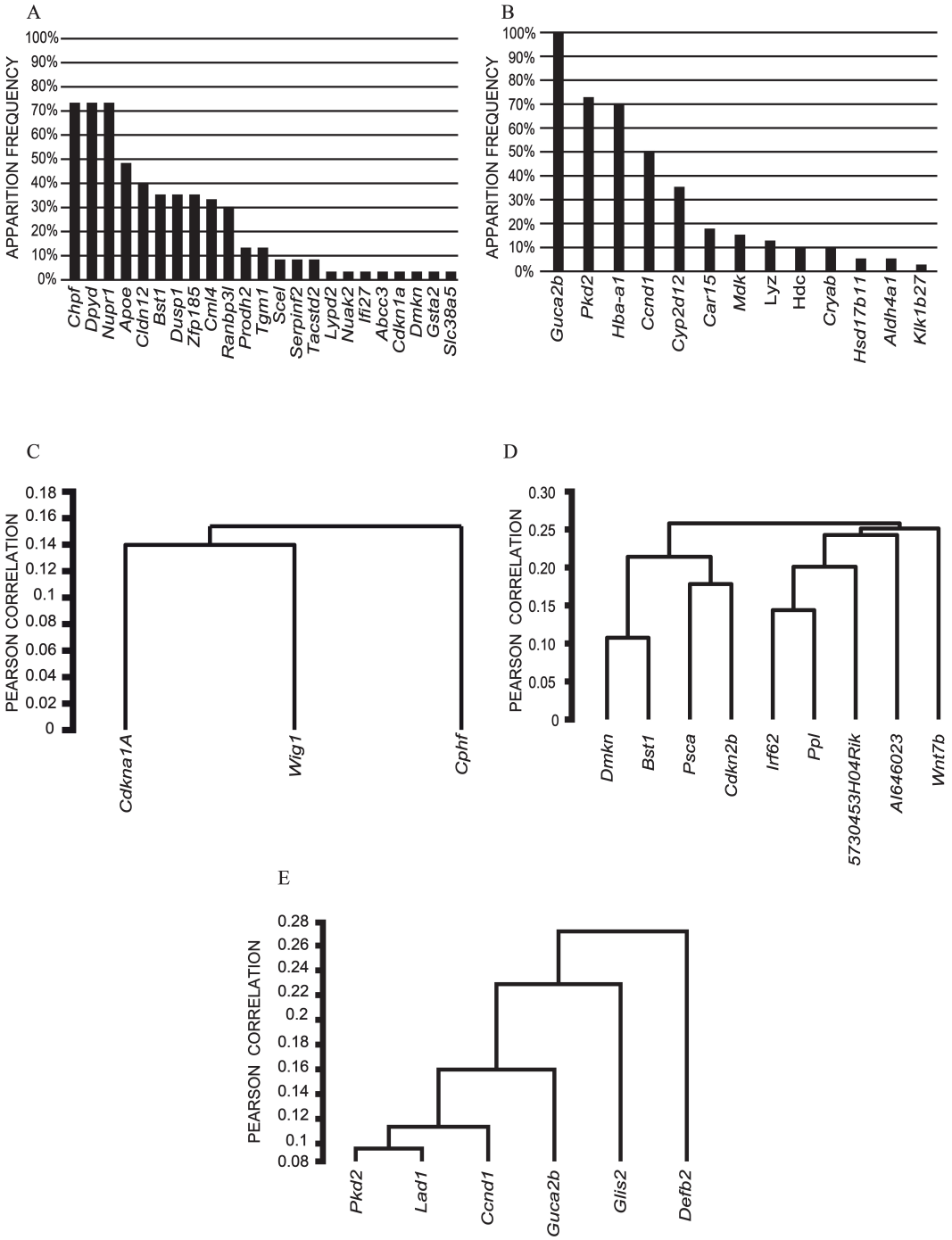

Figure 5.3: Results for gene selection and clustering analysis. A) Frequency of apparition of genes selected for first dataset, B-Frequency of apparition of genes selected for second dataset, C) Dendrogram for the cluster for *Cphf*, D) Dendrogram for the cluster for *Dmkn*, E) Dendrogram for the cluster for *Guca2b*

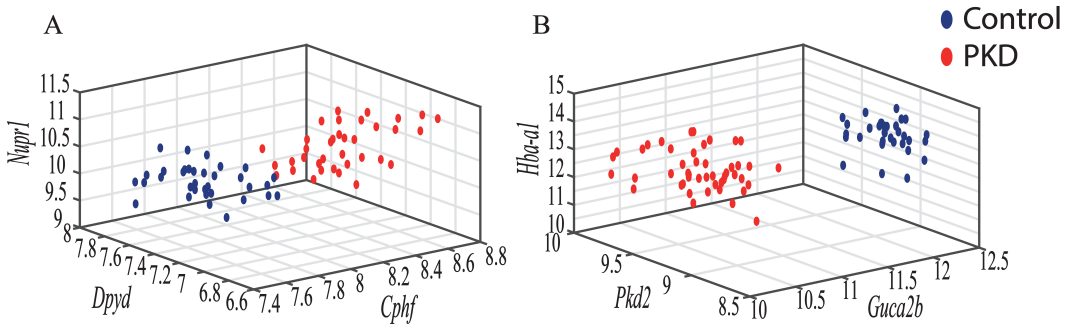

Figure 5.4: Graph for the samples of 3 most selected genes in each dataset. A-Graph for the genes selected for the first dataset B-Graph for the genes selected for the second dataset.

work of García-Suárez et al. [131] and Chowdhury et al. [74]. The gene *Dpyd* was shown to be differentially expressed in ADPKD by another study [71]. Figure 5.4(A) presents a plot for the samples for the 3 genes in which it easy to observe their capacity to separate between normal and disease samples. Also, gene *Cdkn1*, previously shown by Bhunia et al. [36] to play a role in ADPKD appeared as differentially expressed in this study. Other genes that have appeared as differentially expressed or their homologues appeared as differentially expressed in other studies on ADPKD are *Bst1*[193], *Dusp1*[162], *Prodh2*, *Serpinf2* [274] *Abcc3* [63]. In the case of the second dataset, the 3 most selected genes were *Guca2b* whose human homologue is a biomarker for cancer [306], *Pkd2* which is the second gene that is responsible for ADPKD and *Hba-a1* whose human homologue was shown to be upregulated in cancer cells [215]. Figure 5.4B presents a plot for the samples of the 3 genes. Another interesting gene that was found was *Ccnd1* whose human homologue was shown to be expressed in cystic kidneys and is part of the Wnt signalling pathway that is misregulated in ADPKD [204]. Other genes that were selected an. d have been shown to be differentially expressed of their homologues have shown to be differentially expressed in ADPKD are *Lyz*, *Aldh4a1*[71] and *Cryab* [18].

Dataset1		Dataset2	
Selected gene	Gene expression in the mutants vs controls	Selected gene	Gene expression in the mutants vs controls
Chpf	up	*Guca2b*	down
Dpyd	down	*Pkd2*	up
Nupr1	up	*Hba-a1*	down
Apoe	down	*Ccnd1*	up
Cldn12	up	*Cyp2d12*	down
Bst1	up	*Car15*	down
Dusp1	up	*Mdk*	up
Zfp185	up	*Lyz*	up
Cml4	down	*Hdc*	up
Ranbp3l	up	*Cryab*	up
Prodh2	down	*Hsd17b11*	up
Tgm1	down	*Aldh4a1*	up
Scel	up	*Klk1b27*	down
Serpinf2	down		
Tacstd2	up		
Lypd2	up		
Nuak2	down		
Ifi27	up		
Abcc3	up		
Cdkn1a	up		
Dmkn	up		
Gsta2	down		
Slc38a5	down		

Table 5.1: Genes selected for the 2 datasets

5.8.2 Discovery of Genes Similar to the Selected Ones

In the original article for the first dataset, WGCNA was used to cluster the genes while Gene Ontology (GO) analysis and Ingenuity Pathway Analysis (IPA) were further used to find genes biologically related to the selected ones. One cluster with 629 genes represented by an eigengene that correlated well with the genotype that contained 67 of previously reported genes was discovered. IPA revealed 4 genes related to the selected ones, 3 of them, *Tnf*, *Agt* and *Avp* previously connected to the disease and the 4th one, *Hnf4*α was experimentally proven to play a role in the disease. In the article for the second dataset, WGCNA was used just for network topology analysis.

In the present analysis, stage 2 of the proposed method was applied on the pre-processed datasets and the clusters containing the genes selected at stage 1 were isolated. As it can be seen in Figure 5.3C, in the first dataset, *Cphf* was clustered together with *Cdkn1* and *Wig1*, a gene whose human homologue was proven to regulate the human homologue of *Cdkn1*[188]. Also gene *Wnt7b* which was clustered together with *Dmkn* (Figure 5.3D) appears to be of interest since it was proven to play a role in cystogenesis in polycystic kidney disease [289]. Interesting enough, in the original article, Wig1 and *Wnt7b* have appeared in the final set of selected

genes. In the case of the second dataset, the 3 most selected genes, *Guca2b*, *Pkd2* and *Ccnd1* were clustered together (Figure 5.3E). Another gene that might be related to the disease is *Glis2* which was clustered with them and has been proven to be responsible for nephronophthisis [14].

In order to select a final list of genes relevant to the disease for network analysis, stage 3 of the proposed framework was used on both datasets. In order to reduce the number of genes on which expert knowledge is applied, just the clusters containing the 3 most selected genes were analysed. More precisely, STRING was used on a set of genes containing those selected by the algorithm, the genes in the clusters containing the first 3 most selected genes and *Pkd1* in order to find possible protein interactions. To explore a high number of possible connections, the level of confidence set for displaying a connection was set to a minimum. An interaction map of genes that potentially interact with each other and that also contains *Pkd1* was found for both datasets as shown in Figure 5.5.

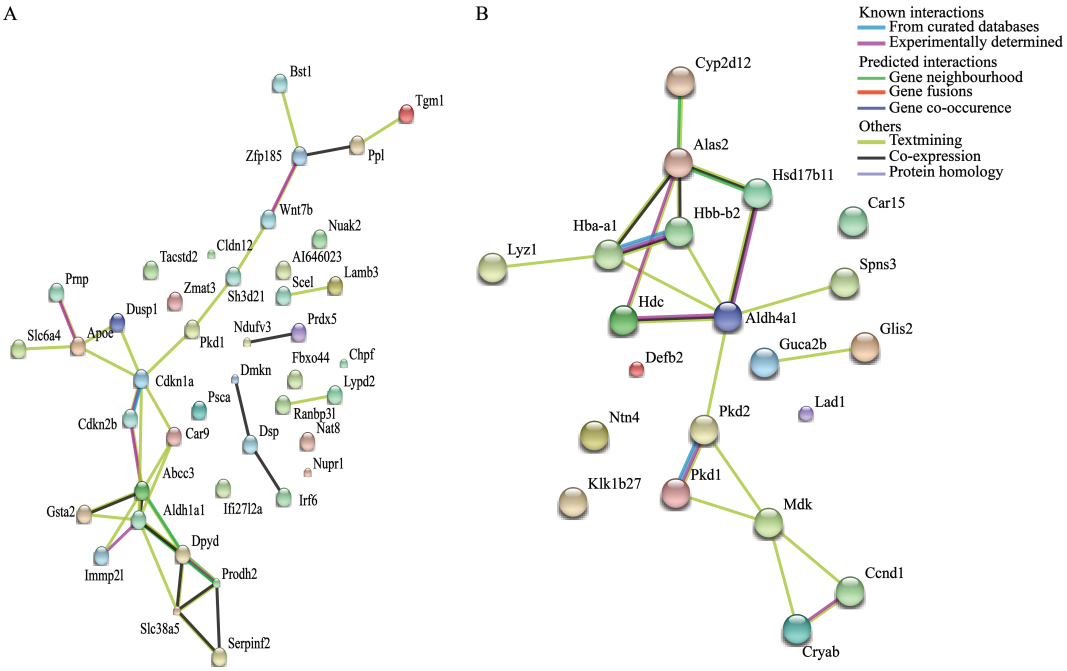

Figure 5.5: Protein interaction map for the 2 datasets. A-Dataset 1, B-Dataset 2

In the case of the first dataset, the final set of genes selected contained the genes in the path determined through STRING plus *Cphf* and *Nupr1*. *Cphf* was added as it was the gene most selected by the first stage of the framework. Combined with the fact that it is not a very well studied gene, it seems to be a good candidate for novel research in the field. Similarly *Nupr1* was frequently selected by the

algorithm and its effects on PKD have not been researched up to now although it is important in cancer biology research. For the second dataset since there were few genes to be analysed, all the genes selected by the algorithm plus the ones that have been clustered with the 3 most selected genes had their regulatory network approximated.

5.8.3 Network Analysis

In the original article describing the first dataset, the purpose of network analysis was to find if there was preservation of gene correlation networks in both conditions. The same 32 samples used for finding differentially expressed genes were analysed. Gene correlation network comparison was done between 4 conditions: P12 vs P14 and mutant vs control. The results showed that there was little change in gene topology between mutant and control subjects. A more significant difference appeared between P12 and P14 with one cluster changing its position in the network. The authors of the study theorize that the genes in that cluster might be responsible for a trigger that greatly affects the speed of the evolution of the disease which they have found in another study [242].

The purpose of the present analysis was to dynamically model the regulatory network that exists between possible relevant genes in the case of mutant and control subjects. The last stage of the proposed framework was applied on the final set of genes for the 2 datasets.

In order to study the dynamical behaviour of a system, time-points have to be defined for the measurements. In the 2 datasets, two time points were provided for each measurement: the day at which tamoxifen started to be administered to the animal leading to *Pkd1* being inactivated and the day at which the animal was killed and a sample of its gene measurements was taken. In order to unify the 2 measurements in a way relevant for the present study, a time-point for a sample was considered to be the difference in days between Pkd1 inactivation and the terminal sample.

To model individual gene expression, the models used were the ones described in the subsection 5.5.1. In the case of the first dataset as only 8 time-points existed, the value for p was 3 so that the number of parameters, 7 would not exceed the number of time points. In the case of the second model, 21 time points were available so a more flexible model with 13 parameters, thus with p = 6 was used

for interpolation. For interpolation the minimum timepoint was subtracted from the rest of the timepoints so that the dynamics of the systems will be modelled starting with t_0=0.

On top of the regularization described in subsection 5.5.1, linear constraints where imposed when approximating the parameters. This happened because, as presented in section 5.6 describing the dataset, if Pkd1 was inactivated before 13 days age, they develop large cysts within a 3-week interval but after 6 months if the gene is inactivated later. This suggests that the system representing kidney gene expression reaches a steady state as cyst growth reaches a plateau. In the case of the first dataset, the last time point corresponds to 12 days while in the case of the second dataset it corresponds to 108 days. Further biological research on the disease may lead to a better approximation of the time when gene values get to a steady state.

The linear constraints for the parameters of the first dataset were $\lambda_{1...3} > 0$ and $w_0 \in (y(t_8) - 3\sigma_m, y(t_8) + 3\sigma_m)$ where σ_m represents the maximum variance for the measurements at any time point and t_8 represents the last timepoint. In the case of the second dataset the conditions were $\lambda_{1...6} > 0$ and $w_0 \in (y(t_{21}) - 3\sigma_{21}, y(t_{21}) + 3\sigma_{21})$ where σ_{21} represents the variance in measurements for the last time point and t_{21} represents the last time point.

As it can be observed in Figure 5.6, in the case of the first dataset, *Cdkn1a* appears to up-regulate *Cphf* in the network for control subjects, a connection that is broken in the case of the mutant subjects. In the case of the second dataset, *Pkd2* seems to downregulate *Guca2b* in the case of the mutant subjects (Figure 5.7). These results open a research field on biological investigation to assess the connection between the 2 pairs of genes.

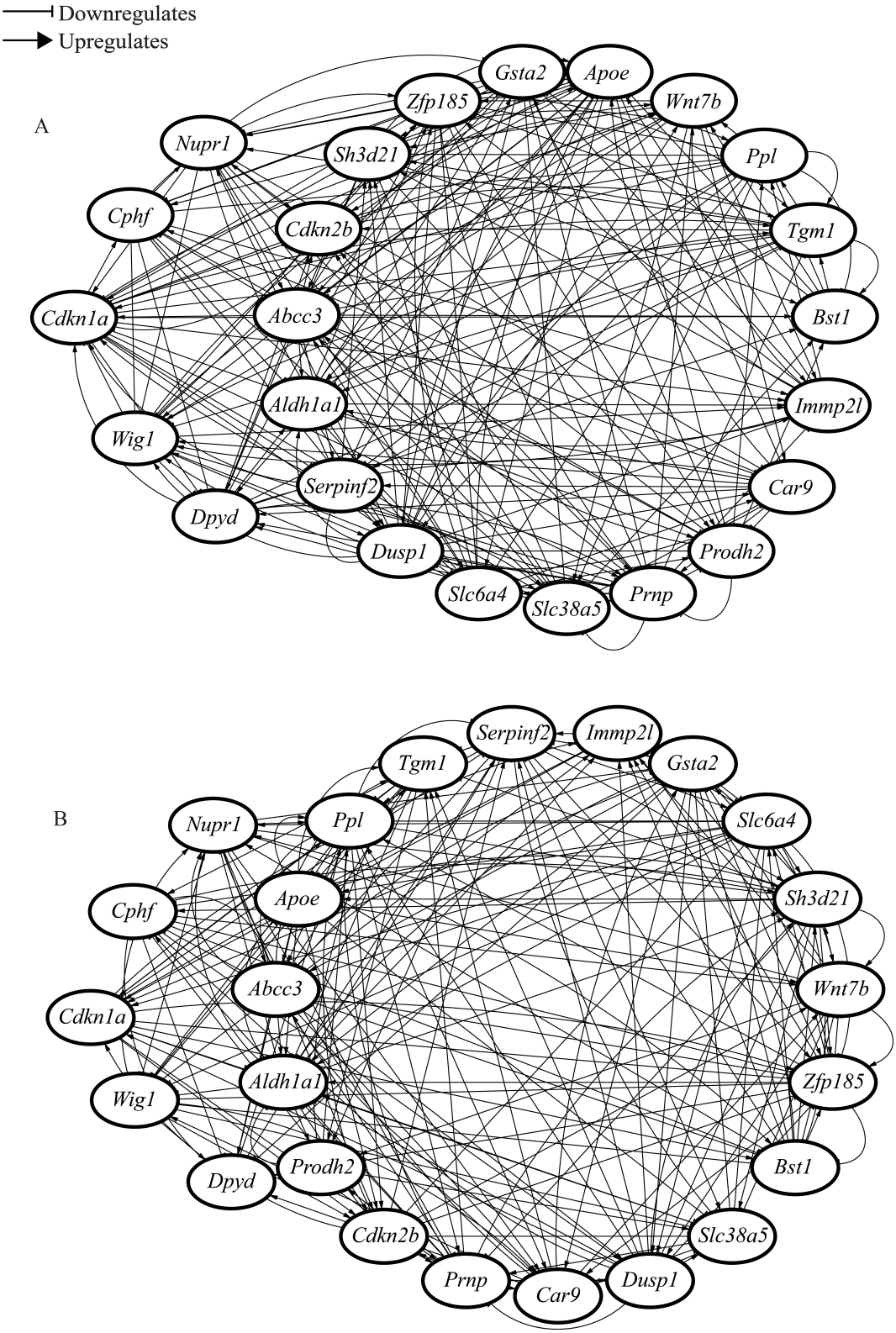

Figure 5.6: Modelled gene network for the first dataset. A) Network for control samples B) Network for mutant samples

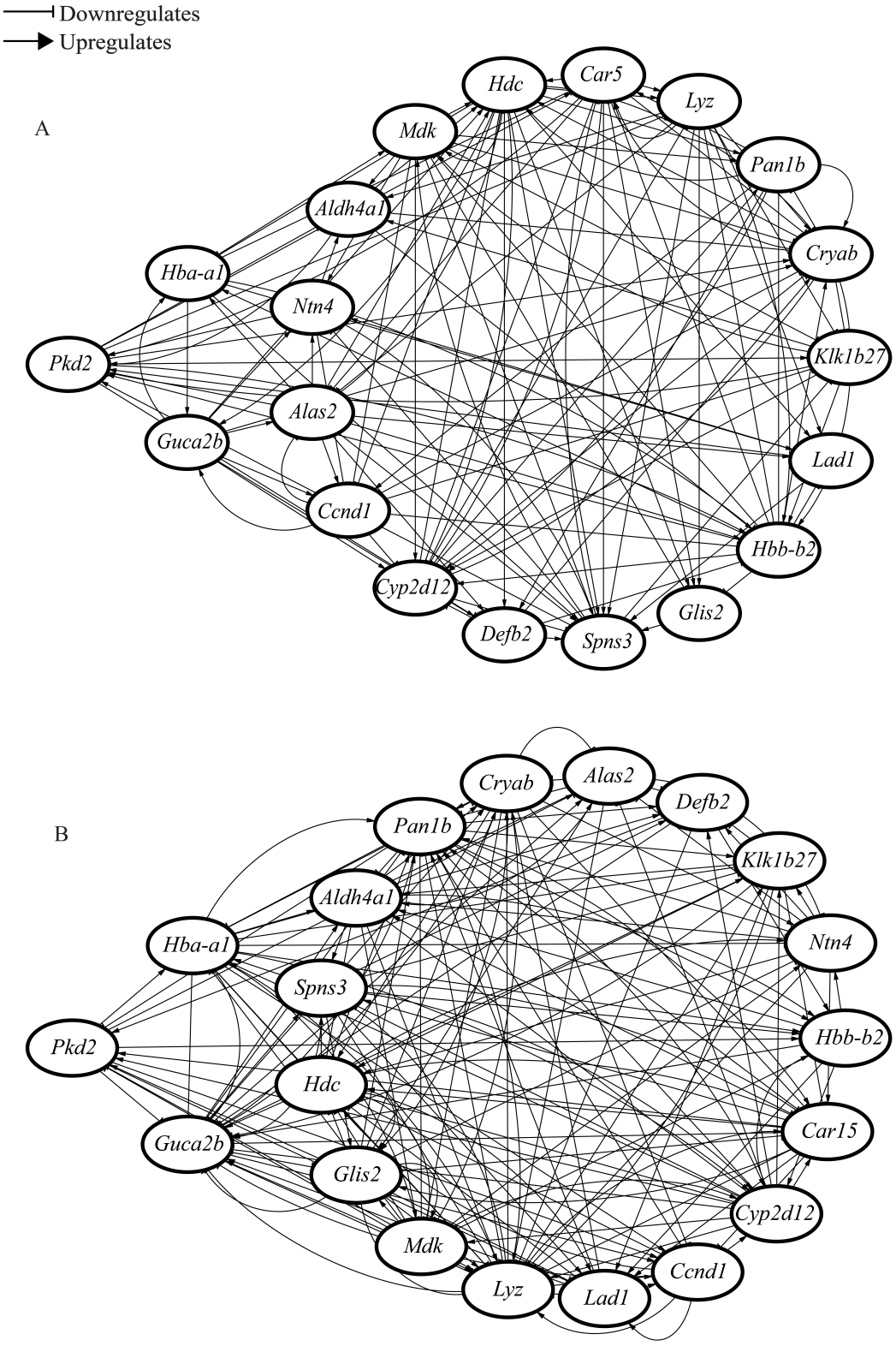

Figure 5.7: Modelled gene network for the second dataset. A) Network for control samples B) Network for mutant samples

5.9 Summary

This chapter proposes a new multi-stage framework for the analysis of time-series microarray data coming from case-control studies. The framework aims to achieve 2 important goals in gene expression data studies, biomarker selection and gene regulatory network identification.

In stage 1, a supervised feature selection method, based on the $l1$-Star algorithm is employed to identify a significantly reduced set of genes that discriminate between control and affected samples. The filtering method of the original algorithm was replaced, from a single hypothesis statistical t-test to a multiple hypothesis corrected method. The strategy is employed on a number of different initializations of a k-fold cross validation scheme. The features are ranked based on the frequency with which they appear in the classifiers obtained this way.

Stage 2 consists in the use of clustering methods for identifying features correlated to the ones selected at stage I that might have been eliminated by the supervised feature selection algorithm. A pFDR filter is used to eliminate features with a statistically insignificant difference between conditions and hierarchical clustering is used with Pearson absolute value as a measure of distance and average distance as a linking rule. In order to obtain clusters, the tree is cut using the DynamicTreeCut method. The genes in the clusters containing previously selected features form a set that can be used for further analysis.

Stage 3 marks the final step in selecting a set of potentially relevant genes. At this point the researchers can combine 3 strategies for choosing the genes. The first one is to use their expert knowledge to isolate genes from those found during clustering and feature selection that could be relevant for mechanisms of the disease. The second strategy is the use of the STRING database to detect previously reported or predicted connections between the genes. The third one is to pick genes frequently selected at step one, for which previous biological knowledge is reduced.

Stage 4 is the stage at which the identification of the gene regulatory network of the selected genes takes place. The employed model is a system of linear differential equation that approximates the response of the system represented by the regulatory network of the selected genes around an equilibrium point. The identification of the model parameters is done in two steps. First, individual models representing sum of exponentials are fitted through the measurements of each individual genes. Second, the individual models are sampled to create enough measurements and the model for

the regulatory network is identified using a semi-definite programming technique available in the literature that allows imposition of the biological characteristics for GRN's, sparsity and stability.

All the presented methods have been previously used in microarray analysis studies but to the author's knowledge it is the first time when they have been combined in a complete framework for analysing gene expression data, by selecting a set of features of interest and creating a dynamical model for their regulatory network.

The proposed framework has been applied to 2 microarray datasets coming from studies corresponding to fast and slow progression models of ADPKD. The selection method has identified a set of genes previously connected with ADPKD, either experimentally proven to play a role in the disease or found as differentially expressed in other ADPKD studies. The analysis has also identified some novel genes of interest, chosen by the feature selection algorithm, the most significant ones being *Cphf* for the fast progression model and *Guca2b* for the slow progression one. Of interest, both genes have been associated with cancer and since both tumours and cysts consist of defective cell turnover, they are promising candidates to explore new mechanisms of the disease. Further biological analysis is needed to confirm their involvement in ADPKD.

Chapter 6

Analysis of Normal vs PKD1,2 Knockdown ciPTEC Cells and the Effects of Octreotide in Low-Density Free Migration Assays

6.1 Introduction

While the previous chapter dealt with the genetic component of ADPKD, the current and the following chapter will describe work done to investigate the disease from the perspective of cell behaviour. This type of analysis is important, as the changes in gene expression do not have too much meaning in themselves when trying to understand a disease, but the way in which they translate into cellular properties is the ultimate goal for genetic research.

By studying both gene expression and cellular behaviour, the scientist can make the connection between genes and the cellular processed they control. Another reason for studying diseases from the perspective of cell behaviour as opposed to microarray studies is that the experiments are cheaper and simpler to set up so that for example the effects of drugs on a disease can be easily quantified without the need of studying the deeper level of gene expression.

Several cellular features altered in polycystic kidney disease have previously been described in a number of studies. These include changes in cell division [292][143], migration [59], [270] and apoptosis [41] [318]. A more thorough de-

scription of these studies has been recorded in Chapter 2 and Chapter 4-section 4.4. A variety of different methods have been employed for these analyses including the use of monoclonal antibodies against proliferating cell nuclear antigen (PCNA) [292] and cell counting [22] at different times to study proliferation, use of modified Boyden chambers [258] and wound healing assays [226] for cell migration and the use of serum starvation and cell counting [377] for assessing cellular resistance to apoptosis.

This chapter presents the analysis of an immortalised human proximal tubular cell line (ciPTEC) [241] using time-lapse microscopy, where initially cells are plated at low density. The present analysis focuses on two specific functions, namely cell division and cell migration. This type of assay has not been used very frequently to study cell movement in the case of ADPKD as modified Boyden chambers and wound healing assays have been more popular . The only similar published analysis was a study by Yao et al. [395] in which the authors study the linearity of cell movement by looking at the confinement ratio of the cell trajectories.This chapter presents a more complex analysis in which random diffusion models are employed to characterize cell migration. The current methodology enables tracking individual cells over the entire period of the experiment, enabling a precise quantification of cell division and motility. In the future the current work of this chapter can be combined with genetic analysis to discover how different genes affect motility and division of ciPTEC cells.

Two main questions were addressed. The first question relates to the effects of disease on cell migration and division by comparing two cell lines: control PTEC cells and PTEC cells with knockdown of PKD 1,2. The second question relates to whether either defect could be corrected in disease cells by octreotide, a somatostatin analogue that has been shown to inhibit secretin-induced cAMP generation in cholangiocytes of animal models of polycystic liver disease [236]. In these experiments, disease cells incubated with octreotide were compared to disease cells treated with DMSO alone.

The remainder of this chapter is organised as follows: section 6.2 provides a detailed description on the experimental materials and methods employed in this analysis, section 6.3 on the analytical ones, section 6.4 presents the results on the analysis done on control vs knockdown cells, section 6.5 the analysis on the effects of octreotide, section 6.6 provides a discussion on the results considering the literature on the subject and further analysis that should be done and section 6.7

summarises the results and draws the conclusion of the chapter.

6.2 Experimental Materials and Methods

6.2.1 Materials

6.2.1.1 Cell Lines Description

The cells used in this experiment were derived and described in Mekahli et al. [241]. A brief description of the procedure used to generate these lines is presented below. miRNA-based short hairpins (miR-shRNA) were created to knock down *PKD1* and *PKD2*. The controls were created by using a miR-shRNA which was directed against DsRed.

Lentiviral vector transfers were produced using miR-shRNA sequences which contained a promoter which was driving a gene that induces resistance to blasticidin. Four vectors were created which encoded miR-shRNA directed against:

- DsRed
- *PKD1*
- *PKD2*
- both *PKD1* and *PKD2*

For producing the ciPTEC cell lines, primary cells coming from patients urine were immortalized using SV40T vector containing geneticin (G418) resistance and a hTERT vector containing hygromycin resistance (hygromycin B). The lentiviral vectors were added to the culture medium in order to create the different types of cells. 10 μg/ml blasticidin were used for the selection of the transduced cells.

In the present study the cells infected with the miR-shRNA sequences designed against DsRed were used as control cells while the *PKD*1,2 knockdowns were used as disease cells. Through the rest of this document the former ones will be called healthy or control while the later will be termed disease, PKD (in figures) or *PKD1,2* knockdown.

6.2.2 Methods

6.2.2.1 Mammalian Cell Culture

PTEC control/ PKD1,2 knockdowns were cultured in DMEM-F12 media (Gibco, 717 Grovemont Cir, Gaithersburg, MD 20877, USA) supplemented with 10% FBS (Biosera, 117 General Ordonez Ave, Marikina, 1800 Metro Manila, Philippines), and penicillin and streptomycin solution (BioWhittaker®, 8830 Biggs Ford Rd, Walkersville, MD 21793, USA). Cells were plated in T25 flasks until they reached 80% confluence with 10 ml media and 10 μl blasticidin (Invitrogen, 168 Third Avenue Waltham, MA USA 02451) and then passaged to T75 flasks where they were grown in 13 ml of media with 13 μl blasticidin. The culturing of cells was done in incubators with 33° temperature and 5% CO_2.The media was replaced the first day after plating in the T 25 flasks and subsequently replaced every 2 days except weekends either in the T25 or later in the T75 flasks.

Transfer to the T75 flasks was performed by removing media, washing them with PBS and incubating them with 2 ml Trypsin at 33° C for 3-5 minutes until they have detached. Next, 8 ml culture media was added to neutralize the effects of trypsin. The cell suspension was then transferred to a 30 ml universal for further processing. Finally 1 or 2 ml of cell suspension were moved to the T75 flasks and media was added so that the final solution would contain 13 ml of cell suspension. Basticidin would be added as described. The same procedure was used to passage the cells in the T75 flasks whenever they would reach 80% confluency.

6.2.2.2 Freezing and Thawing of Cultured Cells

Cells were cryogenically preserved at a temperature of -80° C. The following procedure was used for freezing. After the cells were grown in T25 plates and reached 80% confluence, the passaging procedure described above was applied, with 1 ml of cell suspension being transferred to the T75 flasks. The remaining cells were centrifuged and resuspended in a solution containing 1.8 ml media and 0.2 ml Dimethyl sulfoxide(DMSO). Next the cell suspension was aliquoted to 2 cryogenic vials and moved to the -80° C freezer. For plating in the T25 flasks the cells were thawed by gently manually swirling the vials in a 37° C water bath 60-90 seconds until the cell suspension was defrosted. Next, the suspension was transferred to T25 flasks with 9 ml media and 10 μl blasticidin was added to the mix.

6.2.2.3 Cell Motility Experiment

Two days before the experiment, the media in the T75 flask was replaced with serum free media for cell cycle synchronization. Next, the cells in the T75 flasks were replated as explained above and the remaining cells in the 30 ml universals were used for the motility experiment. First, they were counted using a haemocytometer to assess their density in the cell suspension and the solution was diluted through successive transfers and dilutions to achieve a density of 10000 cells/ml corresponding to 2222 cells/cm^2 in each well. The obtained suspension was replated in 6-well plates with 2 ml cell suspension per plate and 2μl blasticidin.

For the 2 types of experiments, different setups were used. For the control vs disease experiments, control or disease cells were plated into 3 wells each on the same plate, left to attach for 5.5-6 hours and then placed under a microscope. For the octreotide study, all 6 wells were filled with knockdown cells. In the case of the octreotide experiment, after 5.5 hour the media was replaced, in the case of the octreotide group with a solution containing culture media with blasticidin and octreotide at a final concentration of 1 μmol in the media and in the case of control wells with culture media with blasticidin and 1 μl DMSO/ 1 ml of culture media. 3 wells were used for control (DMSO) and 3 wells for octreotide per plate. For imaging, one frame was taken every 10 min over 96h using an Olympus Ix70 microscope controlled by Micro-Manager v1.4 [106]. During imaging, cells were enclosed in a chamber maintained at 33° C under a humidified atmosphere of 5% CO2 in air. For each well, 9 fields were taken.

The concentration of the octreotide was chosen based on an article by Macaulay et al. [228]. In it the authors have an experimental setup similar to the present analysis in which they incubate cells with octreotride or saline solution for the controls for 70 hours. The concentrations for which octreotide shows the most effects on division are 1 nmol and 1 μmol concentration on small-cell lung cancer cell line HX149. Although in the original study the effects for 1nmol were slightly higher than for 1 μmol, the higher concentration was picked for this analysis as it was considered that it would have better chances of producing a more significant effect considering that the cells are different. The present concentration is not to be considered the absolute best for ciPTEC cells, further optimizations could be carried in the future. For each experimental setup two replicates were performed.

6.3 Analytical Methods

Two types of measurements were used to characterize the cells in under between both conditions. First, the average and standard deviation of the division time for cells in under different conditions was calculated to find out if the disease produces produced any changes. Next, cell motility of cells was quantified by calculating the MSD and then fitting a model thorough it. The following subsections describe the procedures used to extract the 2 measurements.

6.3.1 Cell Division Time

In order to quantify the division time, once a mother cell divided, the daughter cells were tracked during the movie by visual inspection and the number of frames between the onset of cell division and separation of the mother cell into two cells was considered the division time for the cells. In order to have enough representatives for a reliable approximation, 30 cells had their division time calculated for each condition in the first replicate and 80-100 per condition in the other replicates. To ensure uniform sampling, the cells were randomly selected to be around 10-35 cells per well.

6.3.2 Cell Motility Analysis

For the motility analysis, individual cells were tracked for the first 72 frames which corresponds to the first 12 hours. The cells chosen had not divided within this time interval and this duration was chosen so that most cells would not have divided based on their expected doubling time. This was important to allow free cell movement.The duration chosen was also in the range used for similar experiments ie, 8 hours in the case of Barbaric et al. [20] and approximately 16 hours in the case of Dieterich et al. [94] although different cell types were studied (MDCK, H7 and H14). Tracking was performed manually using the manual tracking plugin for Image J by Cordelières [79] using the centre of each cell as its tracking point. The origin of the axes system used is the top left corner of the image with the values on the x-axis increasing from left to right and the values on the y-axis from top to bottom. The same strategy for ensuring uniformity as in the case of the division time was applied ie 100 cells were tracked per condition, with around 30-35 cells from each well.

For the MSD calculation, the standard formula that was presented in subsection 4.3.1.3 was used.

6.3.3 Cell Motility Analysis Modelling

The diffusion model proposed in Dieterich et al. [94] :

$$MSD(t) = 4v_{th}^2 t^2 E_{\alpha,3}(-\gamma_\alpha t^\alpha) + (2\eta)^2 \tag{6.1}$$

was used to characterise the mean squared displacement of cells estimated from the individual cell trajectories obtained from the time-lapse imaging data.

The model parameters were estimated as described in Chapter 4 section 4.3.2. Because the procedure is sensitive to initialisation and in order to avoid being trapped in a local minimum, multiple estimation runs were performed, each run starting from different initial conditions. The main steps of the algorithm used for fitting the parameters are summarised below:

Step 1:Initialization

v_{th} parameter: The initial value of the vth parameter was estimated directly from the trajectory data. Specifically, for each cell, the second moment of its distribution of speeds is computed as:

$$< S^2 >= \frac{\sum_{i=1}^{N-1} s_i^2}{N-1} \tag{6.2}$$

where $s_i = \Delta p_i / T$ with Δp_i the displacement between frames i and i+1, T the period between frames and N the total number of frames for which the cell is tracked.

As a result a population of 100-105 measurements was obtained for each condition. Next, for each population a Weibull distribution was fitted to data [47]. 50 random samples drawn from the distribution were used to initialise vth^2 .

For the remaining parameters, 50 initial values were generated by random sampling from a uniform distribution on the interval [0,1]. The choice of interval is motivated by the fact that in the case of normal/super-diffusion movement, which is the type of movement observed experimentally, the parameters

α and γ belong to this interval, as shown in previous studies [20][94]. As no previous information can be used about the last parameter, η as it is a measure of noise it was also initialized in the interval [0,1] .

Step 2: Model Fitting

Model were fit using the trust region reflective algorithm [76] implemented in the lsqcurvefit function in MATLAB by minimizing:

$$\sum_{t_i}(F(x,t_i) - MSD(t_i))^2 \tag{6.3}$$

where MSD(t_i) is the real value for the MSD at time t_i, x the vector of parameters for the diffusion model $\begin{bmatrix} v_{th}^2 \\ \alpha \\ \gamma \\ \eta \end{bmatrix}$ and $F(x,t_i)$ the diffusion model proposed by Dietriech evaluated at timepoint t_i for the values of the parameters in x with 0 as the lower limit for the parameters and infinity as the upper limit except for α where the upper limit is 1.

Step 3: Model Selection

The selection of models was done in 2 stages.

The first stage consisted in a threshold on the error values of the models. The formula used for calculating the errors was:

$$\sum_{i=1}^{72} \frac{\hat{y}_i - y_i}{y_i}^2 \tag{6.4}$$

corresponding to the squared relative errors where $\hat{y}_i$ is the model approximate for time-point i and y_i the real value for MSD at time-point i. The threshold selected was 3.6 corresponding to an average 5% squared relative error for each time-point.

After this step it was observed that most of the values obtained for the parameters seem to be in a similar range although a few significant outliers exist. In order to eliminate the outliers, the interquartile range (IQR) outlier elimination procedure proposed in Tukey [362].

In the case of the present study, the procedure was applied to the set of values obtained for each parameter of the models that remained after the first model selection stage step. The models which appear in all the outlier-free sets for each parameter were selected to the final set of models.

6.4 A Comparison Between the Phenotype in Healthy and Disease Cells

6.4.1 Division time

The division time between the two cell types was compared by looking at the number of frames between 2 consecutive divisions as discussed in subsection 6.3.1. The results are presented in Figure 6.1.

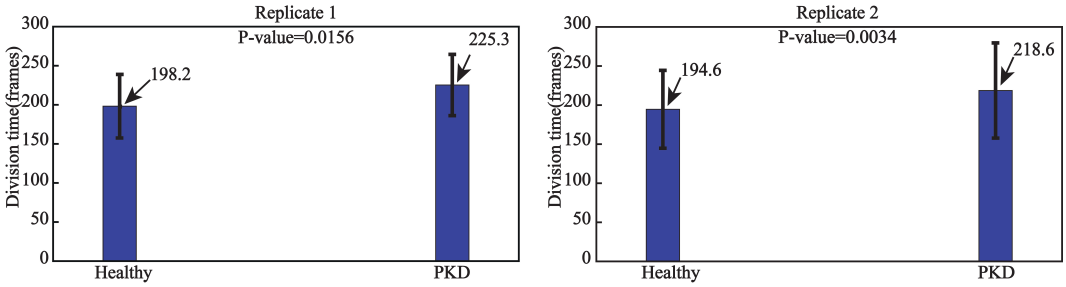

Figure 6.1: Means, standard deviations and the results of a Welch t-test for the division times of healthy vs disease cells

As it can be observed in both replicates, the division time was significantly increased in the case of the disease cells with around 25-30 frames which translates to 4-5 hours. This was an interesting result since the general consensus in the literature is that proliferation is increased in the case of the disease cells. Also it is important to note that there is great agreement between replicates, showing that 30 cells per condition were sufficient to obtain a good average estimate of the division time.

6.4.2 Motility Analysis

The MSD for each condition in an experiment was calculated as described in subsection 6.3.2 and then 50 models were fitted to it as described in subsection 6.3.3. For the first replicate 96 healthy and 101 disease cells were tracked while for the second replicate 101 healthy vs 100 disease cells were tracked.

Figures 6.2 and Figure 6.4 present the MSDs with the most accurate model fit. All the parameters of the models that remained after the 2 filtering steps have been taken into consideration to characterize the conditions (Figure 6.3, Figure 6.5).

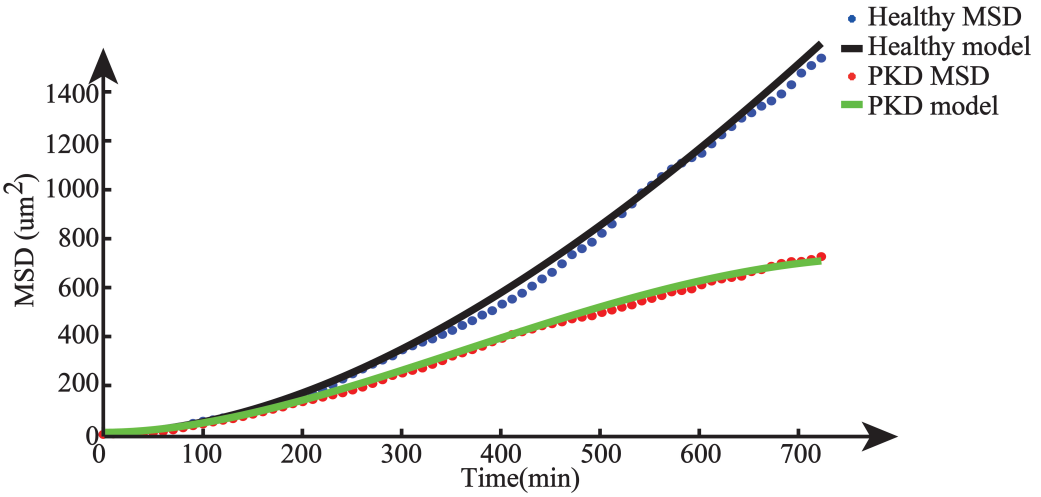

Figure 6.2: MSDs and best model fits for the migration results of the first replicate of the healthy vs disease cells experiment.

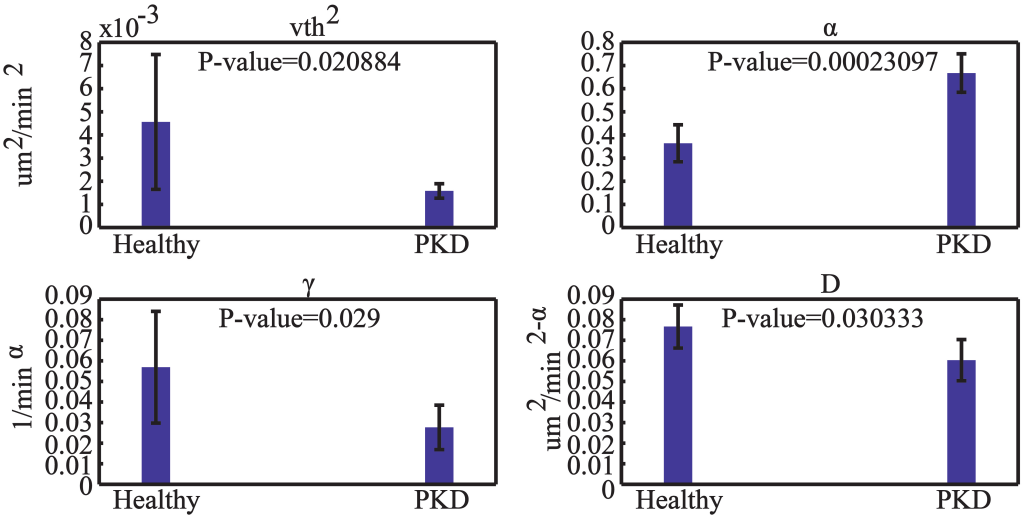

Figure 6.3: Means, standard deviations and the results of a Welch t-test for the parameters of the models for the 2 conditions in the first replicate of the healthy vs disease cells experiment.

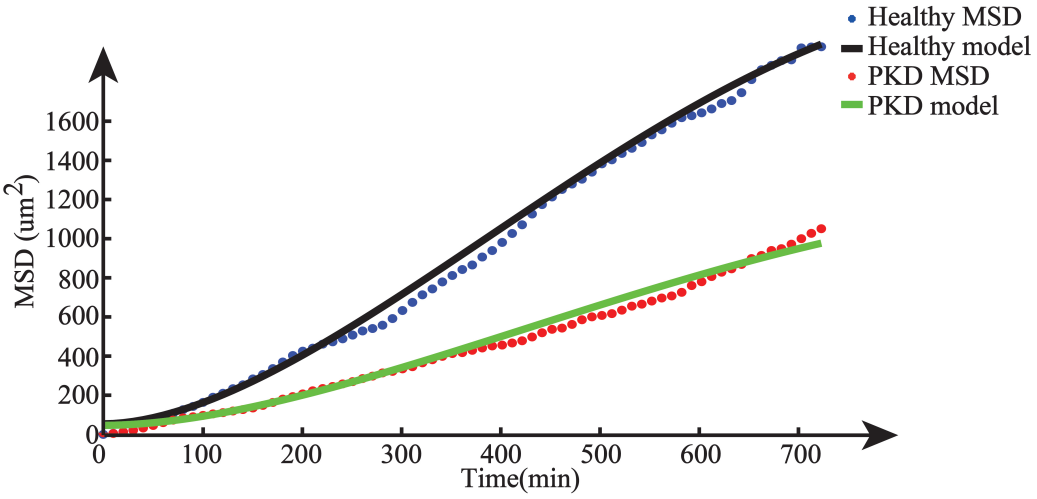

Figure 6.4: MSDs and best model fits for the migration results of the second replicate of the healthy vs disease cells experiment.

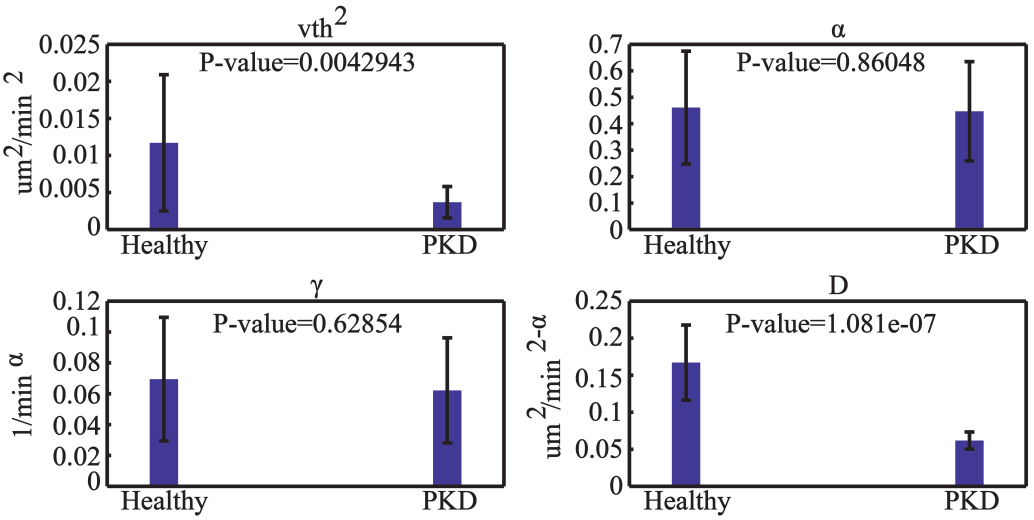

Figure 6.5: Means, standard deviations and the results of a Welch t-test for the parameters of the models for the 2 conditions in the second replicate of the healthy vs disease cells experiment.

As the figures show, the MSD seems to show lower values in the case of the disease cells. This is consistent with previous studies using different assays such as Boyden chambers[258], free migration assays[270] or wound closing assays[40] showing that the migration capability of disease cells is reduced.

The first conclusion based on these results is that thermal speed (vth^2) is lower in the case of the PKD. In both replicates, the difference was significant both statistically and in terms of magnitude, with the thermal speed of the affected group being less than half of that of the control. A second parameter with a statistical difference was the diffusion coefficient D, indicating that the surface covered by the healthy cells increases faster.

6.5 A Comparison Between DMSO and Octreotide Treated Disease Cells

6.5.1 Division Time

The division time of the cells was compared between the 2 growing conditions (Figure 6.6).

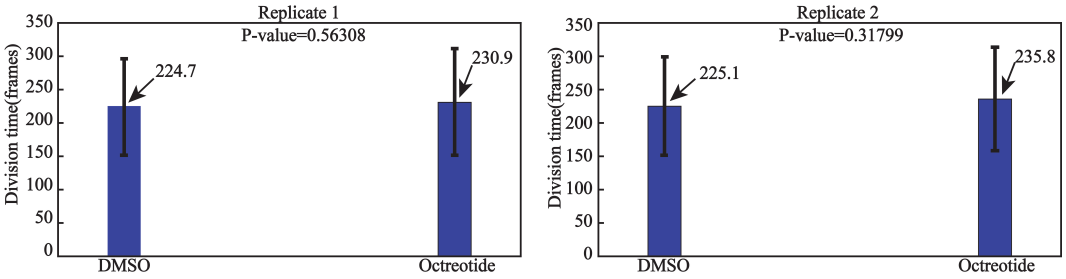

Figure 6.6: Means, standard deviations and the results of a Welch t-test for the division times of DMSO vs octreotide treated cells

As can be observed in both replicates, division time was slightly increased in cells treated with octreotide by around 6-20 frames, which translates to 1-4 hours. The difference was small in magnitude considering that on average, cells take 240-260 frames to divide and the results did not reach statistical significance. The result differs from published data where Octreotide has been shown to reduce the division time of small cell lung cancer(SLCL)cell lines [228] colon cancer cells,[17] or neuroblastomas [43].

6.5.2 Motility Analysis

The same procedure as in the case of the healthy vs disease samples was applied to analyse the data coming from cells treated with DMSO vs cell treated with octreotide. For the first replicate 101 DMSO and 102 octeotride treated cells were tracked while for the second replicate 101 DMSO vs 100 octreotide treated cells were tracked.

The MSD for each condition is presented in Figures 6.7A and 6.9A, as well as the best model fit.

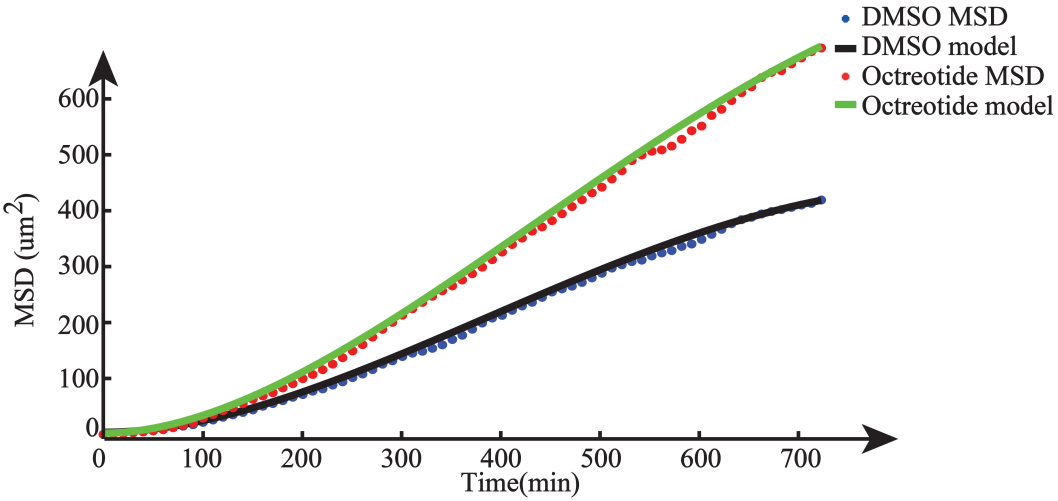

Figure 6.7: MSD's and the best model fit for the migration results of the first replicate of the DMSO vs Octreotide treatment experiment

The first conclusion based on Figures 6.7 and 6.9 is that cell motility in the disease cells was increased by octreotide. However, comparing these results with the previous differences observed between healthy and disease cells, it is apparent that cell motility in disease cells was only partially restored by octreotide.

The next step in the analysis was to look at the parameters describing the models. Figures 6.8 and 6.10 present these results. As in the case of healthy vs disease cells, the plots present the mean and standard deviation of each parameter in the models remaining after the 2 filtering steps.

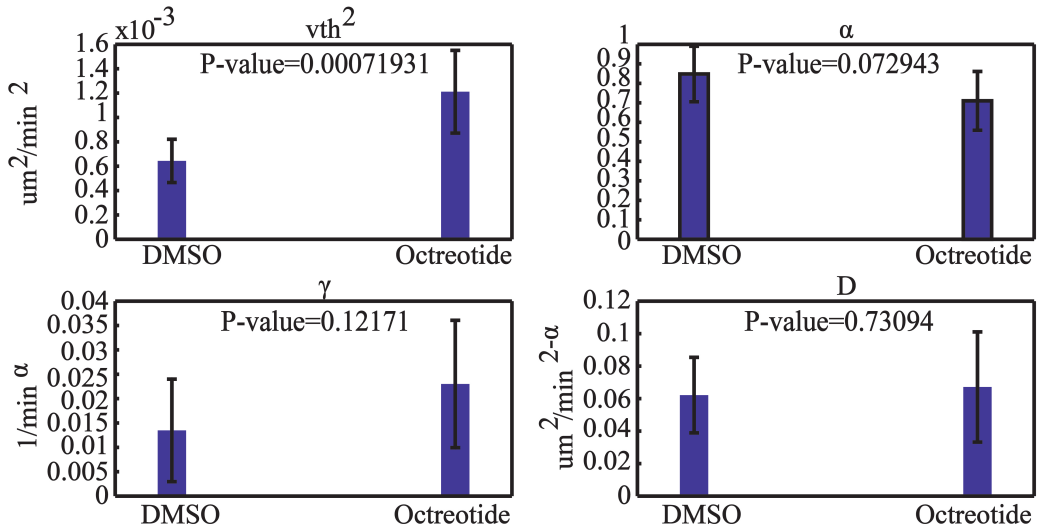

Figure 6.8: Means, standard deviations and the results of a Welch t-test for the parameters of the models for the 2 conditions in the first replicate of the DMSO vs Octreotide cells experiment.

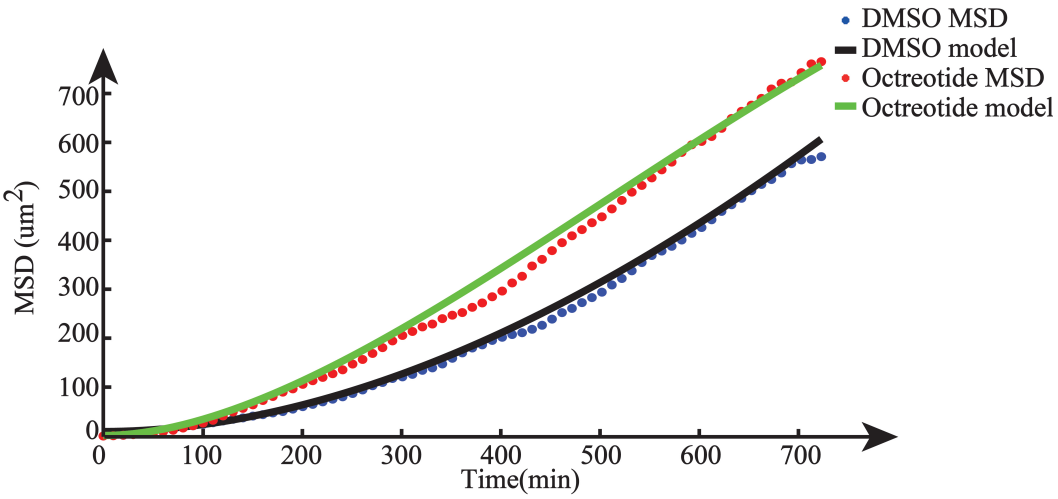

Figure 6.9: MSD's and the best model fit for the migration results of the second replicate of the DMSO vs Octreotide treatment experiment

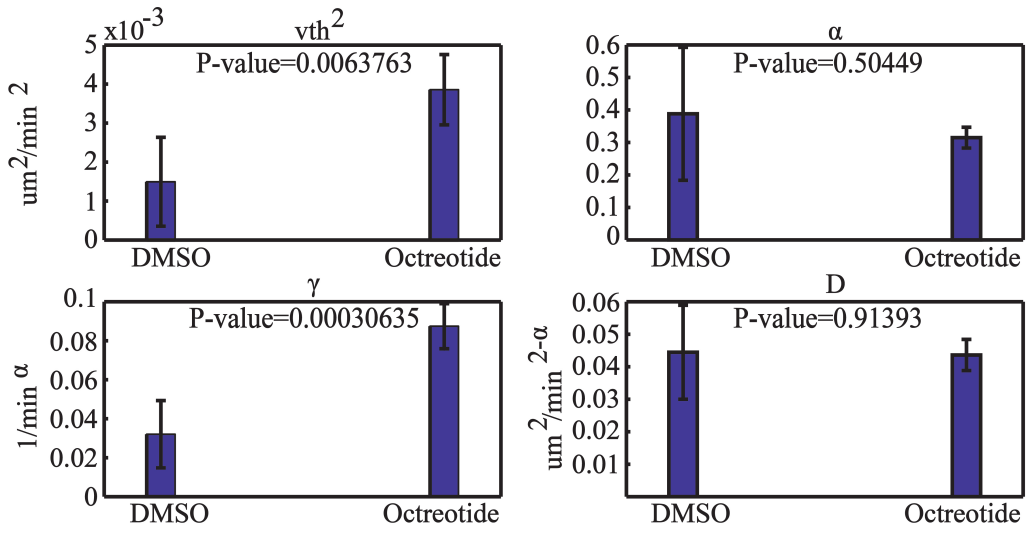

Figure 6.10: Means, standard deviations and the results of a Welch t-test for the parameters of the models for the 2 conditions in the second replicate of the DMSO vs Octreotide cells experiment.

In the case of the parameters obtained in the DMSO vs octreotide comparison, it seems that the thermal speed was increased in the cells treated with octreotide. The difference was statistically significant in both replicates while the difference in magnitude was more than double in both cases. For the other parameters including D, there was no consistent or statistical difference in both replicates.

6.6 Discussion

By time-lapse microscopy, significant differences were observed in the cellular phenotype of control compared to disease human proximal tubular cell lines. First, the cell division time was significantly prolonged in the disease cells. Second, cell migration was reduced in the disease model. Third, a partial correction of the migration defect in disease cells was observed following octreotide.

The increase in cell division time observed in disease was unexpected since other studies have generally reported that PKD1 and PKD2 affect cell proliferation negatively. In a study on rat proximal tubule cells, Ramasubbu et al. [292] found that disease cells had higher proliferation indexes. Similar results were found by Nadasdy et al. [256] in which the authors show increased proliferation in proximal and distal tubule tissue of human kidneys with ADPKD. Bhunia et al. [36]) showed

that expression of human PKD1 in MDCK cells led to cell cycle arrest in the G0/G1 phase. Similarly, Liu and Yu [219] reported that overexpression of PKD2 slowed cell growth in HEK293T and mouse IMCD cells while PKD2 knockdown lead to a recovery of proliferation rates. By contrast, a study by Hanaoka and Guggino [143] in which the authors were investigating the cAMP pathway role in cell proliferation in ADPKD observed that after 4 days, control primary human kidney cells showed higher numbers than primary ADPKD cells; however, only cell counts were reported in this study so it is impossible to know if the differences relate to reduced proliferation or are to increased apoptosis [105] [349].

Several possible explanations could account for these differences. First, it is conceivable that the immortalisation procedure could have altered cell cycle regulation differently in control and disease cells. Third, the 'disease' cells were created by knockdown of PKD1 and 2 in healthy cells and this may not truly reproduce the cystic disease phenotype. Future studies should seek to test which of these possibilities is correct.

In the case of the octreotide experiment, the results are different from published studies using other cell types. In general, octreotide has been shown to decrease cell proliferation. These include bile duct epithelial cells [360], rat pituitary tumor cells [73], liver cancer cells [218] and colon cancer cells [17]. An in-vivo study, Masyuk et al. [236] showed that octreotide seemed to decrease cell proliferation in PCK rat kidney epithelial cells. It is important to note that in their study, the decrease in proliferation evolved with the treatment time from around 25% after 4 weeks of treatment with the drug to around 50% after 16 weeks and also that dosage played a role in the observed results. Based on these the results, a few possible explanations could explain the current findings. First, that PTEC cells do not respond to octreotide. Second, the dose used was too low to produce a significant effect. Third, the treatment duration was too short for octreotide to have an effect on the cells. Further tests with higher concentrations of octreotide or with cells treated with octreotide for a longer period of time prior to imaging could answer these questions.

In the motility experiments, there was a clear reduction in the migration capacities of the disease cells compared to control cells when measuring their MSD. The difference could be attributed to an increased vth and diffusion coefficient. These results are consistent with previous studies which demonstrate that deleting either gene inhibits cells migration and overexpressing them enhances cell motility. Ex-

amples of these studies include one by Boca et al. [40] in which PC1 overexpression was shown to enhance migration in MDCK cells and PC1 deletion shown to reduce migration in MEF cells and another by Luyten et al. [226] in which HEK293T cells overexpressing PC1 were shown to move faster than their control counterparts. Although the current study shows decreased migration in the double knock down PTEC cells due largely to changes in cell speed and randomness of movement, it does not clarify whether the directionality of movement is affected. A limitation of the free cell migration assay is that there is no clear direction in which the cells are expected to move. To further investigate this property, a wound closure assay was used to further investigate the effects of the disease on the movement of the cells. The results of these assays will be reported in the next chapter.

Following octreotide, the MSD plots showed enhanced motility in the disease cells treated with octreotide although the drug did not restore cell motility to normal. The difference observed seems to be explained by an increase in the vth. The effect of octreotide on cell migration has only been examined in previous studies, on bovine retinal muscle cells [332] ESCs and T HESCs cells [10]. In these cases, octreotide inhibited cell motility. These results are the opposite of what was seen in the current experiments. The effect octreotide has on the linearity of cell migration remains was not tested in this study and could be the subject of future experimental work.

6.7 Summary

In this chapter, two specific cell phenotypes were analysed ie cell proliferation and cells migration and compared between control and disease cells. I also studied whether octreotide, a drug in clinical trials for ADPKD, could restore either of these cell phenotypes in disease cells. The assay used was a free-migration time-lapse assay in which the cells where plated at low density in a 6-well plate and imaged so that the trajectories and division time of single cells could be observed. Division time was measured by counting the number of frames between a single cell dividing to form daughter cells. In the case of migration, cell characteristics were studied using a Fractional Klein Kramer model to extract parameters from the cells' MSD.

The findings on the division of the cells were that the double gene knockdown increases division time while octreotride does not have a statistically significant

effect on it. In the case of the migration studies, the double knockdown cells show decreased migration which seems to be explained by a reduction in velocity as wells as in the diffusion coefficient characterizing their movement. Octeotride seems to salvage some of the migration capacities of the cells although not at the level of the healthy ones. The increase seems to be explained by an increase in the velocity of the cells.

Chapter 7

Analysis of Normal vs PKD1,2 Knockdown ciPTEC Cells in Wound Healing Assays

7.1 Introduction

Single cell migration experiments can be useful for calculating the spontaneous speed at which cells are moving [200] or extracting parameters for diffusion models [94]. The results in the previous chapter indicate that at single cell level, normal cells move faster compared to cells with knockdown of PKD1 and PKD2. Given that abnormal genes involved in ADPKD have an adverse effect on normal tissue organisation in the kidney, it is of interest to know if the differential motility at single cell level has any impact on the collective cell movement. The results of this research could be combined with findings on genetic deregulation in order to create a complex picture on how gene expression affect behaviour in groups of cells and how this translates into dysfunction in organs affected by the disease.

To address the question on the movement of a group of cells in the case of ADPKD, this chapter presents a comprehensive quantitative analysis of data obtained from wound closure experiments [178] involving normal and abnormal PKD1/PKD2 knockdown cell lines. Briefly, healthy and PKD1,2 knockdown cells were grown in 6-well plates with 3 plates for each condition. Once the cells reached 100% confluence, the cell layer was scratched using 200 μl pipettes. Next time-lapse microscopy was used to record the closing of the wound. The analysis pro-

vides information about the movement directionality and speed dynamics of collective migration of normal and abnormal renal epithelial cells. Specifically, the data generated by the time-lapse, would-closing experiments was used to derive three characteristics of collective migration namely the rate of wound closure, directionality of movement and temporal texture characterisation of contiguous cell layers.

The first analysis was performed to assess the capacity for wound closure of wild-type compared to disease PTEC cells. Previous studies had showed this parameter to be impaired in different disease cells [59] [60] [395] [177]. The closest study to the present one, Joly et al. [177] compared primary mixed human renal tubular cells (proximal and distal) from healthy and diseased patients but no study to date has studied a pure proximal tubular cell line.

The second analysis was performed to address the unanswered question from the previous chapter about the capacity of the disease cells to maintain directionality. Studies such as by [59] have reported that there is a greater randomness in the movement of disease cells but in this study, the cells analysed were MEF cells.

Finally, Haralick features were used to characterise texture properties of the cell layers. Texture features are typically used to implement automatic classification algorithms [145][46][402] used, for example, to assist disease diagnosis [381].

The remainder of this chapter is organised as follows: Section 7.2 presents the experimental methods employed. Section 7.3 describes the methods used to analyse the data. Sections 7.4, 7.5 and 7.6 provide the comparative analysis results of the wound closing rates, linearity of individual cell trajectories and temporal texture features, respectively. Section 7.7 discusses the biological significance of the results and possible interpretations of the data. Section 7.8 summarises the conclusions of the chapter.

7.2 Experimental Methods

The materials and cell lines used for single cell motility analysis, described in Chapter 6 section 6.2.1, were used to carry out the wound scratch assay. The cell culture conditions were the same. The following subsection will describe the experimental procedures used in the wound/scratch assays.

7.2.1 Scratch Wound Healing Assay

When cells in the T75 flasks reached 80% confluency, the normal growth media was replaced with serum free media for cell cycle synchronization [205]. This facilitates both cell motility and cell scratch experiment to be established following trypsinization of a single cell suspension. Cell cultures were trypsinized using the procedure described in Chapter 6, subsection 6.2.2.1. Next, single cells were replated in 6 well plates at a density of 100,000 cells/ml ie 2 ml cell suspension/well. 2 μl blasticidin were also added to each well. Each 6 well plate was divided equally between conditions: three wells were plated with normal cells and three wells with PC1,2 knockdown (disease) cells. The plated cells were allowed to reach 100% confluency which typically took approximately one week. Culture media was replenished every 2 days using 2ml culture media and 2 μl blasticidin per well. Once the plated cells reached 100% confluence, they were serum starved for 2 days by using 2 ml of serum free media per well. After serum starvation, each monolayer was scratched vertically using a 200 μl pipette tip. Next, the wells were rinsed using fresh culture media and 2 ml culture media and 2 μl blasticidin was added to each well. As in the case of the motility experiments, cell imaging was performed using an Olympus Ix70 microscope with a moving stage. For each well, 8 non-overlapping fields (10 x magnification) were taken along the wound. Images for each field were acquired every 10 minutes. For each position, the cells were recorded for between 2-4 days (one replicate was recorded for 2 days while the other 3 for 4 days).

7.3 Analytical Methods

A number of quantitative spatial and temporal characteristics were extracted from the time-lapse microscopy image data of the scratch wound healing assays. The first characteristic is the rate of 'healing' the gap, from the time the wound is inflicted until it closes, which reflects the effects of factors that alter the motility and growth of cells. The time-dependent rated is calculated based on the ratio between the area in the image occupied by the cells and the total area of the image at successive time points. Furthermore, to characterise and compare the directionality of cell movement within the cell layer during the wound healing, the linearity of individual cell trajectories was calculated for the two experimental conditions involving normal and abnormal cell lines. Finally, Haralick texture features [145] were calcu-

lated for each of the image frames generated over the course of the experiment. The following sections describe in more detail the analytic methods used in each case.

7.3.1 Wound Closing Rate

In order to quantify the wound closure rate, a standard image segmentation strategy [353] was performed on the frames in each video using standard MATLAB procedures. For each frame the following operations were carried out sequentially:

1. entropyfilt function in MATLAB [235] was applied to the image in order to represent the pixels based on the smoothness of the textures they come from. The picture representation is now better for segmentation as the background represented by the wound will be very dark coming from a very homogeneous area of the picture and the cell layer will appear very bright.rather the original representation where the intensity of the pixels show high variance due to noise produced by the illumination of the plate.

2. The filtered image was converted to a grayscale image by applying the mat2gray function. This is important for the next step where the segmentation threshold function needs as input a matrix in which the values of the elements are in the range of a grayscale image.

3. The nonparametric and unsupervised method for automatic threshold selection for image segmentation [269], implemented by the Matlab function graythresh, was used to select the optimal threshold of gray level for extracting the area represented by the wound and the area represented by the cell layer. Once the threshold was computed it was used to convert the image to a black and white one using im2bw.

4. Small connected objects were removed from images using bwareaopen to close areas less than 900 pixels.

5. Small open areas and holes were corrected using an algorithm in which morphological reconstruction is used that is implemented by the imfill function in MATLAB [329]. Once these operations were done, a black and white image is obtained in which the black pixels represent the wound and the white ones the cell layer.

6. The ratio between the number of white pixels and the total number of pixels was calculated for each frame, representing the ratio between the are covered by the cell layer and the total area of the frame.

In order to characterize each condition, the ratio obtained for the first frame in each video is subtracted from the ratios in the following frames. The obtained quantity for a imaging position p will be the time series:

$$cr_p(t_n), n \in 0, \ldots, N-1 \tag{7.1}$$

with t_n the timepoint corresponding to frame number n+1.

Next the average ratios and standard deviations for the frames taken at the same time-point in all positions for a condition are calculated and the results plotted to create a visual description of the wound closing progression.

To further analyse the way in which the wound closing rate evolves with time, the difference between the average ratio of the cell layer in a frame across a condition and the average ratio 4 hours later was computed for the first 100 frames, covering the first day of the experiment. More formally for 2 time series are calculated for each condition:

$$d_{healthy}(t_n) = m_{healthy}(t_n + 24) - m_{healthy}(t_n) \tag{7.2}$$

$$d_{PKD}(t_n) = m_{PKD}(t_n + 24) - m_{PKD}(t_n) \tag{7.3}$$

where $m_c(t_n) = \sum_{p \in C} cr_p(t_n)$ with C the set of positions for a condition c.

In order to compare the rate at which the cells in a condition close the wound after reaching the same average gap width as the initial state of the other condition, the following procedure was applied to shift the series to having same initial conditions:

1. Find which condition has a lower $m_c(t_0)$. This condition will be called c_1 and the other condition c_2.

2. Find the first timepoint, t_f for which

$$m_{c_1}(t_f) \geq m_{c_2}(t_0) \tag{7.4}$$

3. Define the shifted difference time series:

$$sd_{c_1}(t_{i_1-f}) = d_{c_1}(t_{i_1}), i_1 \in \{f, f+1, ..., 99\} \tag{7.5}$$

$$sd_{c_2}(t_{i_2}) = d_{c_2}(t_{i_2}), i_2 \in \{0, ..., 100-f\} \tag{7.6}$$

to be used for further analysis.

7.3.2 Individual Cell Tracking in Scratch Healing Assays

The individual cells were tracked using the method previously described in chapter6, subsection 6.3.2, using the manual tracking plugin for Image J by Cordelières [79]. Cells were followed for 72 frames (corresponding to 12 h) and the obtained trajectories were then analysed using different measurements. In order to give the cells time to be less tight and recover from the scratching, the tracking was started with frame 50, corresponding to 8 more than 8 hours from the initial scratch. The cells analysed for tracking were randomly selected in each movie corresponding to a field, as long as they formed the first cell layer facing the wound. For each well, 30-35 cells were tracked to ensure an equal contributions to the final measurements with a total of 100-105 cells being tracked per condition. The trajectories of the cells represent their centerpoint positions (C(x),C(y)). The system of axes is the same as in the previous chapter.

The first quantity used to assess the directionality of the cells was the confinement ratio defined in Chapter 4 section 4.3.1. In the present analysis, the focus was on the movement of the cells in the direction perpendicular to the scratch, that is along the x-axis of the images, so two other quantities were also calculated.

The first quantity was the distance travelled by each cell along the x-axis over the given time interval, defined as

$$x_{dist} = sign(256 - x(t_0)) \times (x(t_N) - x(t_0)) \tag{7.7}$$

with $x(t_0)$ the initial position of the cell on the x axis and $x(t_N)$ the final position of the cell on the x-axis. The sign function is used to distinguish between the cells located to the left (x < 256) and to the right (x > 256) of the wound, which move in opposite directions.

The second quantity was similar to the confinement ratio, defined as the ratio

between the distance travelled along the x-axis and the total length of the trajectory

$$x_{norm} = x_{dist}/d_{tot} \tag{7.8}$$

where d_{tot} the total length of the trajectory distance defined in Chapter 4-section 4.3.1, was also used to quantify the directionality of cell movement with respect to the x-axis.

An interesting observation that was made during the tracking and analysis of cell movement was that some cells moved away from (rather than toward) the wound during the tracking period. This was quantified by calculating for each cell the difference between its initial and final position along the x axis on the path travelled for 60 minutes (6 successive image frames)

For one cell the measure is defined as:

$$x_{prop} = \frac{card\{sign(256 - x(t_0)) \times (x(t_{i+T}) - x(t_i)) < 0 | i \in \{0, ...N - T - 1\}\}}{N - T} \tag{7.9}$$

where $x(t_i)$ is the cell's position on the x axis at frame i+1, N the total number of frames for which the cell is tracked and T the number of frames for the time window.

7.3.3 Haralick Texture Features

Haralick texture features are used here to characterize the overall texture of the cell layers during the wound healing for the two types of cells. Once image segmentation has been performed, each frame was processed to eliminate the gap before the Haralick texture features were computed.

For each image frame, consider the associated matrix $A \in \mathbb{N}^{mxn}$, where m, n are the number of pixels along the horizontal and vertical axes respectively and A(i,j)=1 if (i,j) belongs to the segmented cell layer and A(i,j) = 0 if the pixel belongs to the gap region.

Given that the wound is approximately rectangular, eliminating the pixels in the image corresponding to every column j in the matrix A satisfying A(i,j)=0, for all i=1,m, most of the area corresponding to the wound in an image is eliminated as shown in Figure 7.1.

For every image frame I_k , the matrix A_k was generated and used to calculate

$$Ic_k = I_k \circ A_k \tag{7.10}$$

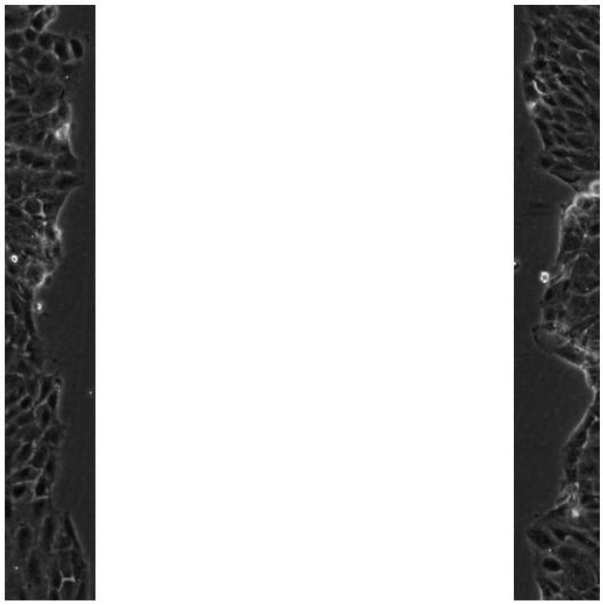

Figure 7.1: Example of the elimination of the wound in a picture. The white area in the middle represents the area eliminated by the algorithm

where $\circ$ is the Hadamard or element-wise product [248].

The final image $I'c_k$ was obtained by eliminating all columns containing only zero entries. The Haralick features described in Chapter 4 section 4.5 were calculated for every $I'c_k$ using the standard Matlab function for calculating the grey-level co-occurrence matrix and the free toolbox available online[366] for extracting the Haralick features.

Pixel intensities were discretised into 32 levels and the co-occurrence matrix was assumed to be symmetrical. In order to make the texture characterization rotationally invariant, the standard method proposed by Haralick et al. [145] which averages the features taken at angles of 0°,45°,90°,135° was applied to each frame.

This resulted in one value for each Haralick feature for each frame of the movie of a specific position.

7.4 Comparison of Wound Healing Rates in Scratch Healing Assays

The first step taken to analyse the wound scratch assay was to compare the rate of wound closure in healthy cells compared to disease cells. As described in Chapter 4 section 4.4, the wound closing rate has been reported to be altered in other diseased cells ie HEK [226], MEF [59] and MEK [395] cells. Four technical replicates were performed. Figure 7.2 presents each experiment individually, as a plot of the mean and standard deviation of the ratios across all fields for each condition.

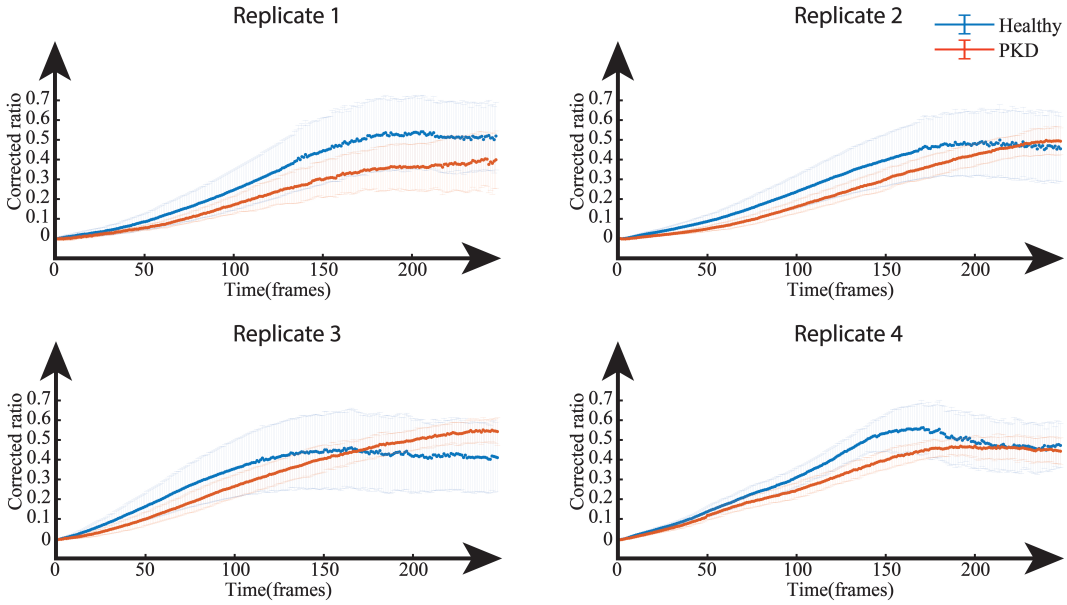

Figure 7.2: Evolution in time of the ratio of the frames occupied by the cells in the wound healing assay

The results are displayed only up to frame 248 corresponding to 2 days of cell tracking. The reason why is this done is that in most movies by this frame cells made contact. Once this happens the rest of the wound is not closed as 2 sheets moving to each other so the data was not considered in this analysis.

The segmentation algorithm gave good results when the wound was closing, correctly identifying background and cell layer areas. Figure 7.3 displays examples in this sense. These results make the first part of the graph where there is constant increase in the ratio, the most important for comparing the wound closing rates of the cells coming from the 2 conditions.

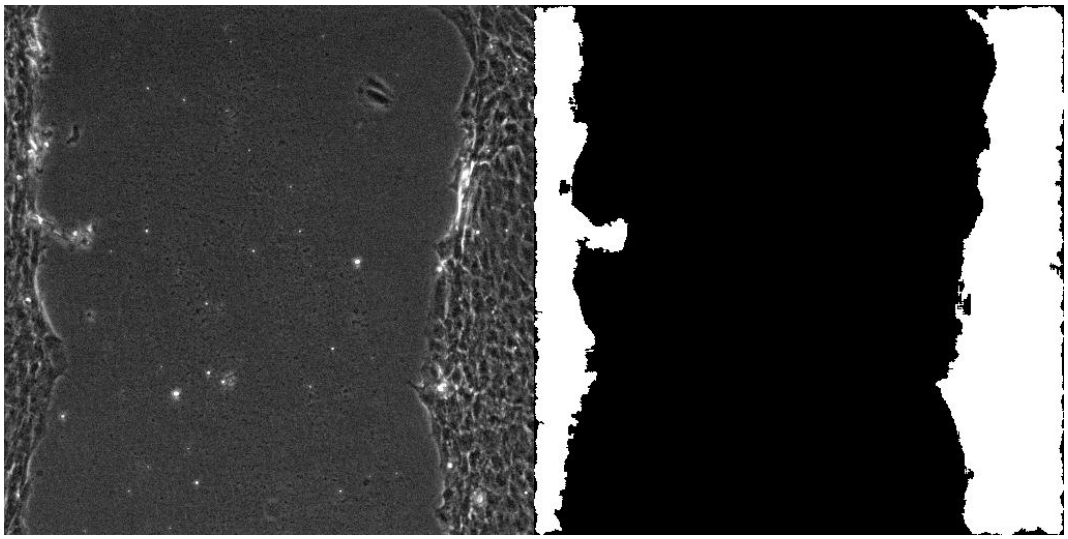

Figure 7.3: Snapshots of the wound closing and the results given by the segmentation algorithm (white - area detected to be covered by the cell layer, black-area detected as cell free)

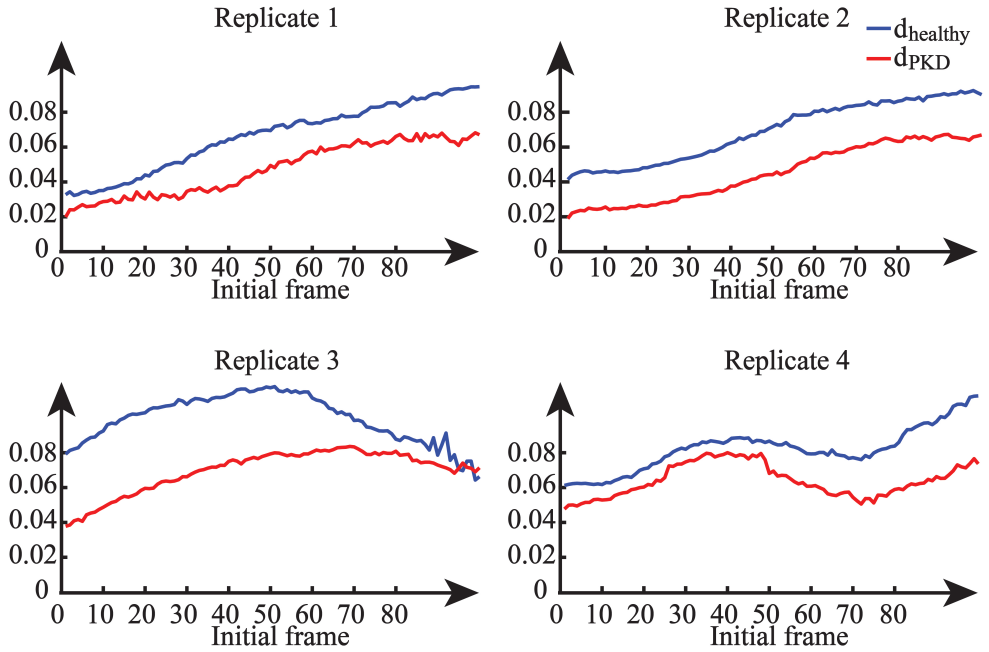

Figure 7.4: Plot of the values of the two timeseries $d_{health}(t_n)$ and $d_{PKD}(t_n)$ for each replicate.

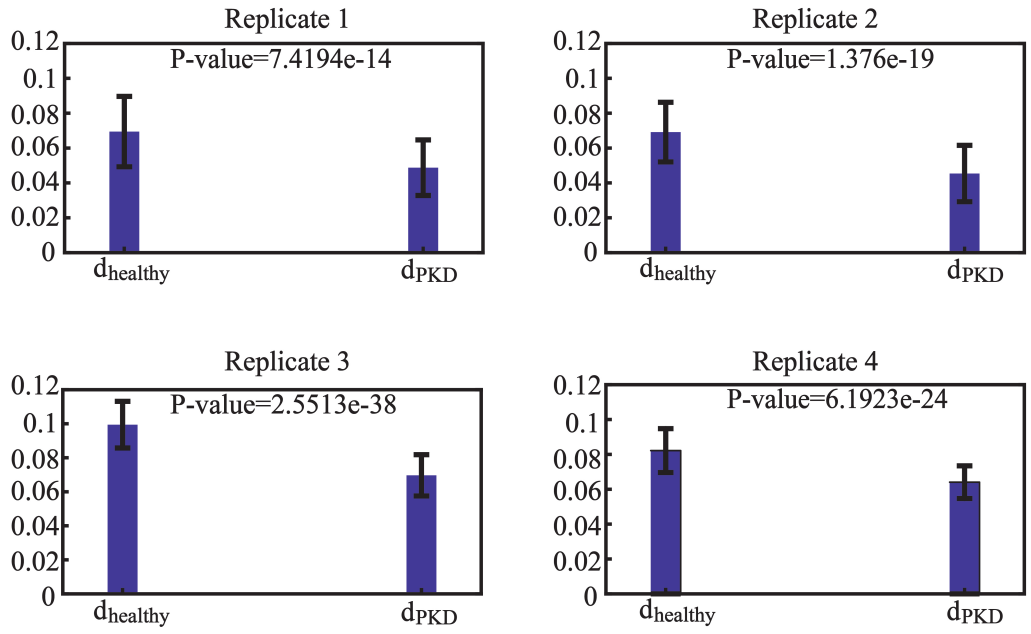

Figure 7.5: Means, standard deviations and the results of a Welch t-test for $d_{health}(t_n)$ and $d_{PKD}(t_n)$ for each replicate

As the results show, there seems to be an increase in the rate at which the healthy cells close the wound in all replicates as the average ratios seem to show a steeper increase in time. In order to further investigate this, the time series $d_{health}(t_n)$ and $d_{PKD}(t_n)$ for the first 100 frames were calculated.

The means and standard deviations for the two time series were calculated and a Welch t-test was applied on their values to quantify the differences between them. Figure 7.4 shows the plots of the time series and Figure 7.5 the calculated statistics.

Two observation can be made, the rates increase with time and the healthy cells show clear higher wound closing rates. This confirms the literature on the subject which indicates that as the wounds narrow, the rate at which they heal increases, making the initial width of the gap an important factor in the cells healing rate [200]. In order to investigate how this affects the present analysis, the mean and standard deviations of the initial ratios across the 2 conditions were calculated for the 4 replicates and a Welch t-test employed to assess the significance of the observed differences. The results are available in Figure 7.6.

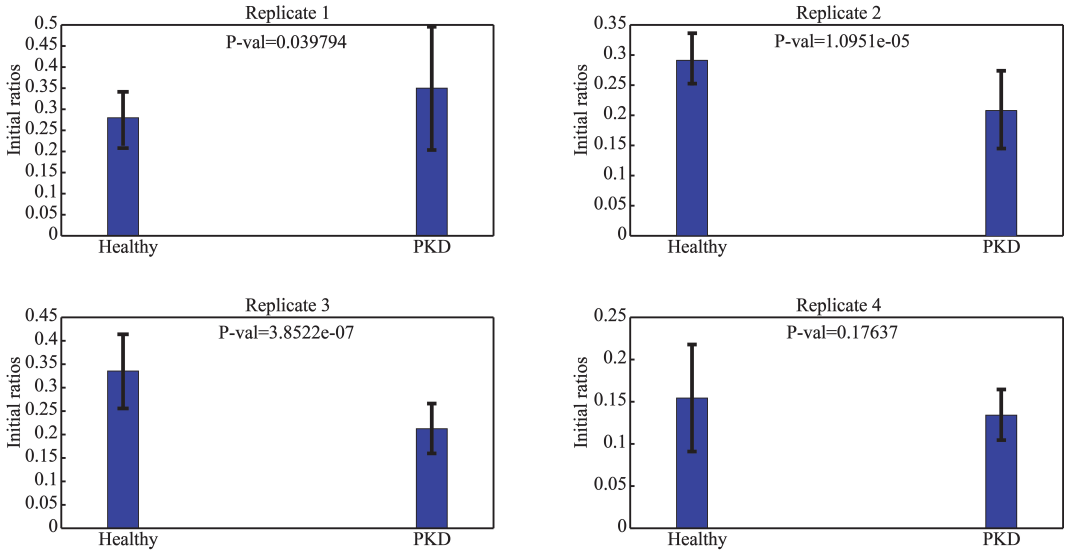

Figure 7.6: Initial ratios for the four replicates

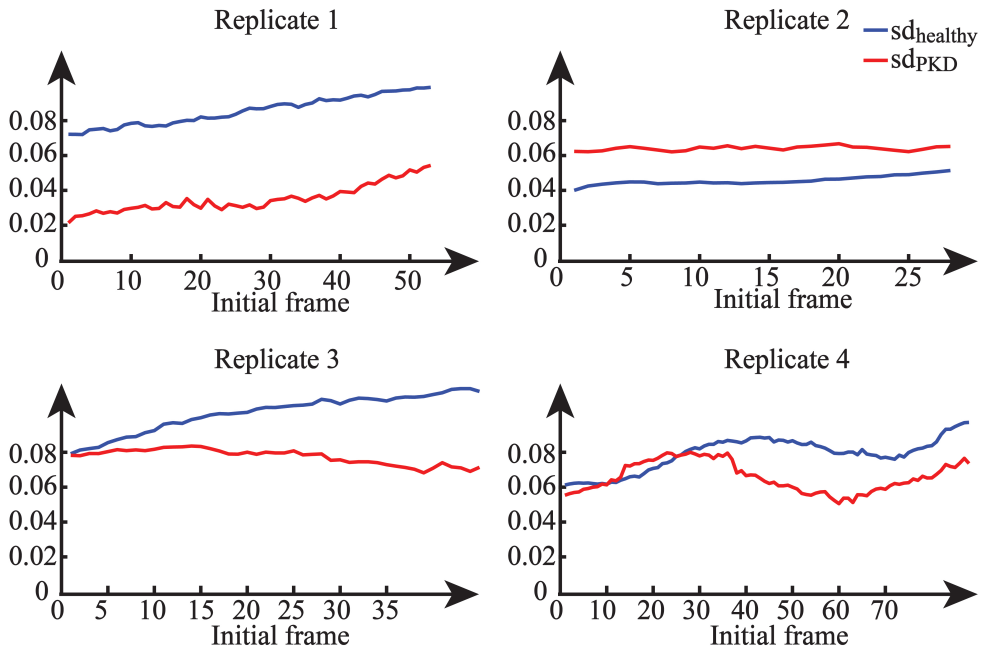

Figure 7.7: Plot of the values of the two timeseries $sd_{health}(t_n)$ and $sd_{PKD}(t_n)$ for each replicate.

As it can be observed, there seems to be quite a uniform distribution in the average ratio of the cell layer in the first frame across replicates, with 2 replicates in which the healthy cell layer occupies more space than the disease layer, 1 replicate in which this situation is reversed and one in which the observed differences are statistically insignificant. In order to eliminate the initial wound width as a factor in cell behaviour, a comparison was done starting in the moment when the average wound width in the condition with higher initial width becomes roughly equal to the average initial width in the other condition. The time series $sd_{health}(t_n)$ and $sd_{PKD}(t_n)$ were calculated, their plots as well as means and standard deviations and the results of a Welch t-test being available in Figure 7.7 and Figure 7.8.

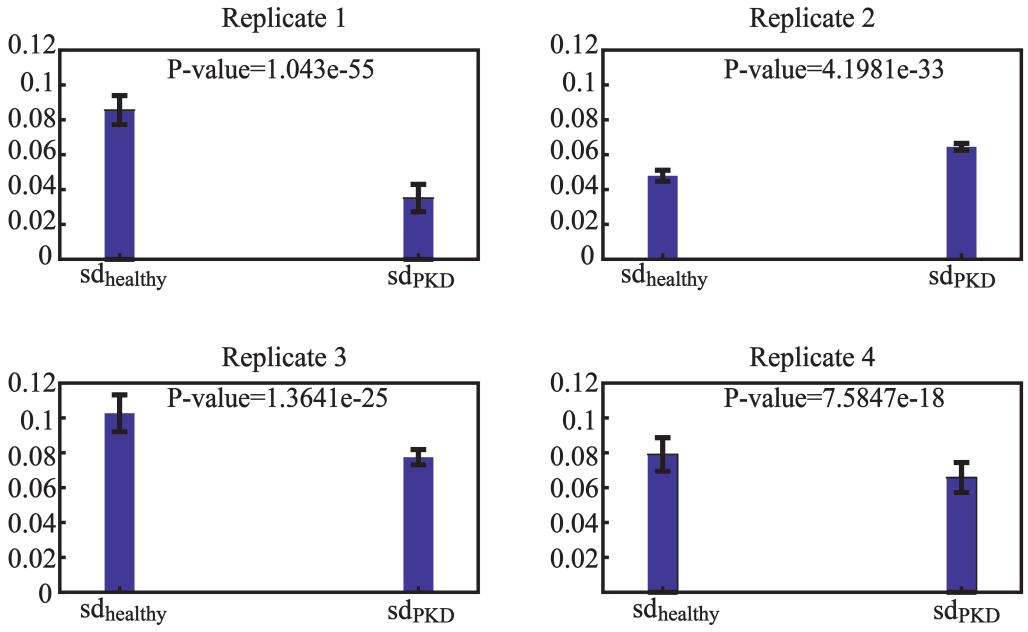

Figure 7.8: Means, standard deviations and the results of a Welch t-test for $sd_{health}(t_n)$ and $sd_{PKD}(t_n)$ for each replicate

The first observation that can be made is that in three of the replicates, each corresponding to a different initial condition the healthy cells close the wound at a faster rate even when shifting the time-series to start at the same average gap width. In the case of replicate number 2 this is inverted, with the disease cells closing the wound faster. The results suggest that the observed differences are due to another factor present in that replicate. As by shifting the time-series the comparison is done between cells that have been travelling for a while (around 10 hours in this

case) and cells that have just underwent scratching which is bound to produce cell damage [178] a reasonable explanation is that the higher migration capacities of the healthy cells were not enough to overcome the stress produced by scratching when comparing to the disease cells which had time to recover.

The results strongly suggest that the healthy cells have a better capacity to close the wound even when accounting for the differences in its initial width. This result is further explored in the following section by analysing the cells capacity for directed movement.

7.5 Directionality of Cell Movement During Wound Healing

For linearity analysis, cells were tracked individually. In one of the experiments, a technical problem led to immobility of the microscope stage for an undetermined period. When the problem was detected and solved, the positions of the imaging fields were slightly moved. The period of the defection and the difference in positioning were small enough not to affect significantly the data on wound closing ratio and on texture analysis but unfortunately the error made linear cell tracking analysis impossible. As a result, for this section of the analysis, only 3 replicates are reported.

The linearity of cell movement was assessed by plotting the trajectories of individual cells starting from the same point as shown in Figure 7.9. These plots do not allow for a clear conclusion of the effects of disease on cell movement but an interesting observation was that some cells seem to be moving away from the wound.

In order to be able to further investigate the effects of PKD1 and PKD2 knockdown on the directionality cell movement, the confinement ratio was calculated for the two experimental . Figure 7.10 shows the mean and variance of the confinement ratio in between conditions for each replicate. A p-value resulting from a Welch t-test was calculated in each case to assess the statistical significance of the difference.

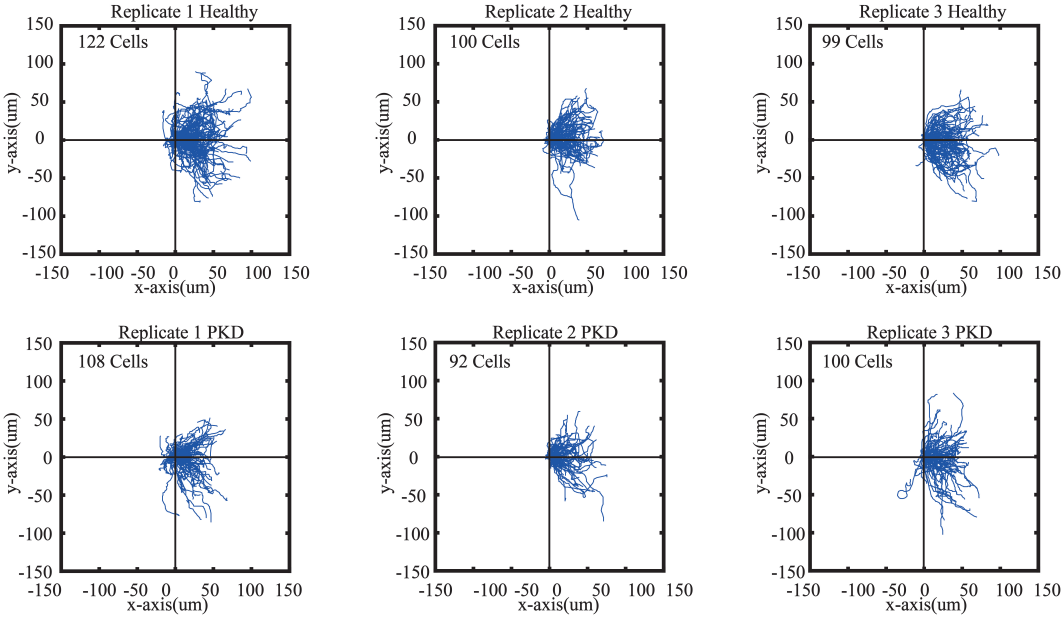

Figure 7.9: Trajectories of the cells in each replicate for each condition. The number of cells whose trajectories were plotted in each case is displayed

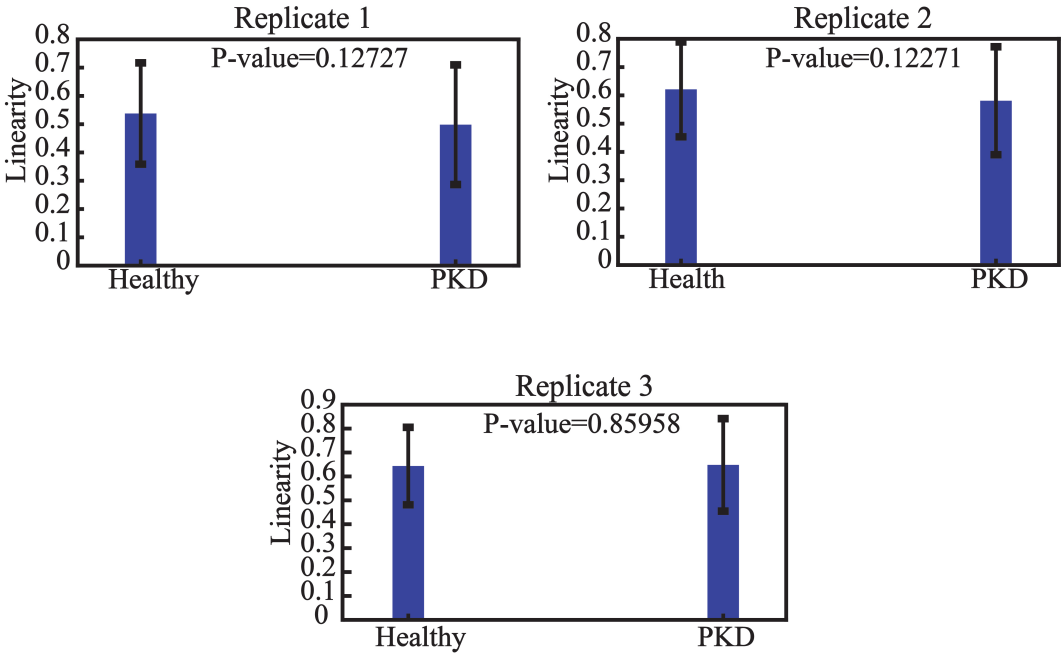

Figure 7.10: Linearity of cell movement (calculated using the confinement ratio)

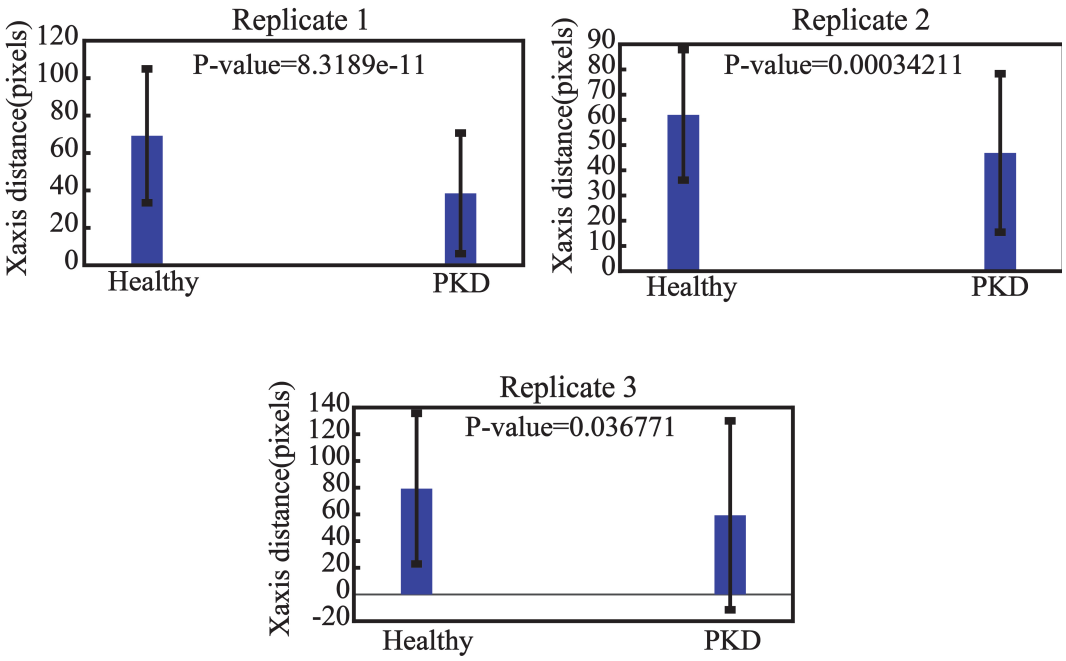

Figure 7.11: Distance travelled by the cells towards the wound along the x-axis.

Figure 7.10 indicates that there might be a reduction in the directionality of the movement of the abnormal cells, which can be observed in the first 2 replicates while in the case of the 3rd replicate the averages seem to be approximately equal. However, the differences in the 3 replicates were not statistically significant. This suggests that this is not the most representative measure as it does not prioritize a specific direction for cell movement. In a wound closing experiment, the cells have a directional movement, which is perpendicular to the direction of the wound. To take this information into consideration, 2 additional measures were employed, the distance travelled by a cell on the x-axis, which is perpendicular to the direction of the wound and the distance travelled on the x-axis relative to the total distance travelled, which will be called the x-axis proportional distance. The results for the 2 measures on the 3 replicates are summarised in Figures 7.11 and 7.12.

As can be observed, the distance travelled on the x-axis is clearly reduced in the case of disease cells. This was expected considering the slower closing rate of the wound. More interesting, the same reduction was observed in the proportional x-axis distance showing that the disease cells seemed to lose some of their capacity for directional movement. In 2 of the replicates the difference is clearly statistical significant, with a p-value below the generally accepted threshold of 0.05. In the case of the 3rd replicate, the p-value was slightly above 0.05. The final analysis of

cell movement was motivated by the observation that some cells seemed to move away from the wound at least for a specific interval of time. This could also explain some of the reduction in the movement on the x-axis. To quantify this, the position of each cell at the beginning and the end of a one hour time window was observed and the proportion of time in which the cell was closer to the wound was recorded. The results are illustrated in Figure 7.13 More information on the procedure can be found in subsection 7.3.2.

As the graphs show, the percentage of time a cell is closer to the wound is consistently lower in the disease cells in all 3 replicates, a difference that was statistically significant.

These results suggest that the difference in the distance travelled by the cells on the x-axis has 2 factors, a loss in the linearity of movement and a decrease in cellular persistence to move towards the wound. Out of the two factors, the second one seems to be the dominant as it was significantly different in all 3 replicates.

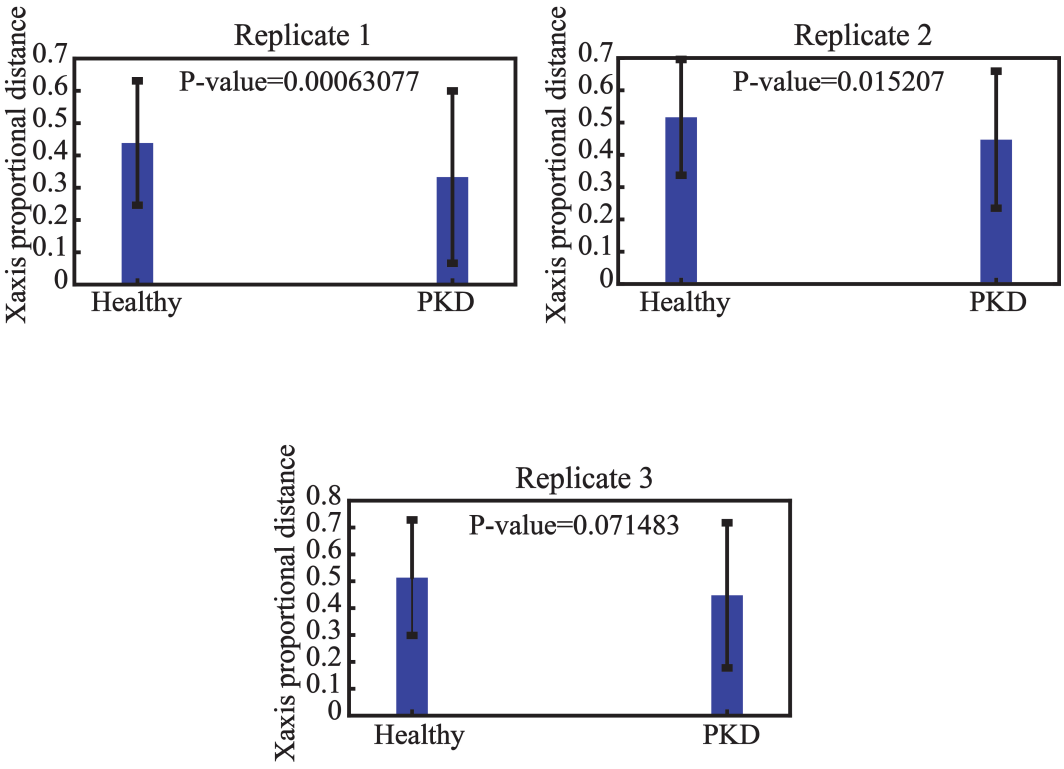

Figure 7.12: Distance travelled towards the wound along the x-axis relative to the total length of the path of the cell

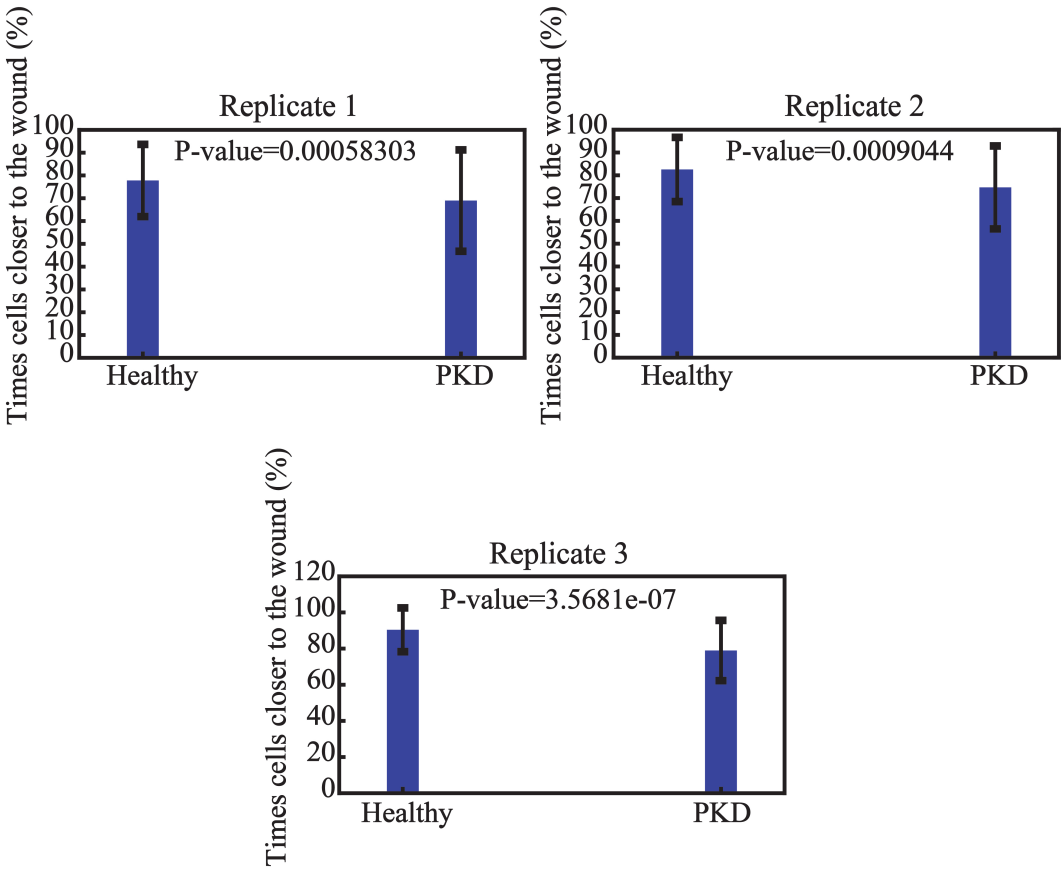

Figure 7.13: Proportion of frames in which a cell is closer to the wound after 1 hour of travelling

7.6 Comparative Texture Analysis

The Haralick textural features of the contiguous cell layers were computed for each frame as described in section 7.3.3. In his original work, Haralick proposed 14 features. In practice, 5 features are generally used as they have shown to produce good results with a varied range of image types such as magnetic resonance imaging (MRI) [381], multispectral bio-images [62] or computed tomography (CT) images[402]. These features are: contrast, energy, entropy, homogeneity and correlation and have been employed in the present analysis. Figures 7.14-7.17 show the data for the 4 replicates.

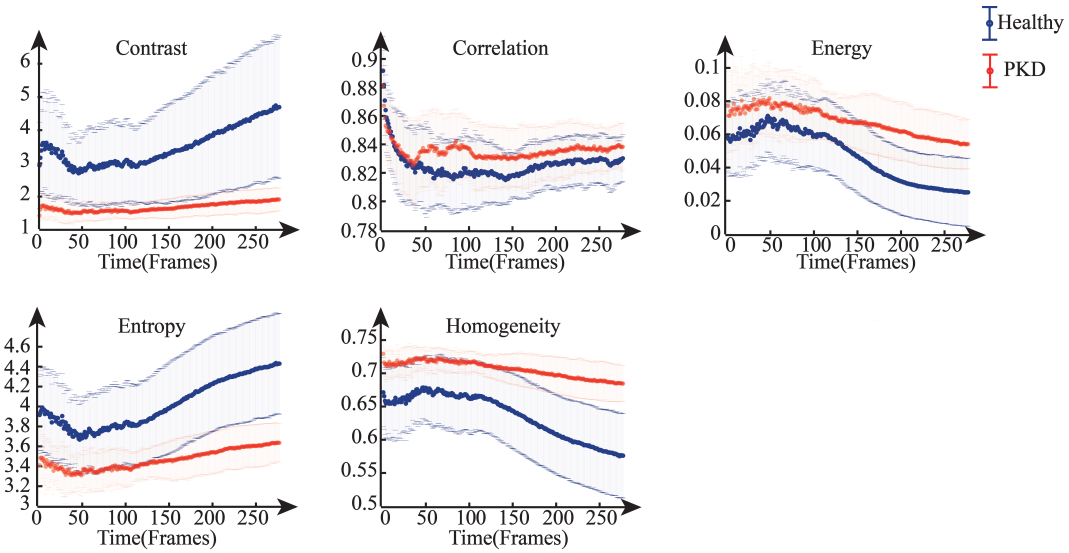

Figure 7.14: Trajectories in time represented through means and standard deviations across all the positions in a condition for the 5 Haralick features in the first replicate of the wound closing experiment

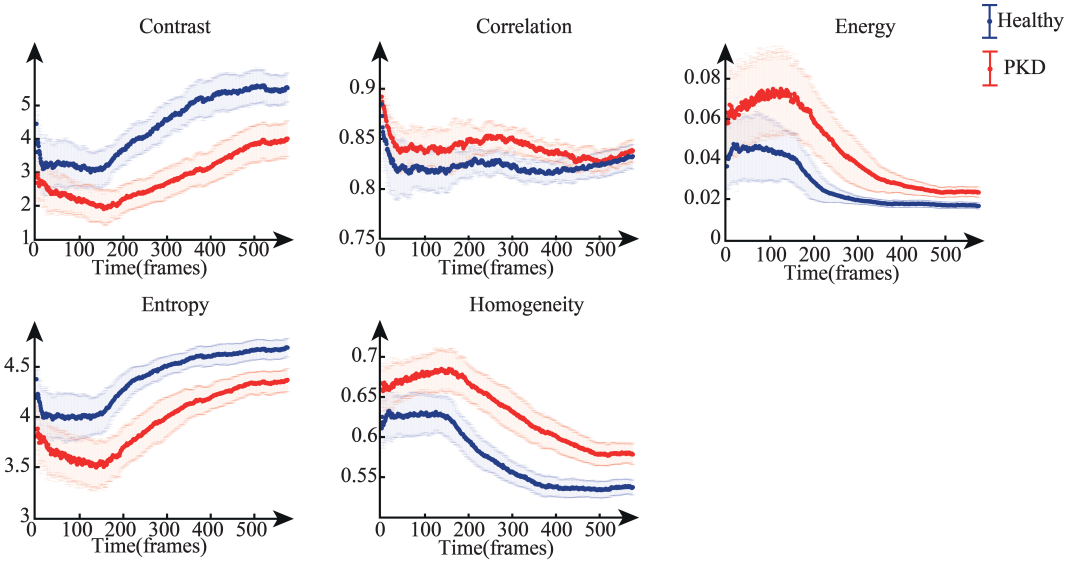

Figure 7.15: Trajectories in time represented through means and standard deviations across all the positions in a condition for the 5 Haralick features in the second replicate of the wound closing experiment

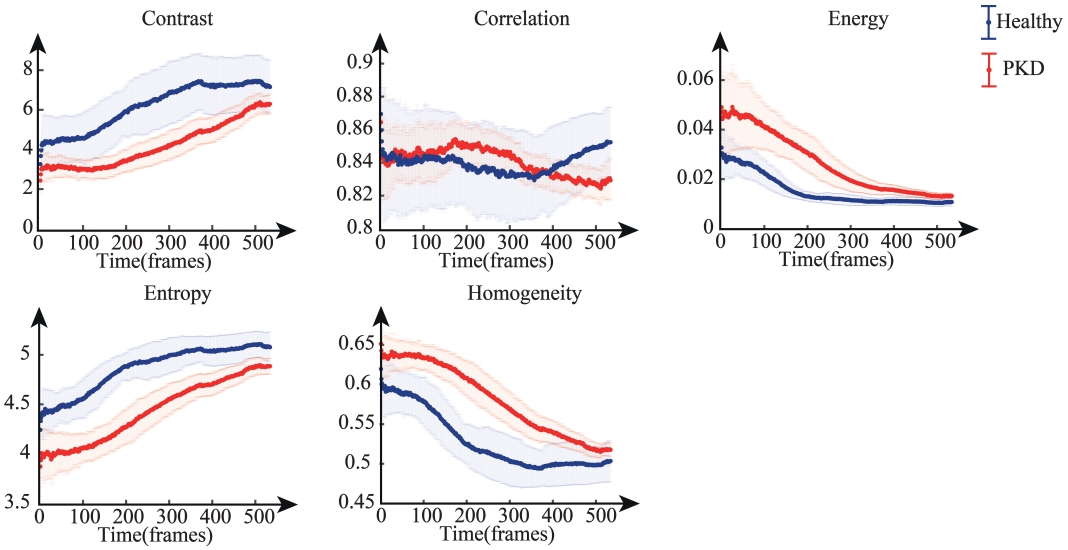

Figure 7.16: Trajectories in time represented through means and standard deviations across all the positions in a condition for the 5 Haralick features in the third replicate of the wound closing experiment

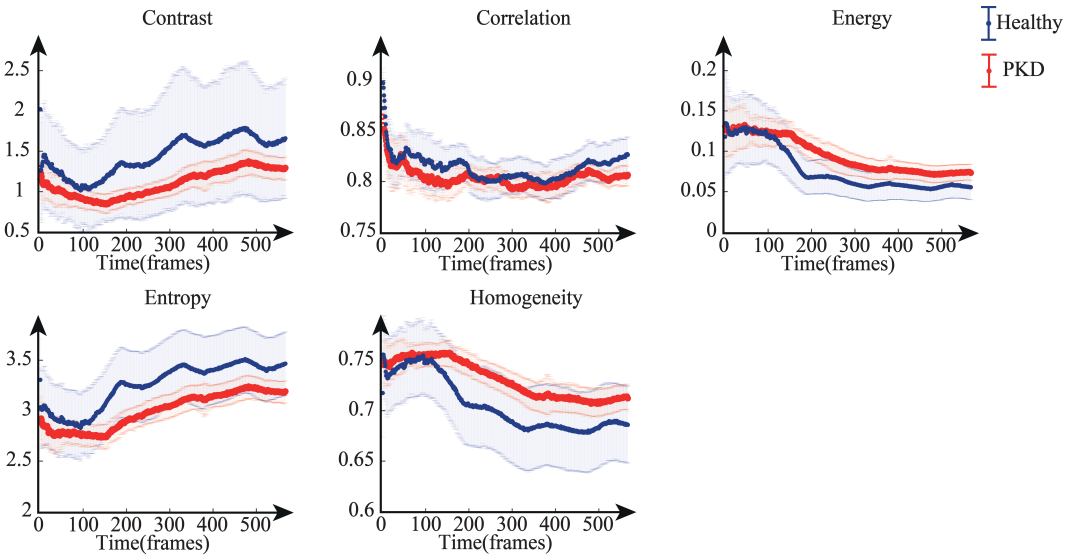

Figure 7.17: Trajectories in time represented through means and standard deviations across all the positions in a condition for the 5 Haralick features in the fourth replicate of the wound closing experiment

With the exception of correlation, the features had a clear separation between the 2 conditions. It is important to note that Contrast which is a measure of differences between neighbouring pixels appeared to be lower in the case of the disease cells. The same pattern can be observed in Entropy which as a measure of randomness becomes lower when pixels have similar values. As expected, the measures of homogeneity of the picture are raised in the case of disease cells. The general information that can be extracted from these results is that the frames containing healthy cells seem to show more differences in the values of the pixels. Another interesting observation is the pattern of progression over time that can be best seen looking at Contrast. The Contrast level seems to be descending in both conditions until around the frame 100-200 after which is starts to increase until around frame 350-400 when it seems to be reaching some stability. Looking at the videos, the cells are quite tight in the beginning when the cell layer was freshly wounded (Figure 7.18A). Then, as the cells start migrating, they get flatter and their tightness relaxes. Between frames 100-200 (Figure 7.18B) is when cells seem to make contact between the 2 sides of the gap. From this moment on, they start to fill the gaps by multiplication so it is expected that the tightness will increase. By frame 400 in most movies, 100% confluence can be observed (Figure 7.18C). These observations suggest that Contrast is correlated with the tightness of the cells. The other conclusion that could be reached if this is true is that disease cells lose some of their cell adhesion characteristics.

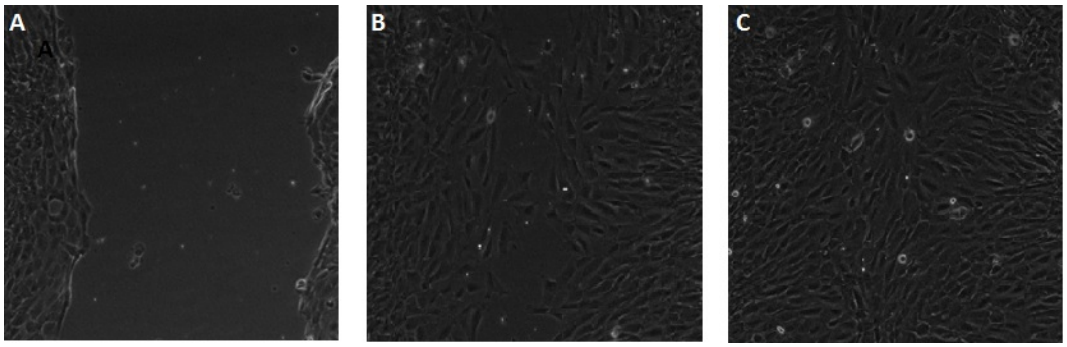

Figure 7.18: Snapshots of the cell layer in a movie. A-frame 1, B-frame 170, C-frame 380

7.7 Discussion

The results presented in this chapter strongly indicate that the knockdown of PKD 1 and 2 genes:

1. disrupts the loss of the directional migration toward extracellular matrix and

2. alters their spatial organisation

The directionality of disease cells migration was impaired most clearly when measuring their movement along the x-axis. The most obvious difference was observed in some cells moving in the opposite direction to the wound, an extreme example where the trajectory of movement was greater than 90° from the desired trajectory with the healthy cells being at positions further from the wound after a period of time in around 20% of the frames while the double knockdown ones in around 30% of the frames. Similar results were reported in MEF cells [59] where cells with deleted PKD1 travelled at an angle greater than 90° to the direction of the wound while the wild-type cells never crossed this threshold.

In respect to the changes in spatial organization, the Contrast was the Haralick texture feature showing the largest variation between the two conditions. To interpret the results it is important to summarise first the basic principle underpinning phase contrast microscopy. In a phase contrast image, the intensity of pixels is calculated based on the shift of phase between light wave-fronts travelling through different portions of the specimen [301]. In a normal phase contrast image of a cell culture, thicker portions of the cell appear darker while the thinner portions appear lighter [301]. Also, specific to the phase contrast microscopy is the so-called halo effect in which the margins of a cell are represented by high intensity pixels. This information is important when looking at shape of cells when they are travelling freely versus when they are stationary within a monolayer. A simplified representation of the 2 cases is shown in Figure 7.19.

Moving cell vs cell in monolayer representation. When the cell are moving their height gradually increases from margins to the nucleus. When they are in the monolayer the height is approximately constant.

As it can be observed from Figure 10, the thickness of a moving cell seems to be gradually descending from the nucleus to its margins while in a monolayer the cells seem to have constant high thickness. In the case of a phase contrast image, the 2 cases should produce a gradual increase in pixel intensity from nucleus to the margins in the case of the moving cell while in the case of the monolayer cell it should produce many pixels with very low intensity representing the cell body surrounded by very high intensity pixels representing the halo effect.

As shown in Chapter 4-subsection 7.3.3 the formula for Contrast is:

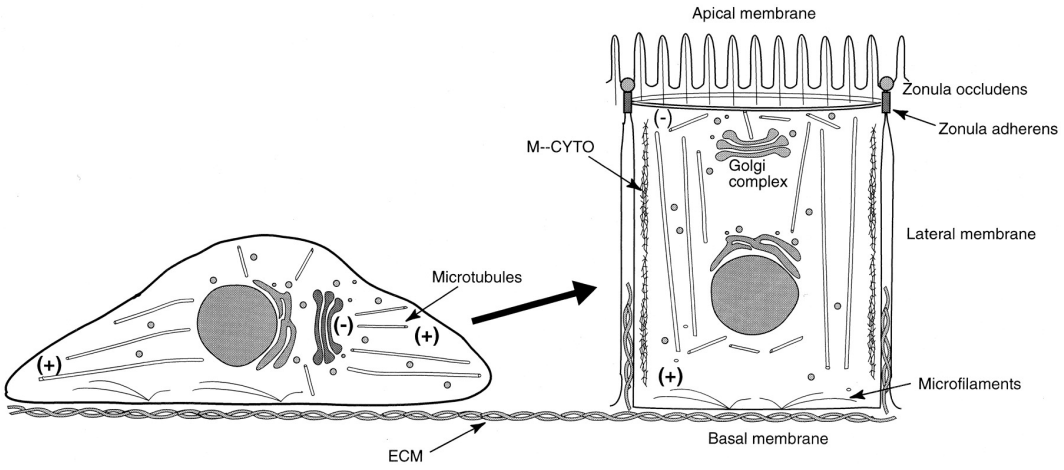

Figure 7.19: Moving cell vs cell in monolayer representation. When the cell are moving their height gradually increases from margins to the nucleus. When they are in the monolayer the height is approximately constant. Taken from: Yeaman, C., Grindstaff, K. K., & Nelson, W. J. (1999). New perspectives on mechanisms involved in generating epithelial cell polarity. Physiological reviews, 79(1), 73-98. [396]

$$Contrast = \sum_{n=0}^{N-1} n^2 \sum_{|i-j|=n} G_{i,j} \tag{7.11}$$

The formula can be simplified by replacing $\sum_{|i-j|=n} G_{i,j}$ with a term p_n which represents the proportion of pixel pairs whose difference in intensity is n and N the maximum difference in level between 2 neighbouring pixels after discretisation. The formula becomes:

$$Contrast = \sum_{n=0}^{N-1} n^2 p_n \tag{7.12}$$

with $\sum_{n=0}^{N-1} p_n = 1$

In the first case, of moving cells when the difference in intensity is gradual it is expected that p_n will have high values for small values of n, representing the pixels in the cell. Also high values of p will exist for the n corresponding to the difference in intensity level between the pixels representing the margin of the cells and the pixels in the halo.

In the second case, it is expected that all the p_n's except p_0 and p_{N-1} will be zero or at most have insignificant values as neighbouring pixels inside the cells will have nearly the same values while the pixels at the margin of the cell will produce significant differences from very dark pixels corresponding to the cell to very bright

ones corresponding to the halo. By increasing the density of the cells, the area of each cell reduces which also increases the ratio between pixels on the border and pixels inside the cell thus lowering p_0 while increasing p_{N-1}. As a result, when 100% confluence is reached, the contrast is expected to grow with the tightness of the cells.

The differences observed in this study are similar to other studies which indicate that the adhesion capacities of cells with PKD1 mutations are reduced. Examples in this sense are represented by an analysis done by Rocco et al. [295] in which they show that adhesion molecules are reduced in kidney cells coming from mice with PKD, or a study by Silberberg et al. [324] in which kidney cells coming from healthy humans were compared to cells coming from human kidney cysts and the authors show the formers form monolayers which are more resistant to the action of trypsin.

Another study [27], this time using computer simulations show that if the parameter responsible for describing cell adhesion is lower than normal, the cells start to create formations that look like cysts.

Further testing is needed to confirm the possibility of using contrast as a measure for cell tightness, but the technology for creating phase contrast images and the way in which the feature is computed combined with the results obtained in this study and the literature available on tightness of cells affected by disease seem to indicate it as a good candidate for this application. At the moment other studies have proposed methods for approximating cell density from images of cells but they rely on computing the average area of each cell after solving the more complex problem of segmentation [171] [53]. A method based on Haralick features will provide a way to do it directly from pixel intensity, circumventing accurate image segmentation.

7.8 Summary

In this chapter, different analysis methods were applied to a wound healing assay to compare the properties of control PTEC cells and those with knockdown of PKD1, 2 during cell migration.

The experiment was set up by growing cells in a 6-well plate until confluence using 3 wells for each condition, then creating a vertical wound in each well using a 200μl pipette tip and taking images using phase contrast microscopy every 10 minutes for a period of 2-4 days, with 7-8 fields per well.

The first measure employed to study the differences between the 2 conditions was the rate at which both types of cells were able to close the wound. In order to do this, the percentage of an image covered by cells was calculated for each frame of the movie taken at a specific location using automatic segmentation. The results show that the healthy cells were faster in closing the wound.

To investigate why disease cells were slower in wound closure, further analysis was performed on the trajectories of individual cells. Although the confinement ratio, a conventional measure of directionality of cell movement, shows some differences between the two conditions, the differences were not statistically significant and thus inconclusive. Further analysis in which the movement of cells along the x-axis and a new quantity consisting in the confinement ratio where the initial and final positions of the cells are evaluated only along the x-axis was employed to reveal differences between the two conditions.

Another analysis carried out involved computing the percentage of time a cell is closer to the wound than further away from it. The results show that the normal cells travel significantly more along the x axis toward the opposite edge of the wound whilst abnormal cells show a significantly higher degree of randomness.

The results of texture analysis using Haralick features show that the texture is smoother in the case of the cells affected by the disease than in the case of healthy cells. The Haralick feature Contrast, appears to be well correlated to cell packing density of the cells suggesting that adhesion characteristics is altered in cells affected by ADPKD . Further studies are needed to confirm this observation.

Chapter 8

Conclusions and Future Work

In this thesis, a study of ADPKD pathogenesis was conducted using microarray data and time-lapse cell imaging. The thesis sought to apply the latest techniques in feature selection, system identification and modelling in order to reveal new information about the disease.

Firstly, a new framework using state of the art techniques for the mathematical analysis of time-course microarray gene expression data has been created during the current work. This serves two purposes: selection of potentially relevant genes for a specific condition and modelling of the regulatory network governing their interactions. By applying the proposed framework to two different sets of microarray data derived from studies of Pkd1 knockout mice, a set of genes previously connected to the disease were identified, the most important being *Cdkn1a* for the first dataset and *Pkd2* for the second, proving its capacity to select relevant features. Also, two genes, *Cphf* (upregulated) and *Guca2b* (downregulated) were indicated as the most relevant by analytical feature selection methods. Of interest, both of them have also been linked to cancer, suggesting they might be biologically relevant for ADPKD. Mathematical models of gene interaction have show that for the first dataset *Cdkn1a* appears to up-regulate *Cphf* in the network for control subjects, a connection that is broken in the case of the mutant subjects. In the case of the second dataset, *Pkd2* seems to downregulate *Guca2b* in the case of the mutant subjects. In future studies, it would be interesting to compare the expression of *Cphf* and *Guca2b* in other disease models, both in vitro (cellular) and in vivo (mouse, rat). If the differences are consistent, the effect of knocking out *Chpf* or over-expressing *Guca2b* in disease cells or tissues could be analysed. The later could be performed as described by Menezes et al. [242] and Menezes et al. [243] to assess their effects on cystogenesis

in vivo. To test the regulations between genes, *Pkd2* and *Cdkn1* could be under and over expressed in diseased and healthy biological models and the effects on *Cphf* and *Guca2b* could be observed.

Secondly, it was observed that on ciPTEC cells, ADPKD, simulated by knocking down Pkd1 and Pkd2, seems to increase division time. Also, the use of a drug, octreotide, on the diseased cells does not seem to affect their division capacity. This is at odds with the literature on the subject where most studies [256][292][349][350] show that proliferation is increased by disease and decreased by octreotide [360][73] [218][17]. The observed results could be a feature of lentiviral transduction, cell immortalisation or tubule of origin so future studies with other cell models including primary cultures is needed to resolve this question. Second, the observed effects of octreotide on cell division could be dose-dependent so more doses should be tested.

Thirdly, quantification of migration using state of the art random motion mathematical models suggests that the disease lowers speed of ciPTEC cells. Measures for straightness of the motion, some never used before on ADPKD studies also suggest that the cells affected by the disease lose their capacity to maintain direction.These results are similar to the literature on the subject. Octreotide used in the low-density healing assays show some increase in the capacity of the cells to migrate but not to the level of the healthy ones. This is an interesting result as the literature shows that octreotide seems to slow down cell migration [332][10]. Again, studies on more cell lines, using various concentrations of octreotide could be used to better understand the effect of the drug on cell migration. Also a study on octreotide effects could be run using the wound-healing assay.

Finally, the last important result is the interesting finding that the Haralick feature Contrast was correlated with the tightness of the cell layer, a feature altered in the disease cells. This observation could be used in future as a non-invasive, easy to compute measure to assess cell confluence and in the case of wound closure analysis, as a measure for cell-cell adhesion. Future work could include analysing a wider range of epithelial cells at different states of confluence. A weakness in the current analysis was a high variance among the same time points in the same experimental conditions. The current analysis used raw images obtained from a phase contrast microscope so future analysis could be performed after image pre-processing to eliminate some of the possible noise in the measurements.

In conclusion, this thesis has utilised new approaches to study ADPKD pathogenesis and created some theoretical instruments for analysing cells and diseases.

Several findings in gene expression, network connections and cell behaviour have been identified which open up new directions for future research.

Bibliography

[1] Statistical algorithms description document. http://tools.thermofisher.com/content/sfs/brochures/sadd_whitepaper.pdf, 2002. [Online; accessed 23–2017].

[2] Assessment report. http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002788/WC500187923.pdf, 2015. [Online; accessed 10-September-2017].

[3] N. Aggarwal and R. Agrawal. First and second order statistics features for classification of magnetic resonance brain images. *Journal of Signal and Information Processing*, 3(02):146, 2012.

[4] G. Aguiari, M. Banzi, S. Gessi, Y. Cai, E. Zeggio, E. Manzati, R. Piva, E. Lambertini, L. Ferrari, D.J. Peters, et al. Deficiency of polycystin-2 reduces ca2+ channel activity and cell proliferation in adpkd lymphoblastoid cells. *The FASEB journal*, 18(7):884–886, 2004.

[5] M.E. Ahsen, N.K. Singh, T. Boren, M. Vidyasagar, and M.A. White. A new feature selection algorithm for two-class classification problems and application to endometrial cancer. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 2976–2982. IEEE, 2012.

[6] T. Akutsu, S. Miyano, S. Kuhara, et al. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific symposium on biocomputing*, volume 4, pages 17–28, 1999.

[7] A. Alecu. *Feature selection and modelling methods for microarray data from acute coronary syndrome*. PhD thesis, University of Sheffield, 2015.

[8] W.E. Allen, D. Zicha, A.J. Ridley, and G.E. Jones. A role for cdc42 in macrophage chemotaxis. *The Journal of cell biology*, 141(5):1147–1157, 1998.

[9] W. Alt. Correlation analysis of two-dimensional locomotion paths. In *Biological motion*, pages 254–268. Springer, 1990.

[10] M. Annunziata, R.M. Luque, M. Durán-Prado, A. Baragli, C. Grande, M. Volante, M.D. Gahete, F. Deltetto, M. Camanni, E. Ghigo, et al. Somatostatin and somatostatin analogues reduce pdgf-induced endometrial cell proliferation and motility. *Human reproduction*, 27(7):2117–2129, 2012.

[11] B. Arambepola. Fast computation of multidimensional discrete fourier transforms. In *IEE Proceedings F-Communications, Radar and Signal Processing*, volume 127, pages 49–52. IET, 1980.

[12] S. Aruna, L. Nandakishore, and S. Rajagopalan. A hybrid feature selection method based on igsbfs and naïve bayes for the diagnosis of erythemato-squamous diseases. *International Journal of Computer Applications*, 41(7), 2012.

[13] R. Atkinson, C. Rhodes, D. Macdonald, and R. Anderson. Scale-free dynamics in the movement patterns of jackals. *Oikos*, 98(1):134–140, 2002.

[14] M. Attanasio, N.H. Uhlenhaut, V.H. Sousa, J.F. O'Toole, E. Otto, K. Anlag, C. Klugmann, A.C. Treier, J. Helou, J.A. Sayer, et al. Loss of glis2 causes nephronophthisis in humans and mice by increased apoptosis and fibrosis. *Nature genetics*, 39(8):1018, 2007.

[15] H. Auer, D.L. Newsom, and K. Kornacker. Expression profiling using affymetrix genechip microarrays. *Microchip Methods in Diagnostics*, pages 35–46, 2009.

[16] O.T. Avery, C.M. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of experimental medicine*, 79(2):137–158, 1944.

[17] G.D. Ayiomamitis, G. Notas, A. Zaravinos, I. Drygiannakis, M. Georgiadou, O. Sfakianaki, N. Mastrodimou, K. Thermos, and E. Kouroumalis. Effects of

octreotide and insulin on colon cancer cellular proliferation and correlation with htert activity. *Oncoscience*, 1(6):457, 2014.

[18] M. Bakun, M. Niemczyk, D. Domanski, R. Jazwiec, A. Perzanowska, S. Niemczyk, M. Kistowski, A. Fabijanska, A. Borowiec, L. Paczek, et al. Urine proteome of autosomal dominant polycystic kidney disease patients. *Clinical proteomics*, 9(1):13, 2012.

[19] M. Bansal, G.D. Gatta, and D. Di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

[20] I. Barbaric, V. Biga, P.J. Gokhale, M. Jones, D. Stavish, A. Glen, D. Coca, and P.W. Andrews. Time-lapse analysis of human embryonic stem cells reveals multiple bottlenecks restricting colony formation and their relief upon culture adaptation. *Stem cell reports*, 3(1):142–155, 2014.

[21] H. Bateman. Higher transcendental functions [volumes i-iii], 1953.

[22] L. Battini, E. Fedorova, S. Macip, X. Li, P.D. Wilson, and G.L. Gusella. Stable knockdown of polycystin-1 confers integrin-$\alpha 2\beta 1$–mediated anoikis resistance. *Journal of the American Society of Nephrology*, 17(11):3049–3058, 2006.

[23] M. Beil, T. Irinopoulou, J. Vassy, and J.P. Rigaut. Chromatin texture analysis in three-dimensional images from confocal scanning laser microscopy. *Analytical and quantitative cytology and histology*, 17(5):323–331, 1995.

[24] M. Belge, M.E. Kilmer, and E.L. Miller. Efficient determination of multiple regularization parameters in a generalized l-curve framework. *Inverse Problems*, 18(4):1161, 2002.

[25] L. Bell and J.A. Madri. Influence of the angiotensin system on endothelial and smooth muscle cell migration. *The American journal of pathology*, 137 (1):7, 1990.

[26] R. Bellman. *Dynamic programming*. Courier Corporation, 2013.

[27] J.M. Belmonte, S.G. Clendenon, G.M. Oliveira, M.H. Swat, E.V. Greene, S. Jeyaraman, J.A. Glazier, and R.L. Bacallao. Virtual-tissue computer simulations define the roles of cell adhesion and proliferation in the onset of kidney cystic disease. *Molecular biology of the cell*, 27(22):3673–3685, 2016.

[28] J.B. Beltman, A.F. Marée, J.N. Lynch, M.J. Miller, and R.J. de Boer. Lymph node topology dictates t cell migration behavior. *Journal of Experimental Medicine*, 204(4):771–780, 2007.

[29] J.B. Beltman, A.F. Marée, and R.J. De Boer. Analysing immune cell migration. *Nature reviews. Immunology*, 9(11):789, 2009.

[30] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[31] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C.M. Perou, and J.S. Marron. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114, 2004.

[32] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[33] P. Berkhin et al. A survey of clustering data mining techniques. *Grouping multidimensional data*, 25:71, 2006.

[34] P. Bermejo, J.A. Gámez, and J.M. Puerta. Speeding up incremental wrapper feature subset selection with naive bayes classifier. *Knowledge-Based Systems*, 55:140–147, 2014.

[35] M.H. Bharati, J.J. Liu, and J.F. MacGregor. Image texture analysis: methods and comparisons. *Chemometrics and intelligent laboratory systems*, 72(1): 57–71, 2004.

[36] A.K. Bhunia, K. Piontek, A. Boletta, L. Liu, F. Qian, P.N. Xu, F.J. Germino, and G.G. Germino. Pkd1 induces p21 waf1 and regulation of the cell cycle via direct activation of the jak-stat signaling pathway in a process requiring pkd2. *Cell*, 109(2):157–168, 2002.

[37] C.M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[38] D.M. Blei. Hierarchical clustering. *Lecture Slides, February*, 2008.

[39] A.L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

[40] M. Boca, L. D'Amato, G. Distefano, R.S. Polishchuk, G.G. Germino, and A. Boletta. Polycystin-1 induces cell migration by regulating phosphatidylinositol 3-kinase-dependent cytoskeletal rearrangements and gsk3β-dependent cell–cell mechanical adhesion. *Molecular biology of the cell*, 18 (10):4050–4061, 2007.

[41] A. Boletta, F. Qian, L.F. Onuchic, A.K. Bhunia, B. Phakdeekitcharoen, K. Hanaoka, W. Guggino, L. Monaco, and G.G. Germino. Polycystin-1, the gene product of pkd1, induces resistance to apoptosis and spontaneous tubulogenesis in mdck cells. *Molecular cell*, 6(5):1267–1273, 2000.

[42] B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[43] P. Borgström, M. Hassan, E. Wassberg, E. Refai, C. Jonsson, S.A. Larsson, H. Jacobsson, and P. Kogner. The somatostatin analogue octreotide inhibits neuroblastoma growth in vivo. *Pediatric research*, 46(3):328–332, 1999.

[44] M. Bornens. Organelle positioning and cell polarity. *Nature reviews. Molecular cell biology*, 9(11):874, 2008.

[45] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[46] K. Bovis and S. Singh. Detection of masses in mammograms using texture features. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 2, pages 267–270. IEEE, 2000.

[47] A.W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

[48] S. Boyd. Ee263 lecture notes: Introduction to linear dynamical systems, 2008.

[49] S. Boyden. The chemotactic effect of mixtures of antibody and antigen on polymorphonuclear leucocytes. *Journal of Experimental Medicine*, 115(3): 453–466, 1962.

[50] M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C. Sugnet, M. Ares, and D. Haussler. Support vector machine classification of microarray gene expression data. *University of California, Santa Cruz, Technical Report UCSC-CRL-99-09*, 1999.

[51] P.O. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nature genetics*, 21(1s):33, 1999.

[52] R. Bumgarner. Overview of dna microarrays: types, applications, and their future. *Current protocols in molecular biology*, pages 22–1, 2013.

[53] S. Busschots, S. O'Toole, J.J. O'Leary, and B. Stordal. Non-invasive and non-destructive measurements of confluence in cultured adherent cell lines. *MethodsX*, 2:8–13, 2015.

[54] A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186, 2000.

[55] R.H. Byrd, R.B. Schnabel, and G.A. Shultz. Approximate solution of the trust region problem by minimization over two-dimensional subspaces. *Mathematical programming*, 40(1):247–263, 1988.

[56] G. Cai, J. Lian, S.S. Shapiro, and D.A. Beacham. Evaluation of endothelial cell migration with a novel in vitro assay system. *Methods in cell science*, 22 (2):107–114, 2000.

[57] L.M.d. Campos. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7(Oct):2149–2187, 2006.

[58] A.M. Carvalho. Scoring functions for learning bayesian networks. *Inesc-id Tec. Rep*, 2009.

[59] M. Castelli, M. Boca, M. Chiaravalli, H. Ramalingam, I. Rowe, G. Distefano, T. Carroll, and A. Boletta. Polycystin-1 binds par3/apkc and controls convergent extension during renal tubular morphogenesis. *Nature communications*, 4:2658, 2013.

[60] M. Castelli, C. De Pascalis, G. Distefano, N. Ducano, A. Oldani, L. Lanzetti, and A. Boletta. Regulation of the microtubular cytoskeleton by polycystin-1 favors focal adhesions turnover to modulate cell adhesion and migration. *BMC cell biology*, 16(1):15, 2015.

[61] J.W. Catto, D.A. Linkens, M.F. Abbod, M. Chen, J.L. Burton, K.M. Feeley, and F.C. Hamdy. Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks. *Clinical Cancer Research*, 9(11):4172–4177, 2003.

[62] A. Chaddad, C. Tanougast, A. Dandache, A. Al Houseini, and A. Bouridane. Improving of colon cancer cells detection based on haralick's features on segmented histopathological images. In *Computer Applications and Industrial Electronics (ICCAIE), 2011 IEEE International Conference on*, pages 87–90. IEEE, 2011.

[63] E. Chang, E.Y. Park, Y. mi Woo, D.H. Kang, Y.H. Hwang, C. Ahn, and J.H. Park. Restoring multidrug resistance-associated protein 3 attenuates cell proliferation in the polycystic kidney. *American Journal of Physiology-Renal Physiology*, 308(9):F1004–F1011, 2015.

[64] H.W. Chang, Y.H. Chiu, H.Y. Kao, C.H. Yang, and W.H. Ho. Comparison of classification algorithms with wrapper-based feature selection for predicting osteoporosis outcome based on genetic factors in a taiwanese women population. *International journal of endocrinology*, 2013, 2013.

[65] M.Y. Chang, E. Parker, S. Ibrahim, J.R. Shortland, M.E. Nahas, J.L. Haylor, and A.C. Ong. Haploinsufficiency of pkd2 is associated with increased tubular cell proliferation and interstitial fibrosis in two murine pkd2 models. *Nephrology Dialysis Transplantation*, 21(8):2078–2084, 2006.

[66] H.C. Chapin and M.J. Caplan. The cell biology of polycystic kidney disease. *The Journal of cell biology*, 191(4):701–710, 2010.

[67] S. Chatterjee, W.Y. Shi, P. Wilson, and A. Mazumdar. Role of lactosylceramide and map kinase in the proliferation of proximal tubular cells in human polycystic kidney disease. *Journal of lipid research*, 37(6):1334–1344, 1996.

[68] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2):e17238, 2011.

[69] K.H. Chen, K.J. Wang, M.L. Tsai, K.M. Wang, A.M. Adrian, W.C. Cheng, T.S. Yang, N.C. Teng, K.P. Tan, and K.S. Chang. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. *BMC bioinformatics*, 15(1):49, 2014.

[70] T. Chen, H.L. He, G.M. Church, et al. Modeling gene expression with differential equations. In *Pacific symposium on biocomputing*, volume 4, page 40, 1999.

[71] W.C. Chen, Y.S. Tzeng, and H. Li. Gene expression in early and progression phases of autosomal dominant polycystic kidney disease. *BMC research notes*, 1(1):131, 2008.

[72] D. Cheng, H. Qi, Z. Li, and J.B. Liu. Stability and stabilization of boolean networks. *International Journal of Robust and Nonlinear Control*, 21(2): 134–156, 2011.

[73] N. Cheung and S.C. Boyages. Somatostatin-14 and its analog octreotide exert a cytostatic effect on gh3 rat pituitary tumor cell proliferation via a transient g0/g1 cell cycle block. *Endocrinology*, 136(10):4174–4181, 1995.

[74] U.R. Chowdhury, R.S. Samant, O. Fodstad, and L.A. Shevde. Emerging role of nuclear protein 1 (nupr1) in cancer biology. *Cancer and Metastasis Reviews*, 28(1-2):225–232, 2009.

[75] E.A. Codling, M.J. Plank, and S. Benhamou. Random walk models in biology. *Journal of the Royal Society Interface*, 5(25):813–834, 2008.

[76] T.F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization*, 6(2):418–445, 1996.

[77] I.P.K.D. Consortium et al. Polycystic kidney disease: the complete structure of the pkd1 gene and its protein. *Cell*, 81(2):289–298, 1995.

[78] L.M. Cope, R.A. Irizarry, H.A. Jaffee, Z. Wu, and T.P. Speed. A benchmark for affymetrix genechip expression measures. *Bioinformatics*, 20(3):323–331, 2004.

[79] F.P. Cordelières. Manual tracking. *Institut Curie, Orsay (France)*, 2005.

[80] T.M. Cover and J.A. Thomas. Entropy, relative entropy and mutual information. *Elements of information theory*, 2:1–55, 1991.

[81] M.W. Craven and J.W. Shavlik. Using neural networks for data mining. *Future generation computer systems*, 13(2-3):211–229, 1997.

[82] D.D. Dalma-Weiszhausz, J. Warrington, E.Y. Tanimoto, and C.G. Miyada. [1] the affymetrix genechip® platform: An overview. *Methods in enzymology*, 410:3–28, 2006.

[83] M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.

[84] A. Datta, A. Choudhary, M.L. Bittner, and E.R. Dougherty. External control in markovian genetic regulatory networks. In *American Control Conference, 2003. Proceedings of the 2003*, volume 4, pages 3614–3619. IEEE, 2003.

[85] A. Datta, A. Choudhary, M.L. Bittner, and E.R. Dougherty. External control in markovian genetic regulatory networks: the imperfect information case. *Bioinformatics*, 20(6):924–930, 2004.

[86] L. Davis. Bit-climbing, representational bias, and test suite design. In *ICGA*, pages 18–23, 1991.

[87] K. De Rajat, N.R. Pal, and S.K. Pal. Feature analysis: Neural network and fuzzy set theoretic approaches. *Pattern Recognition*, 30(10):1579–1590, 1997.

[88] K. Deguchi, S. Ishizaka, Y. Kato, R. Takaki, and J. Toriwaki. Two-dimensional auto-regressive model for analysis and sythesis of gray-level textures. *Proc. of the 1st Int. Sym. for Science on Form, General Ed. S. Ishizaka, Eds. Y. Kato, R. Takaki, and J. Toriwaki*, pages 441–449, 1986.

[89] P. Delmas, H. Nomura, X. Li, M. Lakkis, Y. Luo, Y. Segal, J.M. Fernández-Fernández, P. Harris, A.M. Frischauf, D.A. Brown, et al. Constitutive activation of g-proteins by polycystin-1 is antagonized by polycystin-2. *Journal of Biological Chemistry*, 277(13):11276–11283, 2002.

[90] M.P. Derman, M. Cunha, E. Barros, S.K. Nigam, and L.G. Cantley. Hgf-mediated chemotaxis and tubulogenesis require activation of the phosphatidylinositol 3-kinase. *American Journal of Physiology-Renal Physiology*, 268(6):F1211–F1217, 1995.

[91] P. D'haeseleer, X. Wen, S. Fuhrman, R. Somogyi, et al. Linear modeling of mrna expression levels during cns development and injury. In *Pacific symposium on biocomputing*, volume 4, pages 41–52, 1999.

[92] S.M. Dhanasekaran, T.R. Barrette, D. Ghosh, R. Shah, et al. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822, 2001.

[93] S. Di Cataldo and E. Ficarra. Mining textural knowledge in biological images: Applications, methods and trends. *Computational and structural biotechnology journal*, 15:56–67, 2017.

[94] P. Dieterich, R. Klages, R. Preuss, and A. Schwab. Anomalous dynamics of cell migration. *Proceedings of the National Academy of Sciences*, 105(2): 459–463, 2008.

[95] P.A. DiMilla, J.A. Quinn, S.M. Albelda, and D.A. Lauffenburger. Measurement of individual cell migration parameters for human tissue cells. *AIChE Journal*, 38(7):1092–1104, 1992.

[96] J.D. Doecke, S.M. Laws, N.G. Faux, W. Wilson, S.C. Burnham, C.P. Lam, A. Mondal, J. Bedo, A.I. Bush, B. Brown, et al. Blood-based protein biomarkers for diagnosis of alzheimer disease. *Archives of neurology*, 69 (10):1318–1325, 2012.

[97] D.L. Donoho. For most large underdetermined systems of linear equations the minimal $l1$-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

[98] J. Drukala, K. Urbanska, A. Wilk, M. Grabacka, E. Wybieralska, L. Del Valle, Z. Madeja, and K. Reiss. Ros accumulation and igf-ir inhibition contribute to fenofibrate/pparα-mediated inhibition of glioma cell motility in vitro. *Molecular cancer*, 9(1):159, 2010.

[99] I.A. Drummond. Polycystins, focal adhesions and extracellular matrix interactions. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1812(10):1322–1326, 2011.

[100] P. Du, W.A. Kibbe, and S.M. Lin. nuid: a universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays. *Biology direct*, 2(1):16, 2007.

[101] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Wiley, New York, 1973.

[102] S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103, 2003.

[103] G.A. Dunn. Characterising a kinesis response: time averaged measures of cell speed and directional persistence. *Agents and actions. Supplements*, 12: 14, 1983.

[104] B.P. Durbin, J.S. Hardin, D.M. Hawkins, and D.M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(suppl_1):S105–S110, 2002.

[105] T. Ecder, V.Y. Melnikov, M. Stanley, D. Korular, M.S. Lucia, R.W. Schrier, and C.L. Edelstein. Caspases, bcl-2 proteins and apoptosis in autosomal-dominant polycystic kidney disease. *Kidney international*, 61(4):1220–1230, 2002.

[106] A. Edelstein, N. Amodaj, K. Hoover, R. Vale, and N. Stuurman. Computer control of microscopes using μmanager. *Current protocols in molecular biology*, pages 14–20, 2010.

[107] D. Edwards. Non-linear normalization and background correction in one-channel cdna microarray studies. *Bioinformatics*, 19(7):825–833, 2003.

[108] A. Einstein. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der physik*, 322(8):549–560, 1905.

[109] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[110] B. Ekser and P. Rigotti. Autosomal dominant polycystic kidney disease. *New England Journal of Medicine*, 363(1):71–71, 2010.

[111] V. Engineering. History of the microscope. `http://www.history-of-the-microscope.org/`, 2008. [Online; accessed 19-June-2017].

[112] F. Entschladen, T.L. Drell, K. Lang, K. Masur, D. Palm, P. Bastian, B. Niggemann, and K.S. Zaenker. Analysis methods of human cell migration. *Experimental cell research*, 307(2):418–426, 2005.

[113] S. Even-Ram, A.D. Doyle, M.A. Conti, K. Matsumoto, R.S. Adelstein, and K.M. Yamada. Myosin iia regulates cell motility and actomyosin-microtubule crosstalk. *Nature cell biology*, 9(3):299, 2007.

[114] F. Falciani. *Microarray technology through applications*. Taylor & Francis, 2007.

[115] A. Fauré, A. Naldi, C. Chaouiya, and D. Thieffry. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124–e131, 2006.

[116] S. Feil, N. Valtcheva, and R. Feil. Inducible cre mice. *Gene Knockout Protocols: Second Edition*, pages 343–363, 2009.

[117] K.N. Felekkis, P. Koupepidou, E. Kastanos, R. Witzgall, C.X. Bai, L. Li, L. Tsiokas, N. Gretz, and C. Deltas. Mutant polycystin-2 induces proliferation in primary rat tubular epithelial cells in a stat-1/p21-independent fashion accompanied instead by alterations in expression of p57 kip2 and cdk2. *BMC nephrology*, 9(1):10, 2008.

[118] G. Felsenfeld and H.T. Miles. The physical and chemical properties of nucleic acids. *Annual review of biochemistry*, 36(1):407–448, 1967.

[119] R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[120] F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5(Nov):1531–1555, 2004.

[121] I.K. Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.

[122] D.J. Foran, D. Comaniciu, P. Meer, and L.A. Goodell. Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy. *IEEE Transactions on Information Technology in Biomedicine*, 4(4):265–273, 2000.

[123] C.D. Freel, K.O. Gilliland, C.W. Lane, F.J. Giblin, and M.J. Costello. Fourier analysis of cytoplasmic texture in nuclear fiber cells from transparent and cataractous human and animal lenses. *Experimental eye research*, 74(6): 689–702, 2002.

[124] C.W. Frevert, G. Boggy, T.M. Keenan, and A. Folch. Measurement of cell migration in response to an evolving radial chemokine gradient triggered by a microvalve. *Lab on a Chip*, 6(7):849–856, 2006.

[125] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

[126] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 196–205. Morgan Kaufmann Publishers Inc., 1999.

[127] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[128] M. Fronza, B. Heinzmann, M. Hamburger, S. Laufer, and I. Merfort. Determination of the wound healing effect of calendula extracts using the scratch assay with 3t3 fibroblasts. *Journal of ethnopharmacology*, 126(3):463–467, 2009.

[129] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10): 906–914, 2000.

[130] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on multimedia*, 1 (3):264–277, 1999.

[131] O. García-Suárez, B. García, I. Fernández-Vega, A. Astudillo, and L.M. Quirós. Neuroendocrine tumors show altered expression of chondroitin sulfate, glypican 1, glypican 5, and syndecan 2 depending on their differentiation grade. *Frontiers in oncology*, 4, 2014.

[132] U. Gawlik-Dziki, M. Świeca, M. Sułkowski, D. Dziki, B. Baraniak, and J. Czyż. Antioxidant and anticancer activities of chenopodium quinoa leaves extracts–in vitro study. *Food and Chemical Toxicology*, 57:154–160, 2013.

[133] S.P. Gordon, H. Priest, D.L. Des Marais, W. Schackwitz, M. Figueroa, J. Martin, J.N. Bragg, L. Tyler, C.R. Lee, D. Bryant, et al. Genome diversity in brachypodium distachyon: deep sequencing of highly diverse inbred lines. *The Plant Journal*, 79(3):361–374, 2014.

[134] J.J. Grantham, V.E. Torres, A.B. Chapman, L.M. Guay-Woodford, K.T. Bae, B.F. King Jr, L.H. Wetzel, D.A. Baumgarten, P.J. Kenney, P.C. Harris, et al. Volume progression in polycystic kidney disease. *New England Journal of Medicine*, 354(20):2122–2130, 2006.

[135] A. Greenfield, A. Madar, H. Ostrer, and R. Bonneau. Dream4: Combining

genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5(10):e13397, 2010.

[136] J. Gu, M. Tamura, R. Pankov, E.H. Danen, T. Takino, K. Matsumoto, and K.M. Yamada. Shc and fak differentially regulate cell motility and directionality modulated by pten. *The Journal of cell biology*, 146(2):389–404, 1999.

[137] K.L. Gunderson, S. Kruglyak, M.S. Graige, F. Garcia, B.G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, et al. Decoding randomly ordered dna arrays. *Genome research*, 14(5):870–877, 2004.

[138] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

[139] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[140] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[141] H. Haga, C. Irahara, R. Kobayashi, T. Nakagaki, and K. Kawabata. Collective movement of epithelial cells on a collagen gel substrate. *Biophysical journal*, 88(3):2250–2256, 2005.

[142] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

[143] K. Hanaoka and W.B. Guggino. camp regulates cell proliferation and cyst formation in autosomal polycystic kidney disease cells. *Journal of the American Society of Nephrology*, 11(7):1179–1187, 2000.

[144] P.C. Hansen. *The L-curve and its use in the numerical treatment of inverse problems*. IMM, Department of Mathematical Modelling, Technical Universityof Denmark, 1999.

[145] R.M. Haralick, K. Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[146] P.C. Harris and V.E. Torres. Genetic mechanisms and signaling pathways in autosomal dominant polycystic kidney disease. *The Journal of clinical investigation*, 124(6):2315, 2014.

[147] J.W. Haycock. 3d cell culture: a review of current approaches and techniques. *3D cell culture: methods and protocols*, pages 1–15, 2011.

[148] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke. Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems*, 96(1):86–103, 2009.

[149] D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[150] E. Higashihara, K. Nutahara, T. Okegawa, M. Tanbo, H. Mori, I. Miyazaki, T. Nitatori, and K. Kobayashi. Safety study of somatostatin analogue octreotide for autosomal dominant polycystic kidney disease in japan. *Clinical and experimental nephrology*, 19(4):746–752, 2015.

[151] Z.M. Hira and D.F. Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.

[152] T.K. Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[153] S. Hochreiter, D.A. Clevert, and K. Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–949, 2006.

[154] M.C. Hogan, T.V. Masyuk, L.J. Page, V.J. Kubly, E.J. Bergstralh, X. Li, B. Kim, B.F. King, J. Glockner, D.R. Holmes, et al. Randomized clinical trial of long-acting somatostatin for autosomal dominant polycystic kidney

and liver disease. *Journal of the American Society of Nephrology*, 21(6): 1052–1061, 2010.

[155] N.S. Holter, M. Mitra, A. Maritan, M. Cieplak, J.R. Banavar, and N.V. Fedoroff. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97(15): 8409–8414, 2000.

[156] R. Hooke. *Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses, with observations and inquiries thereupon*. Courier Corporation, 2003.

[157] L.C. Huang, S.Y. Hsu, and E. Lin. A comparison of classification methods for predicting chronic fatigue syndrome based on genetic data. *Journal of Translational Medicine*, 7(1):81, 2009.

[158] S. Huang. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *Journal of molecular medicine*, 77(6):469–480, 1999.

[159] W. Huber, A. Von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl_1):S96–S104, 2002.

[160] S. Huh, H. Su, T. Kanade, et al. Apoptosis detection for adherent cell populations in time-lapse phase-contrast microscopy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 331–339. Springer, 2012.

[161] E.B. Hunt, J. Marin, and P.J. Stone. *Experiments in induction.* Academic press, 1966.

[162] H. Husson, P. Manavalan, V.R. Akmaev, R.J. Russo, B. Cook, B. Richards, D. Barberio, D. Liu, X. Cao, G.M. Landes, et al. New insights into adpkd molecular pathways using combination of sage and microarray technologies. *Genomics*, 84(3):497–510, 2004.

[163] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, et al. An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716*, 2015.

[164] P. Igarashi, S. Somlo, et al. Genetics and pathogenesis of polycystic kidney disease. *Journal of the American Society of Nephrology*, 13(9):2384–2398, 2002.

[165] *Whole-Genome Gene Expression Direct Hybridization Assay Guide*. Illumina Inc, June 2010. Part 11322355 Rev.A.

[166] J. Imitola, K. Raddassi, K.I. Park, F.J. Mueller, M. Nieto, Y.D. Teng, D. Frenkel, J. Li, R.L. Sidman, C.A. Walsh, et al. Directed migration of neural stem cells to sites of cns injury by the stromal cell-derived factor 1α/cxc chemokine receptor 4 pathway. *Proceedings of the National Academy of Sciences*, 101(52):18117–18122, 2004.

[167] I. Inza, B. Sierra, R. Blanco, and P. Larrañaga. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *Journal of Intelligent & Fuzzy Systems*, 12(1):25–33, 2002.

[168] M.V. Iorio, M. Ferracin, C.G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, et al. Microrna gene expression deregulation in human breast cancer. *Cancer research*, 65(16):7065–7070, 2005.

[169] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[170] A. Irrthum, L. Wehenkel, P. Geurts, et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[171] N. Jaccard, L.D. Griffin, A. Keser, R.J. Macown, A. Super, F.S. Veraitch, and N. Szita. Automated method for the rapid and precise estimation of adherent cell culture characteristics from phase contrast microscopy images. *Biotechnology and bioengineering*, 111(3):504–517, 2014.

[172] A.K. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. In *Systems, Man and Cybernetics, 1990. Conference Proceedings., IEEE International Conference on*, pages 14–19. IEEE, 1990.

[173] A.K. Jain, J. Mao, and K.M. Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[174] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on knowledge and data engineering*, 16(11): 1370–1386, 2004.

[175] G.H. John, R. Kohavi, K. Pfleger, et al. Irrelevant features and the subset selection problem. In *Machine learning: proceedings of the eleventh international conference*, pages 121–129, 1994.

[176] W.E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[177] D. Joly, V. Morel, A. Hummel, A. Ruello, P. Nusbaum, N. Patey, L.H. Noël, P. Rousselle, and B. Knebelmann. β 4 integrin and laminin 5 are aberrantly expressed in polycystic kidney disease: Role in increased cell adhesion and migration. *The American journal of pathology*, 163(5):1791–1800, 2003.

[178] J.E. Jonkman, J.A. Cathcart, F. Xu, M.E. Bartolini, J.E. Amon, K.M. Stevens, and P. Colarusso. An introduction to the wound healing assay using live-cell microscopy. *Cell adhesion & migration*, 8(5):440–451, 2014.

[179] M.S. Joshi, P.P. Bartakke, and M. Sutaone. Texture representation using autoregressive models. In *Advances in Computational Tools for Engineering Applications, 2009. ACTEA'09. International Conference on*, pages 386–390. IEEE, 2009.

[180] B. Julész, E. Gilbert, and J.D. Victor. Visual discrimination of textures with identical third-order statistics. *Biological Cybernetics*, 31(3):137–140, 1978.

[181] C.R. Justus, N. Leffler, M. Ruiz-Echevarria, and L.V. Yang. In vitro cell migration and invasion assays. *Journal of visualized experiments: JoVE*, (88), 2014.

[182] S. Kapoor, D. Rodriguez, M. Riwanto, I. Edenhofer, S. Segerer, K. Mitchell, and R.P. Wüthrich. Effect of sodium-glucose cotransport inhibition on polycystic kidney disease progression in pck rats. *PloS one*, 10(4):e0125603, 2015.

[183] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews. Molecular cell biology*, 9(10):770, 2008.

[184] T. Kazmar, M. Šmíd, M. Fuchs, B. Luber, and J. Mattes. Learning cellular texture features in microscopic cancer cell images for automated cell-detection. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 49–52. IEEE, 2010.

[185] C.R. Keese, J. Wegener, S.R. Walker, and I. Giaever. Electrical wound-healing assay for cells in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1554–1559, 2004.

[186] A. Khachaturyan, S. Semenovskaya, and B. Vainstein. A statistical-thermodynamic approach to determination of structure amplitude phases. *Sov. Phys. Crystallogr*, 24:519–524, 1979.

[187] A.A. Khalili and M.R. Ahmad. A review of cell adhesion studies for biomedical and biological applications. *International journal of molecular sciences*, 16(8):18149–18184, 2015.

[188] B.C. Kim, H.C. Lee, J.J. Lee, C.M. Choi, D.K. Kim, J.C. Lee, Y.G. Ko, and J.S. Lee. Wig1 prevents cellular senescence by regulating p21 mrna decay through control of risc recruitment. *The EMBO journal*, 31(22):4289–4303, 2012.

[189] S. Kim, J. Kim, and K.H. Cho. Inferring gene regulatory networks from temporal expression profiles under time-delay and noise. *Computational biology and chemistry*, 31(4):239–245, 2007.

[190] T.Y. Kim, N.H. Cho, G.B. Jeong, E. Bengtsson, and H.K. Choi. 3d texture analysis in renal cell carcinoma tissue image grading. *Computational and mathematical methods in medicine*, 2014, 2014.

[191] J. Klafter, S. Lim, and R. Metzler. *Fractional dynamics: recent advances*. World Scientific, 2012.

[192] R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[193] A. Konvalinka, I. Batruch, T. Tokar, A. Dimitromanolakis, S. Reid, X. Song, Y. Pei, A.P. Drabovich, E.P. Diamandis, I. Jurisica, et al. Quantification of angiotensin ii-regulated proteins in urine of patients with polycystic and other chronic kidney diseases by selected reaction monitoring. *Clinical proteomics*, 13(1):16, 2016.

[194] C. Kooperberg, T.G. Fazzio, J.J. Delrow, and T. Tsukiyama. Improved background correction for spotted dna microarrays. *Journal of Computational Biology*, 9(1):55–66, 2002.

[195] T.J. Koshki, E. Hajizadeh, and M. Karimi. A comparison of selective classification methods in dna microarray data of cancer: Some recommendations for application in health promotion. *Health promotion perspectives*, 3(1): 129, 2013.

[196] S.B. Kotsiantis, I.D. Zaharakis, and P.E. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26 (3):159–190, 2006.

[197] S.B. Kotsiantis, I. Zaharakis, and P. Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[198] P. Koupepidou, K.N. Felekkis, B. Kränzlin, C. Sticht, N. Gretz, and C. Deltas. Cyst formation in the pkd2 (1-703) transgenic rat precedes deregulation of proliferation-related pathways. *BMC nephrology*, 11(1):23, 2010.

[199] S. Kouvroukoglou, K.C. Dee, R. Bizios, L.V. McIntire, and K. Zygourakis. Endothelial cell migration on surfaces modified with immobilized adhesive peptides. *Biomaterials*, 21(17):1725–1733, 2000.

[200] N. Kramer, A. Walzl, C. Unger, M. Rosner, G. Krupitza, M. Hengstschläger, and H. Dolznig. In vitro cell migration and invasion assays. *Mutation Research/Reviews in Mutation Research*, 752(1):10–24, 2013.

[201] K. Kuhn, S.C. Baker, E. Chudin, M.H. Lieu, S. Oeser, H. Bennett, P. Rigault, D. Barker, T.K. McDaniel, and M.S. Chee. A novel, high-performance ran-

dom array platform for quantitative gene expression profiling. *Genome research*, 14(11):2347–2356, 2004.

[202] M.B. Kursa, A. Jankowski, and W.R. Rudnicki. Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.

[203] A. Lakatos, R.J. Franklin, S.C. Barnett, et al. Olfactory ensheathing cells and schwann cells differ in their in vitro interactions with astrocytes. *Glia*, 32(3): 214–225, 2000.

[204] M. Lal, X. Song, J.L. Pluznick, V. Di Giovanni, D.M. Merrick, N.D. Rosenblum, V. Chauvet, C.J. Gottardi, Y. Pei, and M.J. Caplan. Polycystin-1 c-terminal tail associates with β-catenin and inhibits canonical wnt signaling. *Human molecular genetics*, 17(20):3105–3117, 2008.

[205] T.J. Langan and R.C. Chou. Synchronization of mammalian cell cultures by serum deprivation. *Cell Cycle Synchronization: Methods and Protocols*, pages 75–83, 2011.

[206] P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5): 719–720, 2007.

[207] P. Langley et al. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, volume 184, pages 245–271, 1994.

[208] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

[209] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9 (4):1106–1119, 2012.

[210] N. Le Novere. Quantitative and logic modelling of gene and molecular networks. *Nature Reviews. Genetics*, 16(3):146, 2015.

[211] J.W. Lee, J.B. Lee, M. Park, and S.H. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.

[212] J.T. Leek and J.D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[213] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, and R.A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics*, 11(10), 2010.

[214] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001.

[215] X. Li, Z. Wu, Y. Wang, Q. Mei, X. Fu, and W. Han. Characterization of adult α-and β-globin elevated by hydrogen peroxide in cervical cancer cells that play a cytoprotective role against oxidative insults. *PLoS One*, 8(1):e54342, 2013.

[216] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. 1998.

[217] S.M. Lin, P. Du, W. Huber, and W.A. Kibbe. Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic acids research*, 36(2):e11–e11, 2008.

[218] H.L. Liu, L. Huo, and L. Wang. Octreotide inhibits proliferation and induces apoptosis of hepatocellular carcinoma cells. *Acta pharmacologica Sinica*, 25 (10):1380–1386, 2004.

[219] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.

[220] L.Z. Liu, F.X. Wu, and W.J. Zhang. A group lasso-based method for robustly inferring gene regulatory networks from multiple time-course datasets. *BMC systems biology*, 8(3):S1, 2014.

[221] S. Livens, P. Scheunders, G. Van de Wouwer, and D. Van Dyck. Wavelets for texture analysis, an overview. 1997.

[222] N. Loewke. Haralick texture analysis for stem cell identification.

[223] L.H. Loo, L.F. Wu, and S.J. Altschuler. Image-based multivariate profiling of drug responses from single cells. *Nature methods*, 4(5):445–453, 2007.

[224] G.A. Losa and C. Castelli. Nuclear patterns of human breast cancer cells during apoptosis: characterisation by fractal dimension and co-occurrence matrix statistics. *Cell and tissue research*, 322(2):257–267, 2005.

[225] B. Lovell, R. Walker, R.F. Walker, and P. Jackway. Cervical cell classification via co-occurrence and markov random field features. In *In Proceedings of DICTA-95, Digital Image Computing: Techniques and Applications*. Citeseer, 1995.

[226] A. Luyten, X. Su, S. Gondela, Y. Chen, S. Rompani, A. Takakura, and J. Zhou. Aberrant regulation of planar cell polarity in polycystic kidney disease. *Journal of the American Society of Nephrology*, 21(9):1521–1532, 2010.

[227] A.M. Lyapunov. The general problem of the stability of motion. *International Journal of Control*, 55(3):531–534, 1992.

[228] V. Macaulay, I. Smith, M. Everard, J. Teale, J. Reubi, and J. Millar. Experimental and clinical studies with somatostatin analogue octreotide in small cell lung cancer. *British journal of cancer*, 64(3):451, 1991.

[229] B.T. MacDonald, K. Tamai, and X. He. Wnt/β-catenin signaling: components, mechanisms, and diseases. *Developmental cell*, 17(1):9–26, 2009.

[230] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[231] G. Malathi and V. Shanthi. Wavelet image fusion approach for classification of ultrasound placenta complicated by gestational diabetes mellitus. In *Pathophysiology and Complications of Diabetes Mellitus*. InTech, 2012.

[232] S. Maldonado and R. Weber. Embedded feature selection for support vector machines: State-of-the-art and future challenges. In *CIARP*, pages 304–311. Springer, 2011.

[233] B. Mandelbrot. Geometrical facets of statistical physics: scaling and fractals. *Annals of the Israel Physical Society*, 2:225, 1978.

[234] K.C. Martin and A. Ephrussi. mrna localization: gene expression in the spatial dimension. *Cell*, 136(4):719–730, 2009.

[235] B.R. Masters, R.C. Gonzalez, and R. Woods. Digital image processing. *Journal of biomedical optics*, 14(2):029901, 2009.

[236] T.V. Masyuk, A.I. Masyuk, V.E. Torres, P.C. Harris, and N.F. Larusso. Octreotide inhibits hepatic cystogenesis in a rodent model of polycystic liver disease by reducing cholangiocyte adenosine 3', 5'-cyclic monophosphate. *Gastroenterology*, 132(3):1104–1116, 2007.

[237] A. Materka, M. Strzelecki, et al. Texture analysis methods–a review. *Technical university of lodz, institute of electronics, COST B11 report, Brussels*, pages 9–11, 1998.

[238] G. Matheron and G. De Marsily. Is transport in porous media always diffusive? a counterexample. *Water Resources Research*, 16(5):901–917, 1980.

[239] E. Méhes, A. Czirók, B. Hegedüs, T. Vicsek, and V. Jancsik. Laminin-1 increases motility, path-searching, and process dynamism of rat and mouse muller glial cells in vitro: Implication of relationship between cell behavior and formation of retinal morphology. *Cytoskeleton*, 53(3):203–213, 2002.

[240] E. Meijering, O. Dzyubachyk, I. Smal, et al. 9 methods for cell and particle tracking. *Methods in enzymology*, 504(9):183–200, 2012.

[241] D. Mekahli, E. Sammels, T. Luyten, K. Welkenhuyzen, L. van den Heuvel, E. Levtchenko, R. Gijsbers, G. Bultynck, J. Parys, H. De Smedt, et al. Polycystin-1 and polycystin-2 are both required to amplify inositol-trisphosphate-induced ca 2+ release. *Cell calcium*, 51(6):452–458, 2012.

[242] L.F. Menezes, F. Zhou, A.D. Patterson, K.B. Piontek, K.W. Krausz, F.J. Gonzalez, and G.G. Germino. Network analysis of a pkd1-mouse model of autosomal dominant polycystic kidney disease identifies hnf4α as a disease modifier. *PLoS genetics*, 8(11):e1003053, 2012.

[243] L.F. Menezes, C.C. Lin, F. Zhou, and G.G. Germino. Fatty acid oxidation is impaired in an orthologous mouse model of autosomal dominant polycystic kidney disease. *EBioMedicine*, 5:183–192, 2016.

[244] R. Metzler and J. Klafter. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics reports*, 339(1):1–77, 2000.

[245] R. Metzler and J. Klafter. The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. *Journal of Physics A: Mathematical and General*, 37(31):R161, 2004.

[246] P.E. Meyer, C. Schretter, and G. Bontempi. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):261–274, 2008.

[247] M.M. Mihaylova and R.J. Shaw. The ampk signalling pathway coordinates cell growth, autophagy and metabolism. *Nature cell biology*, 13(9):1016, 2011.

[248] E. Million. The hadamard product. *Course Notes*, 3:6, 2007.

[249] G. Mittag-Leffler. Une généralisation de l'intégrale de laplace-abel. *CR Acad. Sci. Paris (Ser. II)*, 137:537–539, 1903.

[250] T. Mochizuki, G. Wu, T. Hayashi, S.L. Xenophontos, B. Veldhuisen, J.J. Saris, D.M. Reynolds, Y. Cai, P.A. Gabow, A. Pierides, et al. Pkd2, a gene for polycystic kidney disease that encodes an integral membrane protein. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 1339–1342, 1996.

[251] D.K. Molina and V.J. DiMaio. Normal organ weights in men: part ii-the brain, lungs, liver, spleen, and kidneys. *The American journal of forensic medicine and pathology*, 33(4):368–372, 2012.

[252] C. Morain, A. Segal, D. Walker, and A. Levi. Abnormalities of neutrophil function do not cause the migration defect in crohn's disease. *Gut*, 22(10): 817–822, 1981.

[253] F. Mordelet and J.P. Vert. Sirene: supervised inference of regulatory networks. *Bioinformatics*, 24(16):i76–i82, 2008.

[254] I. Mostinsky. Diffusion coefficient, 1996.

[255] K.P. Murphy and S. Russell. Dynamic bayesian networks: representation, inference and learning. 2002.

[256] T. Nadasdy, Z. Laszik, G. Lajoie, K.E. Blick, D.E. Wheeler, and F.G. Silva. Proliferative activity of cyst epithelium in human renal cystic diseases. *Journal of the American Society of Nephrology*, 5(7):1462–1468, 1995.

[257] L. Nanni and A. Lumini. A reliable method for cell phenotype image classification. *Artificial intelligence in medicine*, 43(2):87–97, 2008.

[258] C. Nickel, T. Benzing, L. Sellin, P. Gerke, A. Karihaloo, Z.X. Liu, L.G. Cantley, and G. Walz. The polycystin-1 c-terminal fragment triggers branching morphogenesis and migration of tubular kidney epithelial cells. *The Journal of clinical investigation*, 109(4):481, 2002.

[259] E.A. Nigro, M. Castelli, and A. Boletta. Role of the polycystins in cell migration, polarity, and tissue morphogenesis. *Cells*, 4(4):687–705, 2015.

[260] C. Nordqvist. Cysts: Causes, types, and treatments. *https://www.medicalnewstoday.com/*, Aug 2017.

[261] R. Nosaka and K. Fukui. Hep-2 cell classification using rotation invariant co-occurrence among local binary patterns. *Pattern Recognition*, 47(7):2428–2436, 2014.

[262] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 582–585. IEEE, 1994.

[263] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[264] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[265] M.H. Oliver, N.K. Harrison, J.E. Bishop, P.J. Cole, and G.J. Laurent. A rapid and convenient assay for counting cells cultured in microwell plates: application for assessment of growth factors. *Journal of cell science*, 92(3): 513–518, 1989.

[266] A.C. Ong, O. Devuyst, B. Knebelmann, G. Walz, and E.E.W.G. for Inherited. Autosomal dominant polycystic kidney disease: the changing face of clinical management. *The Lancet*, 385(9981):1993–2002, 2015.

[267] K. Osafune, T. Ameku, and A. Watanabe. Method for testing for autosomal dominant polycystic kidney disease and method for screening agent for treatment of the disease, July 15 2016. US Patent App. 15/211,131.

[268] H.G. Othmer, S.R. Dunbar, and W. Alt. Models of dispersal in biological systems. *Journal of mathematical biology*, 26(3):263–298, 1988.

[269] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[270] P. Outeda, D.L. Huso, S.A. Fisher, M.K. Halushka, H. Kim, F. Qian, G.G. Germino, and T. Watnick. Polycystin signaling is required for directed endothelial cell migration and lymphatic development. *Cell reports*, 7(3):634–644, 2014.

[271] A. Packard. Jacobian linearizations, equilibrium points. `https://jagger.berkeley.edu/~pack/me132/Section18.pdf`, 2005. [Online; accessed 28-August-2017].

[272] R. Pal, A. Datta, M.L. Bittner, and E.R. Dougherty. Intervention in context-sensitive probabilistic boolean networks. *Bioinformatics*, 21(7):1211–1218, 2004.

[273] W. Pan, J. Lin, and C.T. Le. How many replicates of arrays are required to detect gene expression changes in microarray experiments? a mixture model approach. *Genome biology*, 3(5):research0022–1, 2002.

[274] P. Pandey, S. Qin, J. Ho, J. Zhou, and J.A. Kreidberg. Systems biology approach to identify transcriptome reprogramming and candidate microrna targets during the progression of polycystic kidney disease. *BMC systems biology*, 5(1):56, 2011.

[275] E.Y. Park, Y.H. Sung, M.H. Yang, J.Y. Noh, S.Y. Park, T.Y. Lee, Y.J. Yook, K.H. Yoo, K.J. Roh, I. Kim, et al. Cyst formation in kidney via b-raf signaling in the pkd2 transgenic mice. *Journal of Biological Chemistry*, 284(11):7214–7222, 2009.

[276] E. Parker, L.J. Newby, C.C. Sharpe, S. Rossetti, A.J. Streets, P.C. Harris, M.J. O'Hare, and A.C. Ong. Insulin-like growth factor-1 induces hyperproliferation of pkd1 cystic cells via a ras/raf dependent signalling pathway. *Kidney international*, 72(2):157, 2007.

[277] G. Passucci, M.E. Brasch, N.O. Deakin, C.E. Turner, J.H. Henderson, and M.L. Manning. Superdiffusive cell motility on 2d substrates modeled as a persistent lévy walk. In *APS Meeting Abstracts*, 2016.

[278] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8): 1226–1238, 2005.

[279] A.P. Pentland. Fractal-based description of natural scenes. *IEEE transactions on pattern analysis and machine intelligence*, (6):661–674, 1984.

[280] L. Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

[281] B.E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d ´ Alche-Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19(suppl_2):ii138–ii148, 2003.

[282] O. Pharmaceutical. Otsuka pharmaceutical's samsca® approved in japan as the worlds's first drug therapy for adpkd, a rare kidney disease. *https://www.otsuka.co.jp/*, Mar 2014.

[283] A. Phinyomark, S. Jitaree, P. Phukpattaranont, and P. Boonyapiphat. Texture analysis of breast cancer cells in microscopic images using critical exponent analysis method. *Procedia Engineering*, 32:232–238, 2012.

[284] K. Piontek, L.F. Menezes, M.A. Garcia-Gonzalez, D.L. Huso, and G.G. Germino. A critical developmental switch defines the kinetics of kidney cyst formation after loss of pkd1. *Nature medicine*, 13(12):1490, 2007.

[285] M. Pirooznia, J.Y. Yang, M.Q. Yang, and Y. Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, 9(1):S13, 2008.

[286] M. Poujade, E. Grasland-Mongrain, A. Hertzog, J. Jouanneau, P. Chavrier, B. Ladoux, A. Buguin, and P. Silberzan. Collective migration of an epithelial monolayer in response to a model wound. *Proceedings of the National Academy of Sciences*, 104(41):15988–15993, 2007.

[287] J. Prados, A. Kalousis, J.C. Sanchez, L. Allard, O. Carrette, and M. Hilario. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics*, 4(8):2320–2332, 2004.

[288] N.D. Price and I. Shmulevich. Biochemical and statistical network models for systems biology. *Current opinion in biotechnology*, 18(4):365–370, 2007.

[289] S. Qin, M. Taglienti, L. Cai, J. Zhou, and J.A. Kreidberg. c-met and nf-κb–dependent overexpression of wnt7a and-7b and pax2 promotes cystogenesis in polycystic kidney disease. *Journal of the American Society of Nephrology*, pages ASN–2011030277, 2012.

[290] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[291] J.R. Quinlan. C4. 5: Programming for machine learning. *Morgan Kauffmann*, 38, 1993.

[292] K. Ramasubbu, N. Gretz, and S. Bachmann. Increased epithelial cell proliferation and abnormal extracellular matrix in rat polycystic kidney disease. *Journal of the American Society of Nephrology*, 9(6):937–945, 1998.

[293] C. Reta, J. Gonzalez, R. Diaz, and J. Guichard. Leukocytes segmentation using markov random fields. In *Software Tools and Algorithms for Biological Systems*, pages 345–353. Springer, 2011.

[294] M.E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G.K. Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707, 2007.

[295] M. Rocco, E. Neilson, J. Hoyer, and F. Ziyadeh. Attenuated expression of epithelial cell adhesion molecules in murine polycystic kidney disease. *American Journal of Physiology-Renal Physiology*, 262(4):F679–F686, 1992.

[296] A.S. Rodin and E. Boerwinkle. Mining genetic epidemiology data with bayesian networks i: Bayesian networks and example application (plasma apoe levels). *Bioinformatics*, 21(15):3273–3278, 2005.

[297] S. Rodius, G. Androsova, L. Götz, R. Liechti, I. Crespo, S. Merz, P.V. Nazarov, N. de Klein, C. Jeanty, J.M. González-Rosa, et al. Analysis of the dynamic co-expression network of heart regeneration in the zebrafish. *Scientific reports*, 6:26822, 2016.

[298] S. Rossetti, M.B. Consugar, A.B. Chapman, V.E. Torres, L.M. Guay-Woodford, J.J. Grantham, W.M. Bennett, C.M. Meyers, D.L. Walker, K. Bae, et al. Comprehensive molecular diagnostics in autosomal dominant polycystic kidney disease. *Journal of the American Society of Nephrology*, 18(7): 2143–2160, 2007.

[299] P.C. Roth, D.C. Arnold, and B.P. Miller. Mrnet: A software-based multicast/reduction network for scalable tools. In *Supercomputing, 2003 ACM/IEEE Conference*, pages 21–21. IEEE, 2003.

[300] I. Rowe and A. Boletta. Defective metabolism in polycystic kidney disease: potential for therapy and open questions. *Nephrology Dialysis Transplantation*, 29(8):1480–1486, 2014.

[301] M.W.D. Rudi Rottenfusser, Erin E. Wilson. Education in microscopy and digital imaging. `http://zeiss-campus.magnet.fsu.edu/articles/basics/contrast.html`. [Online; accessed 24-September-2017].

[302] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

[303] G.S. Sadesky. Cluster analysis and its application in standard setting. In *Halifax, Nova Scotia: Paper Presented at the Annual Meeting of Canadian Society for the Study of Higher Education*, 2003.

[304] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[305] J. Saikumar, D. Hoffmann, T.M. Kim, V.R. Gonzalez, Q. Zhang, P.L. Goering, R.P. Brown, V. Bijol, P.J. Park, S.S. Waikar, et al. Expression, circulation, and excretion profile of microrna-21,-155, and-18a following acute kidney injury. *Toxicological Sciences*, 129(2):256–267, 2012.

[306] J.M. Sayagués, L.A. Corchete, M.L. Gutiérrez, M.E. Sarasquete, M. del Mar Abad, O. Bengoechea, E. Fermiñán, M.F. Anduaga, S. del Carmen, M. Iglesias, et al. Genomic characterization of liver metastases from colorectal cancer patients. *Oncotarget*, 7(45):72908, 2016.

[307] R.J. Schalkoff. *Artificial neural networks*, volume 1. McGraw-Hill New York, 1997.

[308] E. Schiffmann, B.A. Corcoran, and S.M. Wahl. N-formylmethionyl peptides as chemoattractants for leucocytes. *Proceedings of the National Academy of Sciences*, 72(3):1059–1062, 1975.

[309] B. Scholkopf and A.J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[310] A. Schützenmeister and H.P. Piepho. Background correction of two-colour cdna microarray data using spatial smoothing methods. *Theoretical and applied genetics*, 120(2):475, 2010.

[311] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[312] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535, 2008.

[313] H. Seppä, G. Grotendorst, S. Seppä, E. Schiffmann, and G.R. Martin. Platelet-derived growth factor in chemotactic for fibroblasts. *The Journal of Cell Biology*, 92(2):584–588, 1982.

[314] R. Setiono and H. Liu. Neural-network feature selector. *IEEE transactions on neural networks*, 8(3):654–662, 1997.

[315] J.P. Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46 (1):561–584, 1995.

[316] D. Shalon, S.J. Smith, and P.O. Brown. A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome research*, 6(7):639–645, 1996.

[317] S. Shibazaki, Z. Yu, S. Nishio, X. Tian, R.B. Thomson, M. Mitobe, A. Louvi, H. Velazquez, S. Ishibe, L.G. Cantley, et al. Cyst formation and activation of the extracellular regulated kinase pathway after kidney specific inactivation of pkd1. *Human molecular genetics*, 17(11):1505–1516, 2008.

[318] J.M. Shillingford, N.S. Murcia, C.H. Larson, S.H. Low, R. Hedgepeth, N. Brown, C.A. Flask, A.C. Novick, D.A. Goldfarb, A. Kramer-Zucker, et al. The mtor pathway is regulated by polycystin-1, and its inhibition reverses renal cystogenesis in polycystic kidney disease. *Proceedings of the National Academy of Sciences*, 103(14):5466–5471, 2006.

[319] J.M. Shillingford, K.B. Piontek, G.G. Germino, and T. Weimbs. Rapamycin ameliorates pkd resulting from conditional inactivation of pkd1. *Journal of the American Society of Nephrology*, 21(3):489–497, 2010.

[320] I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 2002.

[321] I. Shmulevich, E.R. Dougherty, and W. Zhang. Control of stationary behavior in probabilistic boolean networks by means of structural intervention. *Journal of Biological Systems*, 10(04):431–445, 2002.

[322] I. Shmulevich, E.R. Dougherty, and W. Zhang. Gene perturbation and intervention in probabilistic boolean networks. *Bioinformatics*, 18(10):1319–1331, 2002.

[323] L.H. Siew, R.M. Hodgson, and E.J. Wood. Texture measures for carpet wear assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(1):92–105, 1988.

[324] M. Silberberg, A.J. Charron, R. Bacallao, and A. Wandinger-Ness. Mispolarization of desmosomal proteins and altered intercellular adhesion in autosomal dominant polycystic kidney disease. *American Journal of Physiology-Renal Physiology*, 288(6):F1153–F1163, 2005.

[325] K.J. Simpson, L.M. Selfors, J. Bui, A. Reynolds, D. Leake, A. Khvorova, and J.S. Brugge. Identification of genes that regulate epithelial cell migration using an sirna screening approach. *Nature cell biology*, 10(9):1027, 2008.

[326] S.S.S. Sindhu, S. Geetha, and A. Kannan. Decision tree based light weight intrusion detection using a wrapper approach. *Expert Systems with applications*, 39(1):129–141, 2012.

[327] D. Smith and M.A. Geeves. Cooperative regulation of myosin-actin interactions by a continuous flexible chain ii: actin-tropomyosin-troponin and regulation by calcium. *Biophysical journal*, 84(5):3168–3180, 2003.

[328] T. Smith, W. Marks, G. Lange, W. Sheriff, and E. Neale. A fractal analysis of cell images. *Journal of neuroscience methods*, 27(2):173–180, 1989.

[329] P. Soille. *Morphological image analysis: principles and applications*. Springer Science & Business Media, 2013.

[330] R.L. Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19(12):1484–1491, 2003.

[331] X. Song, V. Di Giovanni, N. He, K. Wang, A. Ingram, N.D. Rosenblum, and Y. Pei. Systems biology of autosomal dominant polycystic kidney disease (adpkd): computational identification of gene expression pathways and integrated regulatory networks. *Human molecular genetics*, 18(13):2328–2343, 2009.

[332] C.W. Spraul, C. Kaven, J. Kampmeier, G.K. Lang, and G.E. Lang. Effect of thalidomide, octreotide, and prednisolone on the migration and proliferation of rpe cells in vitro. *Current eye research*, 19(6):483–490, 1999.

[333] F. Staal, M. van der Burg, L. Wessels, B. Barendregt, M. Baert, C. van den Burg, C. Van Huffel, A. Langerak, V. van der Velden, M. Reinders, et al. Dna microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-b acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia*, 17(7):1324, 2003.

[334] G. Stein, B. Chen, A.S. Wu, and K.A. Hua. Decision tree classifier for network intrusion detection with ga-based feature selection. In *Proceedings of the 43rd annual Southeast regional conference-Volume 2*, pages 136–141. ACM, 2005.

[335] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[336] J.D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[337] J.D. Storey et al. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

[338] A.J. Streets, L.J. Newby, M.J. O'Hare, N.O. Bukanov, O. Ibraghimov-Beskrovnaya, and A.C. Ong. Functional analysis of pkd1 transgenic lines reveals a direct role for polycystin-1 in mediating cell-cell adhesion. *Journal of the American Society of Nephrology*, 14(7):1804–1815, 2003.

[339] A. Stroope, B. Radtke, B. Huang, T. Masyuk, V. Torres, E. Ritman, and N. LaRusso. Hepato-renal pathology in pkd2 ws25/- mice, an animal model

of autosomal dominant polycystic kidney disease. *The American journal of pathology*, 176(3):1282–1291, 2010.

[340] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[341] E.A. Sudicky. A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in the dispersion process. *Water Resources Research*, 22(13):2069–2082, 1986.

[342] R. Suganya and S. Rajaram. Feature extraction and classification of ultrasound liver images using haralick texture-primitive features: Application of svm classifier. In *Recent Trends in Information Technology (ICRTIT), 2013 International Conference on*, pages 596–602. IEEE, 2013.

[343] S.V. Suggs, R.B. Wallace, T. Hirose, E.H. Kawashima, and K. Itakura. Use of synthetic oligonucleotides as hybridization probes: isolation of cloned cdna sequences for human beta 2-microglobulin. *Proceedings of the National Academy of Sciences*, 78(11):6613–6617, 1981.

[344] P. Suraneni, B. Rubinstein, J.R. Unruh, M. Durnin, D. Hanein, and R. Li. The arp2/3 complex is required for lamellipodia extension and directional fibroblast cell migration. *J Cell Biol*, pages jcb–201112113, 2012.

[345] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K.P. Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

[346] V. Takiar, S. Nishio, P. Seo-Mayer, J.D. King, H. Li, L. Zhang, A. Karihaloo, K.R. Hallows, S. Somlo, and M.J. Caplan. Activating amp-activated protein kinase (ampk) slows renal cystogenesis. *Proceedings of the National Academy of Sciences*, 108(6):2462–2467, 2011.

[347] W. Talloen, D.A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H.W. Göhlmann. I/ni-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, 23(21): 2897–2902, 2007.

[348] Y.C. Tan, J. Blumenfeld, and H. Rennert. Autosomal dominant polycystic kidney disease: genetics, mutations and microrna s. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1812(10):1202–1212, 2011.

[349] Y. Tao, J. Kim, S. Faubel, J.C. Wu, S.A. Falk, R.W. Schrier, and C.L. Edelstein. Caspase inhibition reduces tubular apoptosis and proliferation and slows disease progression in polycystic kidney disease. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19): 6954–6959, 2005.

[350] Y. Tao, J. Kim, R.W. Schrier, and C.L. Edelstein. Rapamycin markedly slows disease progression in a rat model of polycystic kidney disease. *Journal of the American Society of Nephrology*, 16(1):46–51, 2005.

[351] A.L. Tarca, R. Romero, and S. Draghici. Analysis of microarray experiments of gene expression profiling. *American journal of obstetrics and gynecology*, 195(2):373–388, 2006.

[352] P. Taylor. Fda turns down otsuka's kidney disease candidate. *http://www.pmlive.com*, Aug 2013.

[353] I. The MathWorks. Texture segmentation using texture filters. `https://uk.mathworks.com/help/images/examples/texture-segmentation-using-texture-filters.html`. [Online; accessed 24-September-2017].

[354] R. Thomas. Boolean formalization of genetic control circuits. *Journal of theoretical biology*, 42(3):563–585, 1973.

[355] R. Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83 (402):394–405, 1988.

[356] A.N. Tikhonov, V.I. Arsenin, and F. John. *Solutions of ill-posed problems*, volume 14. Winston Washington, DC, 1977.

[357] V.E. Torres, A.B. Chapman, O. Devuyst, R.T. Gansevoort, J.J. Grantham, E. Higashihara, R.D. Perrone, H.B. Krasa, J. Ouyang, and F.S. Czerwiec. Tolvaptan in patients with autosomal dominant polycystic kidney disease. *New England Journal of Medicine*, 367(25):2407–2418, 2012.

[358] V.E. Torres and M.L. Watson. Polycystic kidney disease: antiquity to the 20th century. *Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association-European Renal Association*, 13(10):2690–2696, 1998.

[359] G. Totsukawa, Y. Wu, Y. Sasaki, D.J. Hartshorne, Y. Yamakita, S. Yamashiro, and F. Matsumura. Distinct roles of mlck and rock in the regulation of membrane protrusions and focal adhesion dynamics during cell migration of fibroblasts. *The Journal of cell biology*, 164(3):427–439, 2004.

[360] T.F. Tracy Jr, A.J. Tector, M.E. Goerke, S. Kitchen, and D. Lagunoff. Somatostatin analogue (octreotide) inhibits bile duct epithelial cell proliferation and fibrosis after extrahepatic biliary obstruction. *The American journal of pathology*, 143(6):1574, 1993.

[361] Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, 99(22):14031–14036, 2002.

[362] J.W. Tukey. Exploratory data analysis. 1977.

[363] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[364] J.H. Uhm, N.P. Dooley, O. Stuve, G.S. Francis, P. Duquette, J.P. Antel, and V.W. Yong. Migratory behavior of lymphocytes isolated from multiple sclerosis patients: Effects of interferon β-1b therapy. *Annals of neurology*, 46 (3):319–324, 1999.

[365] A. Upadhyaya, J.P. Rieu, J.A. Glazier, and Y. Sawada. Anomalous diffusion and non-gaussian velocity distribution of hydra cells in cellular aggregates. *Physica A: Statistical Mechanics and its Applications*, 293(3):549–558, 2001.

[366] A. Uppuluri. Glcm texture features. *Matlab Central, The Mathworks..(accessed 22.03. 11)*, 2008.

[367] N.A. van Riel and E.D. Sontag. Parameter estimation in models combining signal transduction and metabolic pathways: the dependent input approach. *IEE Proceedings-Systems Biology*, 153(4):263–274, 2006.

[368] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM review*, 38 (1):49–95, 1996.

[369] R.S. Varga. *Geršgorin and his circles*, volume 36. Springer Science & Business Media, 2010.

[370] M. Vasa, S. Fichtlscherer, K. Adler, A. Aicher, H. Martin, A.M. Zeiher, and S. Dimmeler. Increase in circulating endothelial progenitor cells by statin therapy in patients with stable coronary artery disease. *Circulation*, 103(24): 2885–2890, 2001.

[371] M. Vasa, S. Fichtlscherer, A. Aicher, K. Adler, C. Urbich, H. Martin, A.M. Zeiher, and S. Dimmeler. Number and migratory activity of circulating endothelial progenitor cells inversely correlate with risk factors for coronary artery disease. *Circulation research*, 89(1):e1–e7, 2001.

[372] F. Verdeguer, S. Le Corre, E. Fischer, C. Callens, S. Garbay, A. Doyen, P. Igarashi, F. Terzi, and M. Pontoglio. A mitotic transcriptional switch in polycystic kidney disease. *Nature medicine*, 16(1):106–110, 2010.

[373] N. Vijesh, S.K. Chakrabarti, and J. Sreekumar. Modeling of gene regulatory networks: A review. *Journal of Biomedical Science and Engineering*, 6(02): 223, 2013.

[374] M. Walter, K.T. Wright, H. Fuller, S. MacNeil, and W.E.B. Johnson. Mesenchymal stem cell-conditioned medium accelerates skin wound healing: an in vitro study of fibroblast and keratinocyte scratch assays. *Experimental cell research*, 316(7):1271–1281, 2010.

[375] S.J. Wang, W. Saadi, F. Lin, C.M.C. Nguyen, and N.L. Jeon. Differential effects of egf gradient profiles on mda-mb-231 breast cancer cell chemotaxis. *Experimental cell research*, 300(1):180–189, 2004.

[376] M.F. Ware, A. Wells, and D.A. Lauffenburger. Epidermal growth factor alters fibroblast migration speed and directional persistence reciprocally and

in a matrix-dependent manner. *Journal of Cell Science*, 111(16):2423–2432, 1998.

[377] F. Wei, A. Karihaloo, Z. Yu, A. Marlier, P. Seth, S. Shibazaki, T. Wang, S. Somlo, L.G. Cantley, and V.P. Sukhatme. Neutrophil gelatinase-associated lipocalin suppresses cyst growth by pkd1 null cells in vitro and in vivo. *Kidney international*, 74(10):1310–1318, 2008.

[378] B.L. Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[379] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In *Advances in neural information processing systems*, pages 668–674, 2001.

[380] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck, and W. Jacob. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry*, 33(1):32–40, 1998.

[381] A. Wibmer, H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr, J. Zheng, D. Goldman, C. Moskowitz, S.W. Fine, et al. Haralick texture analysis of prostate mri: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different gleason scores. *European radiology*, 25(10):2840–2850, 2015.

[382] M.A.B.C. William M Bennett, Frederic F Rahbari-Oskoui. Patient education: Polycystic kidney disease (beyond the basics). *http://www.uptodate.com*, Feb 2016.

[383] P.D. Wilson. Polycystic kidney disease. *New England Journal of Medicine*, 350(2):151–164, 2004.

[384] S.J. Wilson, K. Amsler, D.P. Hyink, X. Li, W. Lu, J. Zhou, C.R. Burrow, and P.D. Wilson. Inhibition of her-2 (neu/erbb2) restores normal function and structure to polycystic kidney disease (pkd) epithelia. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1762(7):647–655, 2006.

[385] B. Wojciak-Stothard and A.J. Ridley. Shear stress–induced endothelial cell polarization is mediated by rho and rac but not cdc42 or pi 3-kinases. *J Cell Biol*, 161(2):429–439, 2003.

[386] D. Woo. Apoptosis and loss of renal tissue in polycystic kidney diseases. *New England Journal of Medicine*, 333(1):18–25, 1995.

[387] F.X. Wu, L.Z. Liu, and Z.H. Xia. Identification of gene regulatory networks from time course gene expression data. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 795–798. IEEE, 2010.

[388] S. Wu, Z.P. Liu, X. Qiu, and H. Wu. Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. *PloS one*, 9(5):e95276, 2014.

[389] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, S.Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[390] Y. Wu, J.X. Xu, W. El-Jouni, T. Lu, S. Li, Q. Wang, M. Tran, W. Yu, M. Wu, I.E. Barrera, et al. Gα12 is required for renal cystogenesis induced by pkd1 inactivation. *J Cell Sci*, 129(19):3675–3684, 2016.

[391] Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American statistical Association*, 99(468):909–917, 2004.

[392] C.C. Xiang and Y. Chen. cdna microarray technology and its applications. *Biotechnology Advances*, 18(1):35–46, 2000.

[393] Y. Xiao. A tutorial on analysis and simulation of boolean gene regulatory network models. *Current genomics*, 10(7):511–525, 2009.

[394] T. Yamaguchi, S. Nagao, D.P. Wallace, F.A. Belibi, B.D. Cowley, J.C. Pelling, and J.J. Grantham. Cyclic amp activates b-raf and erk in cyst epithelial cells from autosomal-dominant polycystic kidneys. *Kidney international*, 63(6):1983–1994, 2003.

[395] G. Yao, X. Su, V. Nguyen, K. Roberts, X. Li, A. Takakura, M. Plomann, and J. Zhou. Polycystin-1 regulates actin cytoskeleton organization and directional cell migration through a novel pc1-pacsin 2-n-wasp complex. *Human molecular genetics*, 23(10):2769–2779, 2014.

[396] C. Yeaman, K.K. Grindstaff, and W.J. Nelson. New perspectives on mechanisms involved in generating epithelial cell polarity. *Physiological reviews*, 79(1):73–98, 1999.

[397] W. Yin, T. Chen, S.X. Zhou, and A. Chakraborty. Background correction for cdna microarray images using the tv+ l 1 model. *Bioinformatics*, 21(10): 2410–2416, 2005.

[398] P. Yip and K. Rao. A fast computational algorithm for the discrete sine transform. *IEEE Transactions on Communications*, 28(2):304–307, 1980.

[399] B.K. Yoder. Role of primary cilia in the pathogenesis of polycystic kidney disease. *Journal of the American Society of Nephrology*, 18(5):1381–1388, 2007.

[400] J. Yu, S. Ongarello, R. Fiedler, X. Chen, G. Toffolo, C. Cobelli, and Z. Trajanoski. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics*, 21(10):2200–2209, 2005.

[401] M.M. Zavlanos, A.A. Julius, S.P. Boyd, and G.J. Pappas. Inferring stable genetic networks from steady-state data. *Automatica*, 47(6):1113–1122, 2011.

[402] N. Zayed and H.A. Elnemr. Statistical analysis of haralick texture features to discriminate lung abnormalities. *Journal of Biomedical Imaging*, 2015:12, 2015.

[403] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[404] B. Zhang, C. Gaiteri, L.G. Bodea, Z. Wang, J. McElwee, A.A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, et al. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707–720, 2013.

[405] L. Zhang, L. Wang, P. Tian, and S. Tian. Pathway-based feature selection algorithms identify genes discriminating patients with multiple sclerosis apart from controls. *arXiv preprint arXiv:1508.01509*, 2015.

[406] J. Zhu, S. Rosset, R. Tibshirani, and T.J. Hastie. 1-norm support vector machines. In *Advances in neural information processing systems*, pages 49–56, 2004.

[407] D. Zicha, G.A. Dunn, and A.F. Brown. A new direct-viewing chemotaxis chamber. *Journal of cell science*, 99(4):769–775, 1991.

[408] S.H. Zigmond, J.L. Slonczewski, M.W. Wilde, and M. Carson. Polymorphonuclear leukocyte locomotion is insensitive to lowered cytoplasmic calcium levels. *Cytoskeleton*, 9(2):184–189, 1988.