Copy number alterations and primary melanoma survival

Joey Mark Santiago Diaz

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Medicine

August 2020

This candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 641458.

The collection of samples in the Melanoma Cohort Study was funded by Cancer Research UK (project grant C8216/A6129, Programme awards C588/A4994, C588/A10589 and C588/A19167) and Centre Award (C37059/A11941) and by the NIH (R01 CA83115).

I would like to thank the MELGEN consortium for having selected me as one of the Marie Skłodowska-Curie Early Stage Researchers and for giving me the opportunity to pursue this PhD training in the University of Leeds.

I would like to express my sincere gratitude to my supervisors Prof. David Timothy Bishop for and Prof. Julia Newton-Bishop for their unending support and guidance from the start to the end of my PhD. Tim has been very patient in helping me understand statistical genomics concepts that are not familiar to me. He is always willing to help and listens to my ideas and complement them with his expertise in the field. It was a great pleasure to be mentored by him and I will treasure this for the rest of my life. Julia has been helpful to me making sure that I am on the right track. She personally gave me lectures on the basics of melanoma appreciating that I came from a non-biology background. She has provided significant input to my writings and helped me understand a lot of concepts about melanoma.

I am also grateful to university administration staff for their round the clock assistance.

Regular feedback of progress was provided by other members of the Section of Epidemiology and Biostatistics as part of weekly lab-meetings. For this I thank the following:

Dr. Alastair Droop and Dr. Anastasia Filia for all their initial work on this copy number project. Anastasia did the wet lab works to generate the copy number data while Alastair did the dry lab works to visualise and better understand the copy number data. He created the *mcnv* R package that has been very useful for me in visualising and summarising the copy number data. He also provided me with technical help and guidance during the quality control analysis and generation of the new copy number data during and after applying additional quality control steps.

Dr. Juliette Randerson-Moor for all her extra hard work for MELGEN submissions and requirements, for training us and giving us advice on thesis writing, and personally helping me with formatting my thesis.

Dr. Mark Harland for his help and guidance on a lot of copy number analysis and data preparation that I have done all throughout my PhD, especially in the analysis of the *CDKN2A* region, and for helping me with writing some of my conference abstracts.

Ms. May Chan for providing me with clinical datasets I needed for my study, and sometimes for offering to drive me home after work.

Dr. Jérémie Nsengimana for his help in the analysis of transcriptomic data, as well as with his insights on my presentations during the lab meetings.

Dr. Jon Laye for allowing to use some of his images of stained tumour cores, for the helpful advice and comments regarding my copy number works.

Ms. Katie Cairns for helping me with the admin work during my transition period in Leeds, and for processing transactions related to our academic travels for trainings and conferences.

Dr. Jon Davies for his help with clarifying some of the confusions I had with the dataset I am analysing, and for providing feedback to my work and presentation as well

Dr. Sally O'Shea for her encouraging words and the things I learned on her work about quantifying stroma in relation to tumour purity.

Same appreciation and thanks go to Tracy Mell and Faye Elliot for lunch conversations, and their overall contribution to our lab team.

Mr. Geoff Cross for helping me with assembling my work laptop, setting wifi and printer connection, and all the Mac related stuff where I needed help.

Mr. Martin Callaghan for all the help with HPC, especially for providing personal assistance with the installation of software I needed for my data analysis.

I am grateful to the Jönsson Lab in Lund, Sweden for my secondment where I learned about GISTIC which has been very useful in this study.

I am grateful for the friendship with my MELGEN friends and buddies. Sofia Chen for helping me get around Leiden and Amsterdam during the Melgen interview, during Leiden training, during MELGEN reunion meetings and conferences. My team mates in Essen: Reni and Sonia for being my core support system during my stay in Essen. Reni has been a very supportive colleague and friend, I appreciate all our Facebook calls and catch ups, your advice and your wisdom. I look forward to the realisation of all the plans we discussed together.

To our friends in Lund and Leiden, Buddy Shamik, Adriana, Rita, Catarina, and Eirini for our side trips and bonding moments together.

To Ishani, Kamin, Marina and Adam, for being great companion especially during conferences.

Sathya and Joanna for helping me understand different biological concepts, and for the more than three years of our time together and being supportive colleagues in the lab.

To Rohit Thakur, for helping me settle in Leeds, especially with finding my first accommodation, and making sure I have what I need to start in my new home, for always being a good companion, an encourager, and a reliable brother and best friend.

The Amarga-Weisner family: Ate Ann, Kuya Markus, Tristan, and Mamang for welcoming me in their household for dinners, catch ups, tagging me along in their trips, and making me feel always welcome in their family. It made so much difference during my stay in Essen.

Tita Normie for taking care of me during my stay in her flat and for encouraging me whenever I feel down. I always feel your sincerity and genuine care for me, thank you tita. Same goes to Tita Yulan, Jeffie, Ate Belle, Tita Joy, and my Leeds LG mates led by Tita Fe and Kuya Henry, and to my London LG mates led by Harry and Jen.

I thank my friends who I always hang out with in the Swan with Two Necks, Anjo Lapitan for being a brother and friend, Gesner for his thoughts and wisdom; as well as Fanis, Edward, Ron, Nina for our great times together, and for being such good friends and companion.

To the Filipino Society in Leeds University, and to my FRIENDS: Kathrina, Wik, Francis, Rj, Paul, and Nathan, thank you for the friendship and making me feel that I have a family in Leeds. I cherish all the bonding moments we have together. Same goes to Dana and ate Roda for the friendship and companion during my visits in London.

To my Ambizyosa family: Erika, Estef, Zabeth, Janet, Jessica, Rachel, Jennielyn, and Jeffson thank you always being there to support me.

To Dr. Aldrin Sales, Ate Jeru Abella, Camille Gonzales, and Jossel Escabarte: thank you for being there always to listen and give advice. Especially Ate Jeru for the catch ups we are regularly having, this helped me a lot to cope up with thesis life on a pandemic.

To my new friends in London, Hamza, Charlotte, Phoebe, and Joey Fang thank you for the fun moments we had and are about to have together from ping-pong games, long walks, dinners, and chill evenings or weekends.

To Tita Jocelyn Vendivil and family, to my aunts, uncles, and cousins, thank you all for your love and support.

To my siblings, Kuya Max, Dikong Ardee and Ate Miriam, Lyndon, and Allen and to my nice and nephews Jane, Jed and Johann, to lola Aning, and Tita Perlie, my life becomes extra brighter everytime I think of you all.

To my parents Marissa and Rodulfo: Nanay and Tatay, thank you for bringing me into this world. Thank you for working hard for me and my siblings, for raising us, loving us and supporting us. **Nanay**, this is for you, I love you. I hope you are always smiling in heaven.

And above all, I thank the almighty God for the wisdom and strength, and for carrying me through towards the success of this study.

Abstract

Our previous work reported a large-scale copy number (CN) study of primary melanoma [1]. Next generation sequencing (NGS) data from 303 formalin fixed paraffin embedded (FFPE) samples from the Leeds Melanoma Cohort (LMC) were generated. Libraries were generated by random shearing and then sequenced (1.7x coverage). In this study, problematic regions and common germline variations in the genome were identified and excluded accounting to approximately 13.5 % of the autosomal genome. CN was generated by read count accumulated into 10k bp windows, adjusted jointly for sequence mappability and GC-content and in comparison, with Caucasian genomes (n=312) from the 10k Genome Project [2, 3]. *GISTIC 2.0.23* identified significantly deleted or amplified regions in the genome [4]. Comparisons with the TCGA data showed high similarity between the two datasets in terms of location and proportion of samples with CN changes [5, 6]. Minor difference in terms of frequency which may be due to the type of samples processed (frozen vs FFPE), platform used (NGS vs SNPS Array), or disease stage (primary vs metastatic) offer opportunity for discovery of novel CNA in melanoma

Three measures of overall genome instability namely Fraction of Genome Altered (FGA), Aneuploidy Score (AS), and a devised metric I referred as Mean Weighted Segment Mean (MWSM) were estimated. MWSM showed strongest association to most of the patient clinical characteristics and survival among the three measures. Focal analysis was done using each 10k window level copy number data which allowed detection of small copy number aberrations that are associated with patient clinical characteristics and survival

In conclusion, this study showed the feasibility of extracting and analysing whole genome copy number data from FFPE samples given that satisfactory amount of quality control steps to improve data quality is done. While there were interesting associations identified between copy number and patient clinical and tumour characteristics, validation of these results once similar population-based study cohort becomes available.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABHD16B	Abhydrolase Domain Containing 16B
aCGH	array comparative genome hybridization
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
APOBEC3B	apolipoprotein B MRNA Editing Enzyme Catalytic Subunit 3B
AS	**de novo assembly**
B2M	Beta-2-Microglobulin
BHLHA9	Basic Helix-Loop-Helix Family Member A9
C9ORF53	Chromosome 9 Open Reading Frame 53, also called CDKN2A Antisense RNA 1
CBS	circular binary segmentation
CCND1	Cyclin D1
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
CDKN2B	Cyclin Dependent Kinase Inhibitor 2B
CGH	comparative genome hybridization
CMM	cutaneous malignant melanoma
CNV	copy number variation
CYTH3	Cytohesin 3
DBIL5P	Diazepam Binding Inhibitor-Like 5, Pseudogene
DNA	deoxyribonucleic acid
DPEP1	Dipeptidase 1 (Renal)
ESV	European Bioinformatics Institute (EBI) Structural Variation
ERG	Avian v-ets erythroblastosis virus E26 oncogene
FAM157C	Family with Sequence Similarity 157 Member C
FFPE	formalin fixed paraffin embedded
FGF3	Fibroblast Growth Factor 3
FYN	Proto-oncogene tyrosine-protein kinase Fyn
HULC	Hepatocellular Carcinoma Up-Regulated Long Non-Coding RNA
KCNN3	Potassium Calcium-Activated Channel Subfamily N Member 3
LMC	Leeds Melanoma Cohort
MC1R	Melanocortin 1 Receptor

MYC	V-Myc Avian Myelocytomatosis Viral Oncogene Homolog
NGS	next generation sequencing
NF1	Neurofibromin 1
PARK2	Parkin RBR E3 Ubiquitin Protein Ligase
PCMTD2	Protein-L-Isoaspartate (D-Aspartate) O-Methyltransferase Domain Containing 2
PHIP	Pleckstrin Homology Domain Interacting Protein
PTP4A1	Protein Tyrosine Phosphatase Type IVA, Member 1
PVT1	Pvt1 Oncogene (Non-Protein Coding)
RC	**read count**
RP	**read-pair**
RPL5	Ribosomal Protein L5
RPTOR	Regulatory Associated Protein Of MTOR Complex 1
SCM	Skin cutaneous melanoma
SMYD3	SET And MYND Domain Containing 3
SNORA66	Small Nucleolar RNA, H/ACA Box 66
SNORD2	Small Nucleolar RNA, C/D Box 2
SNORD21	Small Nucleolar RNA, C/D Box 21
SR	split read
SV	structural variants
TCGA	The Cancer Genome Atlas
TERT	Telomerase Reverse Transcriptase
UTF1	Undifferentiated Embryonic Cell Transcription Factor 1
UV	Ultraviolet-rays

Chapter 1
Introduction

In this chapter, I provided the introduction to my study by defining and explaining important terms and concepts such as the skin, its parts, melanoma, the epidemiology of melanoma, causes and risks of melanoma, melanoma staging using the AJCC stage, melanoma survival, cancer development, assessment methods for the detection of copy number alterations as well as the test and reference samples in this study as supported by various literature.

1.1 The Human Skin

The skin is the largest organ in the human body in terms of surface area and weight. It measures about 20 squared feet and weighs about 20 pounds. It functions mainly for protection, sensation, and regulation [7-9]. The skin provides our body with protection against mechanical impacts and pressure, fluctuations in the temperature, microbes, and other harmful external factors such as radiation and chemicals. It regulates the body temperature through sweat and hair, the variation in peripheral circulation and fluid stability via sweat. Additionally, it acts as a venue for the synthesis of Vitamin D. In terms of sensation, the skin has enormous networks of nerves that perceives and communicate changes in the environment. The skin has distinct receptors for heat, cold, touch, and pain [10]. The skin has two major layers namely: epidermis and dermis, and a third closely associated layer called hypodermis or subcutaneous fat (See Figure 1.1).

1.1.1 Epidermis

The epidermis is made up of keratinized, stratified squamous epithelium (See Figure 1.2). Depending on where it is located in the body, the epidermis consists of four or five layers of epithelial cells and no blood vessels can be found in it (avascular). Most of the skin has four layers and is called a “thin skin” composed of stratum basale, stratum spinosum, stratum granulosum, and stratum corneum. “Thick skin” is located only on the soles of the feet and palms of the hands. It has an additional later between the stratum corneum and stratum granulosum called the stratum lucidum [11].

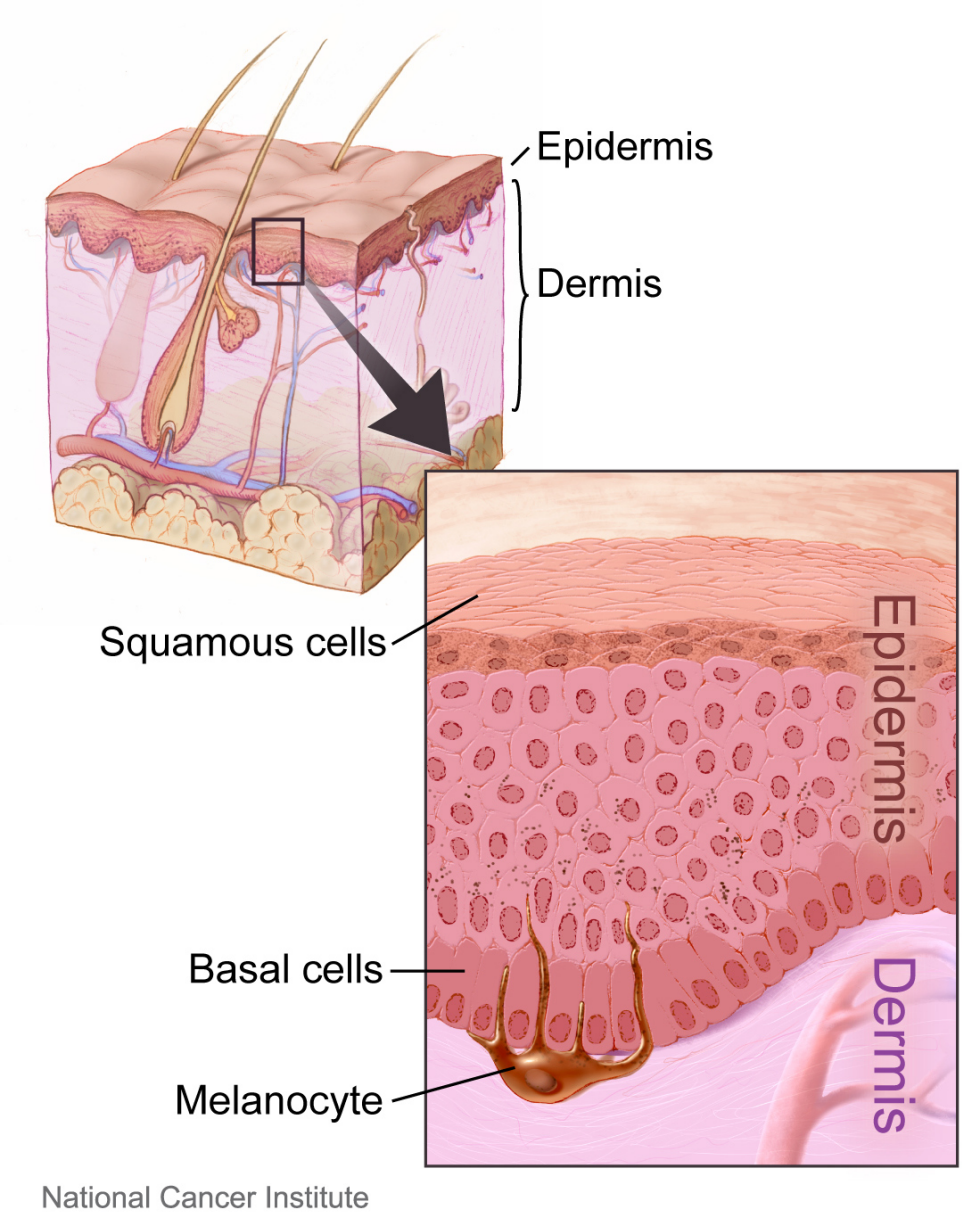


Figure 1.1: Major Layers of the Skin

Taken from National Cancer Institute [12]

The most external layer of the epidermis is the stratum corneum. Its name is derived by the increased keratinization or cornification of the cells in this layer and generally composed of 15 to 30 layers of cells. It is a dry and dead layer that helps prevent liquid loss from underlaying tissues and microbes from entering the skin. It also provides shield against mechanical pressure and abrasion for the softer underlying layers [11].

Stratum lucidum is the epidermal layer next to stratum corneum. It is present only in thick skins of the palms, soles, and digits and composed of thin layer of cells. It is made of keratinocytes that are dead and compacted containing dense amount of eleidin.

Derived from keratohyalin, eleidin is a clear protein rich lipids that gives transparent appearance to the cells and provides shield to water [11].

Stratum granulosom is the third layer of epidermis for thick skin and second layer for the thin skin. This is derived from the keratinocytes that undergo further changes making it grainy in appearance as they are pushed from the stratum spinosum – a deeper layer of the epidermis. It is composed of cells that are 3 to 5 layers that become flatter and their cell membranes thicken. This layer produces huge amounts of fibrous proteins called keratin and keratohyalin which builds up to form lamellar granules within the cells. Keratin and keratohyalin compose the majority of the keratinocyte build up in the stratum granulosum and provide the layer its grainy appearance. The nuclei and the cell components deplete as the cells die but the keratin, keratohyalin, and cell membranes remain and forms the stratum lucidum, and the stratum corneum. The external features of hair and nails are produced in this similar process that involves producing cells full of keratin [11].

Stratum spinosum is the spiny layer of the epidermis because it is formed by protruding cell processes that sticks the cells together using a structure called desmosome – which interlocks with each other to provide strong bond between the cells. The spiny nature of this layer is an indirect product of the staining process and the unstained epidermis does not have this feature. It has 8 to 10 layers of keratinocytes resulted by the cell division in the stratum basale. A type of Dendritic cells called Langerhans cells that functions as macrophage by engulfing foreign materials including bacteria and damaged cells are dispersed among the keratinocytes of this layer. Keratin production is initiated by the keratinocytes in the Stratum spinosum which also releases a water -repelling glycolipid that restrains water loss from the body and causes the skin to be somewhat waterproof. The keratinocytes of the stratum spinosum are pushed to the stratum granulosum as the new keratinocytes are produces on the top of the stratum basale [11].

The deepest epidermal layer is called the stratum basale or stratum germinativum which attaches the epidermis to the basal lamina, under which the layers of dermis can be found. Cells located in the stratum basale attach to the dermis via the basement membrane – which are intertwining collagen fibers. Dermal papilla – a finger like fold or formation located on the topmost part of the dermis, increase the strength of the connection between epidermis and dermis. Greater folding of dermal papilla corresponds to greater strength of connections made. The stratum basale is characterised by one layer of cells predominantly composed of basal cells. Basal cells are cuboidal-shaped stem cells that gives rise to the keratinocytes of the epidermis. This epidermal layer also contains other cells like Merkel cells and melanocytes which

are distributed together with the basal cells. The Merkel cells functions as receptors and plays role in stimulating sensory nerves that is perceived by the brain as touch while melanocytes manufactures the melanin – a pigment that supplies colour to hair and skin as well as provides protection to the living cells of the epidermal layer from being damaged by the exposure to ultraviolet radiation [11].

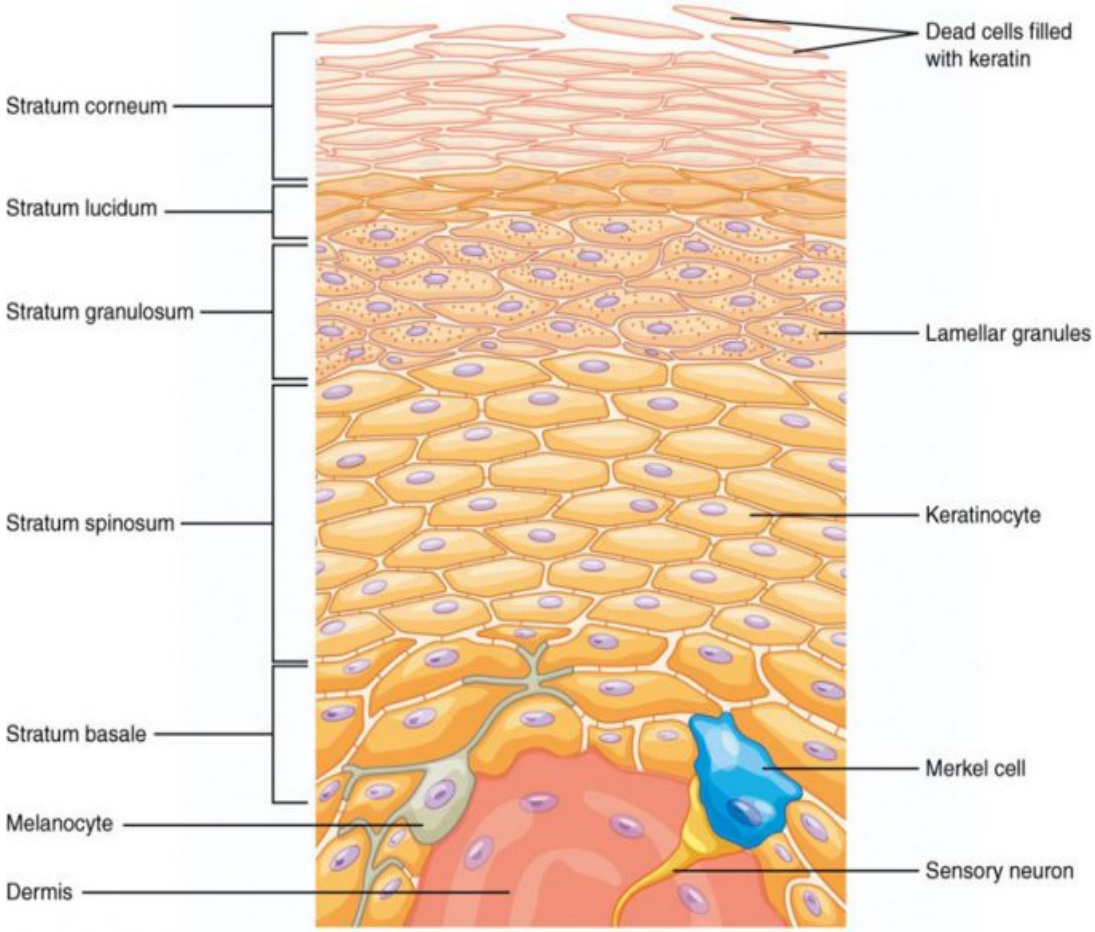


Figure 1.2. Layers of Epidermis

Taken from Oregon State University [11]

1.1.2 Dermis

The second major layer of the skin is called Dermis. This layer contains nerves, blood, lymph vessels, hair follicles, sweat glands and others structures. Capillaries in the dermis supplies oxygen and nutrients to the epidermis. It is made up of two layers of connective tissue that makes up interconnected nets of elastin and collagenous fibres, manufactured by fibroblasts. The shallowest layer in called papillary layer and provides anchorage for the epidermis on its top and is strongly connected to the deeper reticular later.

The papillary layer is composed of loose, areolar, connective tissue. The collagen and elastin fibres of this layer form a loose net with huge amount of ground substance supporting the hydration of the skin. This most external layer of the dermis follows into the epidermis' stratum basale to form papillae with the shape of a fingers. Located on the papillary layer are plenty of small blood vessels, fibroblasts and few fat cells. Additionally, it contains lymphatic capillaries, nerve fibres, touch receptors called Meissner corpuscles, and defensive cells called phagocytes that combats microbes that have entered the skin [11].

The net-like (reticulated) reticular layer is under the papillary layer and consists of compact irregular connective tissue that stops forces in many various directions allowing the skin to be flexible. It is about 80% of the dermis and has plenty of blood vessels and supply of sympathetic and sensory nerves. Elasticity that allows the skin movement is enabled by the elastin fibres while structure and tensile strength are provided by the collagen fibres [11].

1.1.3 Subcutaneous Fat

The layer below the dermis is called hypodermis or subcutaneous layer or superficial fascia acts to join the skin to the underlying fibrous tissues (fascia) surrounding the muscles. This layer is not strictly part of the skin and the boundary between dermis and hypodermis is very difficult to identify. It is composed of well vascularised, loose areolar connective tissue and plenty of adipose tissue acting as fat storage and insulator as well as providing cushioning for the integument. A thick connective tissue coating around skeletal muscles is called fascia [11].

1.1.4 Cells Present in the Skin

There are different types of cells found in the skin. Each type of cell plays role in helping the skin achieve its functions. There major cells present in the skin are

keratinocytes, melanocytes and immune cells. Described below are the different characteristics of each cell type and their functions.

1.1.4.1 Keratinocytes

Keratinocytes (also called prickle cells or skin cells) are the most common cells and make up about 90% of the cells in the epidermis and can be found in each epidermal layer. They are highly specialised epithelial cells that specifically functions to separate an organism from its surroundings. Cells produces precursors and build them into two different structures – the cornified envelope and the keratin intermediate filaments. The cornified envelope is built primarily from involucrin – a highly reactive, soluble protein while intermediate filaments are built from keratin monomers. Keratinocytes become more differentiated and thickens with keratin then eventually wear off [13-15].

1.1.4.2 Melanocytes

Melanocytes are pigment producing cells responsible for making melanin and can be found not only in the epidermis but also in other parts of the human body like hair follicles, mucous membrane, cochlea of the ear, iris of the eye as well as in the mesencephalon of the brain. Skin colour is determined by different pigments such as haemoglobin (red), hemosiderin (brown), carotene (yellow), bilin (yellow), and melanin. Melanin produced by melanocytes can be classified as eumelanin or pheomelanin [16]. Eumelanin is responsible for giving dark brown and black pigmentation of the skin while pheomelanin is responsible for giving yellow, red, and light brown skin pigmentation [17-19]. Melanin pigments are manufactured, stored, and transported through an called melanosome, an organelle synthesised by melanocyte [20].

1.1.4.3 Immune cells

The immune cells in the skin, famously known as Langerhans cells – discovered by a medical doctor Paul Langerhans in 1868, are irregularly shaped dendritic cells found in the stratum spinosum [21, 22] . Langerhans cells embodies the most peripheral basis of the immune system [23]. These are antigen presenting cells that play role in enabling allergic reactions and do not have keratin filament of melanosomes [13].

Langerhans cells gains antigens in peripheral tissues, bring them to the regional lymph nodes, present to naive T cells and promotes adaptive immune response. These cells are strongly immunogenic but may also act as mediators of tolerance as in the case for commensal bacteria. Other functions of these cells are involvement with antimicrobial immunity, immunosurveillance of the skin, induction phase of the contact hypersensitivity as well as in the pathogenesis of chronic inflammatory diseases of the mucosa or the skin[24].

1.2 Melanoma

Cutaneous melanoma (henceforth referred to as melanoma) is a cancer of the melanocyte - the pigment producing cells in the skin. Melanoma is significantly less common than other more harmless forms of skin cancer such as basal cell carcinomas or squamous cell carcinomas, but approximately 20% results in spread through the body (metastasis).

Melanomas are usually visually characterized by a dark, mole-like neoplasm in the skin that grows progressively in size and is irregular in shape but there are clinically described subtypes. The commonest in pale skinned populations is the superficial spreading melanoma, the second commonest nodular melanoma: lentigo maligna melanoma and acral lentiginous melanoma are the least common form [25].

These are categories defined by their appearance to the pathologist examining the tumour section under the microscope but have their clinical correlates when still on the skin. All but nodular melanomas begin in situ and initially grow laterally in the epidermis (the scaly epithelium of the skin, composed of predominantly keratinocytes) and sometimes penetrate deeper into the lower layers of the skin.

The last type - nodular melanoma is invasive but it is more serious as they have penetrated deeper into the skin. Superficial spreading melanoma is the most prevalent type, starting most commonly on the legs for women and on the chest and back for men. This type of melanoma usually initiates growth slowly and spreads out on the surface of the skin.

Figure 1.3 displays the progression of superficial spreading melanoma down the microscope, illustrating the progression of a tumour from single malignant cells in the epidermis to clusters of melanoma cells in the epidermis then clusters of melanoma cells thriving in the dermis [26]. This starts with a benign nevus or a mole that appears are acquired at birth, during childhood or adolescence and may be caused by sun exposure [27, 28].These moles are benign local proliferation of pigment cells (melanocytes) and may be blackish or brownish in colour [29]. Few of these cells further develop and give rise to asymmetric dysplastic naevus or atypical naevus. While most of this further growth dies, some continues to grow into two growth phases namely RGP (radial growth phase) and VGP (vertical growth phase) which are the established malignant melanoma. RGP is characterised by the invasion of the cells in the outer layer of the dermis without forming a node and is predominantly flat. VGP is characterised by a vertical growth of lesion that forms a true tumour and may extend deeper into the tissues. These are invasive and may indicate high likelihood of metastasis [30-32].

The clinical appearance of a superficial spreading melanoma is shown below in Figure 1.4. It has usual characteristics of this type of melanoma being asymmetrical, irregularly pigmented and a somewhat irregular edge.

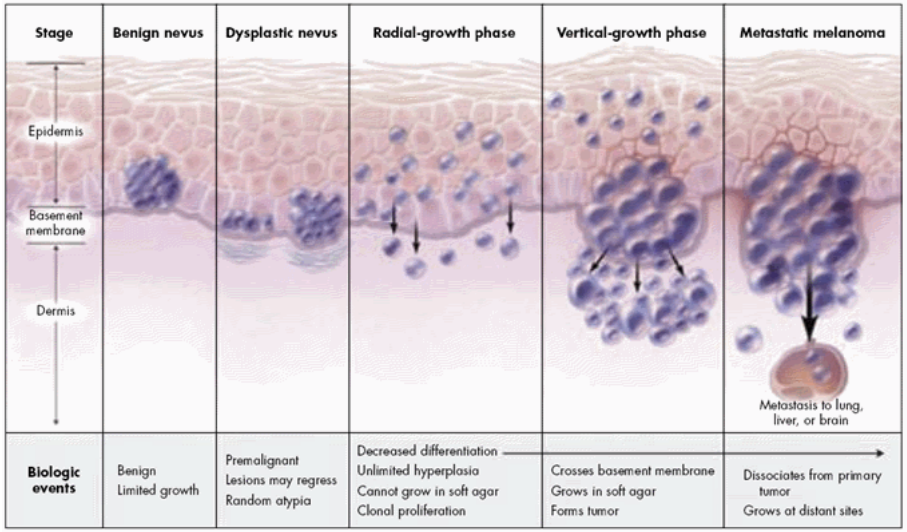


Figure 1.3. Progression of superficial spreading melanoma

Figure 1.4: A superficial spreading melanoma

Nodular melanoma grows quicker than SSM in that these tumours appear to have no precursor phase: the growth appears to be invasive from the beginning. Nodular melanomas are usually seen on the back, head, chest or neck [33].

Lentigo maligna melanoma is usually diagnosed in older people on parts of the skin that are more exposed to sun over many years. This type is usually therefore seen on the neck and face. It is characterised by a long in situ phase when the lesion looks like a dark freckle which slowly increases in size. Over time, the malignant cells proliferate within the epidermis and finally those cells acquire genomic changes sufficient to grow within the dermis i.e. progression occurs to an invasive tumour (Figure 1.5) [34].

Acral melanoma (Figure 1.6) is the rarest type and is usually found on the soles of the feet, palms of the hand, under fingernails or toenails and is more common in populations with darker complexion [25, 35].

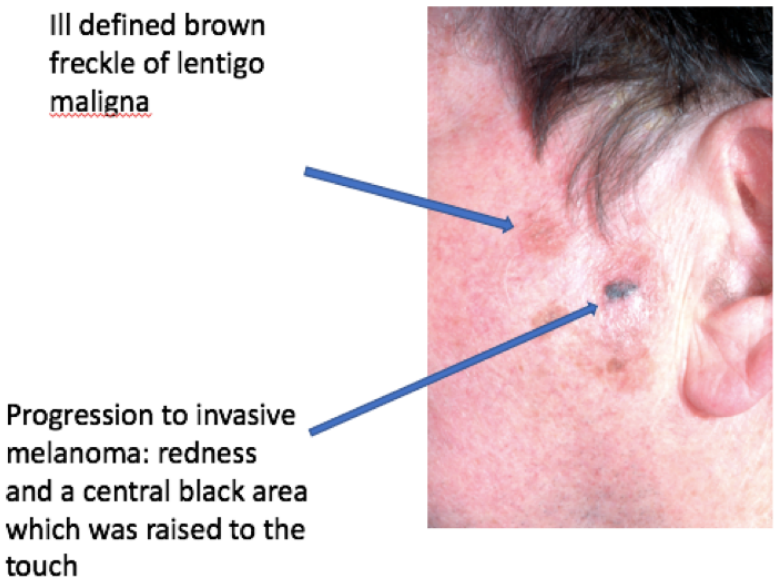


Figure 1.5 . Lentigo melanoma

Figure 1.6. Acral Melanoma

1.3 Epidemiology of Melanoma

Melanoma is an increasingly common form of cancer with increasing incidence attributed to distinct patterns of sun exposure among the genetically susceptible. In 2018, melanoma of the skin was listed (19th) as one of the 20 most common cancers (excluding non-melanoma skin cancer) in the world with 287,000 new cases accounting to 1.7% for all new cancer cases for the year. This ranks as the 15th most common cancer for men with 151,000 new cases accounting for 1.7% of all cancers in men and also the 15th most common cancer in women with 137,000 new cases accounting for 1.7 % of all cancers in women worldwide [36] .

In Europe, a rise in the incidence of melanoma have been observed over the past decades [37]. The estimated age-adjusted new cases of melanoma for 2012 were 11.4 per 100,000 p.a. for males and 11.0 per 100,000 p.a. for females with the lowest incidence rates in Central and Eastern Europe and up to 19 cases per 100,000 p.a. for Northern Europe.

In the UK, melanoma is the fifth most common cancer and accounts for 4% of all new cancer cases in 2015. Three-year UK statistics (2013-2014) indicated 15,419 new melanoma cases (See Table 1.1) with a ratio of 1:1 for both sexes [38]. The incidence of melanoma skin cancer is associated with age with the highest incidence rates in older people. Around 27% of new cases were in people aged 75 years and above. More than 30 % of skin cancer are diagnosed in patients under 50 years. This age at diagnosis is uncommonly early in comparison with many other types of common cancer and therefore melanoma accounts for a significant proportion of years of life lost to cancer. Age specific incidence rates constantly increases from the age 20-24 years and the highest rates are in the age of 90 years and above for males and 85 to 89 years for female (Figure 1.7).

Table 1.1: Melanoma Skin Cancer Number of New Cases, Crude and European Age-Standardised (ASt) Incidence Rates per 100,000 Population p.a. in UK during 2013-2015

LCL: Lower confidence limit UCL: Upper confidence Limit ASt: Age standardised

		England	Wales	Scotland	Northern Ireland	UK
Male	Cases	6490	426	646	155	7717
	Crude Rate	24.2	28.0	24.9	17.2	24.3
	AS Rate	28.0	30.0	28.4	21.9	28.0
	AS Rate - 95% LCL	27.3	27.2	26.2	18.4	27.3
	AS Rate - 95% UCL	28.7	32.9	30.6	25.3	28.6
Female	Cases	6503	396	606	197	7702
	Crude Rate	23.6	25.2	22.0	21.0	23.5
	AS Rate	24.3	24.8	21.8	22.6	24.1
	AS Rate - 95% LCL	23.7	22.4	20.1	19.4	23.5
	AS Rate - 95% UCL	24.9	27.3	23.5	25.7	24.6
Persons	Cases	12993	822	1252	352	15419
	Crude Rate	23.9	26.6	23.4	19.1	23.9
	AS Rate	25.7	26.8	24.4	21.8	25.6
	AS Rate - 95% LCL	25.3	25.0	23.1	19.5	25.2
	AS Rate - 95% UCL	26.2	28.7	25.8	24.1	26.0

95% LCL and 95% UCL are the 95% lower and upper confidence limits around the ASt Rate

Source: cruk.org/cancerstats

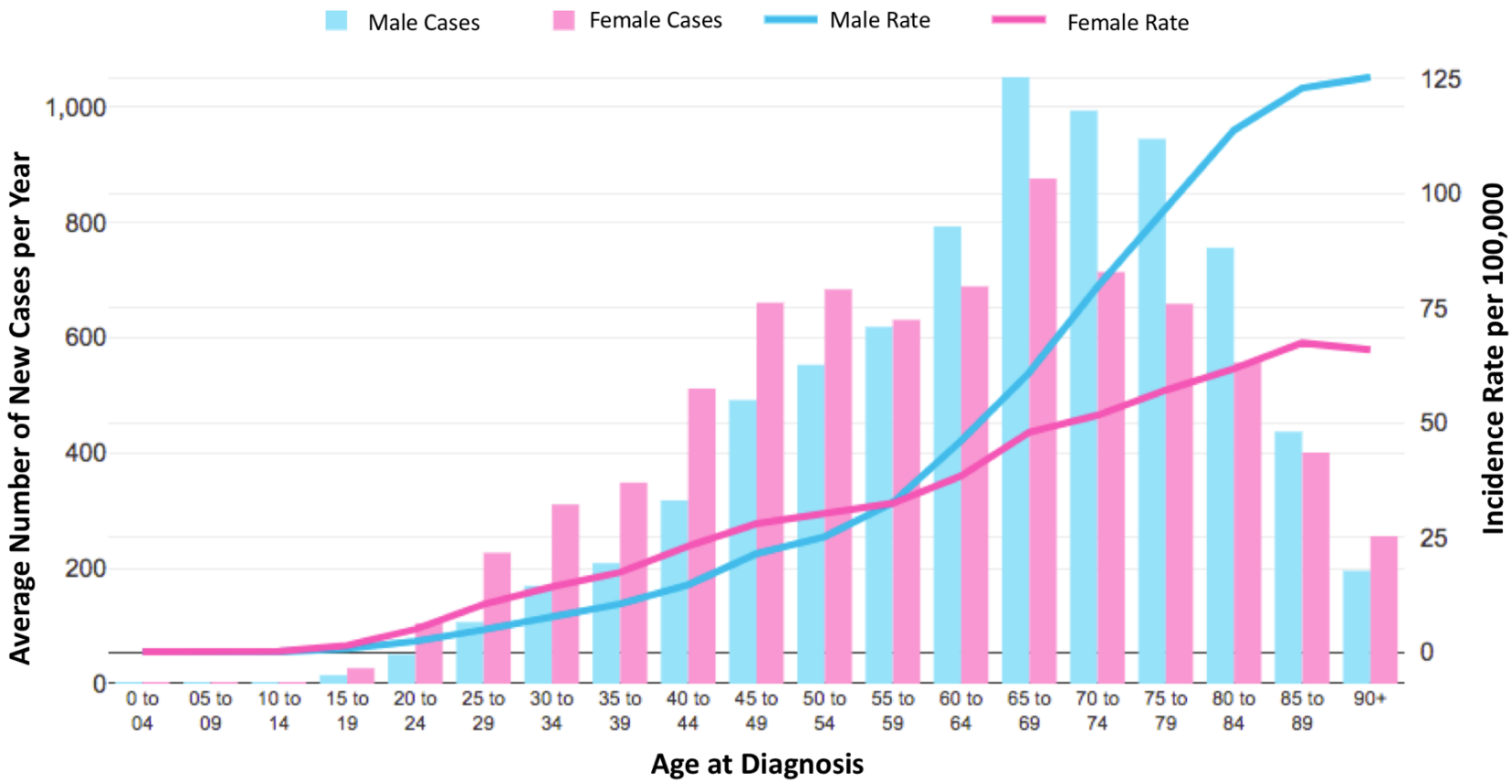


Figure 1.7. Average Number of New Melanoma Skin Cancer Cases per Year and Age-Specific Incidence Rates per 100,000 Population, UK, 2013-2015

Figure taken from Cancer Research UK [38]

1.4 Causes and Risks of Melanoma

A person's risk of melanoma is influenced by their inherited genetic variation, their phenotype (e.g. skin and hair colour or number of melanocytic naevi) and their patterns of UV exposure. It is essentially a disease of pale skinned people but can occur in people of any skin colour. According to the World Health Organization (WHO), the main human risk factors for melanoma are the following [39, 40]:

- Pale skin
- A large number of atypical naevi (moles) is the strongest risk factor for melanoma in pale-skinned populations.
- Melanoma is more common among people with a pale complexion, blue eyes, and red or blond hair.
- The incidence of malignant melanoma in Caucasian populations generally increases with decreasing latitude, with the highest recorded incidence occurring in Australia (33.6 per 100,000 Age Adjusted Rate), where the annual rates are more than what is observed in Europe (11.6 per 100,000) [41].
- High, intermittent exposure to solar UV appears to be a significant risk factor for the development of malignant melanoma [42, 43].
- Most epidemiological studies support a positive association with history of sunburn, particularly sunburn at an early age [44, 45].
- The role of cumulative sun exposure in the development of melanoma is equivocal. However, melanoma risk is higher in people with a history of non-melanoma skin cancers and of solar keratoses, both of which are indicators of cumulative UV exposure.

The study of Fargnoli et. al (2004) among Italian population reported that among the constitutional and environmental factors analysed using logistic regression model, the strongest risk factors are recreational sun exposure (odds ratio [OR] 5.010, 95% confidence interval [CI] 2.110-11.891), the presence of clinically atypical naevi (OR 4.916, 95% CI 2.496-9.995) and the presence of >50 common melanocytic naevi (OR 4.684, 95% CI 2.442-9.231). Additionally, occupational sun exposure (OR 2.573, 95% CI 1.399-4.732), light brown hair (OR 2.336, 95% CI 1.328-4.138), high density of solar lentigos and/or actinic keratoses (OR 1.824, 95% CI 1.0-3.510) and type II, fair skin (OR 1.815, 95% CI 1.031-3.193) and blue eyes (OR 1.757, 95% CI 1.0-3.477) were each significantly associated with cutaneous melanoma risk[46].

Although known factors like sun exposure in the development of melanoma has been known, its aetiology has not been fully understood. A previous study showed that known risk factors of melanoma have different associations by body sites. Patients with melanoma on the head or neck have higher frequency of solar keratoses and lower frequency of nevi compared with those who have melanoma on the trunk [47]. In a prospective study of 152, 949 women and 25, 204 men free from cancer at baseline followed up for up to 14 years, males had a higher risk of developing melanoma on the head/neck and trunk area compared with women. Melanoma located on the upper extremity is most strongly associated with past experience of severe and painful sunburn while a melanoma located on the trunk was most strongly associated with a greater number of moles in both upper and lower extremities[48].

Melanoma may occur in families in which two or more first degree relatives suffer from this disease and is referred as familial melanoma. It is a genetic or inherited condition meaning that the risk of melanoma can be passed from generation to generation within a family. In general, about 8 % of people who are newly diagnosed with melanoma have a first-degree relative with melanoma. About 1 to 2 % has two or more close relatives with melanoma. At present the two primary genes linked to familial melanoma are *CDKN2A* and *CD4* where mutation in these genes increases melanoma risk [49]. It was reported that the frequency of *CDKN2A* mutations in melanoma-prone families are around 5% to 40% [50-52]. The study of Helgadottir et. al (2016) reported that after adjusting for age, sex, and T classification, *CDKN2A* mutated familial melanoma cases had, compared with *CDKN2A* wild type cases, worse survival from melanoma (HR=2.50, 95% CI: 3.65 to 16.51)[53].

1.5 AJCC Stage

The stage of melanoma reflects the severity or the degree of progression of the disease and the prognosis. Thus, the least severe melanomas are those in the earliest stage called Stage 0 (in situ disease) wherein the malignant melanocytes have proliferated only within the epidermis. This can progress to different categories from stage IA to stage IV using the American Joint Committee on Cancer (*AJCC*) *TNM* System (See Table 1.2). T is defined by the characteristics of the tumour in terms of thickness - also known as the Breslow thickness, microscopic ulceration status and mitoses (in AJCC 7th Edition). N is determined by whether melanoma has spread to its nearby or draining lymph nodes. M refers to metastasis of tumours to the other organs in the body [54].

The thickness of the tumour in the AJCC staging system is measured using the Breslow scale to describe how deep has the melanoma invaded the skin. In the Breslow scale, the pathologist uses a micrometre to measure tumour thickness or depth in terms of millimetres (mm) [55]. The thickness of the melanoma tumour is categorised into five namely Tis, T1, T2, T3, and T4 (See Figure 1.8).

Tis (signifying in situ disease) indicates that the melanoma cells are located only in the epidermis. T1 indicates melanoma thickness that is less than 1mm while T2 is characterised by melanoma which thickness of 1mm to 2mm. T3 corresponds to melanoma thickness between 2mm and 4mm while melanoma thicker than 4mm is denoted by T4 [54, 56].

Ulceration in a tumour is manifested by non-intact or broken epidermis overlying a major area of the melanoma [57]. This is a microscopic phenomenon rather than observed by the patient or the dermatologist. Ulceration subdivides T into categories a and b. Ta corresponds to non-ulcerated melanoma while Tb corresponds to an ulcerated melanoma [56]. The presence of microscopic ulceration indicates a higher risk of progression in AJCC. The study of Jewell et al. (2015) reports that ulceration of primary melanomas was associated with more proliferative tumours, tumour vessel invasion, and increased microvessel density. Presence of greater number of macrophages and gene expression pathways associated with wound healing and up-regulation of pro-inflammatory cytokines in the tumour suggests that ulceration is associated with tumour-related inflammation. Modification of signalling pathways involved in inflammation may be reflected by the relative benefit from interferon reported in in patients with ulcerated tumours [58].

Mitotic rate refers to the number of cells that are actively dividing in a sample from melanoma tissue. More number of dividing cells correspond to a higher mitotic rate. This categorises T has to having a mitotic rate of < 1/mm2 or at least 1/mm2. In this staging system, T1a refers to non-ulcerated melanomas with mitotic rate of <1/mm2 while T1b implies that the melanoma is ulcerated and have a mitotic rate of at least 1/mm2 [56].

The N score reflects whether there is evidence that cancer cells from the melanoma have spread to the neighbouring lymph node. Lymph nodes are tiny round organs that are part of the lymphatic system of the human body. The lymphatic system is a component of the body's immune system which consists of network of vessels and organs that contains lymphs. These are clear fluids that carries infection-fighting white blood cells and fluid and waste products from the cells and tissues of the body including cancer cells that have been broken off from the main tumour in case of a person with cancer. The first lymph node/s to which cancer cells are most likely to spread from a

primary tumour is called a sentinel lymph node. Sentinel lymph node biopsy (SLNB) is a procedure used to identify and remove lymph node to examine whether cancer cells are present [59].

The result of SLNB is used to grade the N Score. This can be graded from N0, N1, N2, and N3. N0 implies that the melanoma cells have not spread to the nearby lymph nodes. N1 implies that one nearby lymph node contains melanoma cells. N2 means that two or three nearby lymph nodes contains melanoma cells while N3 indicates 4 or more nearby lymph nodes were contaminated with melanoma cells. Nodes can also be subgroup into Na, Nb and Nc. Na implies that the cancer in the lymph node is visible only with a use of a microscope (micrometastases).

Nb is characterised by clinically detected signs of cancer in the lymph node (macro metastases detected as a lump which can be felt or a mass visible on a scan) while Nc means that tiny areas of the skin that is very close to the primary melanoma (satellite metastases) or in the skin lymphatic channels (in transit metastasis) have detectable melanoma cells [56]. These metastases usually present as a mass or lump in the tissues around or between the site of the primary tumour and the draining nodes.

Table 1.2. *TNM* Staging for Cutaneous Melanoma AJCC 7th Edition [54]

Primary Tumour Characteristics (T)				
T stage			Thickness (mm)	Ulceration
	T1	a	≤ 1.00	Without ulceration and mitosis < 1/mm^2
		b		With ulceration or mitoses ≥ 1/mm^2
	T2	a	1.01-2.00	Without ulceration
		b		With ulceration
	T3	a	2.01-4.00	Without ulceration
		b		With ulceration
	T4	a	> 4.00	Without ulceration
		b		With ulceration
Regional Lymph Nodes (N)				
N stage			Metastatic Burden	Nodal metastatic burden
	N0		0	NA
	N1	a	1	a: Micrometastasis*
		b		b: Macrometastasis†
	N2	a	2-3	a: Micrometastasis*
		b		b: Macrometastasis†
		c		c: In transit metastases/satellites without metastatic nodes
	N3		≥4	metastatic nodes, or matted nodes, or in transit metastases/satellites with metastatic nodes
Distant Metastases (M)				
M stage			Serum LDH	Site
	M0		NA	No distant metastases
	M1	a	Normal	Distant skin, subcutaneous, or nodal metastases
		b	Normal	Lung metastases
		c	Normal	All other visceral metastases^
			Elevated	Any distant metastasis

* Micrometastases are diagnosed after sentinel lymph node biopsy

† Macrometastases are defined as clinically detectable nodal metastases confirmed pathologically

^ Visceral metastases pertains to metastasis to the soft internal organs of the body, including the lungs, heart, and those organs of the reproductive, excretory, circulatory, and digestive systems.

A melanoma that has spread to a different part of the body is called a metastasis (M). Metastasis is classified in two categories: M0 and M1. M0 indicates absence of metastasis while M1 refers to presence of metastasis. M1 can be further subclassified into M1a, M1b and M1c. M1a is described by metastasis to the skin beyond the body part on which the primary tumour was found or lymph nodes that are far away from the primary melanoma. So if a patient presented with a primary tumour on the ankle and developed metastases in the skin of the abdominal wall or in a lymph node in the neck this would be classified as M1a disease. M1b pertains to metastasis in the lung while M1c indicates metastasis in other organs , or there is an increased level of lactate dehydrogenase (LDH) in the blood – an enzyme secreted by the liver and is used to identify the site and severity of tissue damage in the body [56, 60].

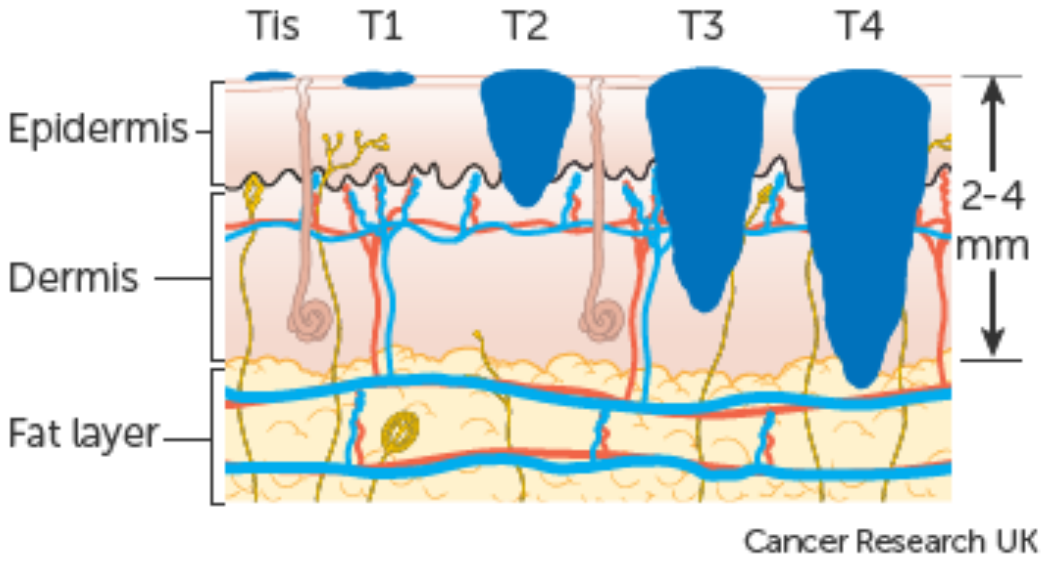


Figure 1.8. Tumour Thickness Categories

Figure taken from Cancer Research UK website [56]

Table 1.3 shows the Clinical Staging for Cutaneous Melanoma based on the work of Balch et al., (2009) from the analysis of 30, 946 patients with stages I, II, and III melanoma and 7.972 patients with stage IV melanoma which reflects the improved understanding of the disease [54].

Table 1.3. Clinical Staging for Cutaneous Melanoma[54]

Stage*	T	N	M
IA	T1a	N0	M0
IB	T1b	N0	M0
	T2a	N0	M0
IIA	T2b	N0	M0
	T3a	N0	M0
IIB	T3b	N0	M0
	T4a	N0	M0
IIC	T4b	N0	M0
III	Any T	N > N0	M0
IV	Any T	Any N	M1

* Clinical staging includes microstaging of the primary melanoma and clinical/radiologic evaluation for metastases. By convention, it should be used after complete excision of the primary melanoma with clinical assessment for regional and distant metastases.

1.6 Melanoma Survival

From 2013 to 2017, almost all (98.2%) of the melanoma patients in the UK survive their disease for at least one year. More than 91 percent (91.3 %) of them survive their disease for at least five years. It is estimated that about 9 in 10 melanoma patients survive their disease for at least ten years [38].

In terms of sex, females tend to have better rate of survival than males for both one, five, and ten-year survival periods. In terms of age, data from England indicates that

95% of the patients diagnosed with melanoma survive their disease for at least five years, compared with 80% for those diagnosed aged 80 and older. When diagnosed immediately (i.e. as early as Stage I), all (100%) of the patients will survive their disease for at least one year as compared to a little more than 1 in 2 (53%) when the disease is diagnosed at the latest stage[38].

In comparison with the European average, the five-year relative survival for men is higher in England, Scotland, and Northern Ireland, but lower in Wales. For women, five year relative survival is higher for England, Scotland, and Northern Ireland, while it is similar for Wales when compared to the European average [38].

1.7 Cancer Development

Cancer results from the uncontrolled growth of abnormal cells in the body, generally characterized by 3 main steps: initiation, promotion and progression [61]. Many cancers initiate by gene mutations, these are single base-pair mutations but can involve additional processes including breaking away of pieces of chromosome from their normal position and joining with different chromosome (translocation). Typically, these changes affect the signalling to genes changing the way they work in the body. Cells have the ability to detect these different changes and destroy themselves (apoptosis) but when they fail to do so, damage cells begin to proliferate and may lead to a malignant tumour that is cancerous. This may be caused or triggered by carcinogenic substances like harmful chemicals, smoking or exposure to radiation or represent simply random chance. This carcinogen driven change is called initiation. It takes repeated damage to the cells before a cancer develops. There are agents that reinforce or further cause damages to the cells called promoters. These can be hormones or drugs that do not independently cause cancer but nurture the initiator cells to become cancerous or it may simply be that the initial mutations give a growth advantage to the cell lineage, again promoting growth. The final step in the cancer development is called progression. It is the step that causes the cell to reproduce and perform its functions differently and become cancerous. The final step in the cancer development is called progression. It is the step that causes the cell to reproduce and perform its functions differently and become cancerous. Doubling time refers to the period required by the cancer cells to grow twice its number. As these cancer cells keep on proliferating, a malignant tumour is formed taking months to years to be felt or detected by imaging test. A summary of how cancer develops is described in the diagram below [61].

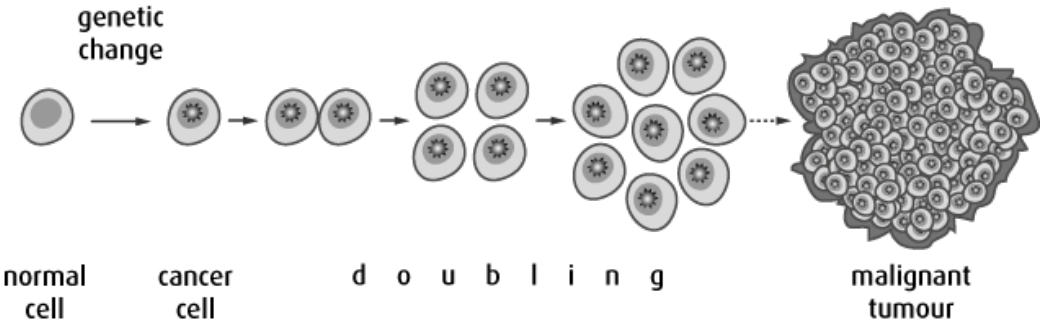


Figure 1.9. Cancer Development.

A Diagram from Canadian Cancer Society (2017).

In terms of the hallmarks of cancer written by Hanahan and Weinberg (2011), it was mentioned that this comprise of six biological capabilities brought by multistep development of human tumours. These hallmarks form an organizing principle for nationalizing the complexities of neoplastic disease. These are composed of sustaining proliferative signalling, evading growth tumour suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. Essential to these hallmarks are genome instability which generates the genetic diversity that facilitates speeding up their acquisition, and inflammation, which promotes multiple hallmark functions[62].

1.7.1 Mutation

Mutation occurs when there is a permanent alteration in the DNA sequence that constitute a gene. This can affect a single base pair in the gene or a large segment of a chromosome that include more than one gene. This can be generally classified in two ways: hereditary or acquired mutations[62]. Hereditary or germline mutations are passed on from parents to offspring and are present in almost all every cell of a person throughout its life. These are from the parent's egg or sperm cells (germ cells). Acquired or somatic mutations occur at any time in a person's life and are present only is some cells in the body. These can be brought by environmental factors such as ultraviolet radiation from the sun or abnormalities that take place in the DNA replication during cell division. Unlike hereditary or germline mutations, acquired or somatic ones are not passed on to the next generation. Below are the different types of mutation that could possibly occur in a cell [63]:

- Missense mutation - This type of mutation that occurs when a single DNA base pair causes the substitution of one amino acid for another in the protein created by a gene.
- Nonsense mutation - This is also a change in a single DNA base pair but instead of causing a substitution of one amino acid for another, it sends an early signal to the cell to stop making the target protein resulting to a shortened one that may not be enough to perform its function.
- Insertion – an insertion mutation changes the number of DNA bases in a gene by adding a portion of the DNA usually resulting to a protein that does not function properly.
- Deletion – This occurs when the number of DNA bases change due to a removal of a portion of DNA which may be small (one or few base pairs of a gene) or large (affecting entire genes or several neighbouring genes) and missing this portion may change the function of the protein(s) manufactured.
- Duplication – This takes place when a piece of DNA is abnormally copied at least once affecting the function of the resulting protein.
- Frameshift mutation – This encompasses insertion, deletion and duplication and occurs when the addition or loss of DNA bases changes a gene's reading frame. The change in the reading frame regroups the codons and changes the codes for the amino acids normally resulting to non-functional protein.
- Repeat expansion- Nucleotide repeats are short DNA sequences that are repeated at least once in a row as exemplified by a trinucleotide (three-base-pair sequences) and a tetranucleotide (four-base-pair sequences) repeats. Repeat expansion occurs when the frequency of repeat of a short DNA sequence is increased and may cause the resulting protein not to function properly.

The commonly reported mutated gene in melanoma were *NRAS* and *BRAF* [64-69].

The study of Hodis et al. (2012) analysed large scale melanoma exome data and discovered six novel melanoma genes such as *PPP6C, RAC1, SNX31, STK19,* and *ARID2* [70]. The Cancer Genome Atlas (TCGA) proposed a genomic classification of melanoma that involves four subtypes based on mutational patterns in *BRAF*, *NRAS*, *NF1*, and absence of mutation in three of these (triple wild type) [71, 72]

1.7.2 Chromothripsis

Chromothripsis is a term that comes from the Greek words *chromo* which means color (which represents chromosomes as they usually undergo staining using particular dyes) and *thripsis* which means “shattering into pieces” [73]. Chromothripsis is a mutational process characterised by up to thousands of clustered chromosomal rearrangements that occur in a single event in localised and bounded genomic regions in one or a few chromosomes, and is linked in both cancer and congenital diseases. It is initiated by one massive genomic rearrangement during a single catastrophic event in the cell's life. It is suggested that in order for a cell to withstand such a destructive event, its occurrence must be the upper limit of what a cell can tolerate [74]. The simplest way to model the occurrence of these rearrangements is through the simultaneous fragmentation of distinct chromosomal regions (breakpoints show a non-random pattern) and then subsequent imperfect reassembly by DNA repair pathways or aberrant DNA replication mechanisms. Chromothripsis happens early in the development of tumour and leads to cellular transformation by loss of tumour suppressors and amplification of oncogenes [75].It has been observed in 2–3% of cancers across all subtypes[75].

Chromothripsis was first reported in a paper in 2011 by Stephens PJ, Greenman CD, Fu B, et al. (2011) in a sequenced genome of a chronic lymphocytic leukaemia. Using paired end sequencing, 55 chromosomal rearrangements were recorded in the long arm of chromosome 8 and a significant number of rearrangements were observed in regions of chromosomes 7, 12, and 15 [73]. Subsequent investigations were done using genome-wide paired-end sequencing and SNP array analysis and found similar patterns of chromothripsis in various human cancers such as melanomas, sarcomas and colorectal, lung and thyroid cancers [76]. In the follow up investigations, about 25% of studied bone cancers showed evidence of chromothripsis. Chromothripsis has been associated with the generation of oncogenic fusions in supratentorial ependymoma, chondromyxoid fibroma, and Ewing sarcoma, the latter two being bone tumours [77, 78].

Chromothripsis has also been reported to be curative. There was study in 2015 about a case of a woman who had WHIM (warts, hypogammaglobulinemia, infections, and myelokathexis) syndrome, an extremely rare autosomal dominant combined immunodeficiency disease, and lost her symptoms during her 30s after the disease allele was deleted by the chromothripsis of chromosome 2 [79].

1.7.3 Kataegis

Kataegis describes a distribution of localized hypermutation identified in some cancer genomes[80]. This term was derived from the Greek word for "thunder", καταιγίς (*kataigis*). Shown to be colocalized with regions of somatic genomic rearrangements, the base mutations in these regions were discovered to be almost exclusively C>T (cytosine to thymine) in the context of a TpC dinucleotide. The study of Nik-Zainal et.al (2012) hypothesized that enzyme of the APOBEC family is responsible for the process of Kataegis. A study showed direct link between the APOBEC deaminases and kataegic clusters of mutations by expressing hyperactive deaminase in yeast cells[81]. Recent evidence has linked the increased expression of the family member APOBEC3B with different human cancers emphasizing its potential contribution to genomic variability and kataegis [82].

1.7.4 Copy Number Variation/Aberration/Alteration

Genomic copy number variation has been implicated in the study of many diseases and is currently one of the promising areas of cancer research. Copy number variation (CNV) is defined as a DNA segment of one kilobase (kb) or larger that is present at a variable copy number as compared to a reference genome[83]. It provides a major contribution to the genomic variation among individuals brought by deletions and duplications of segments in the genome and may have no effect on the phenotype, account for adaptive traits or be associated to a disease [84]. There was a previous confusion between using copy number variation (CNV) and copy number alteration (CNA) or copy number aberrations. As mentioned by Weistra (2016), copy number alterations and copy number aberrations are synonymous. For example, the study of Shah et al. (2007) uses copy number alterations (CNAs) in their paper and both terminologies in the titles of the references while the study of Wu et al. (2014) uses copy number aberrations (CNAs) in paper and both terminologies in titles of their references. [85, 86]. Copy number alterations/aberrations (CNAs) are changes in the copy number of a somatic tissue (i.e. tumour sample) while copy number variations (CNVs) are changes in the copy number of the germline cells [87-89].

Increase in the genomic copy number is termed as a *gain* (potentially resulting in increased expression of an affected gene) a decrease is termed *loss* while no change in the copy number is termed *normal* (diploid). Melanoma has been well associated with non-random breaks and deletions on chromosome 9p21, a region that contains the tumour suppressor gene *CDKN2A* [90]. Aside from *CDKN2A (30.8 % rate of deletion, n=367),* The Cancer Genome Atlas (TCGA) reports *CDKN2B* (28.9%), *MTAP* (25.1%) and *PTEN* (7.1%) as commonly deleted genes in cutaneous melanoma which are also found in chromosome 9p23.1. Commonly amplified genes are *RECQLA* (7.1%, located in 8q24.3), *CCND1* (6.8%, 11q13.3), *BRAF* (6.8%, 7q34), *MITF* (6.8%, 3p13), *NOTCH2* (6.8%, 1p12), *AGO2* (6.8%, 8q24.3), and *MYC* (6.5%, 8q24.21)[5, 6, 71, 91, 92].

1.8 Assessment Methods for CNV

Copy number analysis generally refers to the process of analysing data obtained by an experiment for DNA copy number variation in a patient's sample. This helps identify chromosomal copy number states that may be associated with various diseases, its subtype and more effective treatment [93]. The primary mechanisms that create CNV's include non-homologous end joining, non-allelic homologous recombination,

transposition of transposable elements or pseudogenes, variable numbers of tandem repeats, and replication errors bought about by fork stalling or template switching [94]. Some well-established methods have been used for validation or replication of targeted CNV assessment both for single and multi-locus scale including quantitative PCR (qPCR), paralog-ratio testing (PRT) and molecular copy number counting (MCC). In qPCR, threshold cycles between the test gene and a reference sequence with (assumed) normal copy numbers are compared to derived ratio values that are used for copy number calculation.

A single pair of primers are used by PRT to check for degree of similarity between elements of sequence both in the target locus (with CNV) and a reference locus. With MCC, the aliquots that are positive for a target sequence are counted and compared with those of the other target sequences allowing estimation of the relative copy number of different sequences in a test DNA sample.

For the whole genome CNV profiling, the most commonly used platforms are SNP arrays and comparative genome hybridization (CGH). CGH was initially developed as a method for assessing the copy number of differentially labelled test in relation to a normal reference DNAs using fluorescence *in situ* hybridization (FISH) onto metaphase spreads from a normal sample [95]. This measures the fluorescence ratio along the length of each chromosome specific regions of relative loss and gain in the test sample. A major weakness of this method was the low resolution (typically 5-10 Mb) afforded by metaphase FISH. To address this drawback, large-insert clone libraries were developed and clones assembled to overlapping sequence reads (contigs), as driven by availability of resources created for the Human Genome Project [96, 97]. In this method, test and reference DNA's are differentially fluorescently labeled and hybridized together to an array. This results to a fluorescence ratio that is measured, clone by clone and mapped to the clone's position in the genome. Today, the next generation sequencing transforms biology research [98].

The studies of Teo et al. (2012) and Tattini et al.(2015) described four methods for CNV detection using NGS data including : a) Depth of coverage (DOC) or Read-depth(RD) or Read-count(RC), b) Paired end mapping (PEM) or Read-pair (RP), c) Split read (SR) and d) Assembly based (AS) as illustrated in Figure 1.10 [99, 100].

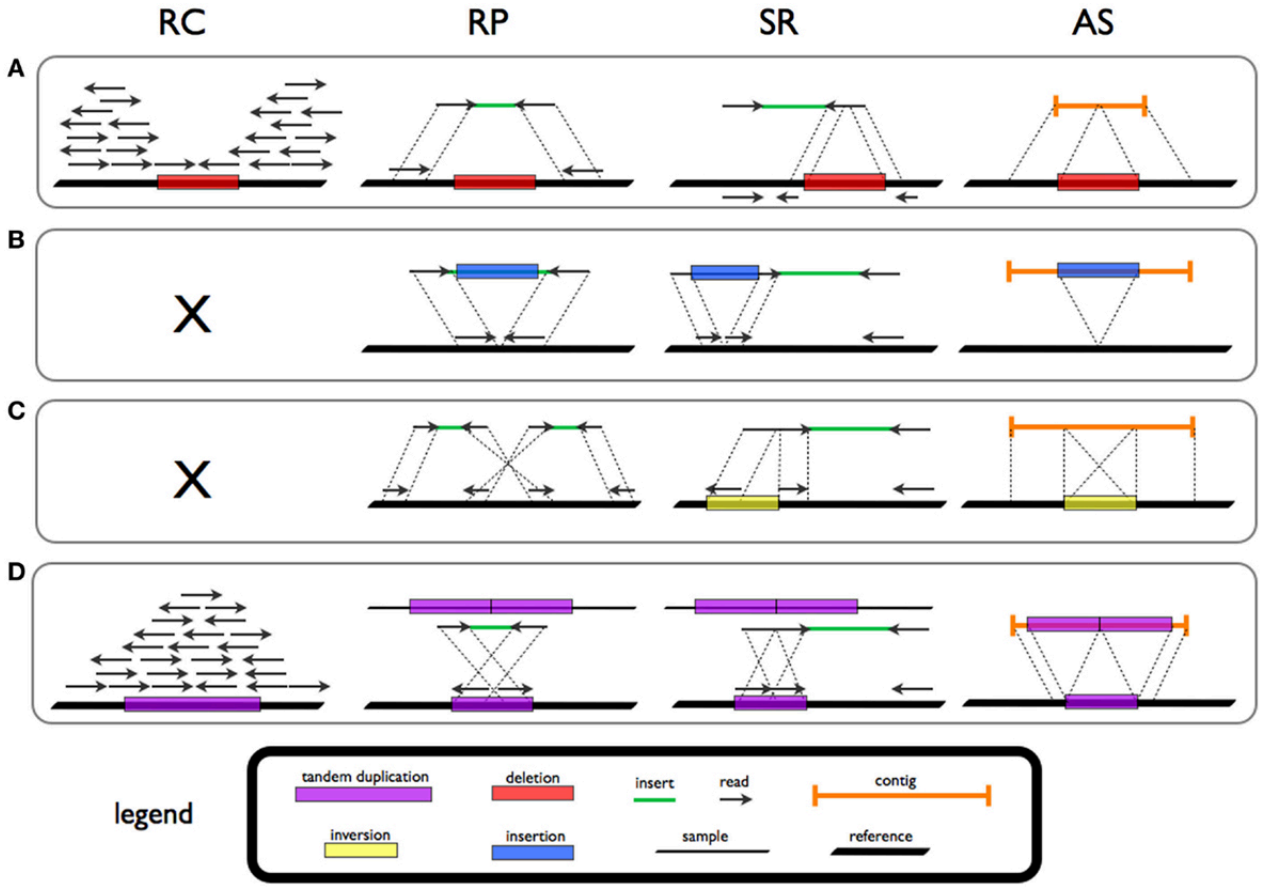


Figure 1.10. Four strategies for the detection of SV signature

Taken from Tattini et al. (2015). deletion (A), novel sequence insertion (B), inversion (C), and tandem duplication (D) in read count (RC), read-pair (RP), split-read (SR), and de novo assembly (AS) methods.

DOC assume a random (Poisson or modified Poisson) distribution in mapping and checks departure from this distribution to identify duplications and deletions - observed DOC/intensity lower than expected indicates a deletion while observed DOC/intensity higher than expected denotes duplication [100, 101]. Tattini et.al. (2012) added that RP methods are based on the length and orientation of pair-ends such that it collects discordant pairs in which the mapping span and/or orientation of the read pairs are inconsistent with the expected insert size. They also mentioned that Split-read methods allow for the discovery of structural variants with single base-pair resolution and investigates presence of variant breakpoints based on the split sequence-read signature breaking the alignment to the reference sequence. This identifies deletion by treating a gap in the read as a marker while stretches in the reference is treated as insertions. *De novo* assembly (AS) methods refer to combining and ordering short segments of the sequence to reassemble the original sequence from which the short segments were sampled [102]. The study of Teo et al. (2012) summarized commonly used software for CNV detection as shown in Table 1.4 below [99].

Table 1.4. Commonly used softwares for CNV detection using NGS data

Taken from Teo et al. (2012)

Programme	Reference	Comments
DOC / RD / RC		
CNVnator*	Abyzov et al., 2011	Uses mean shift approach on fixed window GC-content-adjusted read counts.
Rdxplorer*	Yoon et al., 2009	Uses event-wise testing on fixed window GC-content-adjusted read counts.
SeqCBS	Shen and Zhang, 2012	Gives approximate confidence intervals for assessing confidence in the segmentation.
CNVseq	Xie et al., 2009	Uses ratios between reads from target and reference genome.
SegSeq	Chiang et al., 2009	Segments genomes of a tumour and matched normal sample by a sliding fixed size window. Boundary is refined after
ExomeCNV	Sathirapongsasuti et al., 201	For exome sequencing data. Uses read count ratio to detect CNVs, and B allele frequencies to detect LOH.
Control-FREEC	Boeva et al., 2012	Uses total coverage and B allele frequencies of SNPs to call CNVs and LOH.
PEM / RP		
Variation Hunter*	Hormozdiari et al., 2009	Based on maximum parsimony. Uses soft clustering
BreakDancer*	Chen et al., 2009	Consist of two complementary algorithms: BreakDancerMax predicts insertions, deletions, inversions and inter- and intra-chromosomal translocations; BreakDancerMini predicts small indels.
PEMer*	Korbel et al. 2009	Clusters long and short events separately. Confidence value for each SV. Built in database and simulation programme.
SR		
Pindel*	Ye et al., 2009	Uses pattern growth algorithm. Identifies breakpoints of large deletions and medium sized insertions.
Assembly based Cortex*	Iqbal et al., 2011	Capable of assembling multiple genomes simultaneously.
SOAPdenovo*	Li et al., 2010	Claims faster computation time and longer contig size and assembly accuracy when compared with earlier methods such as ABySS and velvet.
Velvet	Zerbino et al., 2008	—
ABySS	Simpson et al., 2009	—
Combination/others		
Genome STRiP*	Handsaker et al., 2011	Combines DOC, PEM and distribution of evidence across samples and within a genomic locus.
HYDRA	Quinlan et al., 2010	DOC + PEM
ABI tools	McKernan et al., 2009	CBS
Spanner*	Mills et al., 2011	Uses PEM and able to find tandem
SVDetect	Zeitouni et al., 2010	DOC + PEM.Competible with SoLiD and Illumina paired-end reads.

*used in 1000 Genomes Projects

In this study, a shotgun sequencing method called *CNV-Seq* was initially used to detect copy number of DNA from the processed primary melanoma samples. This generates ratios of read counts between the test (X) and the reference (Y) sequence and considers the number of sequence reads and not the length as the key factor in determining the resolution of detection and is fit for NGS methods that can quickly generate large data of short reads [103]. Shown in Figure 1.11 is a comparison of the conceptual steps in *aCGH* and *CNV-Seq* followed by the formulas for computing mean number of reads ($\bar{\lambda}$), and predicted copy number ratio (r) from the work of Xie et al. (2009) [103].

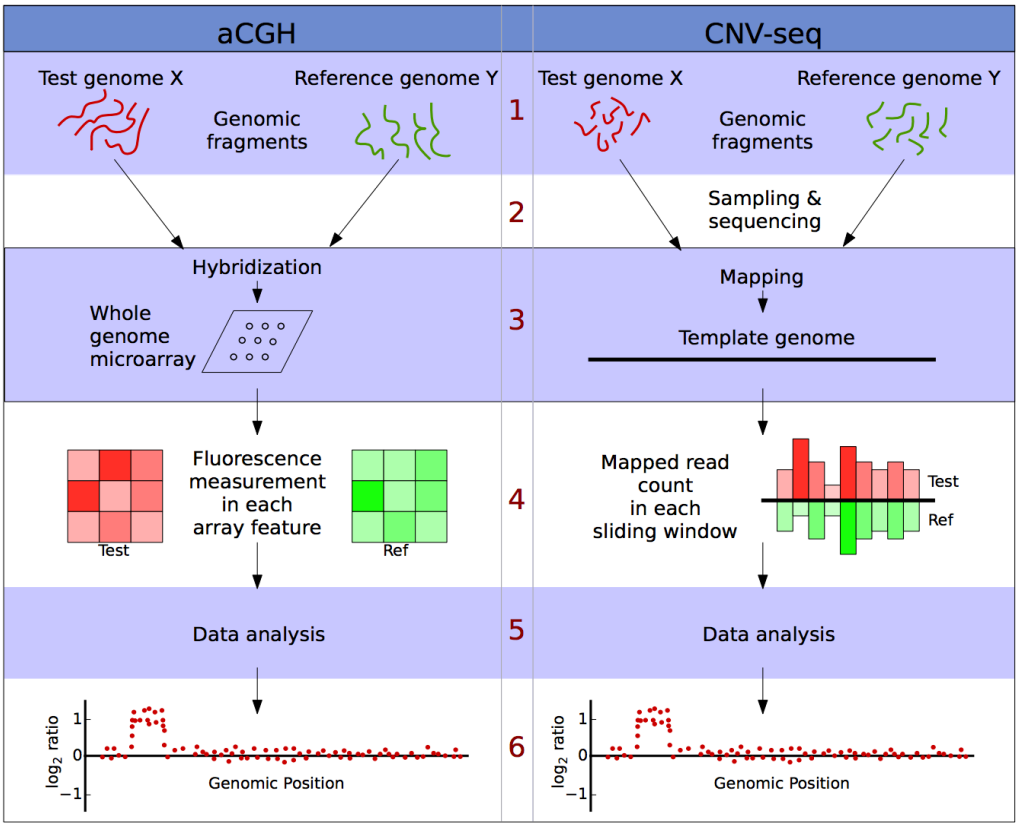


Figure 1.11. Comparison of the conceptual steps in aCGH and CNV-seq methods

Taken from Xie and Tammi (2009)

The mean number of reads for two genomes X and Y in each window determines the distribution of ratios. The mean number of reads (λ) in a window is approximated by Poisson distribution given by:

$$\lambda = NW/G \quad \textbf{Equation 1} \quad \boldsymbol{\lambda = NW/G}$$

where *N* is the total number of sequenced reads and *G* is the size of the genome and W is the size of each window in the genome.

The estimated copy number ratio (*r*) can be calculated as:

$$r = z^{*}(\frac{N_Y}{N_X}) \quad \textbf{Equation 2} \quad \boldsymbol{r = z(\frac{N_Y}{N_X})}$$

where z is the ratio of read counts in the window and N_X and N_Y are the total number of reads in the genomes X and Y.

1.9 The Test and Reference Samples

The test samples in this study are the tumour samples from the Leeds Melanoma Cohort (LMC). There were two reference samples used separately in this study. First is the seven composite normal samples from LMC and second is the 312 Caucasian samples from the 1000 Genomes Project. In this study, the tumour samples were obtained from patients of white British origin and it is important to select similar samples as reference (or control) to account for ethnicity related germline variation. Detailed information on the test and reference samples are found in Chapter 3 (Methodology).

Chapter 2
Research Aims and Objectives

This project was conducted to identify copy number alterations/aberrations (CNAs) in primary melanoma using NGS data derived from the formalin-fixed primary tumour samples taken from participants in the Leeds Melanoma Cohort (LMC) and to test for associations with patient clinical characteristics including survival. Specifically, it aims to:

1. Implement a procedure for analysing copy number data based on NGS,
2. Assess quality control methods by examining the consistency with published literature, especially TCGA
3. Develop and apply additional measures to improve data quality, as required, reassessing consistency with the literature,
4. Identify and characterize novel genomic regions of copy number aberrations,
5. Provide a measure of genomic instability by estimating overall copy number alteration/aberration (CNA) load and investigate the association of these with clinical characteristics including melanoma survival.
6. Associate focal genomic regions with clinical characteristics including melanoma outcome.

Chapter 3
Methods

In this chapter, I discussed the design and the patient sample used in this study. I described how the tumour sampling and replication were done as mentioned in our previous studies. Then, I presented how the copy number data were generated and gave descriptions of the One Thousand Genomes Project. I discussed the copy number windows and factors affecting read counts such as GC content and mappability, and explained the normalization method applied to read counts. This was followed by elaborating the segmentation of the copy number data and finally a discussion on the validation of the *CDKN2A* copy number deletion using MLPA.

3.1 Study Design and Patient Sample

The Leeds Melanoma Cohort (LMC) Study is composed of 2184 population ascertained melanoma patients diagnosed between 2000 and 2012 and invited to participate typically 3 months after diagnosis of primary melanoma; once recruited participants were followed until participant indicated an unwillingness to be contacted further, death or the end of the study (median follow up > 8.6 years as of April 2018) [104]. The LMC has available extensive phenotypic information, biological samples and information describing patterns of UV exposure, the primary risk factor for melanoma as well as measures of pigmentation and naevi which are also associated with risk [105]. Recruitment was categorized by Breslow thickness, the known major predictor of outcome, with under-sampling of lesions thinner than 0.75mm.

For the copy number assessment, tumours from participants with death from melanoma were selected as were tumours from participants who had survived for at least 5 years from diagnosis; these were identified as a comparison group. This study focuses on the 875 participants whose tumours met the criteria.

Additionally, whole genome sequence data of samples from the Phase 3 of 1000 Genomes Project (1KGP) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/) were obtained as normal controls. We selected samples which were similar to the LMC samples in terms of sequencing characteristics (Illumina platform, low coverage, paired end library layout) and ethnicity (Caucasian population). After applying these filters, a total of 312 control samples (British with n=106, Finnish with n=105, and Central Europeans in Utah (CEU) with n=101) were included. These samples were composed of mixed males and females [107-109]. Including the X chromosome in the analysis affects the LOESS correction applied to the read counts unless the samples are all

females [2]. Since the focus of my study is on somatic copy number variations, analysis of the sex chromosomes was excluded.

3.2 Tumour Sampling

There were 875 individuals with primary melanoma selected for this CNV study on the basis of Breslow thickness (lesions > 0.75mm) and survival (survived for at least 5 years from diagnosis at the time of study) as mentioned in Sectioned 3.1. Paraffin embedded blocks (FFPE) from primary tumours were obtained from different NHS Pathology Departments. Tracing of the FFPE blocks was led by the Human Tissue Act manager, Ms. Sandra Tovey. Of the 875 target blocks, 796 samples were available as FFPE primary blocks; some of the blocks were not available from the relevant Pathology laboratory while others were deemed insufficient for sampling. The tissue sampling including sectioning and staining was performed by Dr. Filomena Esteves and Dr. Jonathan Laye.

Staining of the sectioned tissues was done using Mayer's Haematoxylin and 1% Eosin (H&E) to facilitate identification of regions for sampling after examination.

Both Prof. Julia Newton-Bishop and Dr. Jonathan Laye performed the reviewing of the H&E stained slides under a microscope and identified areas to be selected for sampling. A 0.6mm diameter tissue microarray needle was used to obtain tumour cores consistently as previously described and horizontally through the deepest part of the tumour which has the least contamination of stroma or inflammation to increase comparability among tumours [106]. Up to three cores were obtained from a tumour whenever possible while a sampling was not done when it was deemed there would be insufficient tumour left for the following clinical testing, or if the tissue was necrotic.

3.3 Replicates

A standard measure of QC is to examine the extent that replicate samples provide the same information. In this study, there are several types of replicates allowing an examination of the extent to which the results are influenced by statistical and other variations; this a minimal criterion to assess if quality control Is sufficient to allow meaningful comparisons between samples. There was a total of 34 samples from unique patients which were replicated at least twice. Three of these were replicated thrice while the rest are replicated twice resulting to a total of 71 replicated samples. Of the 71 samples, only the top 2 samples of the triplicates were selected in terms of

highest mapped reads. Shown in Table 3.1 below is the summary of the number of the replicates before and after filtering the samples based on mapped read counts (for the triplicates), and after excluding samples which are rejected due to very low alignment rate.

Table 3.1. Commonly used softwares for CNV detection using NGS data

Replicate	Before QC		Third Replicate Excluded*		Rejects Excluded**	
Biological (n=13)	**samples**	**total replicates**	**samples**	**total replicates**	**samples**	**total replicates**
Tumour	2	5 (1 triplicate)	*2*	*4*	2	4
Core	11	22	11	22	11	22
Technical (n=21)		0		0		0
Concentration	2	5 (1 triplicate)	*2*	*4*	2	4
Method	5	10	5	10	5	10
Technical	14	29 (1 triplicate)	*10*	*28*	*10*	*20*
Total	**34**	**71**	**34**	**68**	***30***	***60***

**triplicates with the lowest reads mapped were excluded* ***rejected due to very low alignment rate*

Of the 34 samples, Technical replicates (n=21) that were assessed include: a) 14 libraries which were directly sequenced (“technical”); b) 10 libraries derived from 5 patients of which the same DNA samples were prepared in two ways: manually and using the NEB NEBNext® Ultra DNA Library Prep kit; c) 5 libraries generated using different DNA input from 2 Leeds Chemotherapy Study (LCS) patients (3 libraries using 250ng, 100ng and 25ng DNA from patient 1; 2 libraries using 100ng and 25ng DNA from patient 2) ("concentration"). The biological replicates include a) 22 libraries derived from 11 patients where a second library was prepared using a second core from the same primary tumour block ("core") and b) 5 libraries derived from different primary tumours from the same patient ("tumour", 3 tumours from one patient, two tumours from a second patient).

Of these 34 samples, 4 rejected samples due to very low alignment rate which all belong to the technical replicates were excluded. A total of 30 samples which were replicated twice (60 total replicates) were retained for further analysis.

3.4 Data Generation

Data generation were initially done by Dr. Anastasia Filia and Dr. Alastair Droop, prior to my arrival in Leeds. Whole-genome NGS libraries of the DNA samples were prepared using either manual library preparation method or using the NEB NEBNext® Ultra DNA Library Prep kit for Illumina [107, 108]. Technical replicates were sequenced in order to assess the reproducibility of NGS technology using FFPE derived DNA.

Sequencing of all NGS libraries on Illumina GAII or HiSeq sequencer were done to produce >100bp pair-end reads (either 1 or 5 samples per lane) with a median coverage of 1.8x (single) and 9.1x (multiplexed).

Sequence reads trimming was done using *cutadap*t version 1.8.3: adapters and low quality read tails (quality score<37) were trimmed including reads less than 20 nucleotides to reduce the chances of sequence contamination and improve the speed and quality alignment to the reference [109]. GRCh38 human reference without alternate contigs were used in the alignment of the remaining reads using *bwa mem 0.7.10* [109]. Duplicates were marked using *Picard version 1.119* [110].

To minimize artefacts around known common indels, local realignment was performed using the GATK pipeline with the default parameters added with "filter_bases_not_restored" and "filter_mismatching_base_and_equals" [111]. *samtools calmd* was used to recall BAM file MD tags[112]. Alignments that were unmapped, secondary, QC failed, duplicated or supplementary and with mapping quality of less than 20 were excluded to derive the final trimmed data.

3.5 The GRCh38 Human Genome Reference Build

The 1000 Genomes Project was able to produce more than 100 trillion base pairs of short read sequence from more than 2600 samples coming from 26 populations conducted between 2008 and 2013 [113]. All the processed samples were sequence with two methods: low-coverage whole genome sequencing (WGS) and whole exome sequencing (WES). It was released and aligned on the human reference genome assembly GRCh37 and has more than 85 million genotyped and phased variants [113]. Since then, it has been broadly used by the researchers in the scientific community for studies that commonly involve genotype imputation, mapping expression Quantitative Trait Loci, filtering non-pathogenic variants from exome, whole genome and cancer genome sequencing projects, and genetic analysis of population structure and molecular evolution [114].

In late 2013, The Genome Reference Consortium released the first major update to the reference genome assembly named GRCh38. Zheng-Bradley et al. (2017) realigned the 1000 Genomes Project samples to the GRCh38 reference and listed the following improvements in the human genome assembly as below [114] :

- Correcting erroneous bases, updating the tiling path in highly variable regions, and closing sequence gaps.

- Introducing centromere sequence to replace mega-base stretches of Ns in earlier assemblies. The centromeres are created from a model of the estimated number and order of centromeric repeats.
- Substantially increasing the number of alternative loci associated with the assembly. Following the assembly model introduced with GRCh37 that also supported updates and patches, GRCh38 introduced 261 alternative scaffolds (ALT) to represent diverse haplotypes in 178 chromosomal regions.

The locations of the remaining unclosed gaps (including the short chromosome arms) in the human genome where obtained from the UCSC browser as described in Chapter 5 : Additional Steps to Increase Data Quality.

3.6 CNV Data Windows

To make direct comparison across samples, identical window sizes and window locations were used to bin the data; this has become the standard approach for accumulating information across the genome [115]. Studies have shown acceptability of this method even at window read count as low as 60 per window [2, 116]. Data were binned in different fixed window sizes such as 1M, 100k, 10k, 5k and 1k base-pairs across the whole genome data. This was performed using a software developed by Dr. Alastair Droop (*bamwindow*, available on GitHub at https://github.com/alastair-droop/bamwindow) which divides a reference dataset into fixed size windows, starting at the beginning of each chromosome. Read midpoint was assigned as the focus of the read to each window so that each read falls exactly in one window. In calculating the midpoint, clipped regions are not included. This method allows window to window comparison of adjusted read counts across samples.

3.7 GC Content and Mappability

Despite the great advantages of NGS, genome assembly from its generated data poses some challenges including sequencing bias [117, 118]. This section describes sequencing bias associated with GC content and mappability. GC content is defined as the total number of C and G nucleotides divided by the total number of nucleotides (A, C, T, G) in a given region (e.g. window) of the reference genome.

Several studies have conducted systematic analysis of the influence of GC bias on genome assembly [119, 120]. Extensive GC content reduces the completeness of the genome assembly especially when the amount of GC base pairs exceeds 40%.

Because of this, strong GC bias fragments the assembly due to low coverage of reads in the GC-poor or GC rich regions of a genome [120].

Shotgun and next-generation sequencing (NGS) involve shearing the genome into segments, and sequencing either all or part of the resulting segments; these are termed as *reads*. The overlap between reads serves as the basis of de novo assembly [121]. The reads are then mapped to the reference genome to create a reference assembly. The task of reference assembly is straightforward when the read length is long enough. Long enough reads make the success of creating a reference assembly more achievable.

Mappability indicates the uniqueness of an identified sequence in the reference genome and depends on the length of the sequence and the number of allowed mismatched base pairs. This can be extensively affected by the number of repetitive nucleotides in a sequence. It is known that some regions of the human genome can be sequenced but remains unassembled due to the extensive amount of repetitive sequences present [122]. Unassembled regions can be classified into four types namely: a) telomeres; b) centromeres; c) short-arms of acrocentric chromosomes (chromosomes 13,14,15,21,22 and Y); and d) large heterochromatic regions (in chromosomes 1,9,16 and Y). A specific sequence in the genome is called 'mappable' if the read length (*k-mer*) of the reference genome beginning at the defined location is not exactly repeated at any other part of the genome [123]. The measure of mappability is defined as the proportion of the short-read sequences (typically 24, 36 or 48 bp reads which are uniquely represented in that window); for this analysis, we examined 36 bp reads.

While this project does not involve genome assembly, the deficit in GC reads with shotgun sequencing is an issue for copy number assessment which involves comparing the observed number of reads in a window with the expected number calculated on the basis of a defined genomic profile. For copy number analysis, this bias needs to be taken into account in the analysis.

Adjusting for the read counts to account for GC content and mappability bias has been addressed in several studies. Earlier use of simpler LOESS (Locally Estimated Scatterplot Smoothing) model (which effectively creates local smoothing of the read count) [124-127] in terms of GC content adjustment concentrated on the association between fragment count and GC composition for specified window (or bin) sizes. For a given window, GC content is calculated as the proportion of G and C in that window based on the reference genome as described above. Uniqueness of the sequence was accounted for by estimating a measure of mappability. A python script using *Bowtie* mapper was used to check whether or not a sequence with a known read length is

mappable (i.e. is unique). Then, a GC bias curve is created using a LOESS regression model of count by GC on a random sample of 10000 bins with high mappability (greater than 0.90). Modelling was done using an R package *loess* [125-128]. The work of Benjamini and Speed [123] using two main samples : one sample from the tumour of an ovarian cancer patient and one normal sample from white blood cells, utilized single position models that enables comparison of different possible GC windows and measure the effects of each and compared them in terms of a parameter called Total Variation score (TV score) based on the work of Durrett [129]. Within a library, the fragment length was found to influence the shape of the GC curve. Though not consistent between samples, interaction between GC and fragment length was observe when testing several datasets from different sequencing centres. GC effect was shown to be mostly driven by the GC component of the full fragment. Conditional modelling on the GC of the fragments which defines the strongest bias allows improved correction compared with alternative GC windows [123].

Initially, adjustment of the LMC copy cumber data was performed using the R package *loess* as was used by Benjamini and Speed [123]. Further quality control analysis revealed the need for checking the interaction effects between GC content and mappability to read counts especially in the case of FFPE samples without matched normal.

A newer version of the LMC copy number data was then refined based on the methodology of Scheinin et al. [2] which provides a way of checking for the interaction effects between GC content and mappability using isobar plots as well as to correct for this source of bias using a two-dimensional LOESS model.

In adjusting the read counts, a two-dimensional LOESS model that incorporates the interaction between GC content and mappability was fitted to the median read counts of each 10kb window. The pipeline was available in an R Package called *QDNAseq* produced by Scheinen et al [2]. The LOESS model in this package takes the interaction form of GC content and mappability to predict the median read count in each 10k window bin.

Each median read count per 10kb window was then divided by the LOESS predicted read count to obtain the adjusted read count for each window [2]. This methodology is further discussed in Chapter 5 : Additional Steps to Increase Data Quality.

3.8 Read Count Normalization

Normalization needs to be done in order to remove technical variation (e.g. GC content and sequence mappability) present in the data. GC content for each window was generated from the reference genome excluding bases masked as N. Software named *gem-mappability* was used to calculate for mappability allowing 1 mismatch and a sequence of length 35 bases based on the work of Derrien et al. [130]. They defined mappability as: given some read length of size *k*, the k-frequency $F_k(x)$ of a sequence at a given position x is the number of times the *k-mer* beginning at position *x* appears in the sequence and in its reverse complement, while allowing for some mismatches (e.g. 1 nucleotide base mismatch). Below is the formula used to calculate the *k-mappability* or *k-uniqueness* $M_k(x)$:

Equation 3 $M_k(x) = 1/\ F_k(x)$

The median mappability score for each 35-nucleotide sliding window was then obtained. At first, individual modelling and correction for GC content then for mappability of each chromosome were performed using LOESS before being log transformed.

A separate data set was created that utilized the simultaneous correction for GC content and mappability in the form of interaction using the *QDNAseq* function *estimateCorrection* [2]. This method was found to provide better correction when the two factors adjusted for are found to have interacting effects to read counting. While the initial copy number data adjusted window read counts were based on the total read counts of the chromosome where the window is located, the new data was adjusted based on the whole somatic genome median read count.

Also available for this analysis were seven “normal” samples i.e. paraffin-embedded samples of skin, not from the site of melanoma; these normal samples were chosen as they had been processed in the same manner as the tumour samples. The first composite normal was created from these seven normal samples; the read count for each window was created by taking the median of the adjusted window counts for these seven samples. None of these seven normal samples are matched to any of the LMC tumour samples.

A second composite normal was derived from the 312 samples obtained from the 1000 Genomes Project. Normal samples were adjusted in the same way as the individual tumour samples. Finally, all the individual tumour samples were normalized

to the median corrected consensus normal to derive the corrected read score by taking the log2 of the tumour read counts after being divided by the median read count of the normal samples. Three sets of data for evaluation were derived namely: Data1- read counts were adjusted sequentially for GC content and mappability and normalized to composite normal based on 7 normal skin tissue samples adjusted similarly, Data2 – read counts adjusted concurrently of GC content and mappability then normalized to composite normal based on 7 normal skin tissue samples adjusted similarly, Data3 - read counts adjusted concurrently of GC content and mappability then normalized to composite normal based on 312 1000 Genomes Project samples adjusted similarly.

3.9 Blacklist Windows

Studies have shown that NGS read alignment to specific areas of the genome is poor as exemplified by peri-centromeric and low complexity regions [131]. Copy number data generated in these regions are deemed unreliable. This results to the need to identify and exclude these regions in the analysis.

At the time the initial analysis was done, there were not much available resources about the blacklisted regions in the genome of the reference build GRCh38. Two filters which we utilised in our previous study[1] were initially employed: (1) regions defined as “problematic” (at threshold of top 0.01% of the distribution of coverage at each base (using 500 Mb of sequence) of the sequence data from 1000 Genomes Project data [113] by Pickrell et al. (2011) marking areas in the genome that yield spurious copy number peaks [131] and (2) large regions with low coverage as identified by checking for 10k windows with zero reads in more than 5% of the samples and subsequently excluding any runs of length 150 or less windows between marked windows [1]. The first filter identified 446 out of 308,837 (0.14%) windows at 10k base-pair resolution while the second filter identified 32,852 out of 308,837 (10.64%) that includes the whole of chromosome Y yielding an initial count of 33,096 blacklisted windows (10.72%) at 10k resolution.

A list of modelled centromeres and heterochromatins obtained from *https://www.ncbi.nlm.nih.gov/grc/human* (Genome Reference Consortium) and a list of gaps in the human genome taken from *http://genome.ucsc.edu/cgi-bin/hgTables* were considered as blacklisted regions denoted as Centrogaps. The *QDNAseq* pipeline which is based on residual filter calculation after adjusting for GC content and mappability was used to identify windows that were highly variable in the genome using the 312 control samples from 1000 Genome Project.

Considering only the autosomal genomes, Centrogaps identified 17,904 10k windows while the residual filter identified 35, 856 10k windows to be included in the blacklisted regions. All windows in the Centrogaps list were also found in the list based on the residual filter. A total of 38,215 unique 10k windows from our dataset, including the 2,359 unique windows from the earlier method were considered as the final blacklist accounting for 13 % of the autosomal genome.

3.10 Calculation of Copy Number

Following the exclusion of blacklisted windows and the creation of an adjusted read count for each window, all the adjusted window read counts were normalise to the median of their respective genomes using the *QDNAseq* function *normalizeBins* [2]. Then, the copy number *LR* in each window is calculated as a log2 transformed copy number ratio between a tumour and germline reference as below:

Equation 4 $$\boldsymbol{LR} = \boldsymbol{log_2}(\frac{\boldsymbol{CN_{Tumour}}}{\boldsymbol{CN_{Reference}}})$$

Where *CNTumour* = the adjusted read count of a given sample in a given window, and

CNReference = the adjusted read count of a given sample in a given window of the reference samples which could either be the 7 LMC or the 312 1000 Genomes Project samples. Throughout this thesis, the term copy number that pertains to copy number data used in my analyses refers to *LR*, the log2 ratio between tumour copy number and the reference (normal tissue) copy number.

3.11 Segmentation of CNA data

Copy number segmentation refers to the process of splitting the chromosome into regions of equal copy number (segments) that accounts for noise by smoothing the data. In this study, we used the widely used segmentation algorithm Circular Binary Segmentation (CBS) which was originally developed by Olshen et al. [132]. This method provides a natural way to assign chromosome segments as contiguous regions and bypasses parametric modelling of the data with its use of a permutation reference distribution [132].

Let $W_1, ..., W_m$, be the adjusted read counts corresponding to the m markers or windows for the segment being considered. The test statistic is obtained by the maximal t-statistic given by $T=\max_{1\le i<j\le m}|T_{ij}|$, where $|T_{ij}|$ is the two-sample t- statistic to compare the mean of the observations with index from $i + 1$ to j , to the mean of the rest of the observations. The formula is given below:

Equation 5 $$T_{ij} = \frac{\bar{Y}_{ij} - \bar{Z}_{ij}}{s_{ij}\sqrt{\frac{1}{j-i}+\frac{1}{(m-j+1)}}},$$

Where $\bar{Y}_{ij} = (X_{i+1} + \cdots + X_j)/(j-1)$, $\bar{Z}_{ij} = (X_{i+1} + \cdots + X_i + X_{j+1} + \cdots + X_m)/(m-j+1)$, and s_{ij}^2 is the corresponding mean squared error. Note that if we treat the segment being tested as indexed by a circle by connecting its two endpoints then the method tests whether there are two complementary arcs that have different means. The change is deemed statistically significant if the P-value is smaller than the set significance level $\propto$ (typically 0.01) and identify the locations of the change-points as the i and j (if $j<m$) that maximise the test statistic [136].

The above process is followed by the process of recursively identifying the change points for two arcs. Once the change points are identified, the location of the change point and the average read count (segment mean) between the two consecutive change points are obtained [132]. This method segments data by checking for change-points using a maximal t-test but takes more computational time to evaluate the significance of the change-points [133]. A faster version of this algorithm was implemented in the *DNAcopy* package of the R Bioconductor project. The better computational speed of this method was obtained by using a hybrid approach to obtain the significance of the t-statistic in linear time. Additionally, an early stopping rule was introduced when there is strong evidence for the presence of a significant change [134].

3.12 MLPA analysis

The analysis presented in this thesis focuses on the statistical analysis of FFPE derived melanoma tumours. This approach was adopted because of the paucity of other techniques for examining copy number among such small, formalin fixed tumours. To validate the results is desirable but molecular analysis other than examining a focused region of the genome is prohibitive in terms of DNA content.

With that in mind, we used a known technique termed Multiple Ligation Probe Amplification (MLPA) which uses test and control probes and examines the ratio of test to control probes using qPCR. MRC Holland ™ produce a kit targeting the *CDKN2A* region of the genome ("SALSA MLPA Probemix P419 *CDKN2A/2B-CDK4*"**)**, a region of considerable interest for melanoma.

Tumours with sufficient available DNA which were subject to this analysis used samples taken from as close as possible to the core taken for the copy number assessment. The kit was run as per the guidelines by Dr. Joanna Pozniak within our laboratory and she also performed the statistical analysis; she provided me with estimated copy number at the site of each of the probes. These numbers are incorporated into the comparative analysis of MLPA and NGS sequenced-based analysis.

Chapter 4
Assessment of CNA Data Quality Phase 1

The aim in this chapter is to:

- Provide an initial assessment of CNA data quality and decide whether it merits proceeding with further analysis.

4.1 Introduction

The standard quality measures applied in the generation of the copy number data are detailed in Chapter 3 : Methodology. This chapter aims to assess the quality of the data by, firstly, identifying the appropriate window size to be employed in further analyses. To do this, the replicates are explored in terms of the cases' copy number profiles such as average number of segments, average segmented length, raw counts, and adjusted/corrected read counts. A description of the *mcnv* R package is also provided which is the main tool used in visualising and manipulating the copy number data [135]. Finally, a comparison of LMC CNA data with published TCGA dataset on copy number variation of melanoma was performed.

4.2 Methods

4.2.1 The mcnv R Package

The *mcnv* version 1.5-22 R package created by Dr. Alastair Droop was used in majority of the copy number profile visualization. This package was primarily designed for the analysis and manipulation of the Melanoma CNV Project data. A basic tutorial including installation instruction of this package is available at http://alastair-droop.github.io/MCNV-tutorial/. Visualisation of copy number profiles provided better overview in manually assessing the quality of data, as well as copy number patterns across samples.

4.2.2 Selection of Window Size

To focus the analysis of copy number, the changes generated with different window sizes (1kb, 5kb, 10kb, 100kb, and 1mb) for the *CDKN2A* regions were assessed and compared. Segmented copy number profiles were generated in all the provided data window sizes and the resolution that provided the reasonable compromise between information content and amount of noise was chosen for further analysis. This is done by analysing a focused region in the genome covering the *CDKN2A* region where with the most common form of deletion in melanoma and most cancers is located [71]. I then checked how the known *CDKN2A* deletion (validated using MLPA as discussed in Section 3.12) in this region was clearly identified in different window resolution/sizes. The window size or resolution that gives the clearest depiction of the variation in this region was selected for further analysis. The whole genome copy number profiles where then assessed and showed that the 10k window resolution has the clearest depiction of overall known aberrations in the melanoma genome. Both the creation of copy number windows (or bins), and segmentation of the whole genome copy number data are described in Chapter 3 (Section 3.6 CNV Data Windows, and Section 3.11 Segmentation of the CNA Data).

4.2.3 Assessing similarity of replicates

As previously stated in Chapter 3 (Section 3.3) a total of 60 samples (30 patients) were used in the first set of replicate analysis. Descriptions of these replicates are also in Chapter 3. Randomly selected pairs of replicates were plotted and compared in terms of whole genome copy number profile. Then, a visualisation of the level of similarity among the different types of replicates was done by plotting the mean number of fragments per chromosome for each pair of samples such as core, tumour, concentration, method, and technical. Pearson's r was calculated for samples group as to cores or technical replicates, then for all the replicates. A more linear pattern between the mean number of fragments per chromosome for each pair of replicates indicates better consistency of the data.

A window level analysis of the adjusted read counts for each pair of replicates was also performed using Pearson's r to assess similarity as previously done our initial copy number paper [1].

4.2.4 Calculation of Number of Segments and Segmented Length

Ten out of 355 samples were rejected due to very low alignment rates (3.1×10^6 – 34.3×10^6 aligned reads retained) after filtering, leaving 345 samples which were retained for further quality control steps. For these samples, segments derived from

circular binary segmentation (CBS) reflect information available for each sample. The appropriate metrics for segmented copy number data are: the number of segments for each sample, the segment length and the alignment with existing literature. These metrics represent the copy number profile with lesser noise and facilitates assessment of data quality and copy number patterns. To calculate the average number of segments per chromosome, all segments per chromosome of each sample were counted and divided by the total number of samples analysed (n=345). Similarly, average segmented length was calculated by taking all base pairs aligning to the segments per chromosome and dividing by the number of samples analysed (n=345).

4.2.5 Linearity of Chromosome Segments and Segmented Length

For normal samples, it is expected that the somatic chromosome segment changes uniformly with its size. The expectation is different for cancer samples where segment breaks were observed across the genome. Under the assumption that the majority of chromosomes do not harbour specific genetic changes associated with melanoma development and that the chromosomes generally varies directly with its size (which the exemption of chromosome 21 and 22 where the latter is longer), we would expect there to be a linear relationship between the length of the chromosome and number of fragments (or breaks or segments). Exceptions will apply of course if particular chromosomes contain regions where copy number changes are required for melanoma development. This analysis explores the linearity of the relationship between the average number of segments across all samples (plotted on the Y-axis) and the average total segmented length (plotted on the X-axis) for each chromosome. Simple linear regression was applied to examine goodness of fit.

4.2.6 Examination of the esv3620012

Visual inspection of the copy number traces identified common variation in copy number was identified within the region 9p21 which is very close to the location of the *CDKN2A* region (*9p21.3*). Review of genomic databases suggested that this variation represented a previously documented germline copy number variation (esv3620012; chr9: 23,362,412–23,378,071) which is about 1.4 Mb from *CDKN2A*. This variation consists of the presence or absence of a 16kb region [136] (www.ensembl.org/Homo_sapiens/StructuralVariation/Explore?r=9:23362314-23378180;sv=esv3620012;svf=114158056;vdb=variation). This variation is documented in 1000 Genomes Project dataset. Based on genome-wide genotyping of germline samples from the Leeds Melanoma Cohort, followed by imputation of the presence based on the 1000 Genomes Project panel (conducted by Dr. Mark Iles) identified persons within the Cohort who carried the mutation (the copy number variant

was in complete LD with several SNPs including *rs4977836*). Initial analysis on *esv3620012* within the NGS dataset within the two 10kb windows which covered the sites of the deletion were examined and compared to the number of reads in those two windows with the average adjusted read count from the 10 adjacent windows each side of this copy number variation. This was visualised using histogram grouped by genotype to check for difference in distribution and confirm the linkage disequilibrium with *rs4977836*.

4.2.7 Comparison of NGS versus MLPA data

A more detailed description of the MLPA analysis conducted by Dr. Joanna Pozniak was stated in Chapter 3 (Section 3.12 MLPA Analysis). Results from this experiment were visually compared with the NGS data which are the main data of this project. Two sets of copy number data from these two sources were compared focusing on the *CDKN2A* region. The copy number plot was created using the *mncvplot* function from the *mcnv* package [135]. This is aided with the function *geom_segment* from the *ggplot2* package [137] to draw the MLPA data points annotated with vertical lines where the endpoints represent the boundaries of the 95% confidence interval about the mean. The plotting code created by Dr. Alastair Droop was used as a guide in this analysis.

4.2.8 Comparison between LMC and TCGA CNA Data

The *cBioPortal for Cancer Genomics* has published a list of regions that are deleted or amplified in metastatic skin cutaneous melanoma based on the work of the TCGA Research Network (http://cancergenome.nih.gov/) [5, 6, 138]. Three genomic characteristics were specified in the TCGA file namely:

- region start and region end,
- peak start and peak end, and
- enlarged peak start and enlarged peak end.

Gene level copy number data available for 367 samples were downloaded from http://www.cbioportal.org/study?id=skcm_tcga#summary.

For the initial analysis, the region described by “peak start” and “peak end” and genes covered by these were identified in LMC data. For each gene in each region, the locus was obtained using the R package *bioMart* [139, 140]. The adjusted read count was average across the LMC copy number data windows covered by each given locus. The proportion of samples with deletion or amplification in TCGA and LMC were calculated and their distributions were compared using two-sided bar plots

representing TCGA proportion of samples with aberrations on one side and LMC on the other.

4.3 Results

4.3.1 Selection of Window Size

The copy number profile for the *CDKN2A* region in different window sizes is shown below in Figure 4.1. The x-axis represents the genomic region around the *CDKN2A* region while the y-axis represents the log2 adjusted copy number for the selected sample. The top plot presents the copy number profile of the *CDKN2A* region, known to be commonly deleted in melanoma and most cancers with 1 megabase resolution (1mb window) while the plot at the bottom represents the 1 kilobase (1kb window) resolution. Each dot on the plot represents the log2 copy number of the selected sample in a given window. The colour represents segments; adjacent dots similarly coloured belongs to the same implied segment while dots that have different colours have different copy number and belong to separate segments. It can be seen that copy number plot for 1mb resolution does not identify the *CDKN2A* loss in the sample while a moderate loss was detected in 100kb window. Significant deletion was detected by the segmentation algorithm in 10kb, 5kb, and 1kb windows resolution. To select the resolution for further analysis, a compromise between information content and amount of noise was considered. This suggests that the 10kb resolution should be chosen as the appropriate window size for further analysis of the copy number data as it is able to detect the *CDKN2A* region with limited noise. Additionally, a known common germline variation (dots between the green and pink dots on the 10kb window) called *esv3620012* was observed only in this resolution. A further analysis on this ESV is discussed on this chapter. Results for other samples mimicked Figure 4.1. This is in line with the previous study of Gusnanto et al. (2014) which looked at estimating the optimal window size for the analysis of low-coverage next-generation sequence data based on Akaike's information criterion (AIC) and cross-validation (CV) log-likelihood by plotting the AIC and CV log-likelihood curve as a function of window size. They concluded that in their 2 datasets (LS041, and LS010), the optimal window size is the one which has at least 60 reads per window [119].

Figure 4.1. The *CDKN2A* Region. Copy number profile shown in different window sizes (from bottom to top: 1kb, 5kb, 10kb, 100kb, and 1mb)

4.3.2 Whole Genome Copy Number Visualisation

A visualisation of the whole genome copy number profile in each sample was generated to assess the distribution of the log2 adjusted read counts for each sample. Below is a whole genome copy number plot based on the 10kb window. The whole genome is divided into segments which are regions with equal copy number. A clear separation between 1p segment and 1q segment can be observed for this sample (Figure 4.13). Using zero as the baseline normal copy number, chromosome 1p has lower copy number than expected while chromosome 1q has higher copy number than expected at this resolution.

4.3.3 Assessing similarity of replicates

Each pair of technical replicates (technic, method, and concentration) and biological replicates (tumour, core) were plotted in Figure 4.2 to Figure 4.11. The tumour replicates showed differences in the segmented genome particularly in chromosomes 6p and 9p regions and could more likely be due to the inherent biological variability of the tumours (Figure 4.2). Comparison between these two tumour replicates using scatterplot (Figure 4.3) shows moderate correlation (Pearson's r=0.33, $P<2.2x10^{-16}$) of their respective 10k window copy number ratios. The comparison of paired core replicates as depicted in Figure 4.4 shows overall similarity of the genome with some differences in segmentation patterns that can be noticed in the regions of 2p and 6q. Comparison between these two core replicates using scatterplot (Figure 4.5) shows strong correlation (Pearson's r=0.67, $P<2.2x10^{-16}$) of their respective 10k window copy number ratios. The pair of technical replicates showed highly similar genomic profiles (Figure 4.6). Comparison between these two technical replicates using scatterplot (Figure 4.7) shows very strong correlation (Pearson's r=0.81, $P<2.2x10^{-16}$) of their respective 10k window copy number ratios. The plot for a pair of samples processed using the different laboratory methods for library construction is shown in Figure 4.8. The two samples show highly similar patterns of aberrations across the genomes. Comparison between these two method replicates using scatterplot (Figure 4.9) shows strong correlation (Pearson's r=0.76, $P<2.2x10^{-16}$) of their respective 10k window copy number ratios. Finally, Figure 4.10 shows the plot of paired samples processed using different concentrations. The pair of samples displays highly similar genome plots indicating good consistency and quality of the data. Comparison between these two concentration replicates using scatterplot (Figure 4.11) shows strong correlation (Pearson's r=0.77, $P<2.2x10^{-16}$) of their respective 10k window copy number ratios.

It is visually evident that the genome profiles tend to be heavily affected by noise. Analysis of the replicates was reperformed in Chapter 6, after further data quality

control steps were implemented. An overall visualisation of the level of similarity among the different types of replicates used is displayed in Figure 4.12. This shows the number of fragments of one replicate plotted against of that of another replicate. An overall correlation of 0.74 ($P=3x10^{-6}$ for assessing deviation from randomness) indicate a high level of similarity among the replicates. This measure was heavily affected by the other observations taken from the biological replicates (tumour, core) which were observed to be more variable or heterogenous and less “consistent”. Analysis of replicates including only the technical replicates (technic, method, and concentration) show very high correlations (Pearson’s r=0.78) of the two replicates (P=7.8x10-5) as expected because these are more controlled and homogenous in comparison with the biological replicates.

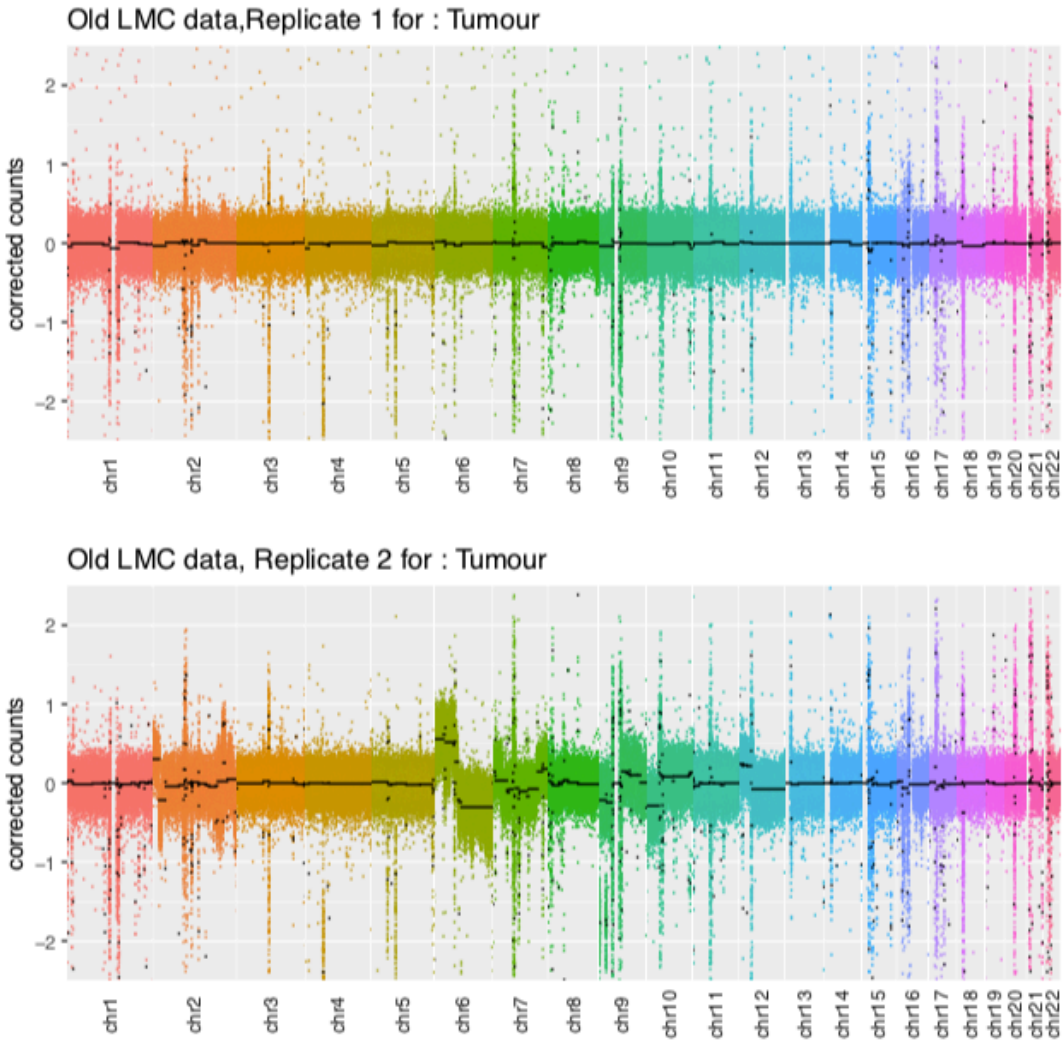


Figure 4.2. Comparison of Analysis of 2 tumours from same case, showing notable differences including the 6p and 9p regions

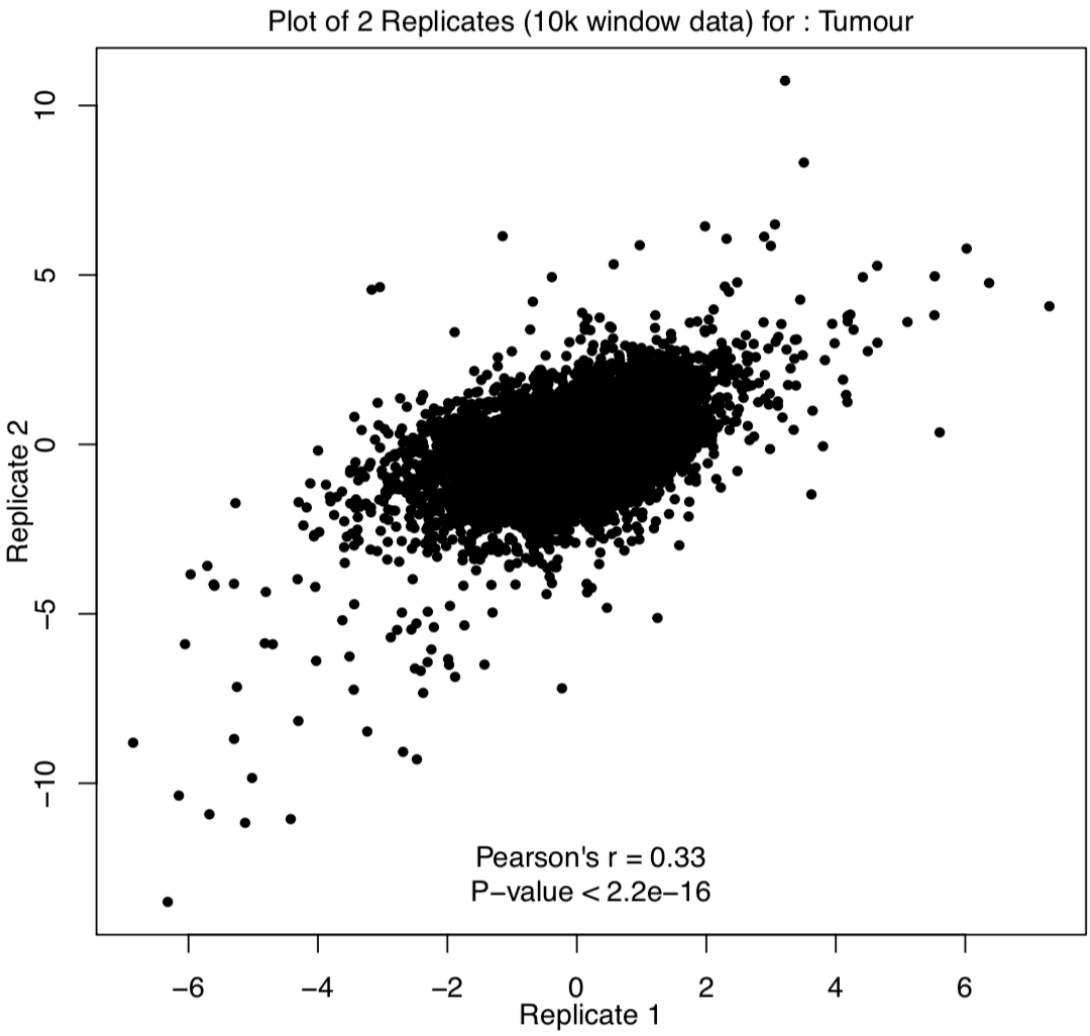


Figure 4.3. Scatterplot of 2 tumours showing moderate correlation of copy number ratios at 10k window resolution

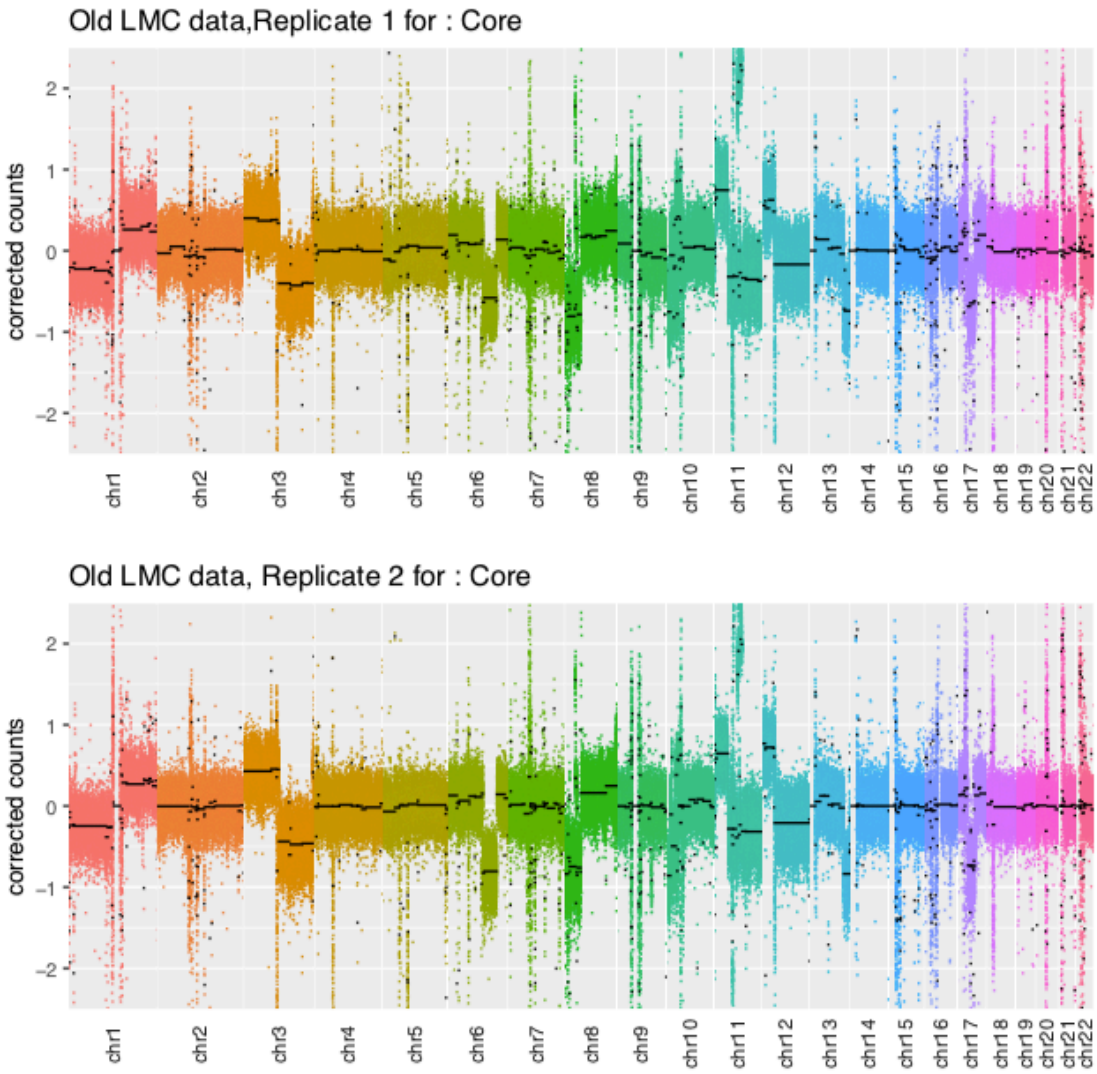


Figure 4.4. Analysis of 2 cores from the same tumour showing overall similarity but with some differences in segmentation pattern (e.g. chromosomes 2p & 6q)

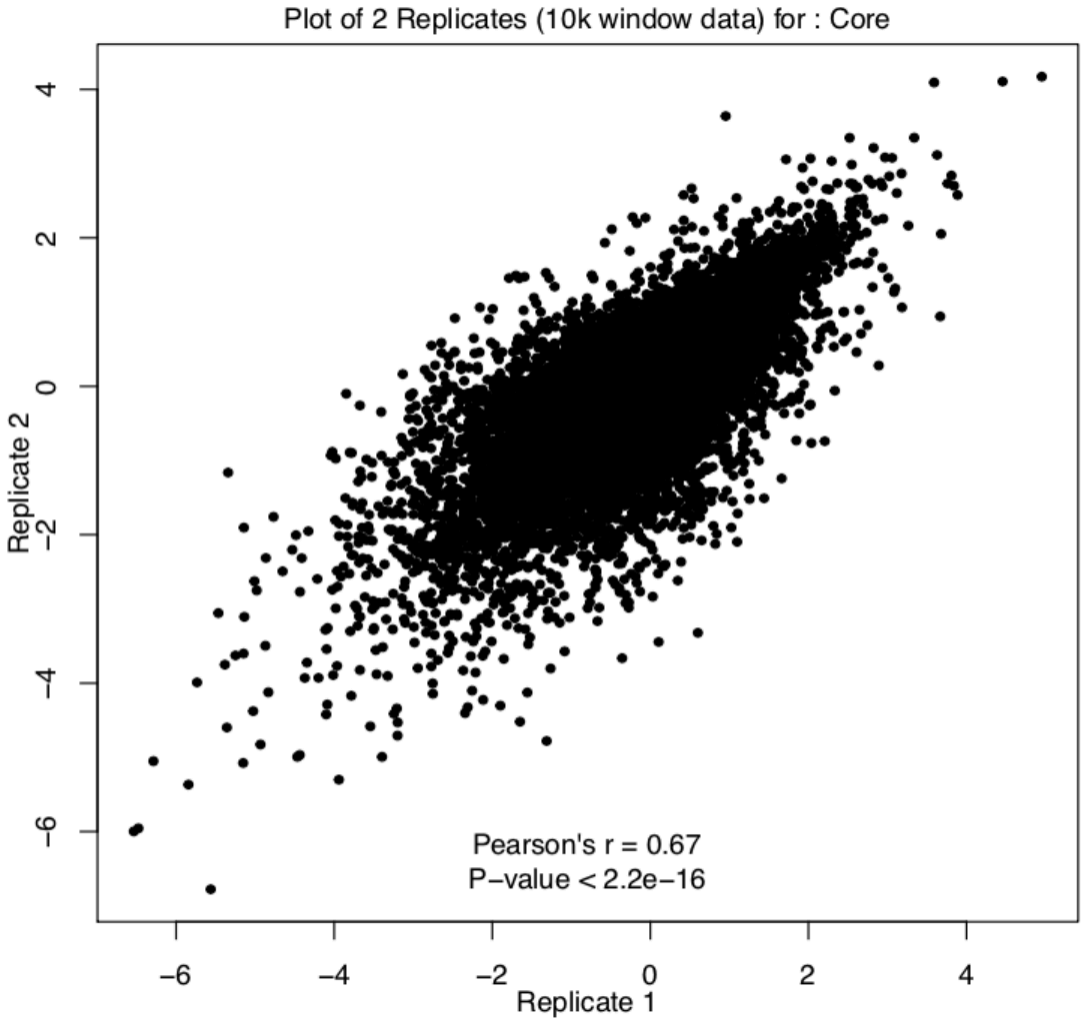


Figure 4.5. Scatterplot of 2 cores from the same tumour showing strong correlation of copy number ratio at 10k window resolution

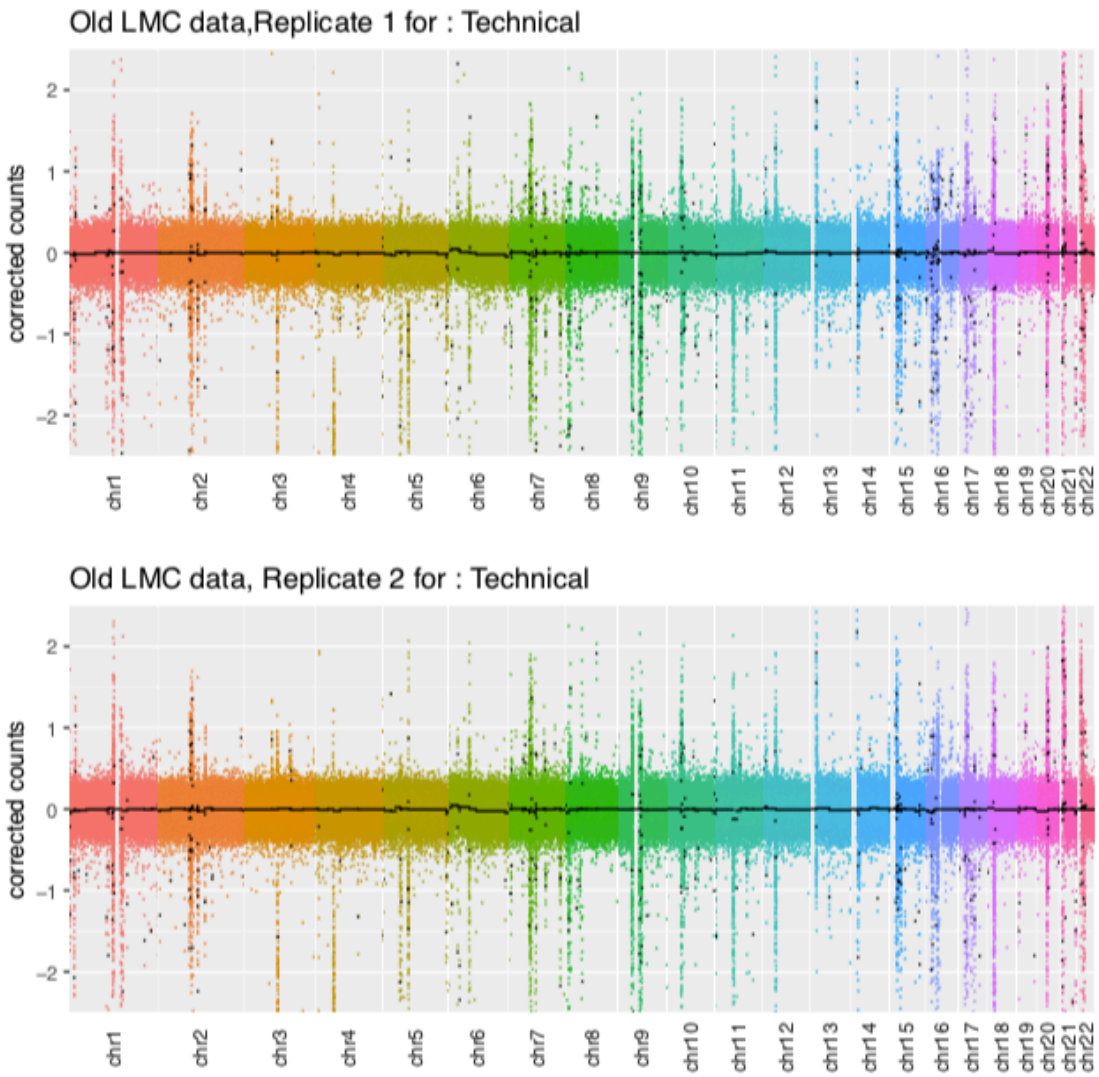


Figure 4.6. Analysis of the same library sequenced twice in this comparatively silent (in copy number terms) tumour showing consistency.

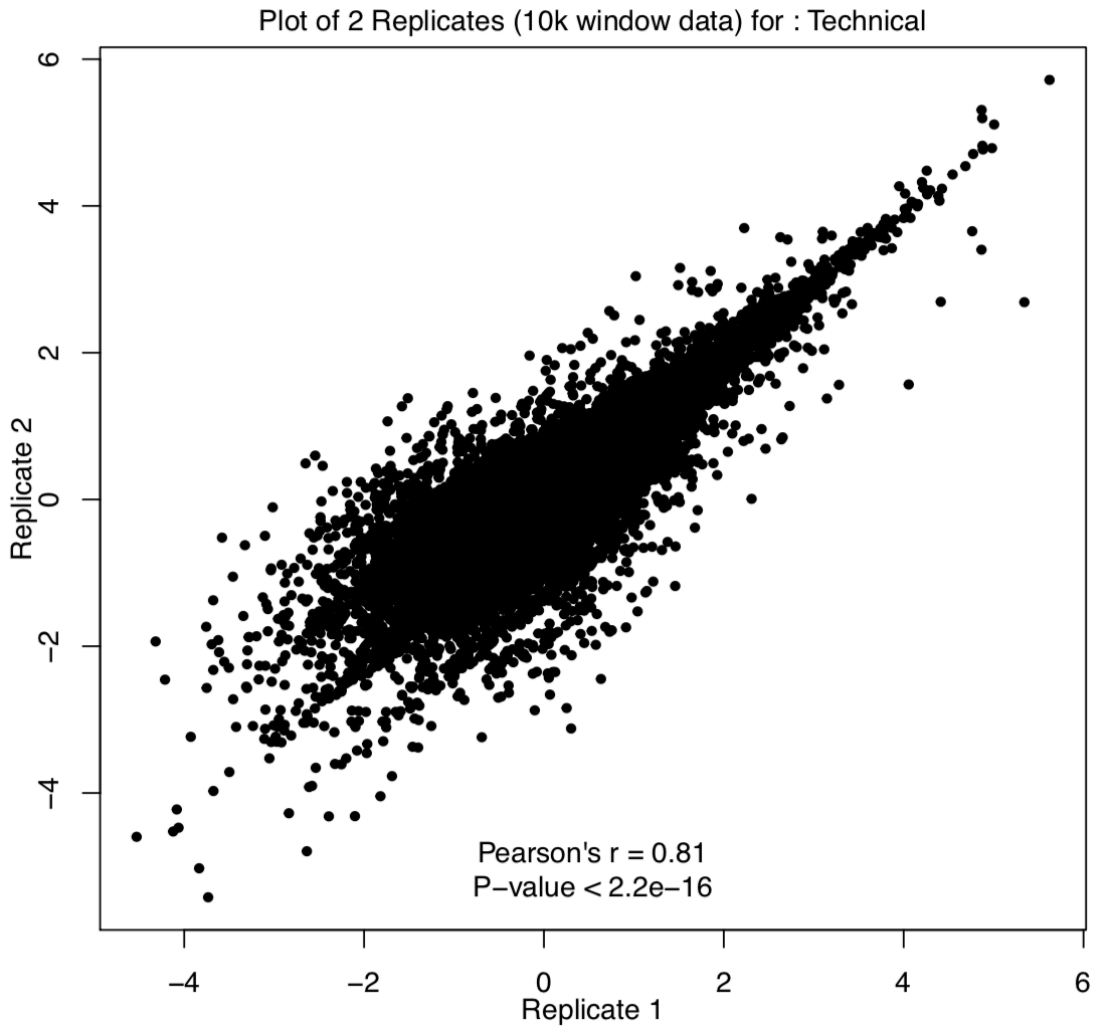


Figure 4.7. Scatterplot of 2 tumours as technical replicates showing very strong correlation of copy number ratios at 10k window resolution

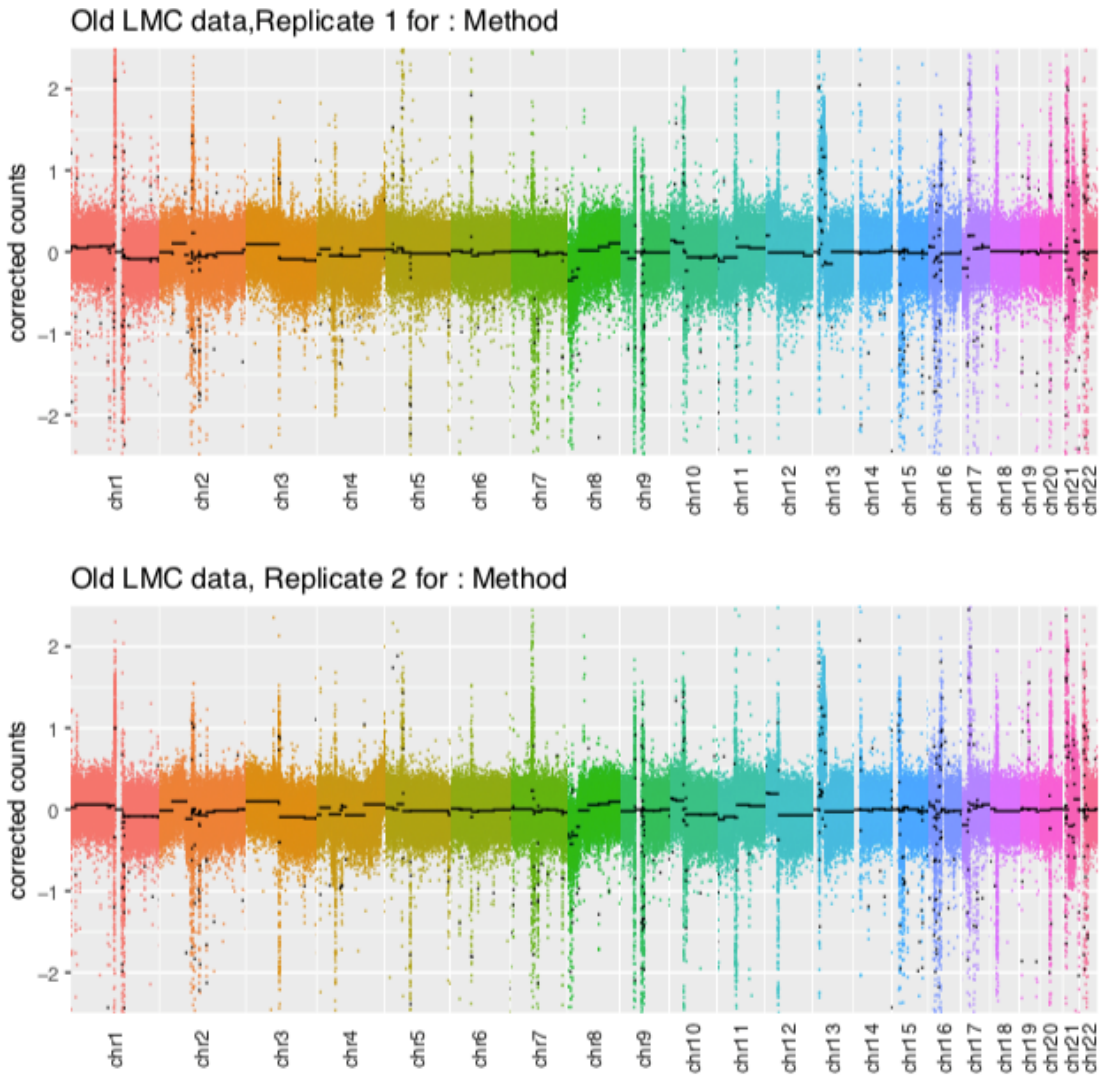


Figure 4.8. Analysis of the same core with libraries prepared by different laboratory methods showing overall similarity in this tumour but some modest differences (e.g. size of segmented regions).

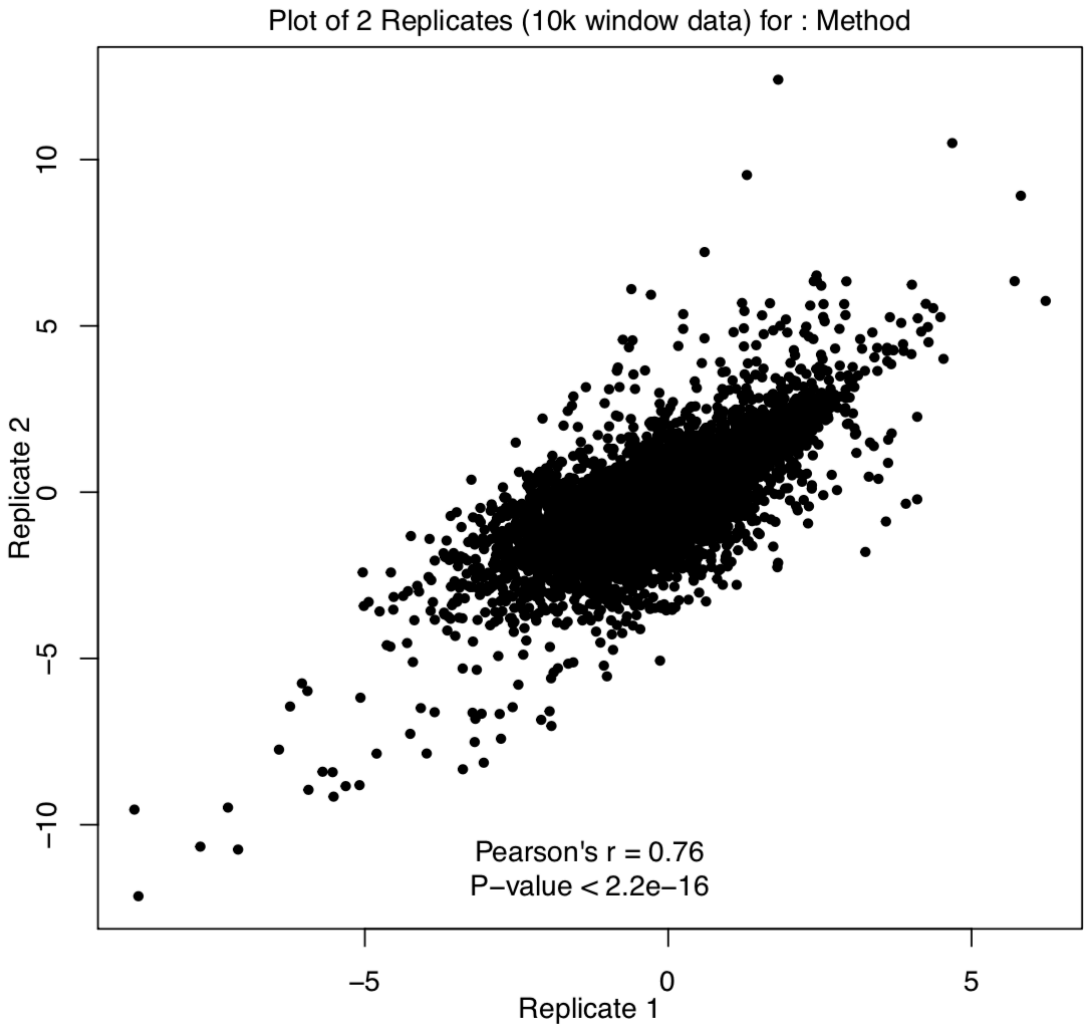


Figure 4.9. Scatterplot of 2 tumours as method replicates showing strong correlation of copy number ratios at 10k window resolution

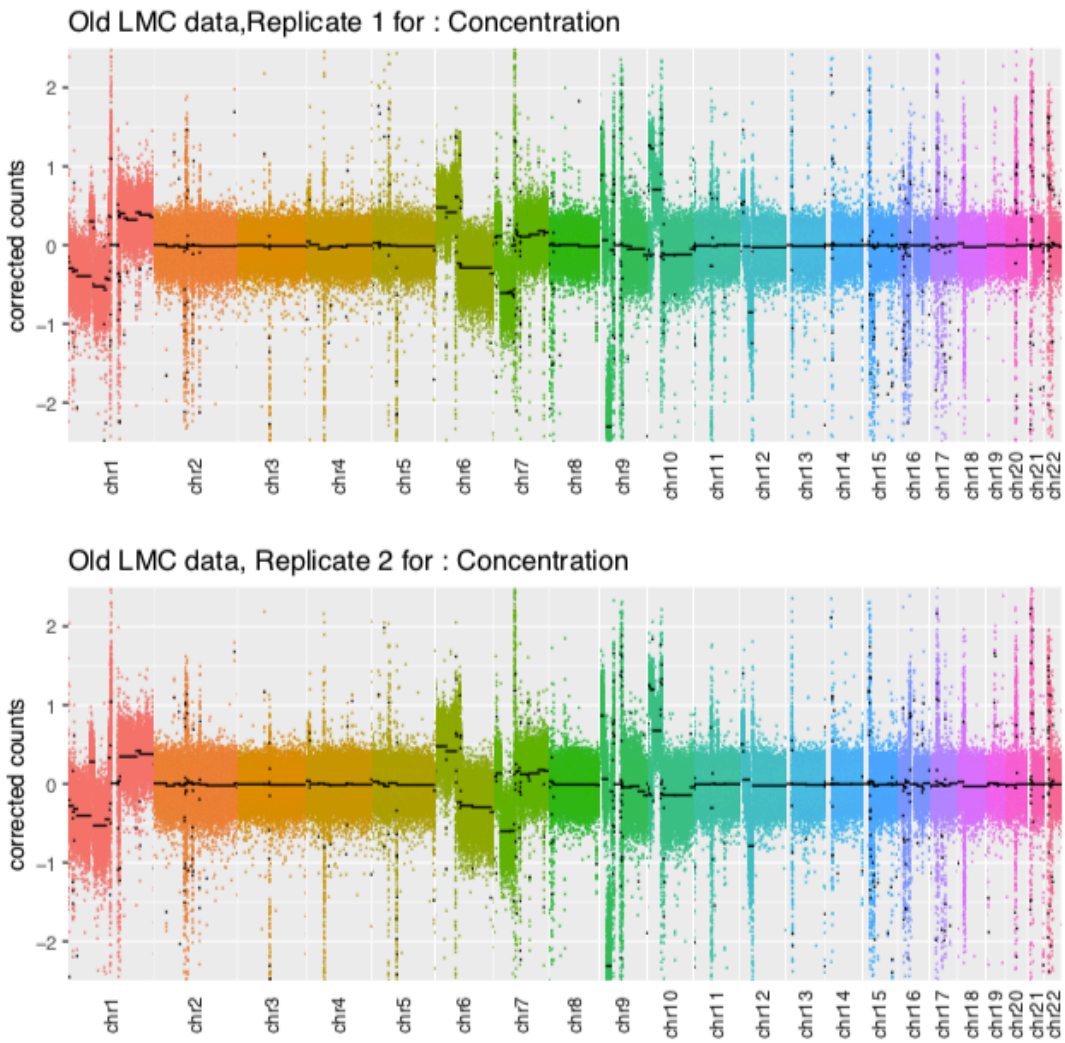


Figure 4.10. Analysis of the same library analysed at two different concentrations showing overall consistency.

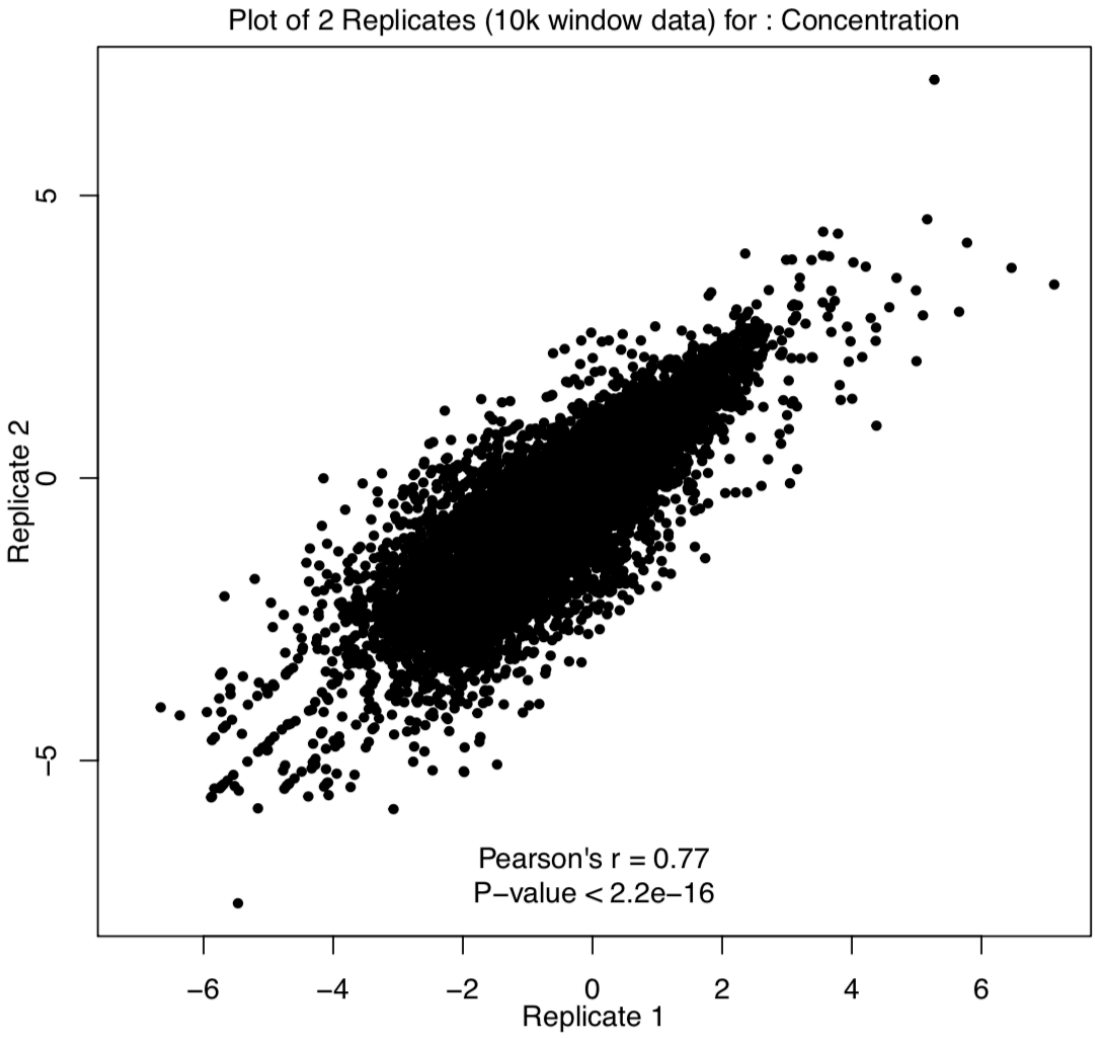


Figure 4.11. Scatterplot of 2 tumours as concentration replicates showing strong correlation of copy number ratios at 10k window resolution

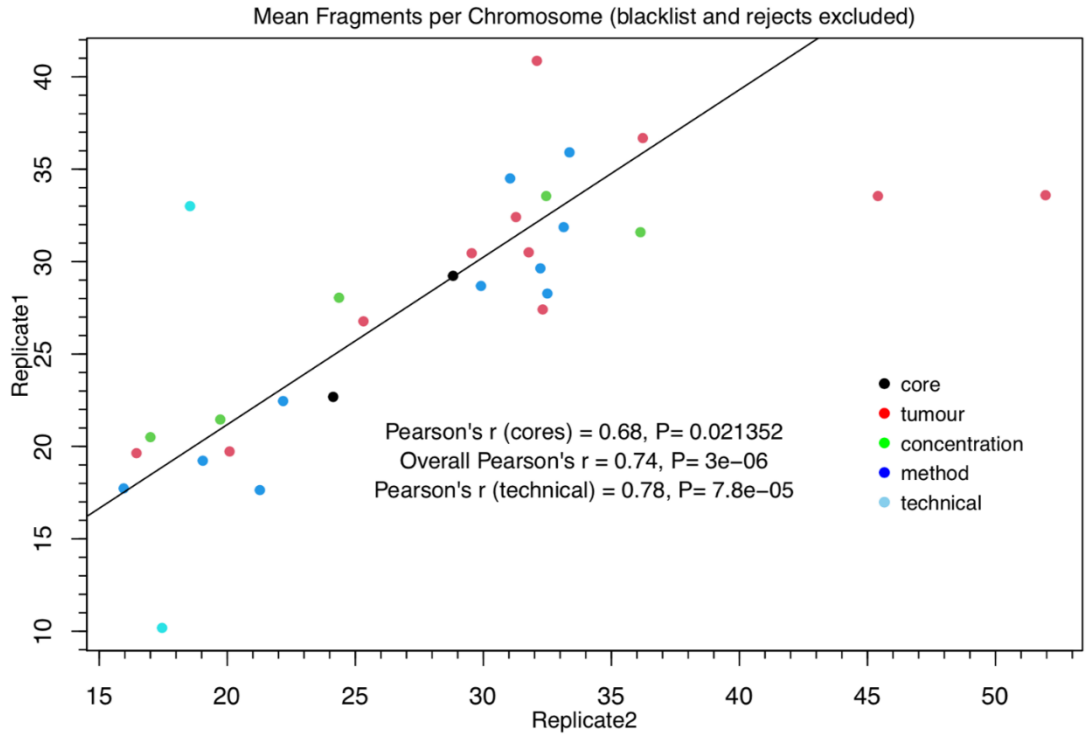


Figure 4.12. Plot of different replicates.

There was a total of 34 samples which were replicated at least twice. Three of these were replicated thrice while the rest are replicated twice resulting to a total of 71 replicated samples. Of the 71 samples, only the top 2 samples of the triplicates were selected in terms of highest mapped reads. Of the remaining 68 samples, 8 samples (from 4 patients) were rejected due to very low coverage. A total of 60 samples (30 patients) were used in the first set of replicate analysis. Of these, 34 were “technical” replicates (technical=20, method=10, concentration=4) and 26 were biological replicates (core= 22, tumour=4)

4.3.4 Chromosome Level Copy Number Visualisation

The whole genome profile consistently presents peaks that are potentially noise based on the fact that most other samples do not have such changes; of course, by the nature of random variation, rare changes of copy number may reflect causal changes but this will not be feasible to disentangle with modest numbers of samples. To better understand this, I checked for the chromosome specific plots of this sample. Below are the plots for chromosomes 1 and 2 (Figure 4.14). This allowed me to inspect the copy number profile in more detail as to the amount of noise present and the location of the noisy regions in the genome. The plots clearly showed the observe peaks of noise are located on or near the telomeric and centromeric regions of each chromosome whose mapped sequence reads are known to be difficult to assemble in the genome. These regions were identified and needs to be masked in the genome as part of the blacklisted regions as done in Chapter 5.

Normal	no
Passed QC	yes
Window Size	10k

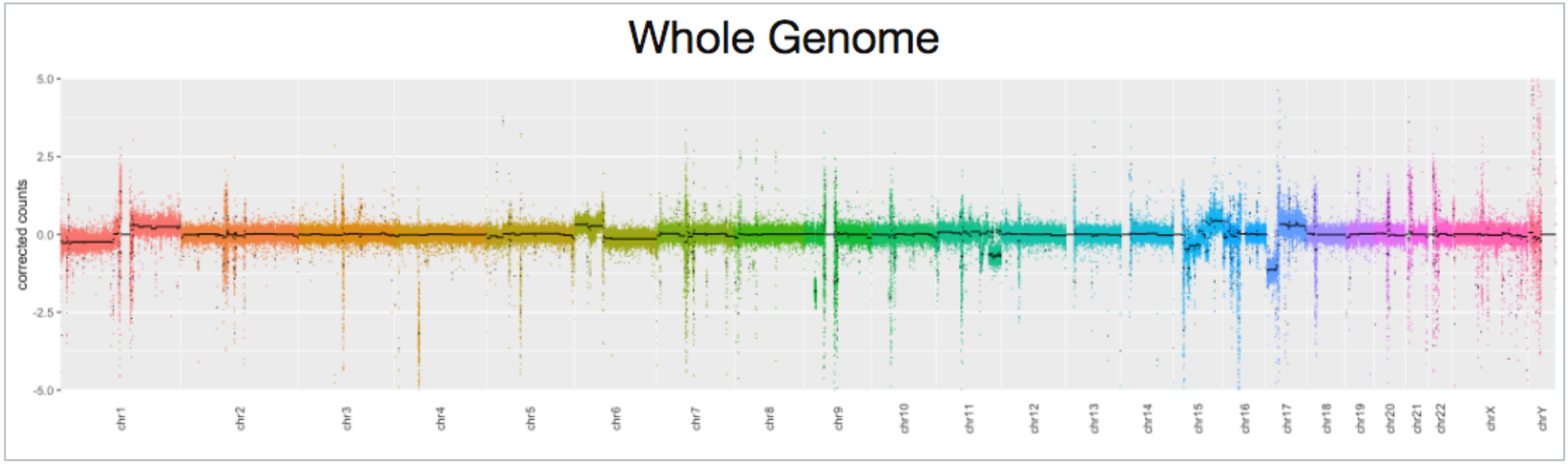


Figure 4.13. Whole genome copy number profile for Sample S1.

The X-axis represents the chromosomal regions while the Y axis represents the log2(adjusted read counts). Hence, 0 on the Y-axis represents no copy number change.

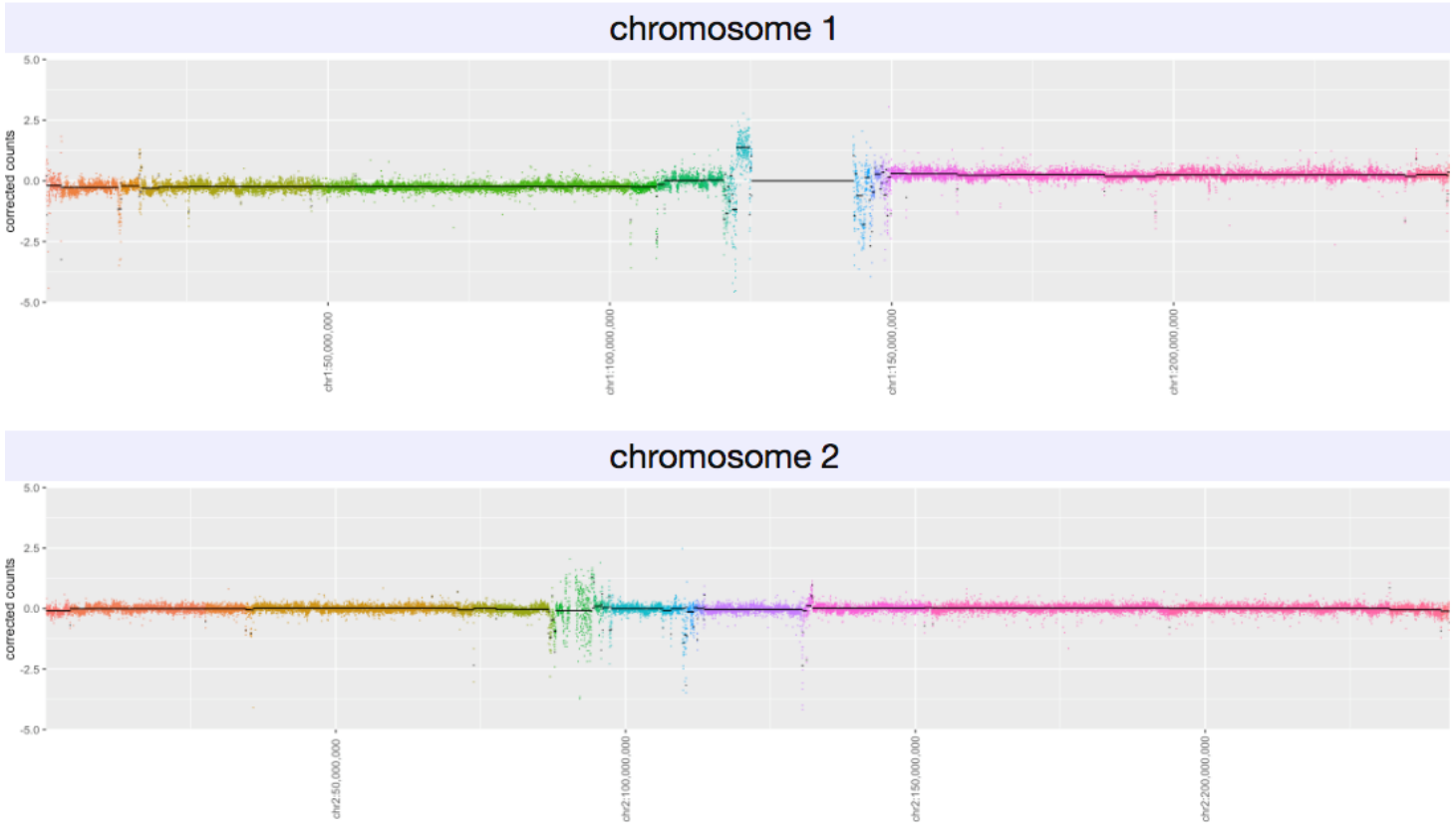


Figure 4.14. Chromosomes 1 and 2 copy number profile for Sample S1.

X axis represents the chromosomal regions while Y axis represents the log2(adjusted read count). Chromosome level copy number profile for Sample S1. X axis represents the chromosomal region while Y axis represents the log2(adjusted read count). Chromosome 1 is plotted on top while in the bottom is chromosome 2.

4.3.5 Calculation of Number of Segments and Segmented Length

The mean segmented length and number of fragments (or segments) per chromosome is summarised in Table 4.1 below across all samples. A higher variability in number of fragments as measured by standard deviation in relation to chromosome size is observed in chromosomes 6, 7, 11, and 16 with a standard deviation of 16.8, 19.5, 15.3 and 13.5 fragments ranging between 1 and 167 fragments. While these chromosomes are known to be associated with melanoma (i.e. chromosome 6 where *G9a* which was shown to drive oncogenesis is located[141] ;chromosome 7 is where the *BRAF* gene is located, a commonly mutated gene in melanoma [67, 69, 93, 142]; chromosome 11 where *CCND1* which is associated with melanoma progression is located [93, 143-146] ; and a locus in chromosome 16 was shown to be associated with melanoma risk[147]), the amount of variability observed in the genome in general merits further investigation.

Table 4.1. Segmented length and Number of Segments/Fragments across all samples.

Chromomsome	Mean Segmented Length(Base pairs)	No. of Fragments				
		Mean	Median	SD*	Min	Max
1	248,555,941	40.2	37	20.7	1	240
2	241,827,330	36.9	36	19.2	1	190
3	198,132,450	16.5	14	11.1	1	70
4	190,068,547	14.9	13	11.6	1	99
5	181,287,637	25.2	22	12.6	1	95
6	170,600,585	20.7	17	16.8	1	167
7	158,902,091	44.6	43	19.5	9	165
8	144,951,119	18.8	17	10.5	1	86
9	138,210,480	18.7	17	11.6	1	92
10	133,587,544	21.1	20	10.2	1	88
11	134,867,112	22.1	19	15.3	1	135
12	133,146,051	13.2	11	11.4	1	120
13	114,237,078	13.0	13	7.4	1	92
14	106,959,904	8.7	7	6.1	1	37
15	101,731,892	26.4	26	10.4	7	98
16	89,965,778	37.3	37	13.5	1	79
17	83,007,195	25.2	24	11.7	1	95
18	80,299,720	7.7	7	4.8	1	31
19	58,502,186	11.7	11	6.4	1	49
20	64,385,737	6.1	5	4.8	1	39
21	46,662,901	4.7	4	3.9	1	46
22	50,676,377	14.3	14	8.5	1	111
X	155,711,898	15.5	15	7.6	1	64
Y	57,070,111	1.0	0	0.0	1	1

*SD: Standard Deviation

s

4.3.6 Testing for linear relationship of replicates

Aside from the linear relationship between average number of segments and average segmented length per chromosome, this section allowed me to assess the influence of different factors (blacklisting, exclusion of rejected samples) in the overall copy number profile of the patients. Figure 4.15A shows the plot without excluding the blacklisted regions and the rejected samples. Figure 4.15B shows the plot excluding the 10 rejected samples. Figure 4.15C and Figure 4.15D show plots excluding the 10 blacklisted regions only, and excluding both the blacklist regions and 10 rejected samples respectively. While removal of the ten poor samples has minimal effects of the overall presentation of the results, the removal of the blacklisted regions has a noticeably more substantial impact. Simple observation of the y-axis scale shows that removal of these regions reduces the range of number of fragments.

In Figure 4.15A and Figure 4.15B a borderline significance of the relationship between the average number of chromosome segments and average segmented chromosome length was observed both for the analysis without excluding the blacklisted region and rejected samples (P=0.0644) and the one with rejected samples only excluded (P=0.0633). In Figure 4.15C, a highly significant linear relationship was observed after excluding the blacklist region (P=0.005282) from the analysis as well as with Figure 4.15D with slight improvement in linearity after excluding both blacklist regions and rejected samples (P=0.005189). The percentages of variance explained by the linear models are 16 %, 16 %, 33 % and 33 % (Figure 4.15A-D respectively). It can also be observed that chromosomes with the most excessive aberrations were chromosomes 3, 4, 7, 15, 16, 17 and 22 which are known to contain melanoma CNA regions [71, 148-152].

Our initial paper on copy number analysis [1] examined different types of paired samples consisting of cores from two separate tumours from the same patient, two cores from the same tumour and repeat analysis of the sample from the same core. A correlation analysis using Pearson's r was performed on each replicate pair in terms of adjusted read count per window to assess the similarity of the replicates. Similar analysis was performed and summarised in Table 4.2 showing that majority (27 out of 30) of the paired technical replicates are significantly correlated ($P < 0.05$ for one pair, $P < 0.01$ for the rest) except for one "core" replicate, one "methods" replicate, and one "technical" replicate. For biological replicates, all four "concentration" replicates are significantly correlated ($P<0.0001$). Three out of four of the "tumour" replicates are not significantly correlated as represented by the Pair ID D17. D17 corresponds to triplicates of a tumour sample. The lower correlations observed were more likely due to the inherent variability of the tumour. In our analysis of samples with triplicates, aside

from the initial quality assessment performed, we chose the sample with highest reads for our further analysis. Inter tumour variability was only assessed based on correlation analysis using different pairs of tumours. Overall, the average of the correlation of the paired replicates is 0.60.

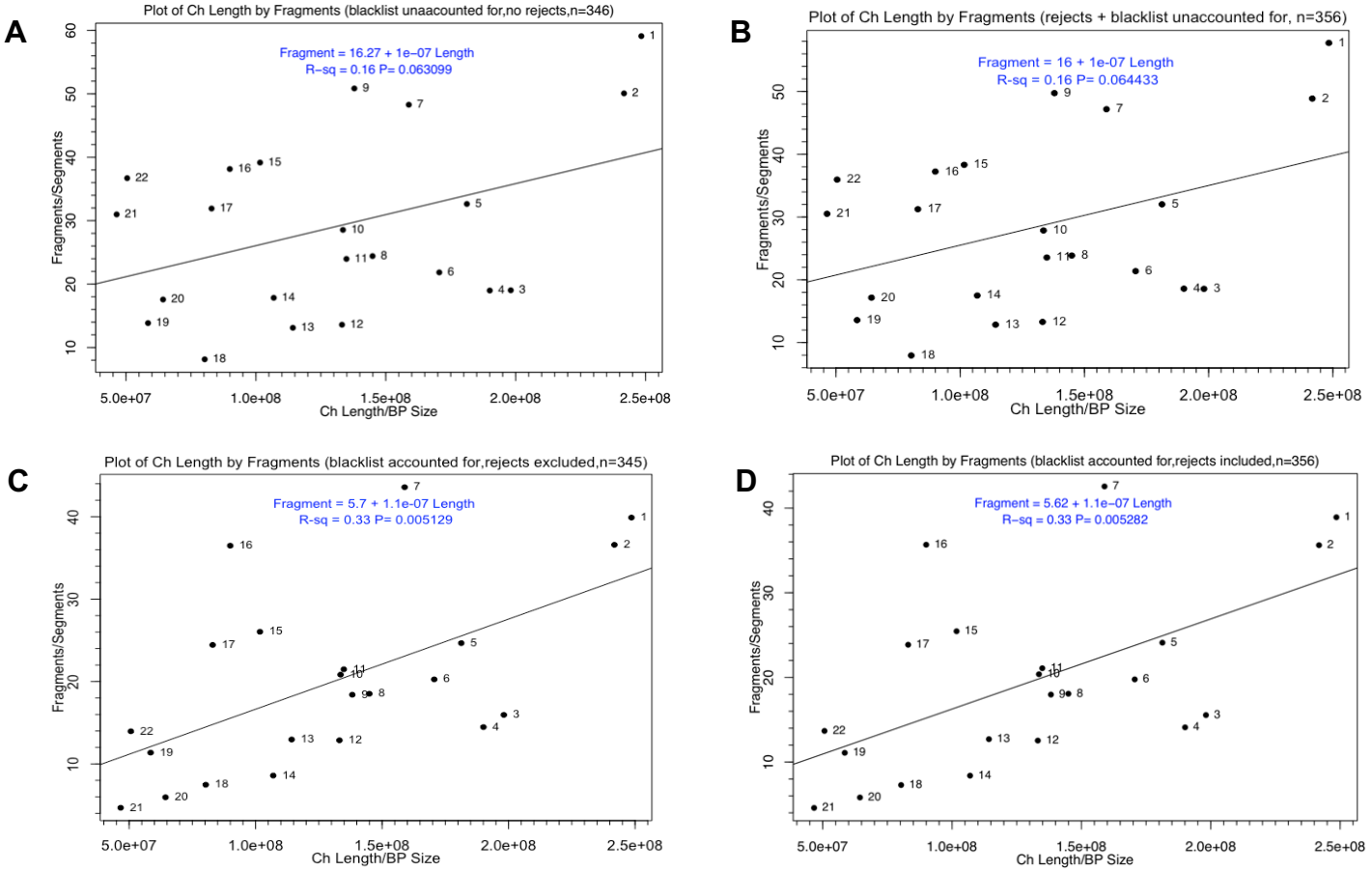


Figure 4.15. Linear plots of average number of segments and average segmented length per chromosome

A, all samples with all regions; B, excluding the blacklisted regions but retaining the rejected regions; C, all regions but excluding the 11 rejected samples; D excluding the rejected samples and the blacklisted regions.

Table 4.2. Correlation of replicates using adjusted read counts

Pair ID	Replicate Type	Pearson's r	Significance
D17	Tumour	0.22	.
D17		0.26	.
D03		0.26	.
D17		0.65	****
D33	Concentration	0.68	****
D34		0.72	****
D34		0.73	****
D34		0.75	****
D26	Core	0.40	*
D01		0.51	**
D02		0.53	**
D11		0.65	****
D18		0.70	****
D29		0.72	****
D05		0.19	.
D07		0.53	**
D13		0.55	**
D14		0.63	****
D08		0.80	****
D16	Method	0.36	.
D23		0.50	**
D31		0.64	****
D04		0.69	****
D12		0.78	****
D06	Technical	0.29	.
D21		0.48	**
D24		0.56	***
D25		0.59	***
D22		0.61	***
D09		0.63	****
D15		0.65	****
D27		0.69	****
D28		0.74	****
D20		0.76	****
D10		0.78	****
D19		0.81	****
D32		0.82	****
D30		0.93	****
Average Correlation		**0.60**	

. not significant at 5% level ** 5 % level*

*** 1 % level* **** 0.1 % level* ***** 0.1 % level*

4.3.7 Examination of the ESV Region: esv3620012

The difference between the adjusted read counts in the windows corresponding to esv3620012 with the adjacent windows showed a 3 peaked distribution (Figure 4.16). This is explained by the 3 genotypes of the SNP *rs4977836* (depicted by distinct colours of each genotype – see figure legend) implying that this location is indeed the proposed ESV but also the methodology sufficed to identify and quantify this variation.

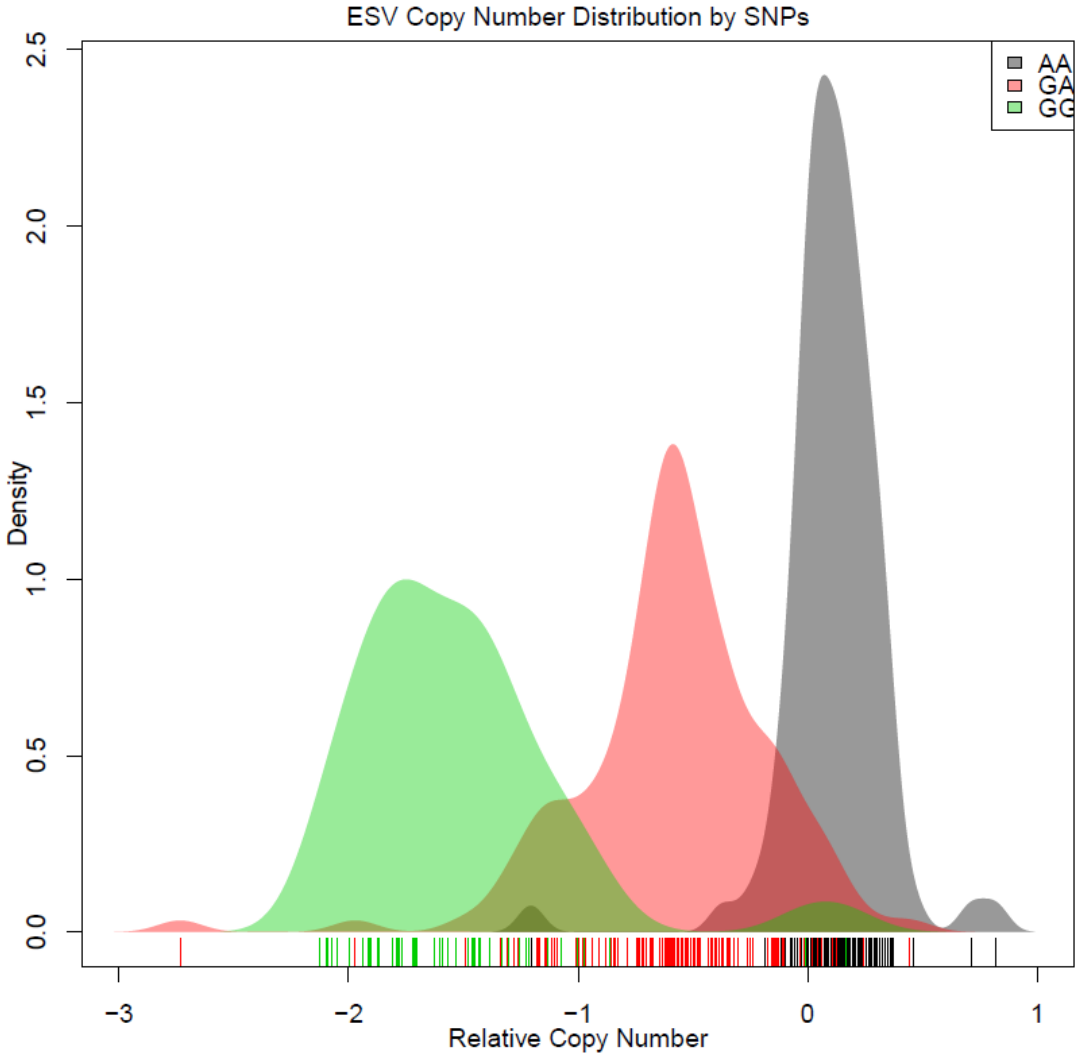


Figure 4.16. Examination of identified common variation in 9p21 by 3 genotypes of the SNP *rs4977836* postulated to be esv3620012.

The distribution of the read count aligned to the germline *esv* with the adjacent windows varies with the predicted genotype at the esv. Distributions are plotted by genotype at rs4977836.

4.3.8 Comparison of NGS Data versus MLPA results

Our lab's first paper on this data also includes comparison of NGS and MLPA copy number analysis focusing on the *CDKN2A* region. Of the 26 samples tested in both methods, 11 samples were deemed incomparable because of not-interpretable MLPA results. Of the 15 comparable samples, 13 have matching results for both methods while the remaining two samples did not match [1]. Shown in Figure 4.17 below is the plot of copy number profile of one sample displaying the copy number profile in the *CDKN2A* region with the light green region covering *MTAP*, light blue covering *CDKN2A*, and dark blue covering *CDKN2B*. Both ends of the plot shows normal copy number profile around zero and in the middle shows deletions of different regions. A total of 11 probes (blue dots) from MLPA analysis was plotted against the copy number segments with 95% confidence limits represented by red vertical lines. In this sample, copy number results in all the MLPA probes matches those of the NGS data as they are consistently close if not coinciding with the NGS copy number segment (black horizontal lines surrounded by yellow green and skyblue dots representing copy number windows) in each region.

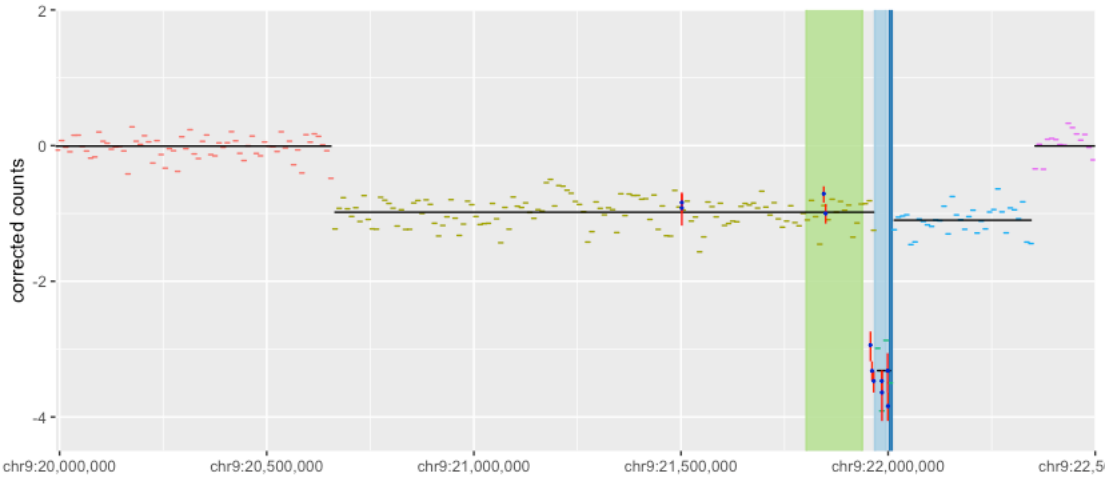


Figure 4.17. MLPA versus NGS CNA data

4.3.9 Comparison with TCGA List of Genes with Aberrations

The adjusted read counts were averaged across the LMC CNV data windows covered for each locus identified within TCGA. Figure 4.18 shows the list of genes with deletions in skin cutaneous melanoma by TCGA. Copy numbers of these genes were checked in LMC data (n=346) and compared with TCGA results (n=367). For LMC

data, an average adjusted (or corrected) read count > -0.3 was set to define a copy number loss. Percentage of samples with deletions were computed for the two sets of samples.

Results from LMC data do not fully resemble the figures from TCGA but there were also clear similarities. The difference might be due to technical variation, nature of the samples as TCGA data predominantly contain metastatic samples while LMC data contains purely primary tumour. On the other hand, the two data show similarly high percentage (< 20%) of samples with deletions in *UTF1, FAM157C, CDKN2A, C9ORF53, FYN, SNORA66, SNORD2, B2M, BHLHA9, PARK2, RPL5, SNORD21, MC1R, DBIL5P and DPEP1*.

The TCGA list of genes with amplifications in skin cutaneous melanoma are shown in Figure 4.19. For LMC data, an average adjusted read count >0.3 was set to define an amplification in copy number. Similarly, percentage of samples with amplification was computed and compared with those of TCGA. Results show that as for deletions, LMC samples with do entirely resemble the TCGA distribution. Genes that similarly show high percentage of samples with amplifications were *HULC, CCND1 (>20%), KCNN3, MYC, RPTOR, PTP4A1, ABHD16B, PVT1, TERT, SMYD3, PCMTD2, CYTH3 (>10%).* A previous study compared *CCND1* amplification in primary and metastasis melanoma and found higher frequency of amplifications of this gene in primaries [153]. This supports what is observed in LMC (24.3%) data in comparison with TCGA (18.8%) data.

Figure 4.18. Copy Number Status of LMC and TCGA Samples (Deletions) based on TCGA identified regions

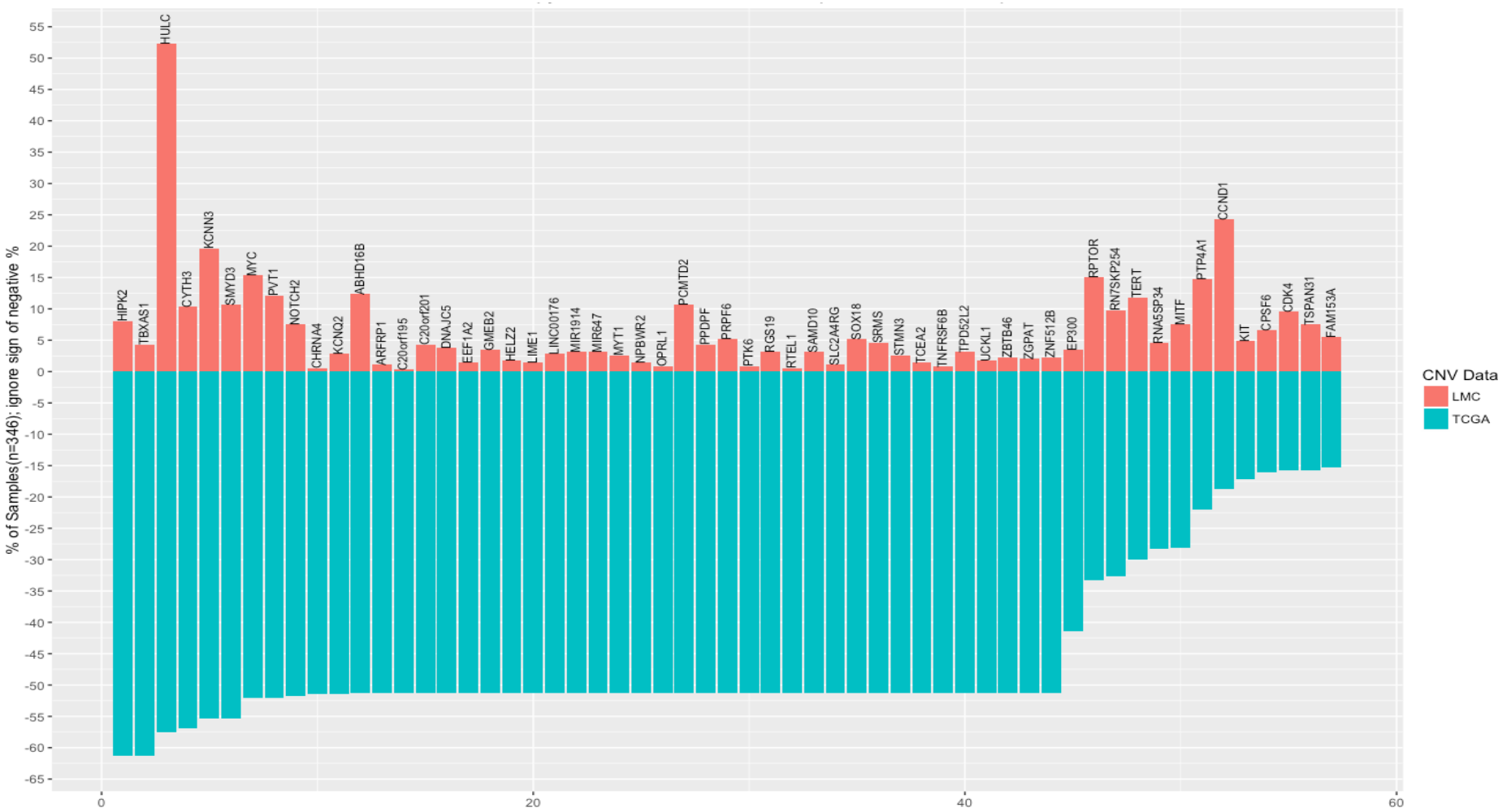


Figure 4.19. Copy Number Status of LMC and TCGA Samples (Amplifications) based on TCGA identified regions

4.4 Discussion

This chapter started with the discussion on the selection of the window size to be used for the analysis based on the ability to capture the *CDKN2A* region which is the main focus in melanoma. Visualisation of the copy number profiles of the samples in this region showed that 10k window provides the optimal compromise between information content and noise as also supported by the analysis of the *esv3620012* which is an identified common germline variation in copy number within the 9p21 region, and is very close to the location of the *CDKN2A* region (*9p21.3*). Other approach to window selection includes the method proposed by Gusnanto et al. (2014) which looked at estimating the optimal window size for the analysis of low-coverage next-generation sequence data based on Akaike's information criterion (AIC) and cross-validation (CV) log-likelihood by plotting the AIC and CV log-likelihood curve as a function of window size[119].

In the absence of a “gold standard” to assess how well our analysis has performed, we must ensure that all quality control is conducted and investigate the consistency of the resulting data, that is both internally (e.g. by looking at replicates) and externally by showing comparability with external data such as TCGA. One of the few approaches to validity is to compare the copy number estimate from NGS with that derived from MLPA applied to as close a DNA sample as is feasible. Because MLPA requires significant quantities of DNA, this means of comparison is restricted, in this instance to the CDKN2A region, a region with a commercially available MLPA kit.

Initial quality control procedures were done on the CNV data, starting with read quality etc. Analysis of replicates were done by plotting replicate samples with one another on the basis of number of chromosome fragments, estimated segmented chromosome length, raw read counts and corrected read counts. Significant correlation between the samples (excluding the biological ones) over and above that is for random pairs of samples provided evidence of high reproducibility of the data and indicates quality control measures produce consistent data.

Plot of chromosome fragments on the chromosome length reveals linearity as would be expected both for data including rejected samples and blacklisted regions. Excluding rejected samples and blacklisted region in the analysis significantly improved linear relationship between chromosome fragments and chromosome length. Excessive copy number aberrations were observed in chromosomes 3, 4, 7, 10, 11, 15, 16, 17 and 22 making these as potential targets for evaluation in subsequent analyses.

Another examination done on the data was checking for known structural variants *esv3620012* (*ESVs* are structural variants named by EBI). Analysis of the read counts on its location across SNP genotypes within the NGS dataset was performed. The two 10kb windows covering the sites of the deletion were considered and compared based on the number of reads in those two windows with the average adjusted read count from the 10 adjacent windows each side of this copy number variation. The difference between the read counts in the putative *esv* windows with the adjacent windows showed a 3 peaked distribution which was explained by the 3 genotypes of the SNP implying that this location is indeed the proposed *esv* but also the methodology sufficed to identify and quantify this variation.

LMC samples were compared with TCGA samples in terms of their published list of genes with deletions and amplifications in skin cutaneous melanoma. Though LMC samples did not appear to fully resemble the distribution of deletions and amplifications in TCGA samples, a similar high percentage of samples with deletions and amplifications in some genes were observed. Major difference in the two samples might be due to technical variation and or the nature of the samples used as TCGA is composed of predominantly metastatic melanoma was LMC is composed purely of primary melanomas.

4.5 Conclusion

Results of the initial copy number data assessment showed variability of the across different regions of the genome especially at chromosome 7. While the methodology performed was able to identify and quantify the know common germline variation: *esv3620012* in human, comparison of the frequency of somatic gains and deletions of the LMC copy number data with those of TCGA showed patterns of similarity but still requires to be improved and suggests that further steps can be done to increase CNA data quality. It has been found out that frequent occurrence of noisy peaks in the copy number profile lies in the centromeric and telomeric region of the genome. This should further be investigated and accounted before proceeding with the further analysis of the data.

Chapter 5
Additional Steps to Improve Data Quality

This chapter discusses the additional steps explored in attempt to improve the quality of the LMC copy number data. It includes new methodologies applied as well as updated information from the online genomic research resources such as UCSC Browser and Genomic Reference Consortium website.

5.1 Introduction

Subsequent to the initial LMC CN data acquisition and QC (conducted by Dr. Alistair Droop and Dr. Anastasia Filia), there have been publications describing studies of relevance which added or updated genomic information.

- The *QDNAseq* (Quantitative DNA sequencing for chromosomal aberrations) pipeline was developed and became available as an R package[2]. This pipeline performs GC content and mappability correction using a two-dimensional LOESS model with an interaction term. It empirically identifies highly variable regions (blacklist) in the genome of a given set of samples.
- Updated published list of gaps including centromeres, telomeres, and regions of heterochromatin were obtained from the UCSC Browser and Genome Reference Consortium Website [154, 155].
- The 1000 Genomes Project read realigned to reference assembly GRCh38 provided more capability to identify common germline variations in normal samples matching the LMC characteristics as mentioned in Chapter 1.1 [114]. This helped identify more highly variable regions on the genome that were not identified in our previous work.

In this Chapter, I describe how these developments were incorporated into my analysis and the impact on the results. Utilisation of the 1000 Genomes Project samples as normal control was initially suggested by Dr. Arief Gusnanto who acted as my examiner during my first-year transfer. The study of Scheinin et al. (2014) in estimating shallow sequencing derived somatic copy number data from FFPE samples was made known to me by Dr. Daniela Robles-Espinoza as part of her feedback in my presentation in Genomel 2017 in Genoa, Italy.

5.2 Methods

5.2.1 The *QDNAseq* Pipeline

The *QDNAseq* (Quantitative DNA sequencing for chromosomal aberrations) pipeline was developed to implement novel copy number profile correction and blacklisting approaches described in the work of Scheinin et al. (2014) [2]; it was developed specifically for low coverage copy number analysis from formalin fixed material. This was implemented in *R* or *RStudio* and uses bam files as input[156, 157]. This can be downloaded in Bioconductor with an available user manual that details how to use the package [158]. This package has several functions which includes simultaneously correcting for GC content and mappability using a two-dimensional LOESS model, identification of highly variable regions in the genome including the common germline variations, copy number segmentation and copy number calling. The application of the package was limited to the first two functions as these were the novelties of this package. The last two functions were done separately in different R packages which will be discussed further in the later sections.

5.2.2 Generation of Additional Blacklist

The phase 1 of blacklist identification done by Dr. Alastair Droop was based on an earlier study that identified spurious copy number peaks in the genome using GRCh37 genome build [131] whose information was derived from the 1000 Genomes Project. This consisted of 33,096 10kb windows which are described in Chapter 3. This set of blacklists successfully reduced noise from the LMC CNA data but was deemed insufficient due to many regions in the genome that still shows high level of noise (a point that was also recognised by Scheinin et al. (2014) [2]). This reflects the need to consider further analytical steps to increase data quality and adjust for other technical variations that may have taken place from sampling to sequencing. This variation may also be due to the normal samples used and its size (7 normal samples from the skin combined to one summary dataset), and difference in the genome builds used in generating the blacklist (GRCh37 vs GRCh38). Though there are existing tools like LiftOver (https://genome.sph.umich.edu/wiki/LiftOver, https://genome.ucsc.edu/cgi-bin/hgLiftOver), *hgLiftOver* (http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgLiftOver), and *liftOver* (https://rdrr.io/bioc/rtracklayer/man/liftOver.html), that “lifts over” one genome build to another, they fail to account for the updates in the newer version specifically for the nonrepetitive regions in the genome. Lifting over is most appropriately only for non-

repetitive sites of the genome where significant change between two genome builds was not seen. It also compromises accuracy in regions with updated information such as identification of variants in any region that was newly added to the newer genome build [114].

For genome builds as updated as GRCh38, the UCSC Browser has a GAPS track that indicates gaps in the human genome. This includes centromeres, telomeres, contigs, heterochromatins, and short arms. Gaps pertain to the "unfinished" genome sequences that could either be heterochromatic (~200 mb) or euchromatic (~24.4 mb) regions [159]. Heterochromatic regions are defined as large regions of the genome that are almost exclusively composed of tandem repeats; this includes centromeric satellite DNA and acrocentric portions of the human chromosomes while euchromatic regions contain genes [159]. The information on the location of UCSC gaps is located on http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgTables?hgsid=6186842_AYlQa1OGEnUoRkQEj02VSdI1vjsw . This can be manually selected by assigning the following values in the search fields:

Clade: Mammal

Genome: Human

assembly: Dec. 2013 (GRCh38/hg38)

group: Mapping and Sequencing

track: Gap

region: genome

A screenshot showing how this step was done is shown in Figure 5.1. The definition of the different types of genome gaps below were mostly derived from the UCSC Browser and the work of Eichler et al. (2004):

Centromere: the modelled location of the centromeric regions in the human genome. There were only 22 autosomal centromeric regions considered in this study.

Telomere: the location of the telomeric regions and spans to a fixed size of 10kb on each side of the chromosomes. There were 44 telomeric regions considered in this study.

Heterochromatin: these consists of large regions that are almost exclusively composed of tandem repeats. There were 11 heterochromatic regions obtained from the UCSC browser and spans from 20kb to 3000kb in length.

Short arms: the location of the short arms of the five chromosomes 13, 14, 15, 21 and 22 were not targeted as part of the Human Genome Project. These contains DNA that are highly repetitive and generally regarded as heterochromatic region. The region length ranges from 500kb to 1699 kb.

Scaffold: these are sections of the genome sequence that are derived from end sequenced whole genome shotgun clones and spans from 10bp to 624 kb in length and may compose of clone gaps or contig gaps as described below:

Clones: these are gaps that were classified as unable to subclone from those that were covered by clones but were difficult to finish/complete thus not bridging the gap. Most clone gaps contain high amount of segmental duplications. There were 98 clone gaps identified in this list.

Contigs: there are sets of contiguous overlapping clones and aggregated in the genome. This is a subset of a scaffold and accounts to 135 regions included in the list spanning from 100bp to 400 kb.

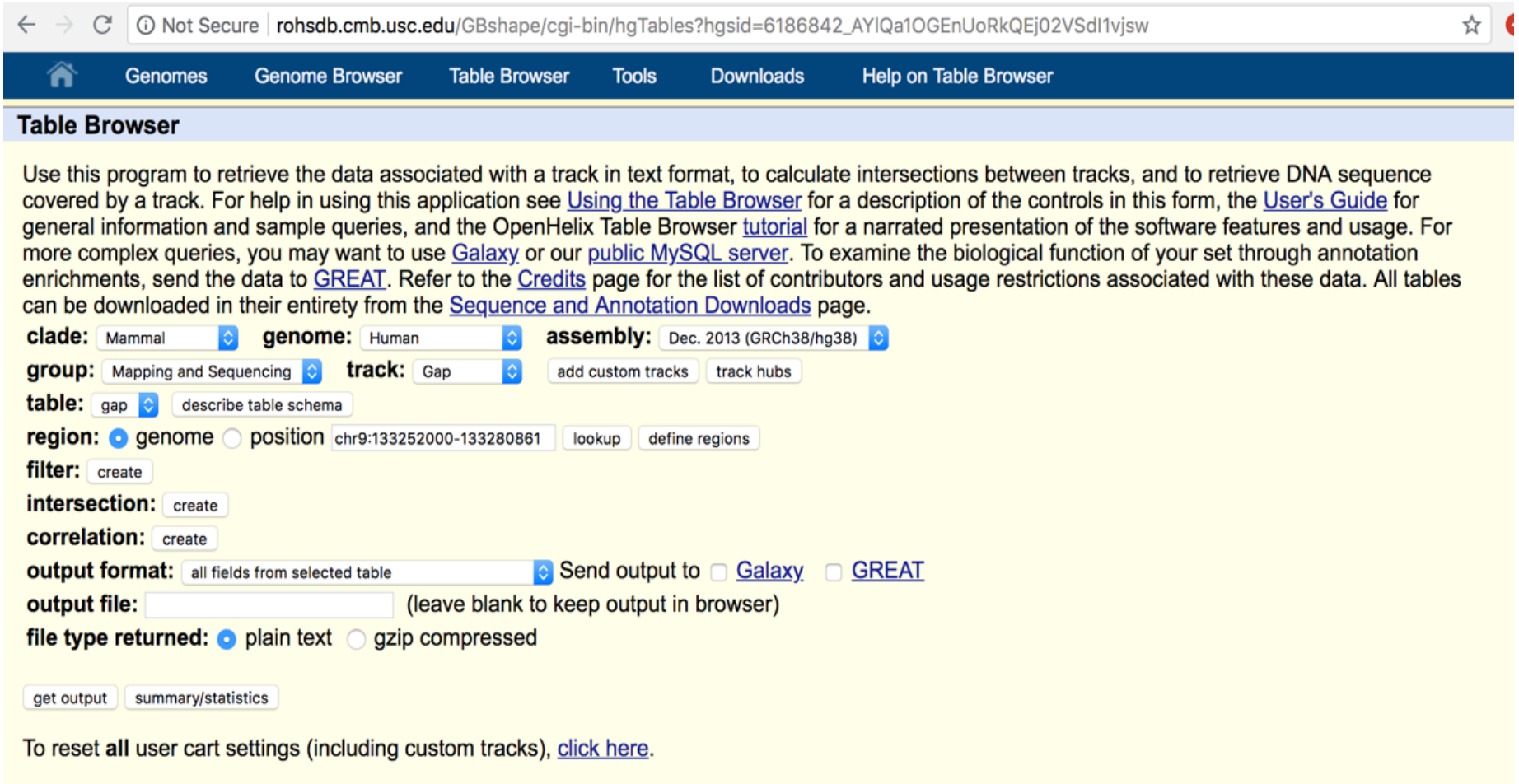


Figure 5.1. Manually selecting the gaps in the UCSC Genome Browser

The information about the location of modelled centromeres and heterochromatins were located in the Genome Reference Consortium website at https://www.ncbi.nlm.nih.gov/grc/human [155]. A. Kundaje (2016) published a list of new blacklists in the hg38 genome build. This was done by redoing their previous work on the hg37 genome build on the hg38 mapped samples. His list identified regions that have anomalous, unstructured, and high peaked in the NGS experiments independent of cell line and experiment type. This list of blacklist is downloadable online in http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38-human/ [160, 161].

All these regions were identified in the LMC 10kb window CNA and 1000 Genomes Project data and were masked before empirically identifying a new set of blacklists based on common variation in the genome of normal samples from 1000 Genomes Project using QDNAseq pipeline.

5.2.3 Obtaining 1000 Genomes Project (1KGP) Samples

Data from the 1000 Genomes Project (1KGP, http://www.internationalgenome.org/about/) were obtained as normal controls. The final version of 1KGP consists of 2504 germline samples from 26 populations from different parts of the world categorised as South Asian (SAS, 20%), European (EUR, 20%), East Asian (EAS, 20%), American (AMR, 14%), African (AFR, 26%) [113]. These samples were sequenced originally in GRCh37 format and were later on realigned to GRCh38 with alternative scaffold-aware BWA-MEM and available as CRAM, a sequence compression format based on a reference [114].

The following criteria were used to select the samples which are processed similarly to the LMC samples: sequencing done with paired-end Illumina equipment, short read length sequencing, low coverage, and those sampled were of Caucasian ancestry. In total, 312 1KGP samples (British with n=106, Finnish with n=105, and Central Europeans living in Utah with n=101) meeting these criteria were included as normal controls. Corresponding sequences for the 312 selected normal samples were downloaded from ftp.sra.ebi.ac.uk/vol1/ and securely stored on the university server.

CRAM files were converted to bam format using *samtools* as required for the windowing package: *bamwindow (*bamwindow available on GitHub at https://github.com/alastair-droop/bamwindow) [112]. The resulting windowed sequences were read into R using *QDNAseq* package and were used to estimate

highly variable copy number regions in the genome of Caucasian population as well as the germline copy number.

5.2.4 Highly Variable Regions in Caucasian Populations

A previous study has shown the presence of highly variable windows in the genome [2]; these windows which are typically isolated are recurrent in terms of differing rates among samples (even germline samples) and still remain after accounting for previously published blacklists in the genome. These were initially hypothesised to be common germline variations and hence detailed analysis of germline samples was conducted.

Similar methodology was applied in this study analysing a significantly increased number of normal samples (38 in the previous study (Scheinin, Sie et al. 2014) versus 312 in this study). Another improvement of this study is a more specific ethnic group alignment to the case samples which addressed the recommendation of the previous study to provide more power in adding more population specificity to detecting common germline variations. The normal samples were processed similarly as the case samples using the *QDNAseq* pipeline. The new set of blacklists were compared with previous sets generated and visualised using the Venn diagram *venn* function of *gplots* package in R [162]. The various sources were then combined to make a broader set of blacklists to reduce the rate of false copy number peak detection reducing the noise in the genomic profiles of samples.

5.2.5 Read Counts Adjustment Phase 2

GC content and mappability of the sequence reads affect the raw read counts [123, 130, 163]. In the initial correction made for these data, adjustments were made for these two factors separately. Although independent correction is effective for many samples, some samples contain artifactual variations and this is likely especially true for formalin fixed samples. Independent or sequential correction of GC content and mappability provides benefit if these two factors do not have an interaction effect on the raw read counts. However, an interactive effect would mean that the separate adjustment is likely non-optimal. I, therefore, checked for the evidence of an interaction by plotting the median read counts at each bin as a function of GC and mappability. With evidence of interaction between GC content and mappability, simultaneous correction was performed as recommended by Scheinin et al. [2]. This involved first fitting a LOESS surface through the medians of the windows with the same combinations of GC and mappability. The raw read count for

each bin is corrected by dividing the raw read count by the LOESS value of its combination of GC and mappability as shown in the equation below.

Equation 6 $CorrectedCount = RawCount/LoessFit$

where

CorrectedCount is the corrected counts for windows with a combination of the GC content and mappability

RawCount is the raw read counts for windows with a combination of the GC content and mappability

LoessFit is the fitted read count using a loess curve for windows with a combination of the GC content and mappability

This is implemented using the *estimateCorrection* function under *QDNAseq* R package [2]. After correction/adjustment, the adjusted read counts were then normalised around the corresponding genome median read count using the function *normalizeBins* which is also under *QDNAseq* R package [2].

5.2.6 Comparison with Germline Copy Number

Ideally, in copy number analysis, a tumour copy number profile is compared with a germline copy number from a matched normal tissue. Lack of matched normal sample is a frequent occurrence for formalin fixed samples, either because of lack of normal samples or because of the expense of generating information from normal samples. In this study seven composite normal samples were available (not all from patients in this study) and used as a reference having been formalin fixed similarly to the tumour samples; these samples were taken from wide local excisions removed during melanoma removal but were so distant from the primary as to be tumour free. In comparison, germline copy number profiles from the available 312 1KGP samples were also included. For a given sample, a log transformed copy number ratio between a tumour and germline reference as described in Chapter 3 (Section 3.10 Calculation of Copy Number).

5.3 Results

5.3.1 Additional Blacklist from Published Resources

A total of 315 autosomal gap regions were extracted from the UCSC browser website (Table 5.1) corresponding to 22 centromeres (6600 kb), 98 clone regions (568 kb), 135 contigs (1472 kb), 11 heterochromatins (4248 kb), 5 short arms (6716 kb) and 44 telomeres (44 kb) accounting for a total of 19648 kb region of the genome (~1% of the genome).

Table 5.1.Distribution of genome gaps by type as obtained using UCSC Browser

(Accessed on December 15, 2018). Includes only autosomal regions

Gap type	No. of regions	Size(kb)
centromere	22	6600.0
clone	98	567.7
contig	135	1472.0
heterochromatin	11	4247.7
short_arm	5	6716.1
telomere	44	44.0
Total	**315**	**19647.5**

Shown below in Table 5.2 is the list of modelled centromeres and heterochromatin published in the GRC website. Centromeric regions measured from 128 kb (chromosome 6) to 580 kb (chromosome 18). Additionally, a heterochromatin region found on chromosome 7 was identified and spans up to 15 kb in size. Another set of blacklists included in this study was identified from the works of Kundaje (2016) that accounts for almost 2 kb of the genome (Table 5.3). According to Kundaje, a considerable proportion of blacklisted regions vanished when the new genome build was released. Despite the small size, inclusion of this known blacklist remains important in the attempt to reduce false positives in identifying regions of copy number alterations. There were only 9 10k window blacklist identified in the study of Kundaje (2016) which are also found in all other sources of blacklist providing more confidence in the inclusion of this list to the final blacklisted regions.

Table 5.2. List Modelled Centromeres and Heterochromatin

Region name	Location	Start	Stop	Size(kb)
CEN1	chr1	122026460	125184587	315.8
CEN2	chr2	92188146	94090557	190.2
CEN3	chr3	90772459	93655574	288.3
CEN4	chr4	49708101	51743951	203.6
CEN5	chr5	46485901	50059807	357.4
CEN6	chr6	58553889	59829934	127.6
CEN7	chr7	58169654	60828234	265.9
HET7	chr7	61377789	61528020	15.0
CEN8	chr8	44033745	45877265	184.4
CEN9	chr9	43236168	45518558	228.2
CEN10	chr10	39686683	41593521	190.7
CEN11	chr11	51078349	54425074	334.7
CEN12	chr12	34769408	37185252	241.6
CEN13	chr13	16000001	18051248	205.1
CEN14	chr14	16000001	18173523	217.4
CEN15	chr15	17000001	19725254	272.5
CEN16	chr16	36311159	38280682	197.0
CEN17	chr17	22813680	26885980	407.2
CEN18	chr18	15460900	20861206	540.0
CEN19	chr19	24498981	27190874	269.2
CEN20	chr20	26436233	30038348	360.2
CEN21	chr21	10864561	12915808	205.1
CEN22	chr22	12954789	15054318	210.0
	Total			**5827.1**

Table 5.3. Kundaje's list of blacklisted regions in hg38

Location	Start	End	Size(kb)
chr1	124450730	124450960	0.02
chr2	90397520	90397900	0.04
chr2	90398120	90398760	0.06
chr3	93470260	93470870	0.06
chr4	49118760	49119010	0.03
chr4	49120790	49121130	0.03
chr5	49601430	49602300	0.09
chr5	49657080	49657690	0.06
chr5	49661330	49661570	0.02
chr10	38528030	38529790	0.18
chr10	42070420	42070660	0.02
chr16	34571420	34571640	0.02
chr16	34572700	34572930	0.02
chr16	34584530	34584840	0.03
chr16	34585000	34585220	0.02
chr16	34585700	34586380	0.07
chr16	34586660	34587100	0.04
chr16	34587060	34587660	0.06
chr16	34587900	34588170	0.03
chr16	34593000	34593590	0.06
chr16	34594490	34594720	0.02
chr16	34594900	34595150	0.03
chr16	34595320	34595570	0.03
chr16	46380910	46381140	0.02
chr16	46386270	46386530	0.03
chr16	46390180	46390930	0.08
chr16	46394370	46395100	0.07
chr16	46395670	46395910	0.02
chr16	46398780	46399020	0.02
chr16	46400700	46400970	0.03
chr20	28513520	28513770	0.03
chr20	31060210	31060770	0.06
chr20	31061050	31061560	0.05
chr20	31063990	31064490	0.05
chr20	31067930	31069060	0.11
chr20	31069000	31069280	0.03
chr21	8219780	8220120	0.03
chr21	8234330	8234620	0.03
	Total		**1.70**

Inclusion of these blacklisted effectively reduced the amount of recurrent noise in the LMC data. Examining the different blacklists after being applied in the LMC data, a significant overlap was noticed among the blacklists used by Droop, Kundaje and

Centrogaps (Figure 5.2). A total of 17,393 windows from CentroGaps were also located in either the Kundaje or the Droop list while only 511 windows were unique. The Kundaje list showed high rate of overlap with Centrogaps and Droop with 20/21 (95%) windows in common and only 1 unique window. Of the 25,310 autosomal windows from Droop's list, 17,402 (70%) windows were also found in Centrogaps and the Kundaje list with 7,908 unique windows. After accounting for the initial set of blacklists from Centrogaps, Kundaje and Droop, my *QDNAseq* pipeline derived blacklists that were inherent in the normal control based on the 312 1KGP samples from Caucasian population. These regions consist of unique 12,392 (32% of the overall blacklists) 10k windows; this has improved the quality of the LMC data by empirically identifying the location of highly variable regions in the genome that were not identified in the previous works and published references. In totality, 38,215 10k windows accounting to more than 13% of the autosomal genome (287, 509 10k windows) were included in the final blacklist. A visualization of the location of these regions in the genome is depicted in Figure 5.3.

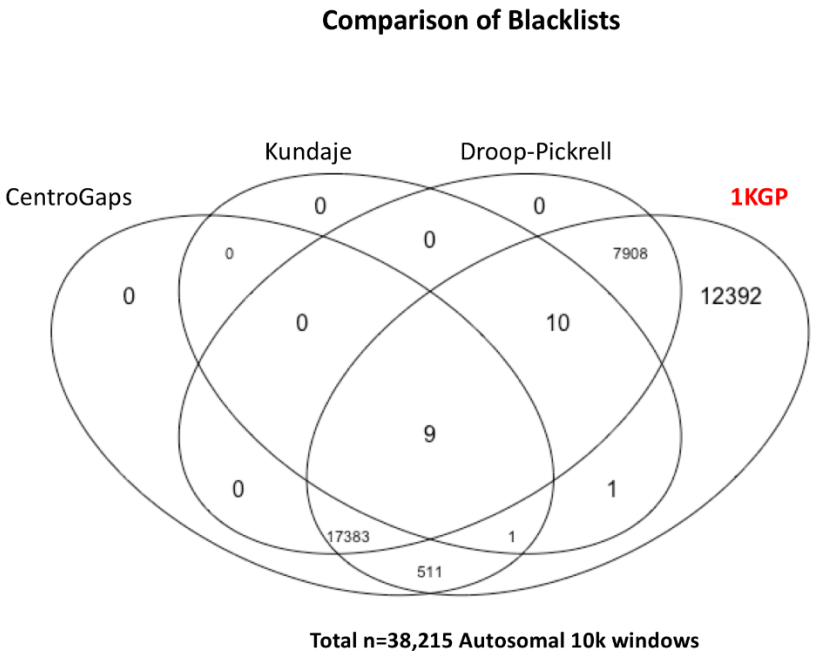


Figure 5.2. Relationship among the different blacklist used

The Droop list is represented by the vertical bars/lines in maroon, Kundaje in black, CentroGaps in green, and the non-highlighted peaks were the new set of blacklists derived based on the noise filter applied to the 312 1KGP samples. The y-axis represents the median read count of the samples for each 10k window in the

genome while x-axis shows the 10k window index from chromosomes 1-22. The upper shows the median adjusted copy number profile of the 7 LMC control samples with the different blacklists and when the read counts covered by these regions were not removed. It is very noticeable that a significant number of regions in the genome that show highly variable copy number profile where located in the blacklisted regions. The lower plot shows the same data after removing the copy number profile located in the blacklisted regions. A significantly cleaner copy number profile was observed showing the importance of accounting for blacklisted regions in somatic copy number analysis and the usefulness of using 1KGP normal samples in identifying highly variable regions in the human genome using the same genome build GRCh38.

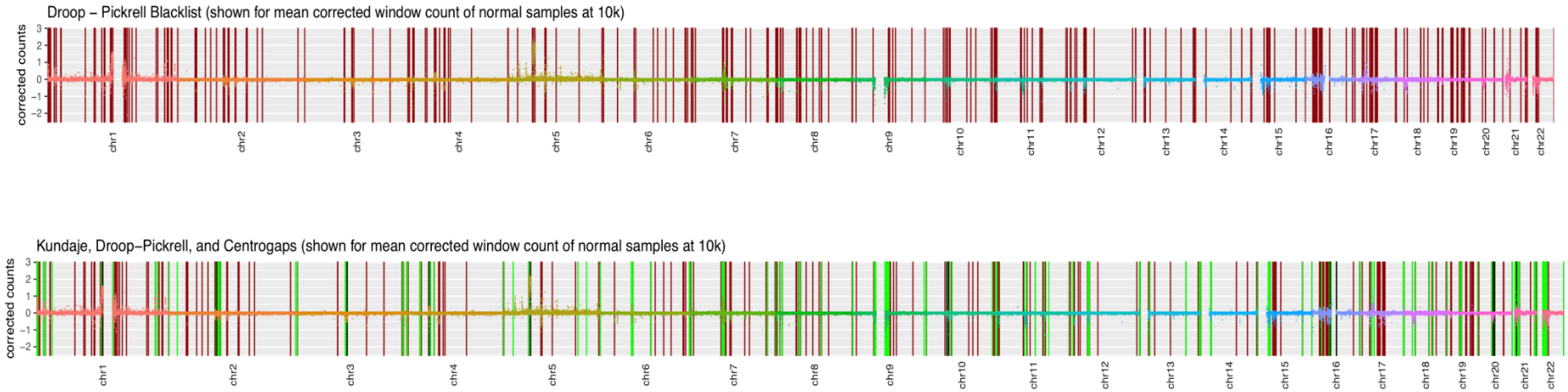


Figure 5.3. Location of blacklists in the genome.

The upper plot displays the initial blacklist used from the work of Dr. Alastair Droop. The lower plot presents the initial blacklist, including the additional ones from the UCSC browser gaps, GRC list of centromeric and telomeric regions, and *QDNAseq* identified list of highly variable regions in the genome

5.3.2 Adjusting for the interaction effect of GC content and mappability

An essential consideration in correcting for GC content and mappability bias is checking for their interaction effect on the read counts. This was done by creating an isobar plot which plots the median read counts for each window as a function of GC and mappability for a given sample.

The isobar plot below in

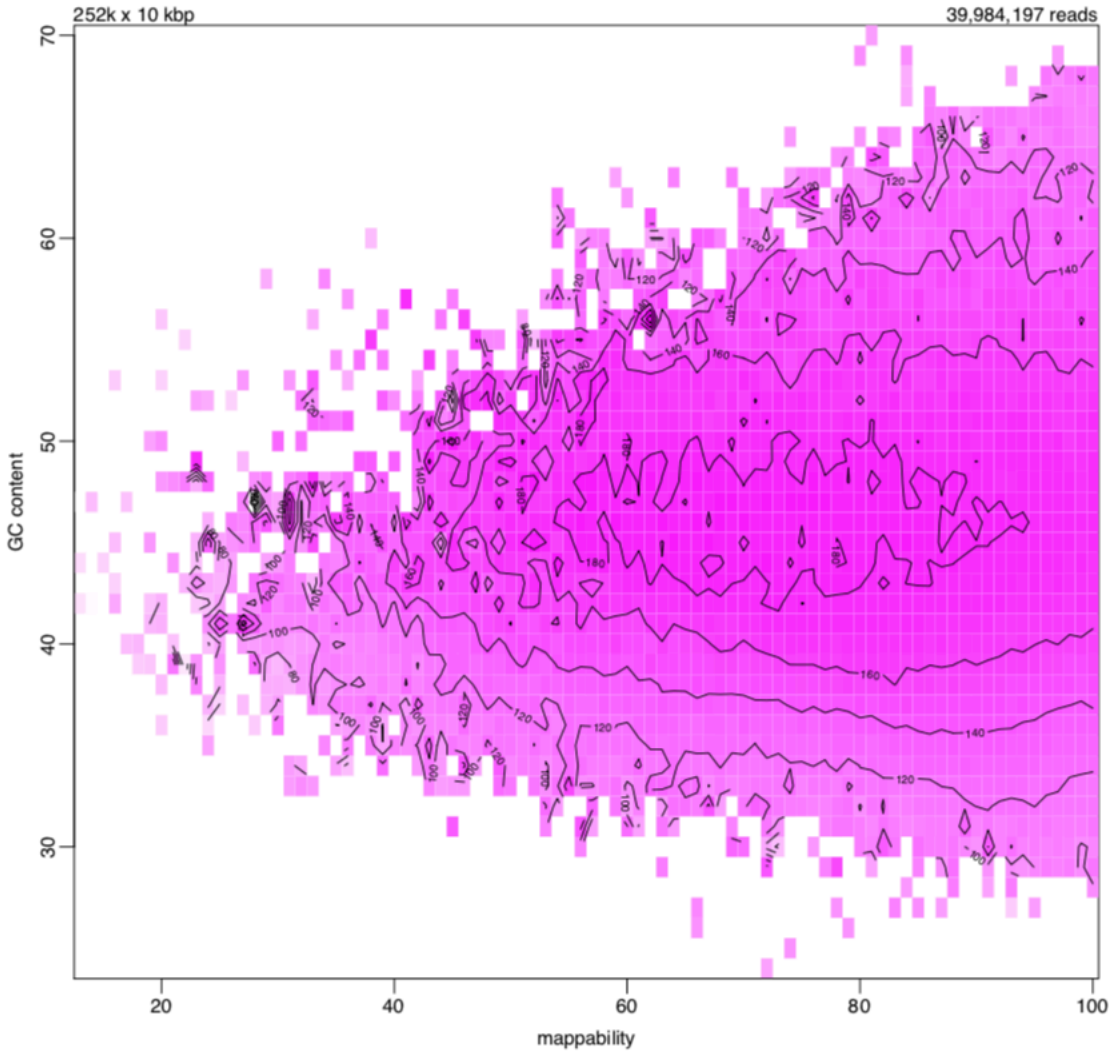


Figure 5.4 shows the median read counts per bin as a function of GC content and mappability showing the interaction between GC content and mappability as generated using QDNAseq package in R. The black lines represent each combination of GC content and mappability. The colour intensity represents the median read count for a given combination of GC content and mappability. It can be observed that both low and high median read counts are observed in a varying GC content at a given fixed value of mappability implying an interaction [2]. The adjusted read counts for each window were calculated by first fitting a LOESS surface through the medians of the windows

with the same combinations of GC and mappability. The corresponding loess fit for the given sample in Figure 5.4 is displayed in Figure 5.5. The raw read count for each bin is corrected by dividing the raw read count by the loess value of its combination of GC and mappability as described in Section 5.2.5 Equation 6.

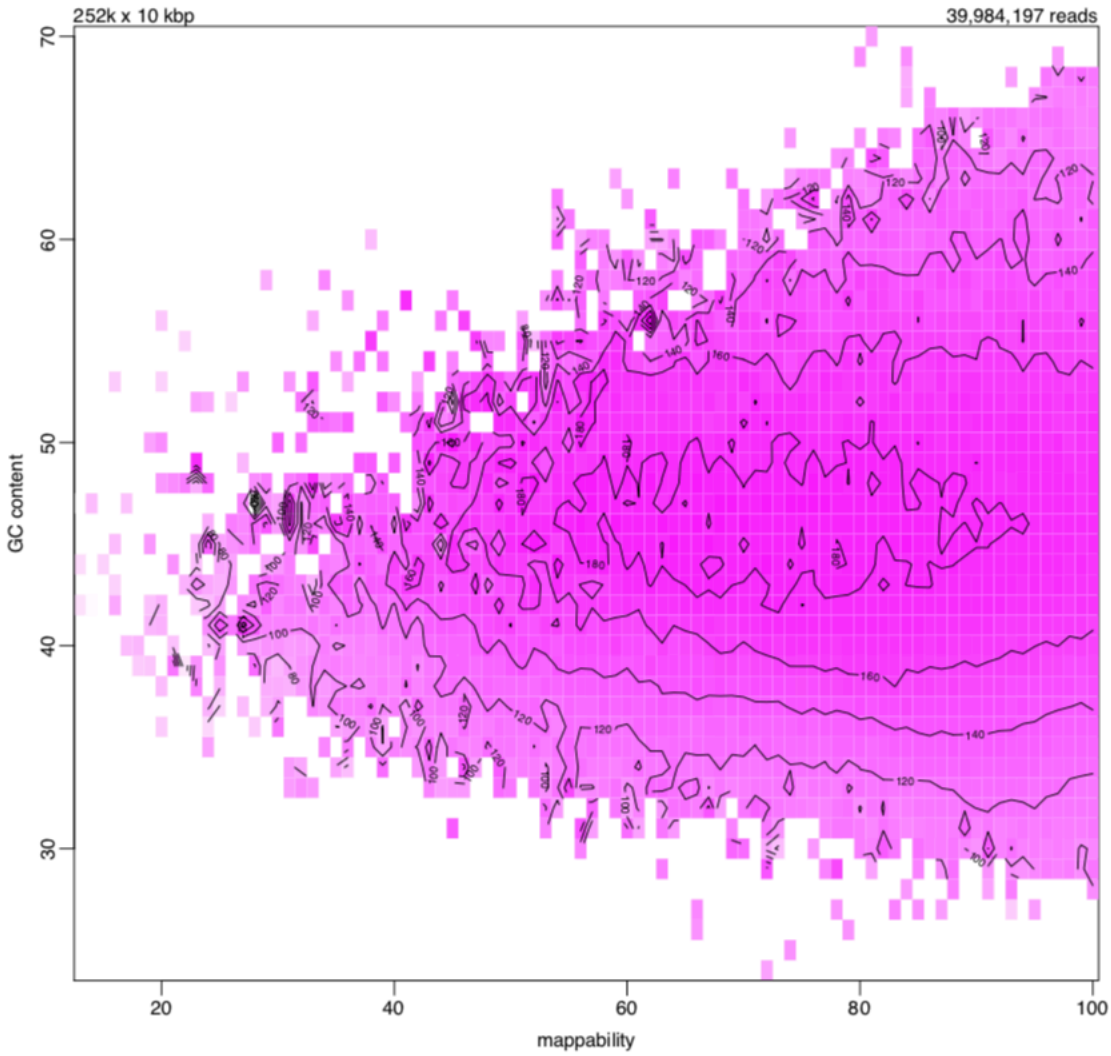


Figure 5.4. Median read counts per bin as a function of GC content and mappability.

Varying median read counts are observed in a varying GC content at a given fixed value of mappability implies an interaction between GC content and mappability

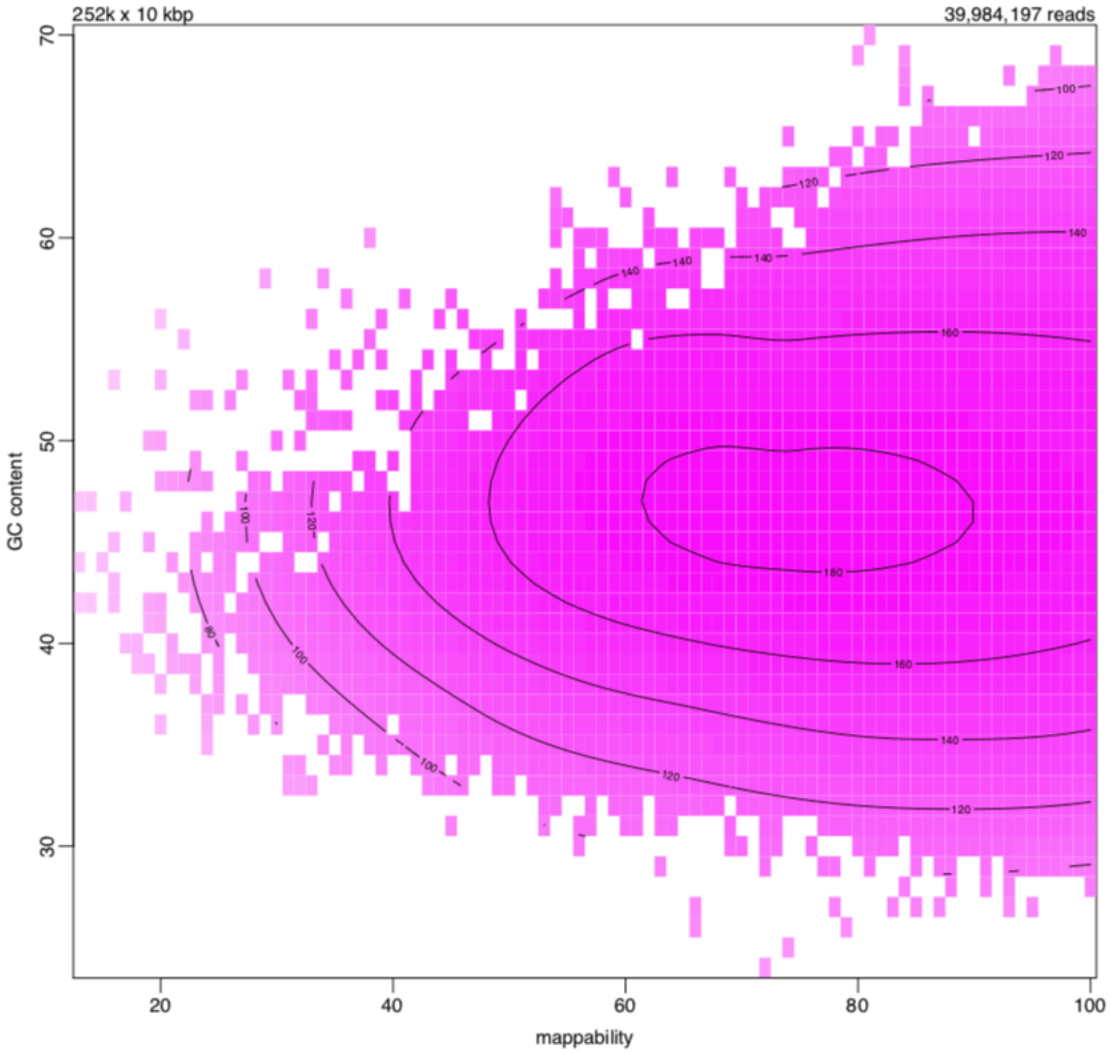


Figure 5.5. Loess fit for a given combination of GC content and mappability.

A loess model is used to estimate a loess fit through the median read count in the windows for each combination of GC content and mappability. The raw read count is divided by the loess fitted value for each combination of GC content and mappability to estimate the corrected counts in each window.

To assess the impact of adjusting for the interaction effect of GC content and mappability to LMC copy number profile, a comparison of the copy number profiles using the different GC content and mappability adjustment methods (sequential versus simultaneous/interaction) is shown in Figure 5.6 below.

The upper plot shows the median copy number profile of the 7 LMC control samples adjusted sequentially for GC content and mappability. It can be seen that the median sample profile is very close to zero reflecting that expected copy number profile for normal samples. Plot of the samples using the mean and median read counts as well as examples of individual sample plot for two samples to show variation of read counts are shown in Appendix F. It can also be noticed that some noise remains visible after removing the copy number profiles that fall within the blacklisted regions. This noise was significantly reduced when a simultaneous correction of sample read counts for GC content and mappability was done (lower plot). The same impact on the copy number profiles was observed in a previous study using FFPE samples [2].

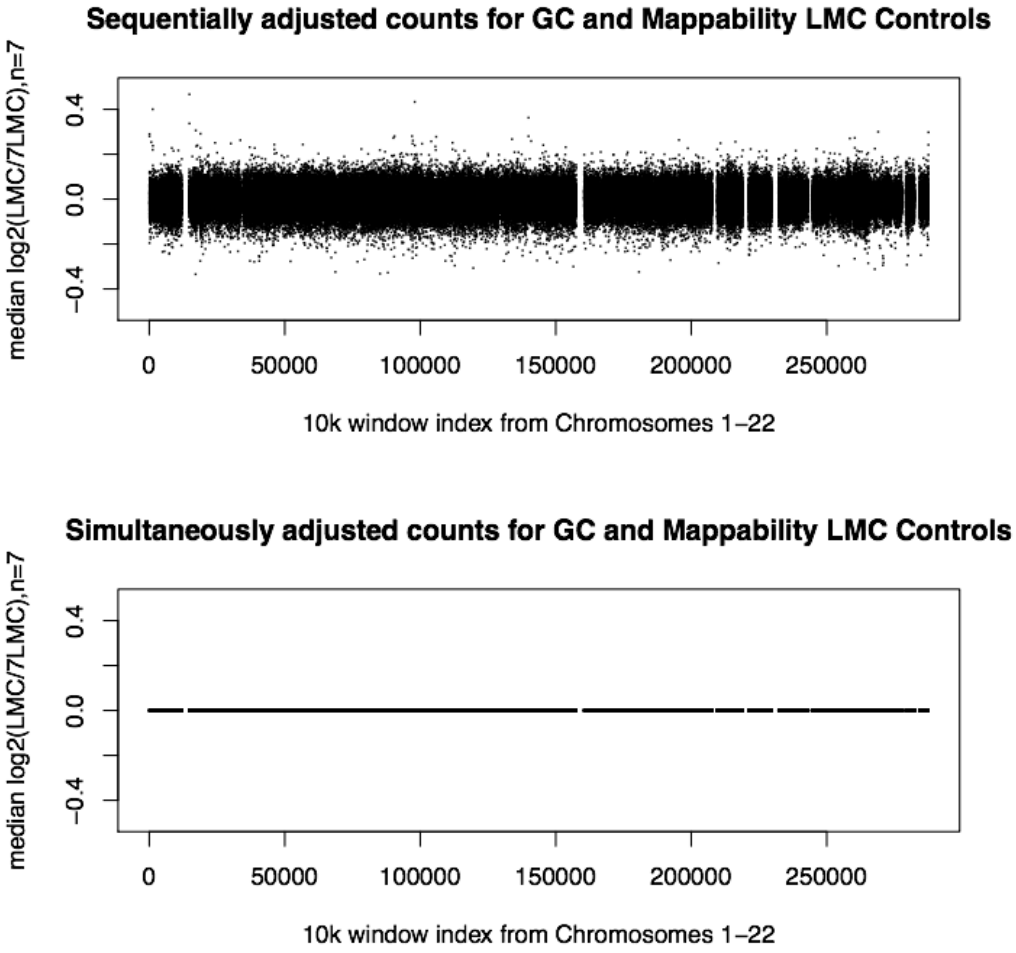


Figure 5.6. Average LMC copy number profile using sequential and simultaneous correction for GC content and mappability

5.3.3 Reference copy number from normal samples

Germline copy number profiles from 7 normal skin LMC samples and 312 1KGP were estimated and presented different strengths and weaknesses. LMC samples were derived from the skin and expected to reflect a more representation of the skin germline but is present only in a very few numbers. Samples from 1KGP were much larger and were derived from blood samples. Figure 5.7 shows the comparison of the two samples after similarly being simultaneously adjusted for GC content and mappability.

Although both samples showed germline copy number profile around zero in the autosomal regions, it is highly noticeable that the 1KGP samples provides a cleaner profile. On the basis of general uniformity across the genome that depicts more a normal population, I decided to use the median read count per window of the 312 1KGP samples as the reference germline copy number for the 303 LMC tumours.

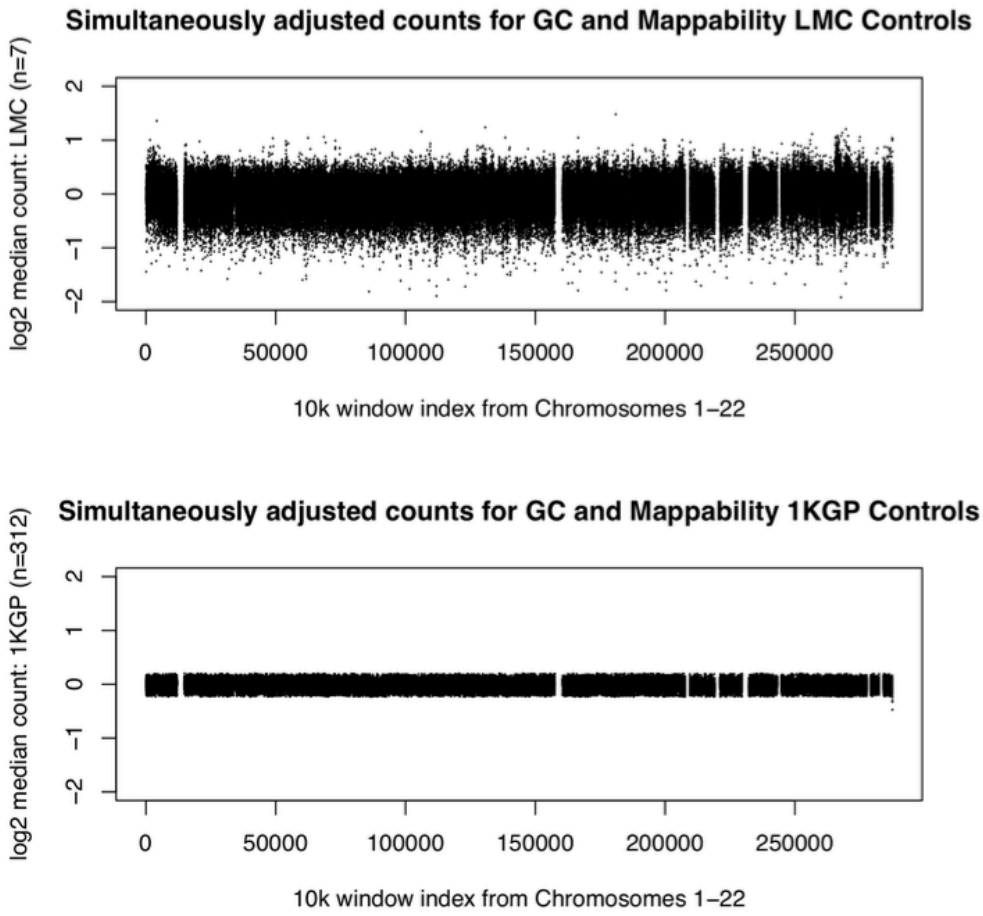


Figure 5.7. Median adjusted read counts per 10K window of LMC samples compared with 312 1KGP samples.

5.4 Discussion

A critical consideration in whole genome copy number analysis of FFPE tumours is quality control during the estimation of the copy number profile. This is likely particularly an issue in dealing with suboptimal DNA quality and low DNA yield and tends to produce noisy copy number profile of the sample [164]. For degradation prone DNA from FFPE, it was recommended to check for the interaction effect of GC content and mappability to the read counts obtained. This can be done by plotting the median read counts at each bin as a function of GC and mappability represented by isobar plot created using available R package *QDNAseq*[2]. After demonstrating the GC content and mappability interaction effect to read counts, adjustment using a two-dimensional (GC and mappability) LOESS model with interaction term was performed. Various authors have investigated the presence of windows which give highly variable results reflecting sequences which are complicated or not unique; the blacklisted windows identified by the various authors have a lot in common reflecting genomic complications rather than technical issues with particular experiments. The combined list that I generated included all previously identified windows from blacklisted regions mentioned in this study and contributed to improve QC even further.

Identifying and masking the additional blacklisted regions in the genome significantly reduced the amount of noise in the data. This study took advantage of the publicly available whole genome sequence data from the 1000 Genomes Project to identify common germline variations and other technical artefacts inherent in the germline using the GRCh38 genome build. These “normal” samples were selected based on Caucasian origin to match the ethnicity of LMC tumour samples making it more able to identify ethnicity specific germline variations.

In May, 2020, Kundaje et al. (2020) published in the ENCODE portal (https://www.encodeproject.org/files/ENCFF356LFX/) a file (identifier:ENCFF356LFX) containing new list of blacklisted regions containing 852 regions (See Appendix E) that maps to 7,182 10K windows in the genome[160, 165]. I cross checked this list with the updated LMC blacklist and found that 6,628 (92.3%) of these windows were common to both lists. The remaining mismatch (554 10k windows) may either be due to overlap to either end of a region (i.e. last few bases at the start or end of a blacklisted region that maps to a window with bigger part outside the blacklisted region but still led to the blacklisting of the entire 10K window) or maybe due to shorter primers used in our methodology making some regions less characterised to be identified as part of a blacklist.

All these mismatches were considered and added in identifying significantly associated regions in the window level analysis yielding to a total of 38,769 blacklisted 10k windows that accounts to 13.5% of the autosomal genome. Matched normal samples were not available for the LMC tumours, a decision having been made to maximise the number of tumours analysed while recognising the implied difficulties of exploring infrequent changes in specific tumours. To provide a germline reference in identifying somatic copy number profile in LMC tumours, two sets of normal samples namely the 7 LMC and the 312 1KGP samples were compared. The 1KGP showed cleaner germline copy number profiles compared with that of 7 LMC normal samples from the skin and was used as the reference of comparison for LMC tumours to identify somatic copy number alterations in subsequent analyses.

Chapter 6
Final Data Quality Assessment

This chapter discusses the assessment of the overall LMC dataset after the additional quality control steps described in previous chapters were performed. It includes replicating analyses from Chapter 4 and incorporating the algorithm *GISTIC* (version 2.023) to identify significant copy number peaks in the genome and identify gene level copy number calls to facilitate the comparison between the LMC and TCGA [4].

6.1 Introduction

Additional quality control steps have shown significant improvement in the data quality as demonstrated in the normal samples. Similar processing was also applied in the copy number data derived from the tumours. This section discusses the analysis repeated mostly from Chapter 4 and compares the results to obtain an objective evidence of data improvement after performing further data quality steps. Replicate analyses were plotted to assess consistency of copy number profiles. The overall genome plots were produced to see the distribution of copy number aberrations across the genome. Similarly, chromosome level plots were generated to provide a higher resolution visualisation of the copy number data. The segments or the number of fragments in the genome were calculated to assess the extent of genomic variation. “Segmented length” defined as the total sum of all identified segments was calculated as a proxy for mappable chromosome size. Consequently, the average number of segments in each chromosome across all the samples were plotted as a function of the segmented length. This has been shown to follow a linear trend in our previous analysis and further quality control steps was expected to enhance this trend. Deviations from linearity could be due to common genetic changes for melanoma.

Further validation included checking for the deletions in the *CDKN2A* region which is the most common form of deletion in most cancers and particularly melanoma. We have previously shown in Chapter 4 the ability of the previous (version prior to additional QC steps) copy number data to detect the known common germline variation esv3620012 (chr9: 23,362,412 - 23,378,071). Following these further QC steps, all identified common germlines variations, including noise due to FFPE processing were masked as I focus on the analysis of the somatic regions.

I also repeated the comparison of TCGA and LMC copy number datasets in terms of TCGA list of genes with deletions or amplifications. This time, a new step to identify

copy number peaks which was included by incorporating *GISTIC* (version 2.023) was employed; this is in keeping with the approach taken within the TCGA analysis. Although there were major differences between the starting material and processes underlying LMC and TCGA copy number data (e.g. FFPE vs fresh samples, NGS vs SNPs Array, primary vs. metastatic), broad similarity between the two datasets was observed as would be expected but is nevertheless reassuring. Finally, I also considered graphically exploring potential influence of normal cells contamination in the copy number profile of the samples. Since normal cells which are mainly composed of stromal cells are mainly diploid, they have the tendency to cause the copy number aberration signal to shrink to diploid state depending on the level of contamination [166].

6.2 Methods

6.2.1 Repeated analyses

For replicate analysis, I generated the plot of the whole genome copy number profile, calculated the number of segments per chromosome and average segmented length, linear plot, window level and segment level correlation analyses using Pearson's r, and assessment of the *CDKN2A* region; these methods are described in Chapter 4. The only difference is the use of the newly adjusted call rate data. The comparison of LMC and TCGA datasets in terms of TCGA list of genes with deletions or amplifications was done by firstly applying *GISTIC* (version 2.023) (described below) analysis to LMC data making it more comparable with TCGA data.

6.2.2 Identification of Significant Copy Number Peaks Using GISTIC 2.023

I used *GISTIC2.023 (Genomic Identification of Significant Targets in Cancer)* to identify regions of significant copy number aberrations in the LMC [4]. This method uses segmented copy number data as an input which are then deconstructed into underlying somatic copy number alterations (SCNAs). It has the ability to identify both focal and arm level aberrations in the genome based on the length of aberrant copy number.

GISTIC2.023 scores each region in the genome according to the probability with which a given list of SCNAs would occur by chance. Regions that yield higher scores are more likely to contain true SCNAs and more likely that they were positively selected. *GISTIC2.023* then identifies independently significant regions as aberrated

regions of the genome. These regions may extend over more than one gene or may contain no gene at all but achieve significance because of its close distance to a target gene. Lastly, *GISTIC2.023* indicates which part of the identified region with aberration is most likely to contain a gene/s being targeted. These regions are called markers. The p-values for each marker is obtained by comparing the score at each locus to a background distribution simulated by random permutation of the marker locations in each sample. This step controls for the sample specific variability in the rate of copy number alteration. The resulting p-values for multiple testing were corrected using Benjamini-Hochberg false discovery rate method and is denoted as q-value [170]

6.2.3 GISTIC Input: Segmented copy number data

The LMC CNV data were segmented using circular binary segmentation (CBS) [134]. An example of one sample segmented using CBS in R mcnv package is shown in Figure 6.1 below. Segments are represented by the black straight lines and the number of its occurrence in each sample is counted. Theoretically, a normal sample compared with a diploid reference would show a somatic copy number profile that is very close to zero measured on the log scale with no breakpoints in the genome, except for the identified blacklisted regions. The more segments or breakpoints a somatic copy number profile has corresponds to more aberration in the genome. As an example, there are at least five segments visible in chromosome 1 below.

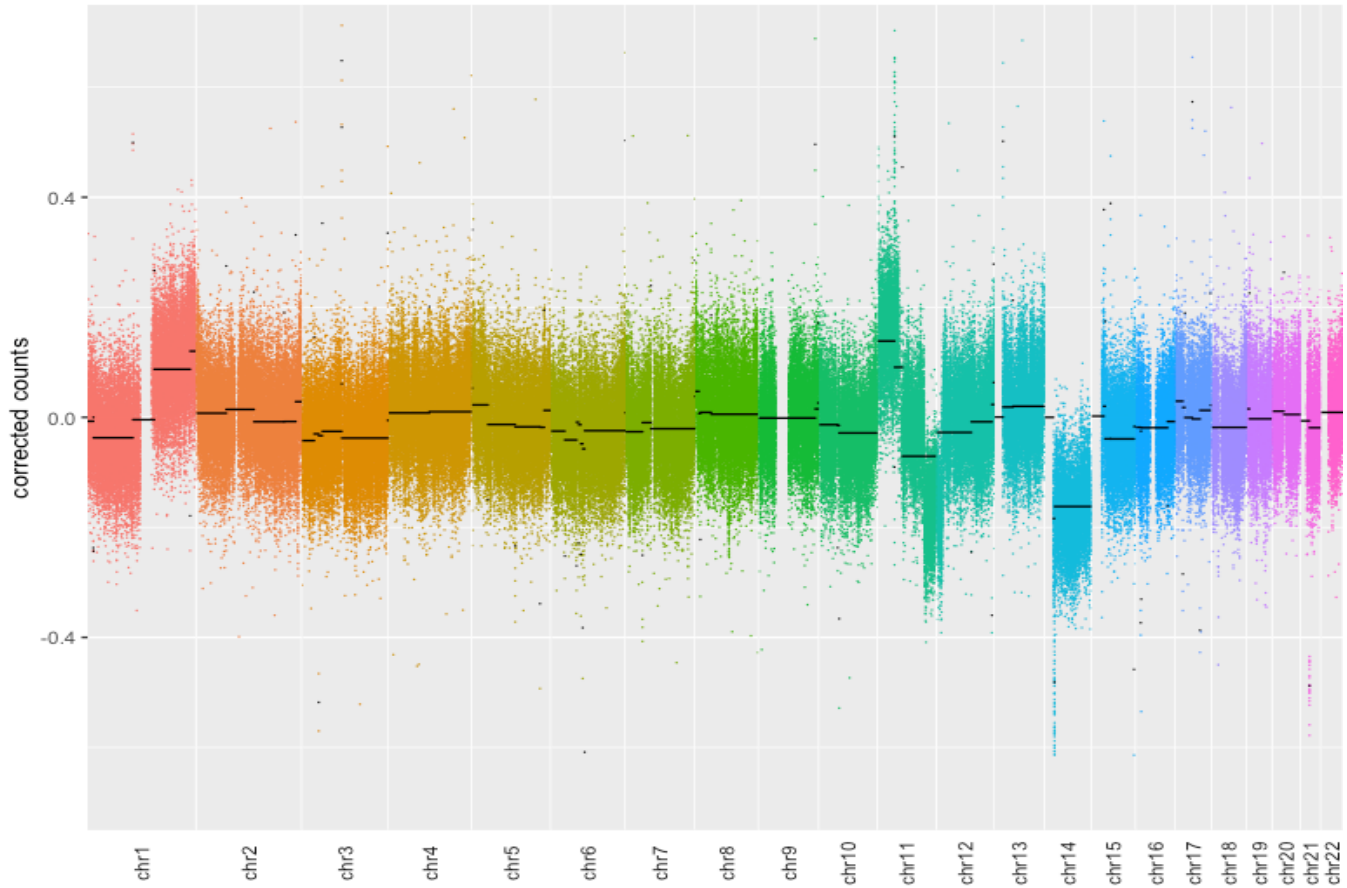


Figure 6.1. Visualization of segments for 1 sample.

Whole of chromosomes Y and X were blacklisted. The y-axis indicates the genomic locations labelled by autosomal chromosomes while the x axis indicates the adjusted read count: log2(LMC/1KGP).

6.3 Results

6.3.1 Whole Genome Plots

Whole genome plots showing copy number profile for all the patients were generated with Figure 6.1 as an example. The plots were manually assessed for quality. Samples showing poor quality copy number profiles were excluded in the further analysis. There were 11 samples removed due to poor quality and are identical to those samples which were previously identified in the old CNA data as having poor quality. Two examples are shown in

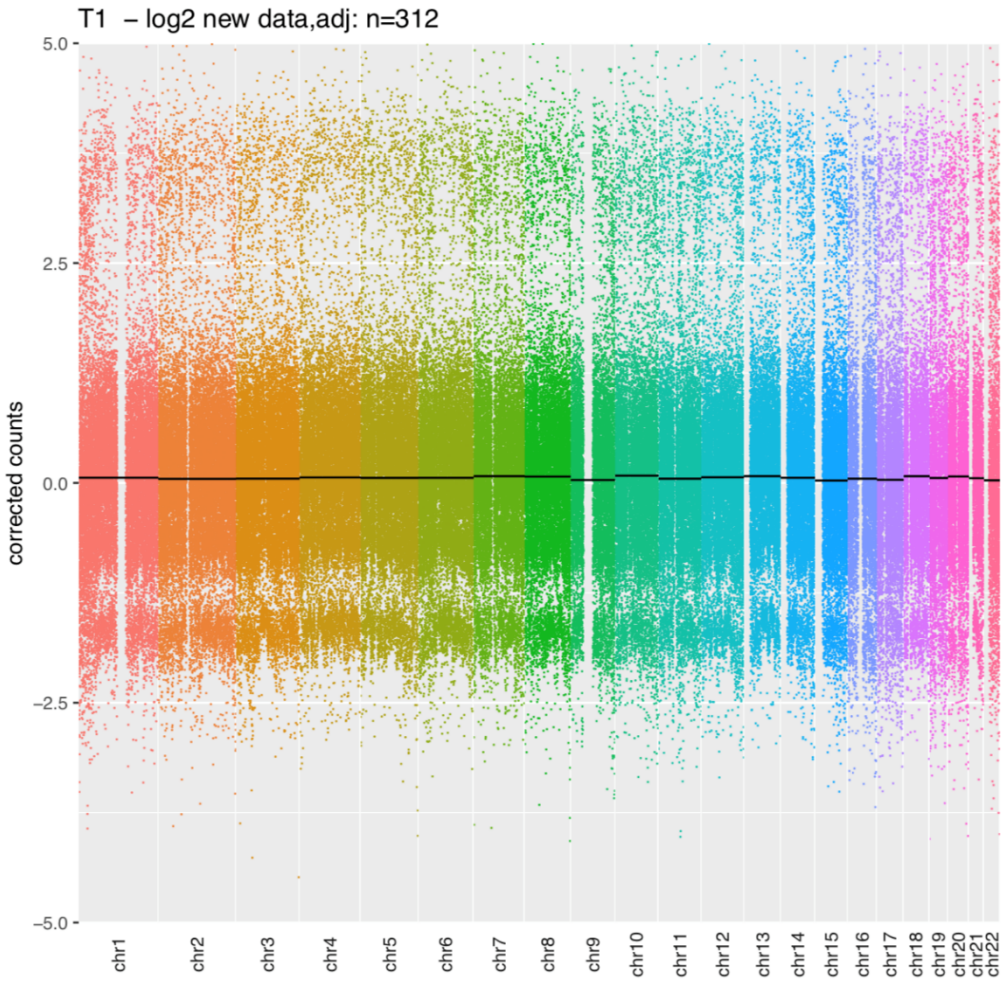


Figure 6.2 and Figure 6.3 (See Appendix A for the rest of the rejected samples) showing highly noisy copy number profiles. The 11 rejected samples have very low alignment rate and were excluded from further analyses. For samples which passed the initial quality assessment, it would be noticed that some segments in the copy number profiles depart from the expected $\log_2$ copy number values of around 0, 0.5, 1, or 1.5. This could be due but not limited to several factors such as: a.) the level of degradation of the DNA sample used in the sequencing as would be expected from FFPE samples, b.) common germline variations, and c.) highly repeated somatic variations. The first two factors considered were addressed in Chapter 5 by accounting for the GC content and mappability interaction in the correction of read counts and utilising publicly available information to remove the highly variable regions in the genome, as well as the common germline variations while the third factor was checked with the literature.

6.3.2 Assessing similarity of replicates

As previously discussed in Chapter 4, the analysis of replicates allows an examination of the extent to which the results are influenced by statistical variation; this analysis of consistency contributes to the determination as to whether quality control is sufficient to allow meaningful comparisons between samples. Each pair of technical replicates (technic, method, and concentration) and biological replicates (tumour, core) were plotted in Figure 6.4 to Figure 6.5. The tumour replicates showed differences in terms of the segmented genome and these are more likely due to the inherent biological variability of the tumour (Figure 6.4). Figure 6.5 shows the genome plot for a pair of core replicates showing high level of similarity with some differences in terms of data quality where the first sample tend to have noisier data. Ten more pairs of core replicates were plotted in Appendix B.2 to B.11 and showed generally similar genomes.

For technical replicates, comparison of paired samples is displayed in Figure 4.6 showing highly similar genomes. Nine more paired samples are plotted in Appendix B and showed highly similar genomes. Figure 4.8 shows the plot for a pair of samples processed using the different laboratory methods for library construction. The two samples show highly similar patterns of aberrations across the genomes. Same level of similarity was observed for the other pairs of samples processed using different methods as plotted in Appendices B.12 to B.20. In Figure 4.10, shows the plot of paired samples processed using different concentrations. The pair of samples displays highly similar genome plots indicating good quality of the data. Same

observation was noticed for the other pair of samples processed in different concentration as plotted in Appendix B.24.

To provide a visualisation of the level of similarity among the different types of replicates used; Figure 6.9 shows the number of fragments of one replicate plotted against of that of another replicate. An overall correlation of 0.50 (P=4.7x10-3 for assessing deviation from randomness) indicate a substantial similarity among the replicates. This measure was heavily affected by the other observations taken from the biological replicates (tumour, core) which were observed to be more variable and less “consistent”. Analysis of replicates including only the technical replicates (technic, method, and concentration) show very high correlations (Pearson’s r=0.91) of the two replicates (P=4.4x10-8).

As previously done in Chapter 4 based on our initial copy number paper [1] , examination of different types of paired samples consisting of cores from two separate tumours from the same patient, two cores from the same tumour and repeat analysis of the sample from the same core was performed as summarised in Table 6.1. This shows that showing that majority (27 out of 30) of the paired technical replicates are significantly correlated ($P < 0.05$ for 2 pairs, $P < 0.01$ for the rest) except for one “core” replicate, and two “technical” replicates. For biological replicates, all four “concentration” replicates are significantly correlated ($P<0.0001$). Three out of four of the “tumour” replicates are not significantly correlated reflecting inherent variability among tumours. These results were highly similar with what was observed in Chapter 4.

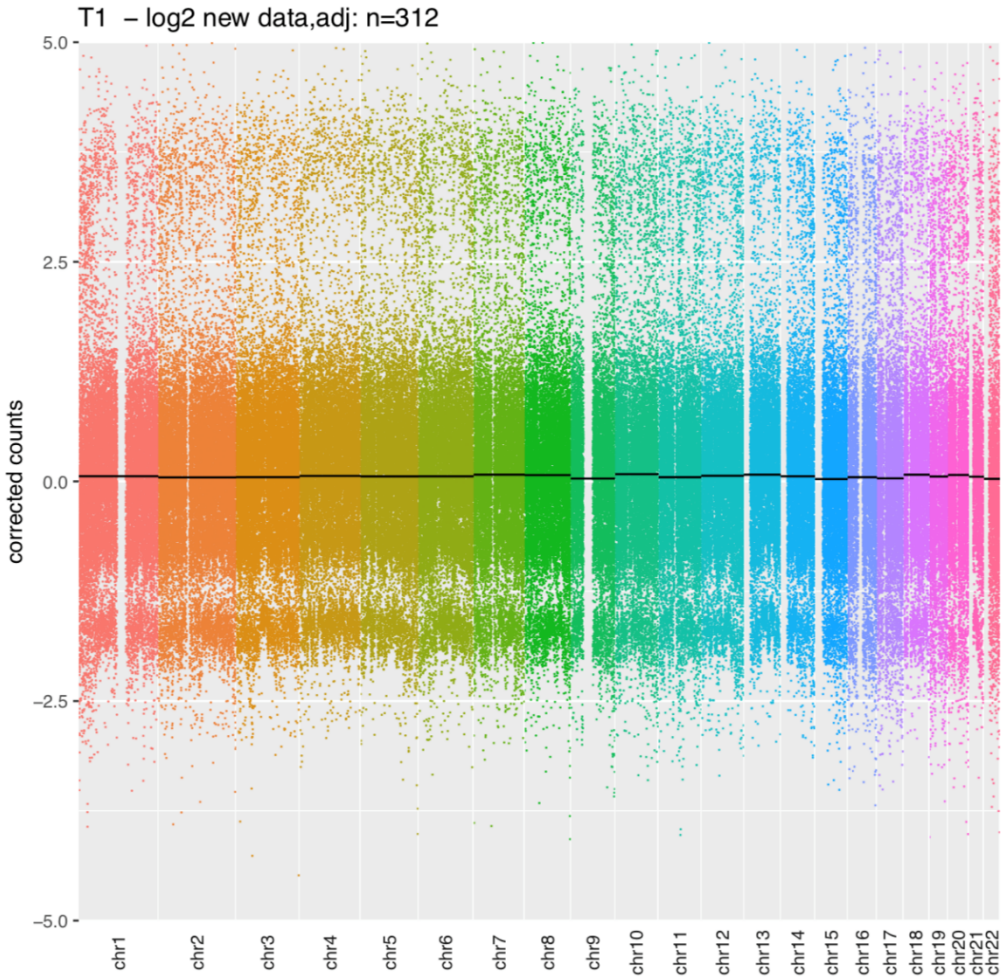


Figure 6.2. Rejected Sample 1. This sample was excluded due to very low alignment rate.

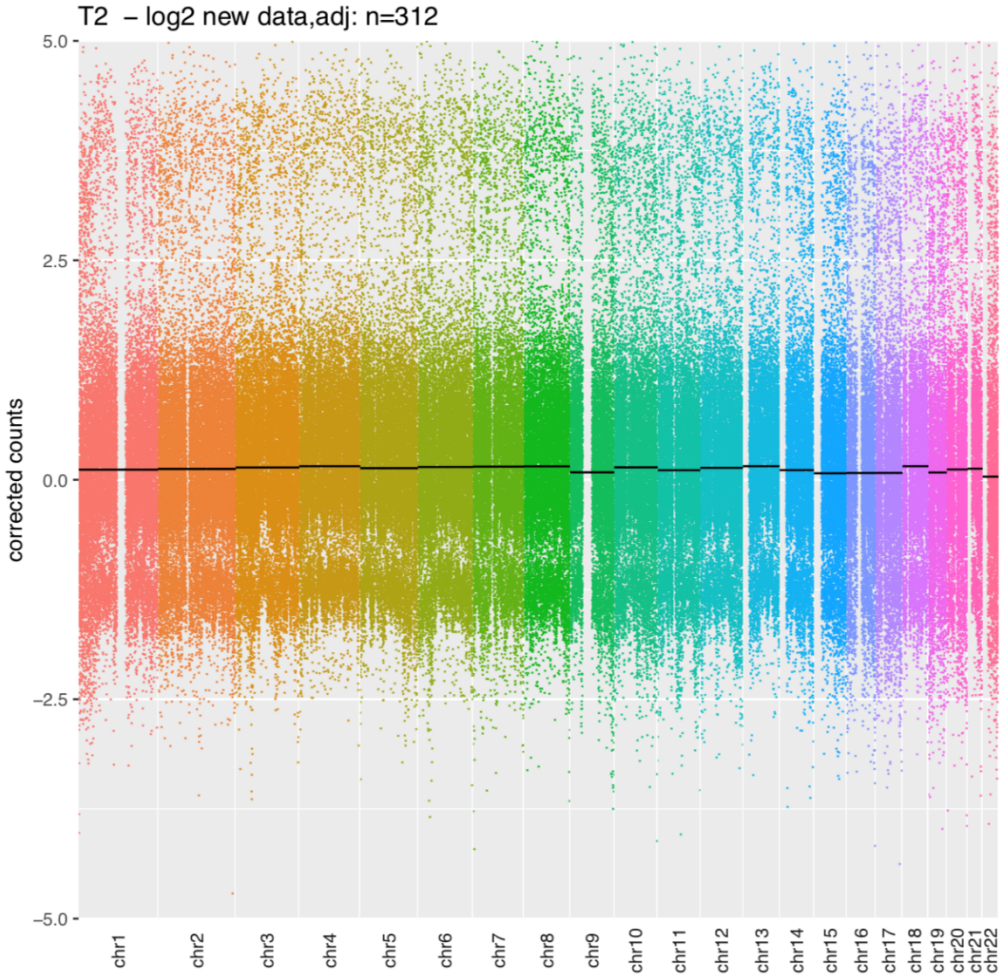


Figure 6.3. Rejected Sample 2. This sample was excluded due to very low alignment rate.

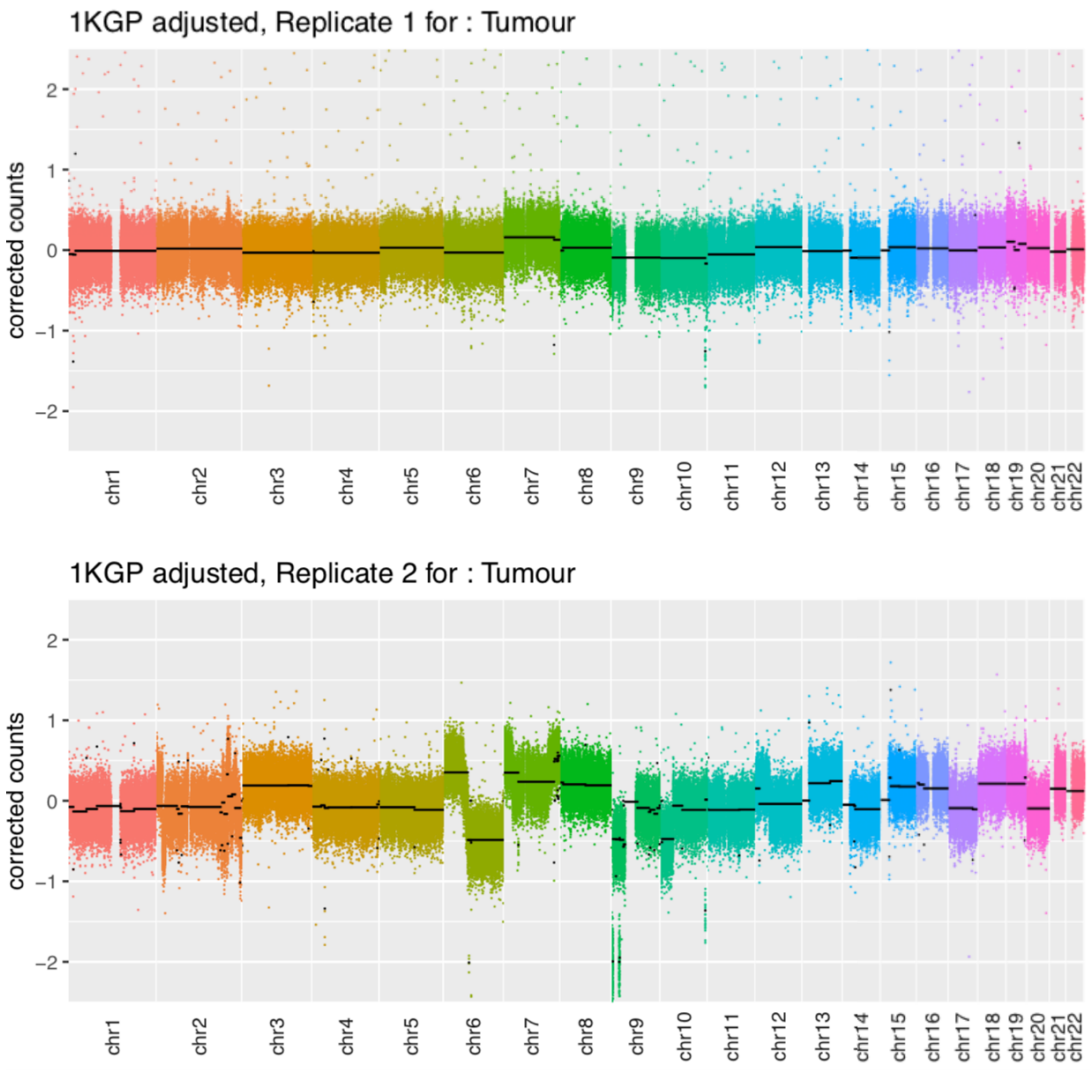


Figure 6.4. Comparison of analyses of 2 tumours from same case, showing notable differences including the 9p region.

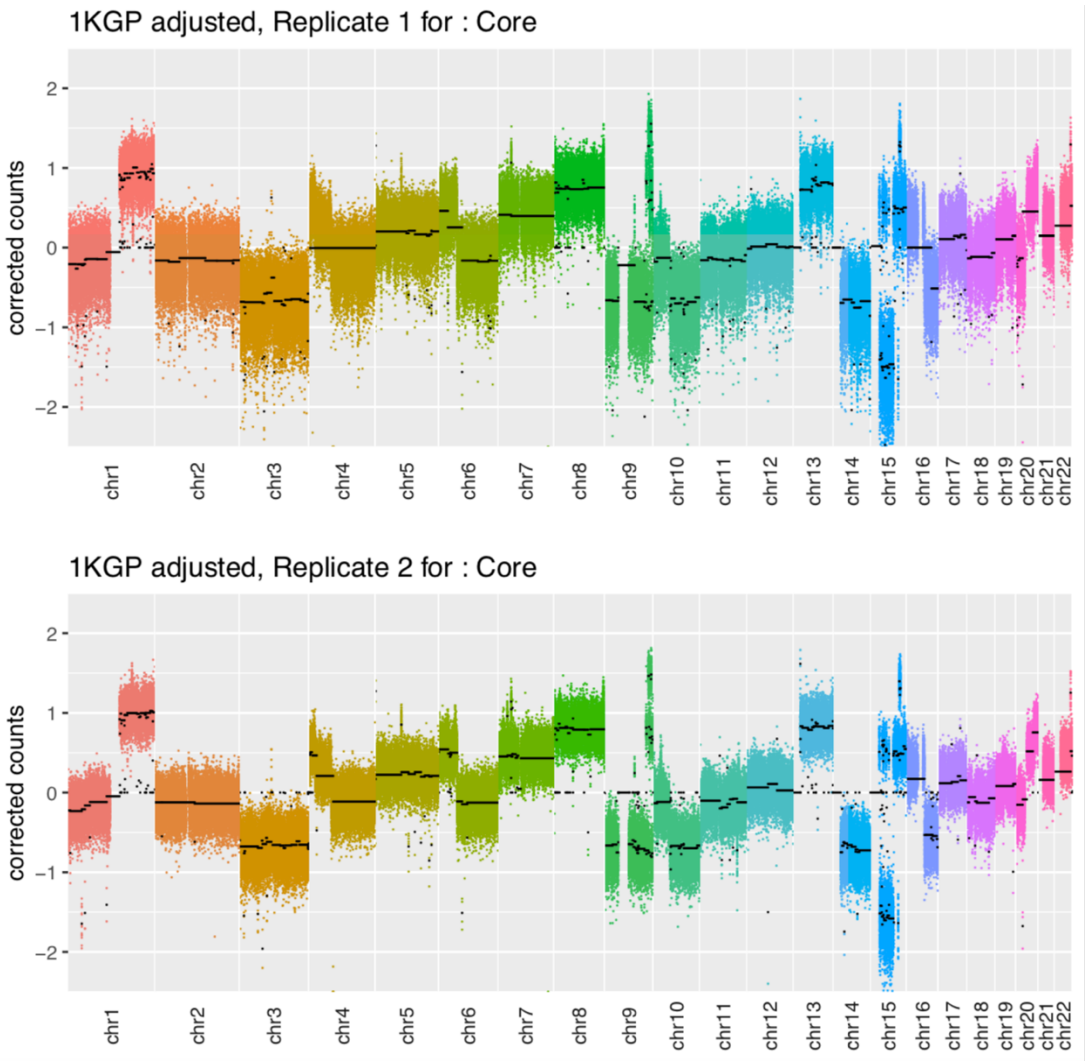


Figure 6.5. Analysis of 2 cores from the same tumour showing overall similarity but with some differences in segmentation pattern (e.g. chromosomes 4p & 6p)

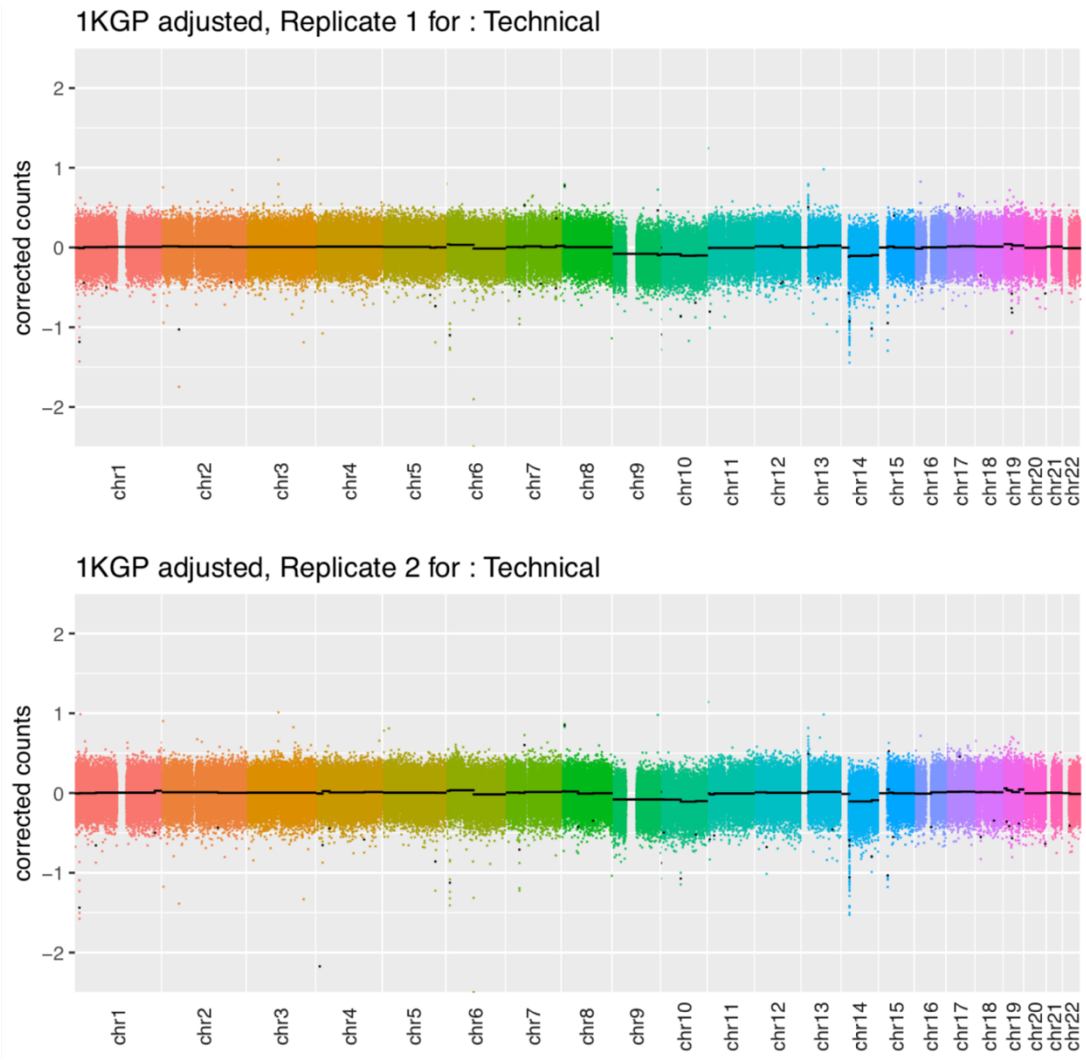


Figure 6.6. Analysis of the same library sequenced twice in this comparatively silent (in copy number terms) tumour showing consistency.

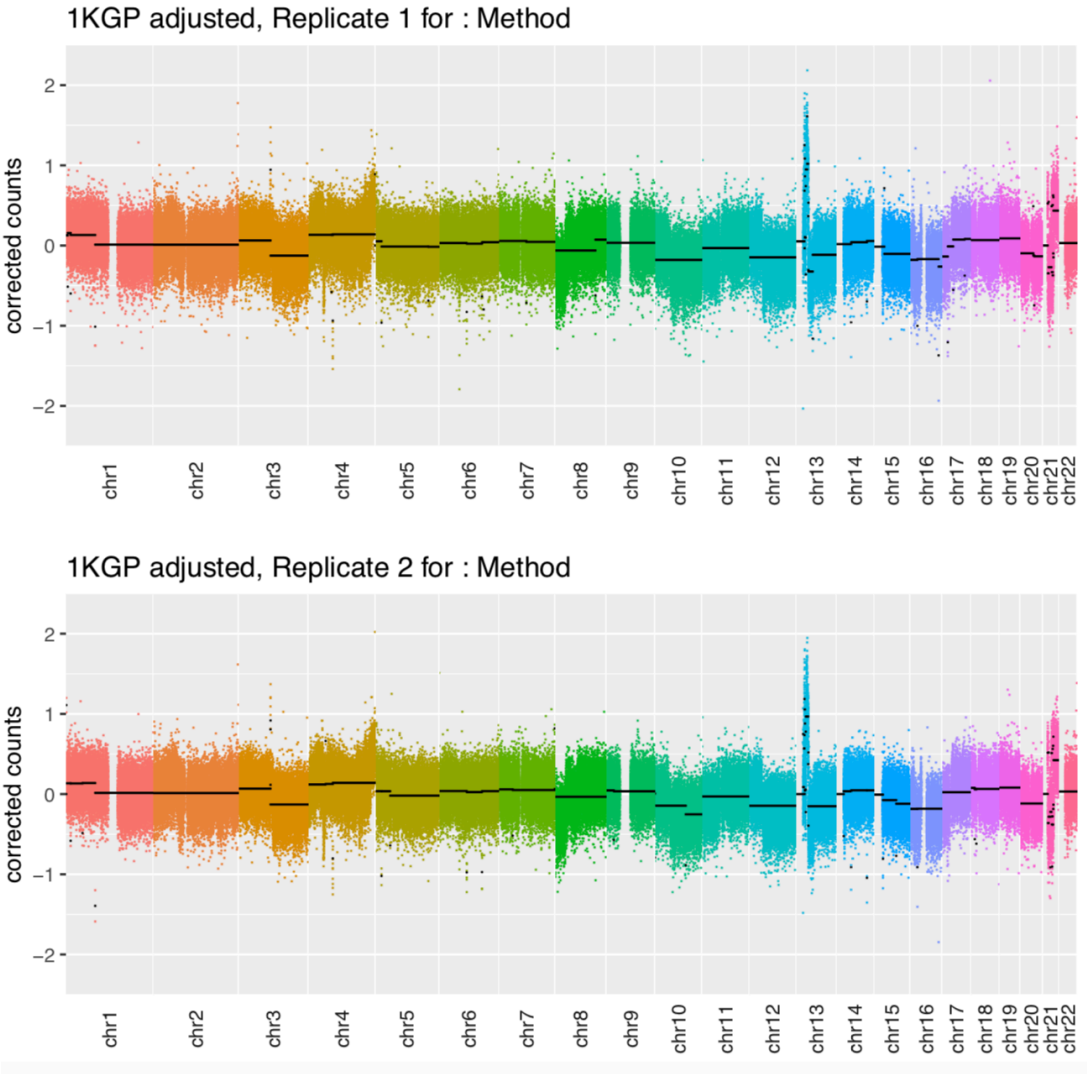


Figure 6.7. Analysis of the same core with libraries prepared by different laboratory methods showing overall similarity in this tumour but some modest differences (e.g. size of segmented regions).

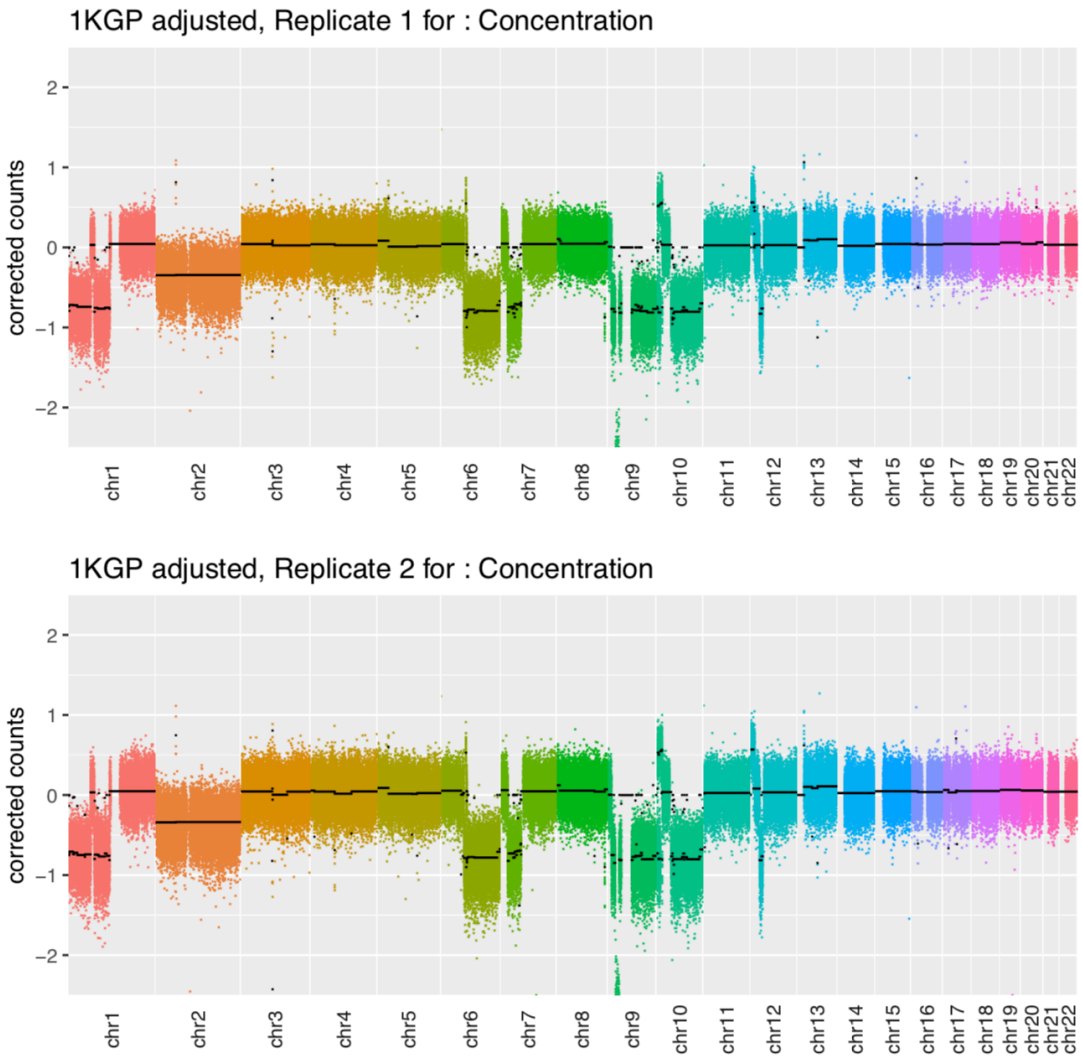


Figure 6.8. Analysis of the same library analysed at two different concentrations showing overall consistency.

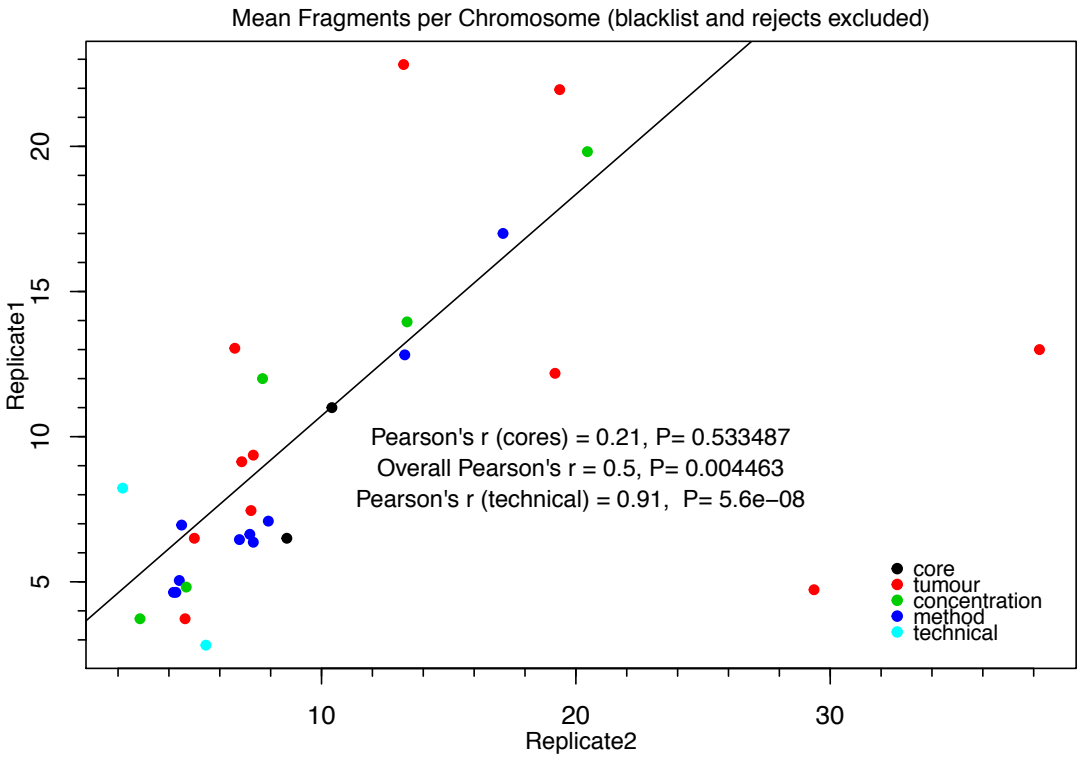


Figure 6.9. Plot of different replicates.

There was a total of 34 samples which were replicated at least twice. Three of these were replicated thrice while the rest are replicated twice resulting to a total of 71 replicated samples. Of the 71 samples, only the top 2 samples of the triplicates were selected in terms of highest mapped reads. Of the remaining 68 samples, 8 samples (from 4 patients) were rejected due to very low coverage. A total of 60 samples (30 patients) were used in the first set of replicate analysis. Of these, 34 were “technical” replicates (technical=20, method=10, concentration=4) and 26 were biological replicates (core= 22, tumour=4)

Table 6.1. Correlation of replications using adjusted read counts

Pair ID	Replicate Type	Pearson's r	Significance
D03	Tumour	0.32	.
D17		0.37	.
D17		0.47	.
D17		0.67	***
D33	Concentration	0.80	****
D34		0.84	****
D34		0.85	****
D34		0.85	****
D26	Core	0.58	*
D02		0.64	**
D11		0.80	****
D01		0.81	****
D18		0.81	****
D29		0.86	****
D05		0.26	.
D07		0.70	***
D13		0.74	****
D08		0.80	****
D14		0.85	****
D16	Method	0.51	*
D23		0.68	***
D31		0.76	****
D04		0.82	****
D12		0.87	****
D06	Technical	0.39	.
D21		0.41	.
D22		0.65	**
D27		0.69	***
D28		0.75	****
D09		0.79	****
D10		0.81	****
D24		0.82	****
D25		0.82	****
D15		0.83	****
D32		0.85	****
D19		0.86	****
D20		0.89	****
D30		0.96	****
Average Correlation		**0.71**	

. not significant at 5% level ** 5 % level*

*** 1 % level* **** 0.1 % level* ***** 0.1 % level*

6.3.3 Calculation of the Average number of Segments and Average Segmented Length per Chromosome

The summary for number of fragments and average segmented length in each chromosome in the genome is presented in Table 6.2 below. The segmented length is from 46.68 mb to 248.77 mb with chromosome 1 as longest and chromosome 21 as the shortest. The number of segments per chromosome in each sample is from 1 to 189 segments. The mean number of segments is from 2.8 to 23.4 with chromosome 21 having the smallest and chromosome 1 having the highest. It is worth noting that after the largest of the chromosomes in terms of size (chromosome 1), chromosome 7 showed most variability in relation to its size, as compared with the other chromosomes.

Table 6.2. Summary of fragments and segmented length in each chromosome

The sex chromosomes X and Y are blacklisted.

Chromosome	Mean Segmented Length (base pairs)	No. of Fragments			
		Mean	SD	Min	Max
1	248,766,514	19.2	23.4	1	189
2	242,079,186	11.8	16.6	1	158
3	198,164,493	13.3	12.2	1	72
4	190,121,512	9.6	11.9	1	100
5	181,395,844	14.3	16.0	1	107
6	170,657,609	15.0	12.6	1	64
7	159,177,311	17.1	22.5	1	109
8	145,029,263	11.0	10.4	1	64
9	138,258,858	13.9	13.8	1	70
10	133,670,968	12.8	15.0	1	92
11	134,953,539	13.5	19.9	1	180
12	133,157,931	12.0	13.2	1	115
13	114,287,205	8.0	9.1	1	78
14	106,979,869	6.7	6.5	1	40
15	101,895,836	10.0	13.4	1	86
16	90,260,006	7.9	7.7	1	47
17	83,154,837	10.4	11.7	1	89
18	80,333,413	4.3	5.1	1	30
19	58,534,790	8.4	6.8	1	48
20	64,406,593	4.0	5.9	1	58
21	46,681,975	2.8	3.9	1	36
22	50,750,957	6.8	13.2	1	117

*SD: standard deviation

To better understand this high variability of segments in chromosome 7, it was compared to a similarly sized chromosome (chromosome 6). Figure 6.10 shows comparison of chromosomes 6 and 7 in terms of the number of segments across the 303 patients. It can be seen that both the two chromosomes have very high number of segments in some patients. A higher frequency of aberration in chromosome 7 is observed as compared to the frequency of aberration in chromosome 6, as well as in terms of amount of variability.

The number of segments from Table 4.1 in Chapter 4 Section 4.3.5 was compared in the number of segments in this chapter in Table 6.2 in terms of mean ranks of the segments using Wilcoxon test. Table 6.3 summarises the results of the comparison showing that the number of segments in the new data significantly decreased in comparison with the old data. The old data have 1.2 (chromosome 12, P=0.008) to 7.4 times ($P=4.12 \times 10^{-88}$) more number of segments on the average than the new data. This is primarily attributed to the improvement in the quality of the data after the additional quality control steps were done as discussed in Chapter 5.

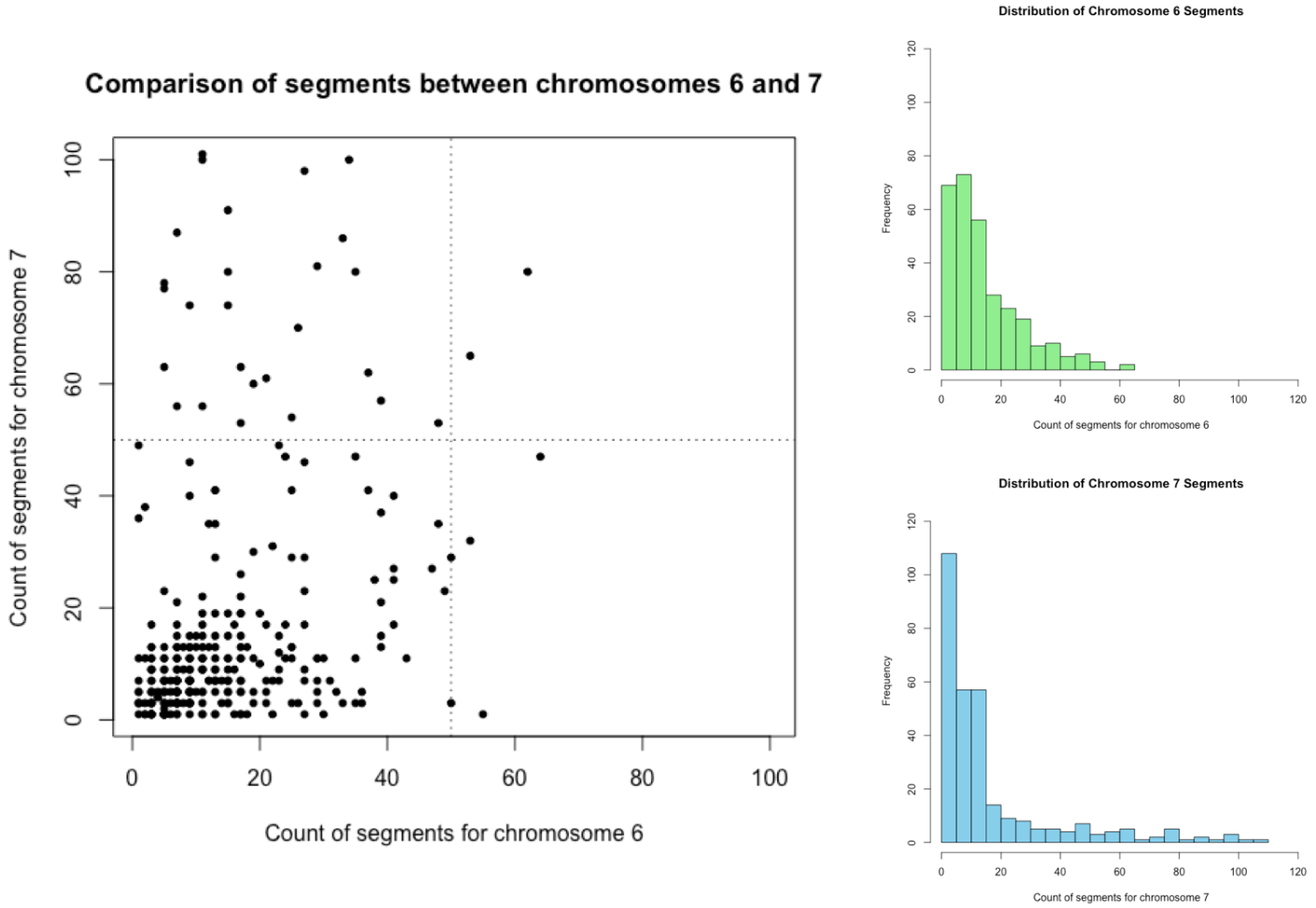


Figure 6.10. Comparison of number of segments of chromosomes 6 and 7 across samples showing similarity across many samples but with notably more segments on chromosome 7

Table 6.3. Summary of comparison of fragments each chromosome between the old and new data.

The sex chromosomes X and Y are blacklisted.

Chromosome	Median(Old)	Median(New)	Fold change(median)	Wilcoxon P-value
1	37	11	3.4	5.05E-47
2	36	9	4.0	3.92E-68
3	14	9	1.6	1.49E-06
4	13	7	1.9	3.90E-18
5	22	9	2.4	1.53E-43
6	17	11	1.5	3.14E-10
7	43	9	4.8	2.90E-55
8	17	7	2.4	4.82E-27
9	17	9	1.9	3.18E-16
10	20	7	2.9	5.38E-29
11	19	7	2.7	3.78E-29
12	11	9	1.2	8.44E-03
13	13	5	2.6	2.42E-33
14	7	5	1.4	3.31E-09
15	26	6	4.3	8.13E-66
16	37	5	7.4	4.12E-88
17	24	7	3.4	4.61E-58
18	7	3	2.3	8.66E-32
19	11	7	1.6	2.43E-13
20	5	3	1.7	5.29E-17
21	4	1	4.0	1.00E-26
22	14	3	4.7	1.02E-49

6.3.4 Linearity of number of segments with segmented length

An initial assessment to check the quality of segmented data is done using linear regression where linearity of number of segments in a chromosome and segmented length (approximate for chromosome length) is shown in Figure 6.12 . A significant linear relationship between number of segments in a chromosome and segmented length is detected ($P=1.9 \times 10^{-5}$). This shows improvement based on the same analysis applied to the old data ($P=5.2 \times 10^{-3}$). A separate plot excluding replicates is shown in Figure 6.12 showing the same trend with very slight decrease ($P=2.2 \times 10^{-5}$) in the significance of the linear regression. A high mean number of segments was observed in chromosome 7 as shown previously.

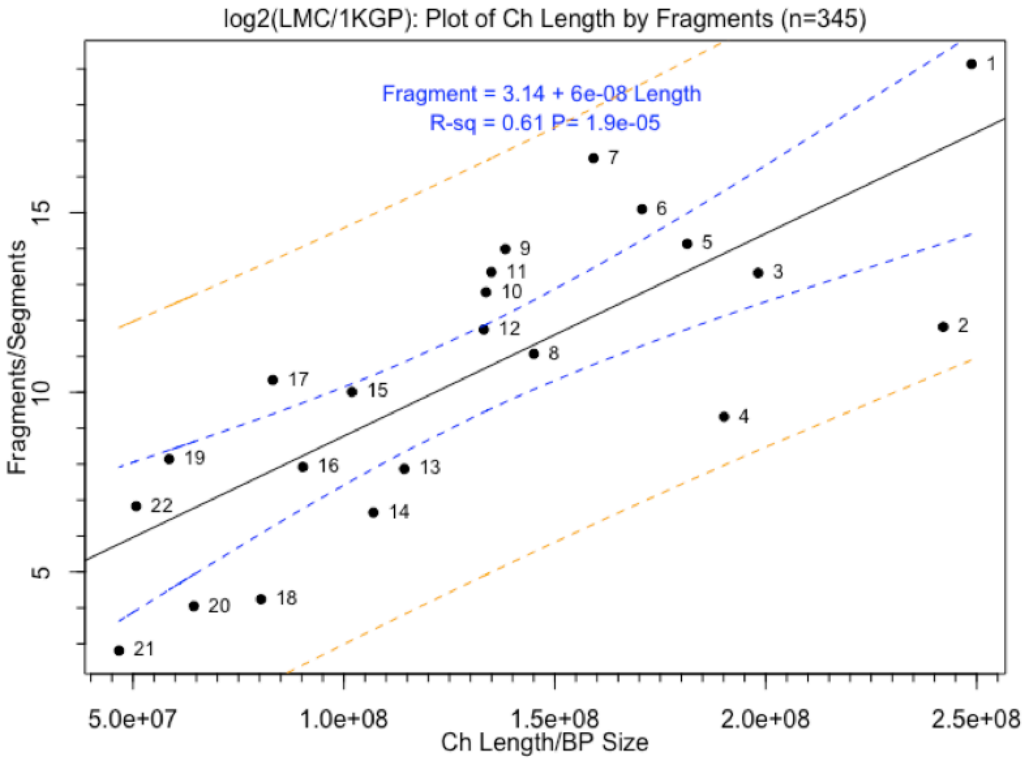


Figure 6.11. Plot of mean segment by mean segmented length by chromosome including replicates

X axis represents the mean number of segmented lengths of in each chromosome while Y axis represents the mean number of segments in each chromosome. The blue lines represent 95% confidence interval around the regression line while the orange line represents the 95% confidence interval around the prediction.

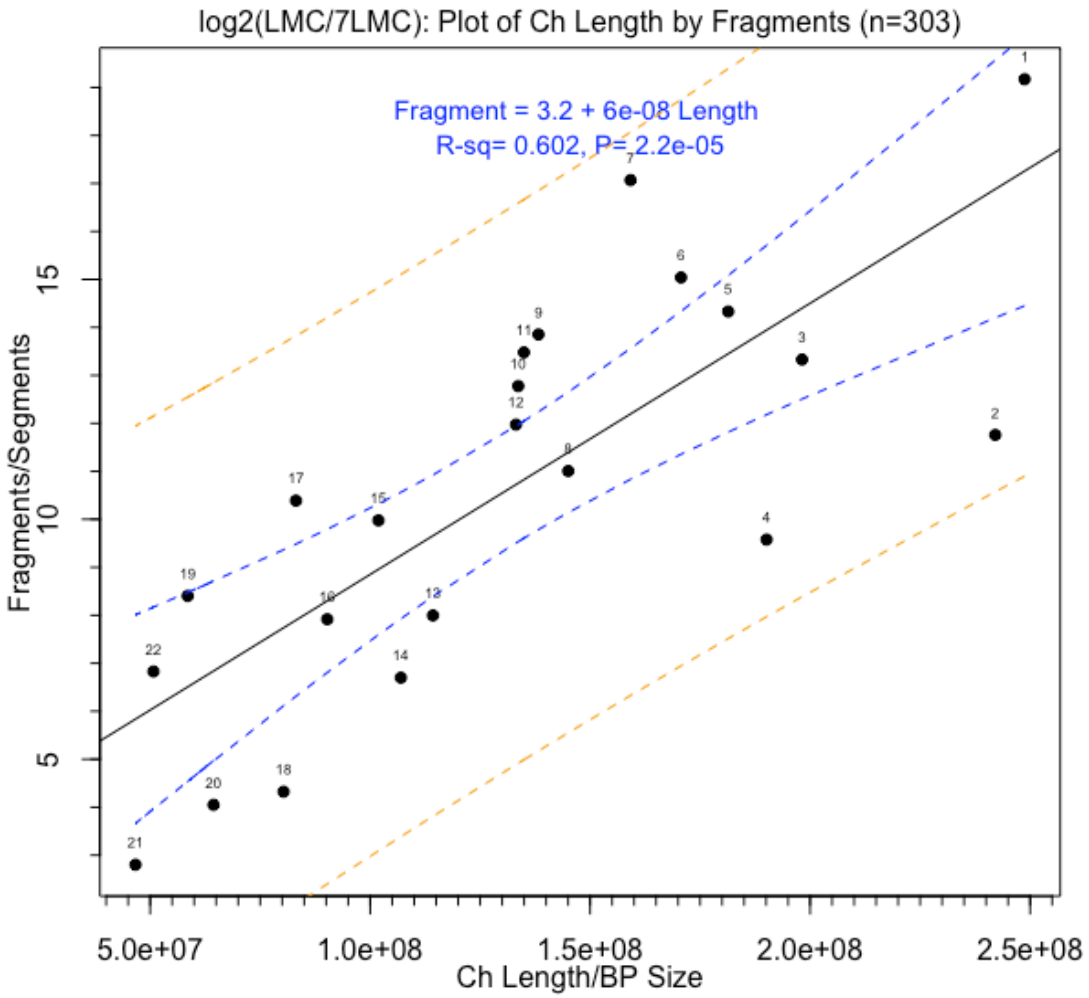


Figure 6.12. Plot of mean segment by mean segmented length by chromosome excluding replicates.

X axis represents the mean number of segmented lengths of in each chromosome while Y axis represents the mean number of segments in each chromosome. The blue lines represent 95% confidence interval around the regression line while the orange line represents the 95% confidence interval around the prediction.

6.3.5 The *CDKN2A* Region

A focused examination of the copy number profile in the *CDKN2A* region was done being the most common copy number alteration in melanoma and indeed in many forms of cancer [71]. The new data has successfully detected the deletions in the samples with known deletion of the *CDKN2A* region. Comparing the plots of *CDKN2A* region from the old data and the new data reveals that the patterns of loss identified by the old data were also identified using the new data with some differences. A

careful comparison and interpretation of the new findings in the data was guided by Dr. Mark Harland, a biologist with particular expertise about this region.

Figure 6.13 to Figure 6.16 show four samples (T27, T46, T52, T192) where the new data was able to identify (via segmentation using CBS) *CDKN2A* deletion which were not previously identified by the previous data. Figure 6.17 shows one sample (T129) where the noise in the copy number data produced was reduced and has led to the interpretation of a single window deletion as compared to a three-window segment deletion in the previous data. Figure 6.18 , Figure 6.19, and Figure 6.20 highlight the important impact of the new data where more complicated patterns of loss or deletion in the *CDKN2A* region were identified and has provided a tidier and better resolution of the aberration in three samples (T152, T180, and T242). Figure 6.21 and Figure 6.22 show two samples (T132 and T215) whose noise was reduced enabling the detection of a larger region of loss or deletion while Figure 6.23 shows reduction in the noise of the copy number data which has led to identification of a more focused region of loss in one sample (T213) as compared to the old data. For Figure 6.24, the sample (T4) presents a classic example of a copy number loss which starts and ends approximately halfway along a window at each end resulting to copy number segments at each end that has copy number value between loss and normal and has been classified by the new data as a separate segment. Figure 6.25 shows a sample (T91) where the resolution of two separate regions of loss was reduced and led to a single region of loss in the new data.

Furthermore, 6 samples (T12, T54, T139, T186, T197, T285) have regions which show loss but were not detected by the segmentation in both old and new data (See Appendix C.1 to C.6).

Although the new data showed some difference with the detection of the *CDKN2A* region as compared with that of the old data, it showed generally more beneficial impact in terms of recognising the *CDKN2A* deletions that were not previously identified as well as revealing more complicated patterns of loss. Overall, the average of the correlation of the paired replicates is 0.71.

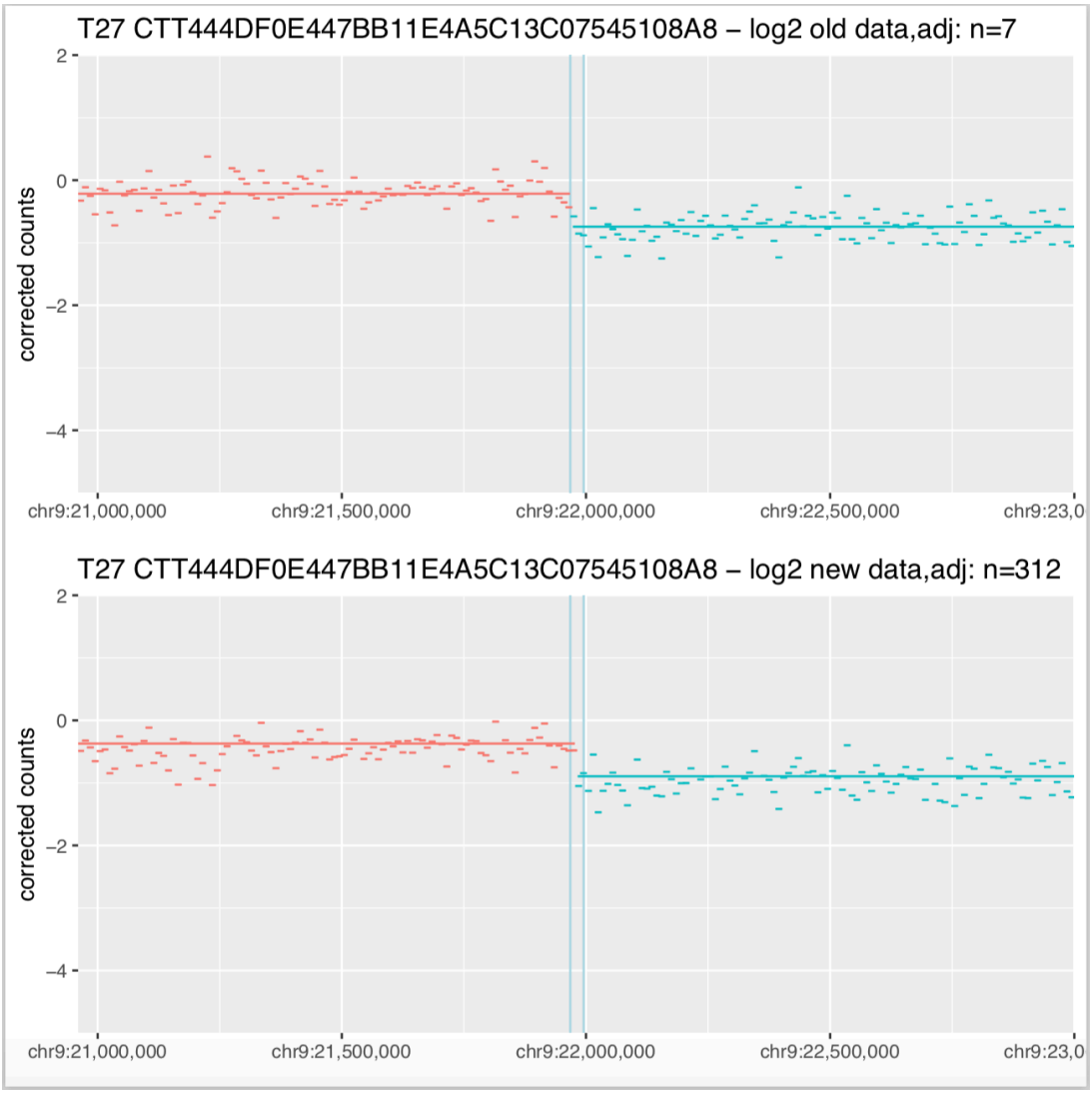


Figure 6.13. The *CDKN2A* region for Sample 27 showing improved resolution of *CDKN2A* region.

The top figure reflects the original analysis and the lower figure following more extensive QC.

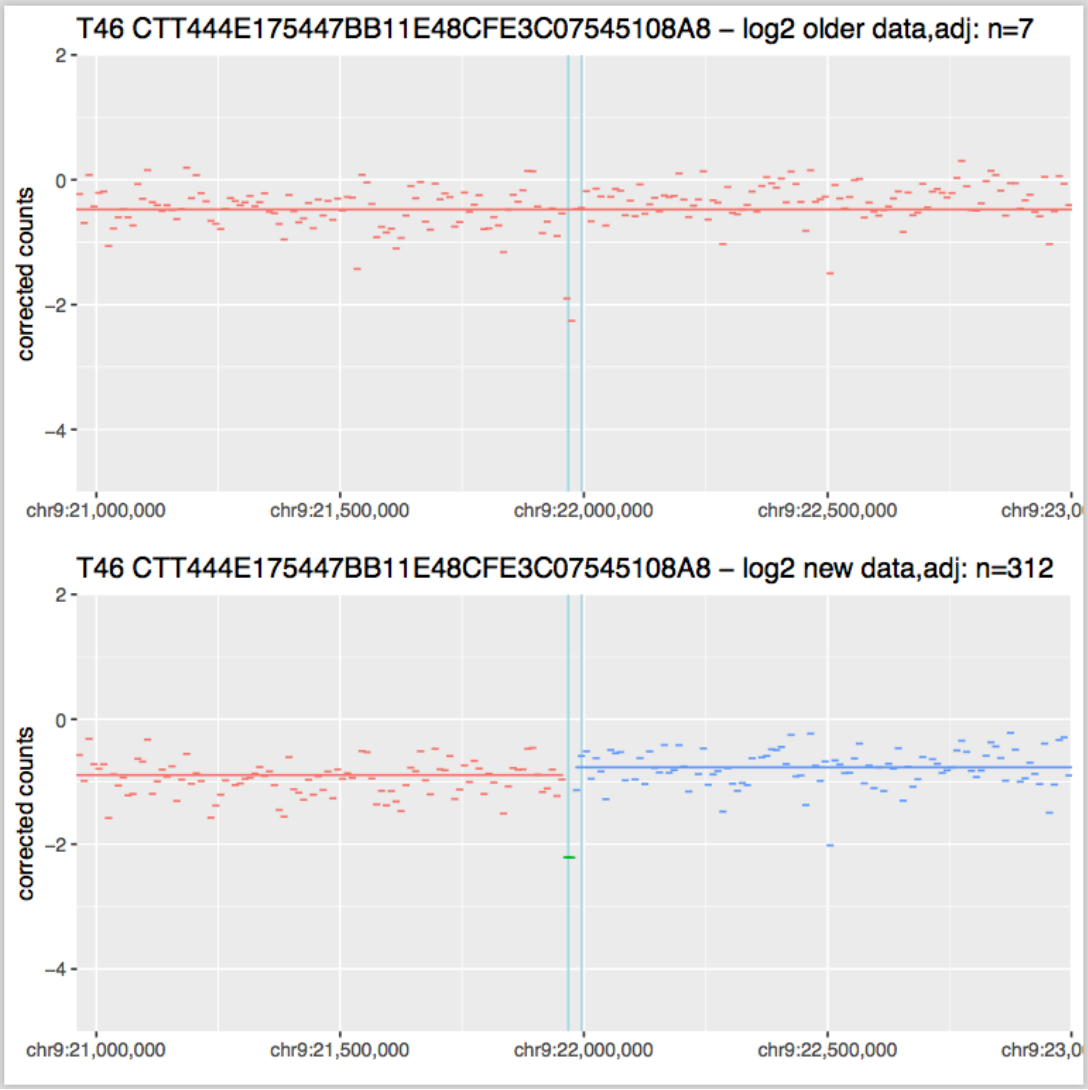


Figure 6.14. The *CDKN2A* region for Sample 46 showing improved resolution of *CDKN2A* region.

The top figure reflects the original analysis and the lower figure following more extensive QC.

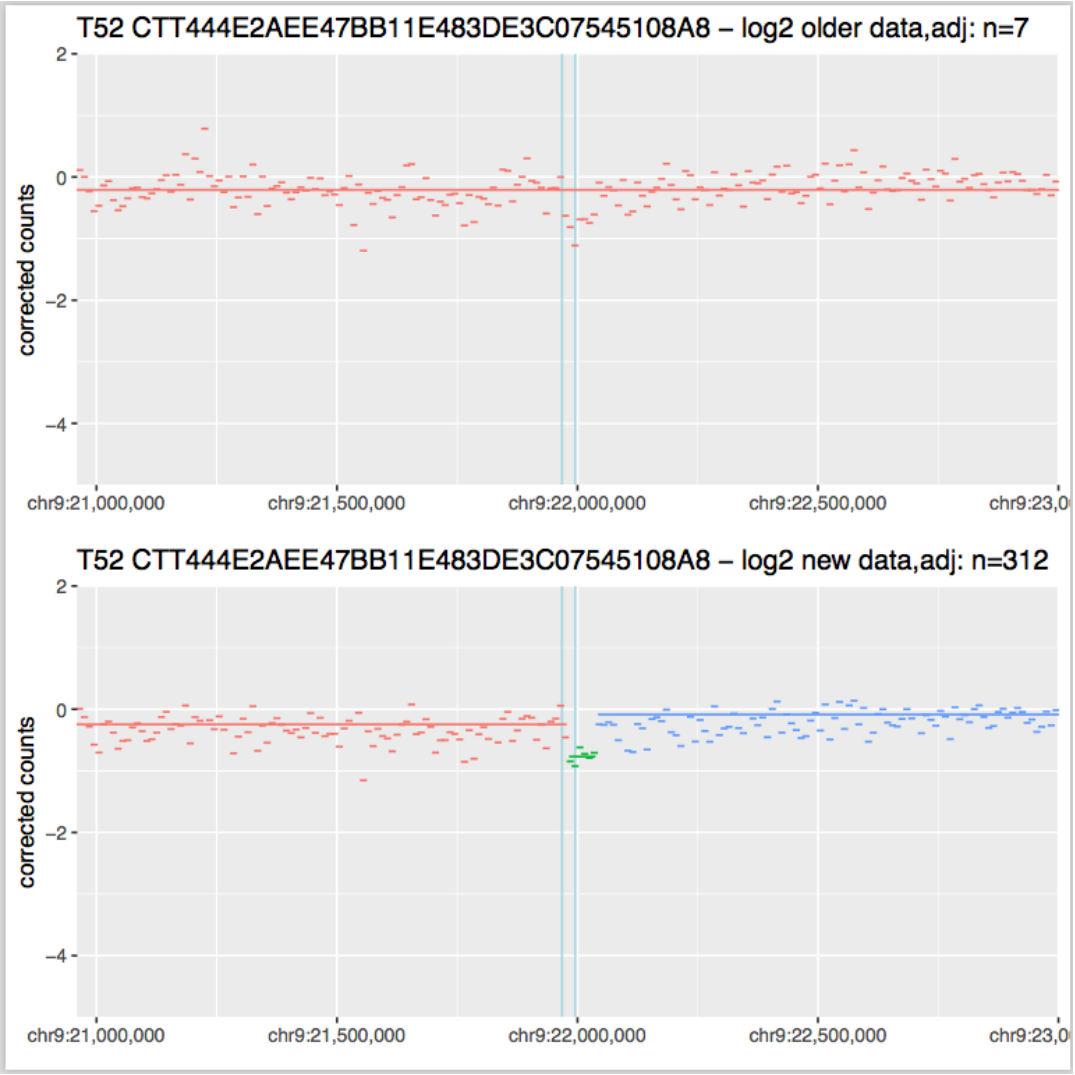


Figure 6.15. The *CDKN2A* region for Sample 52 showing improved resolution of *CDKN2A* region.

The top figure reflects the original analysis and the lower figure following more extensive QC.

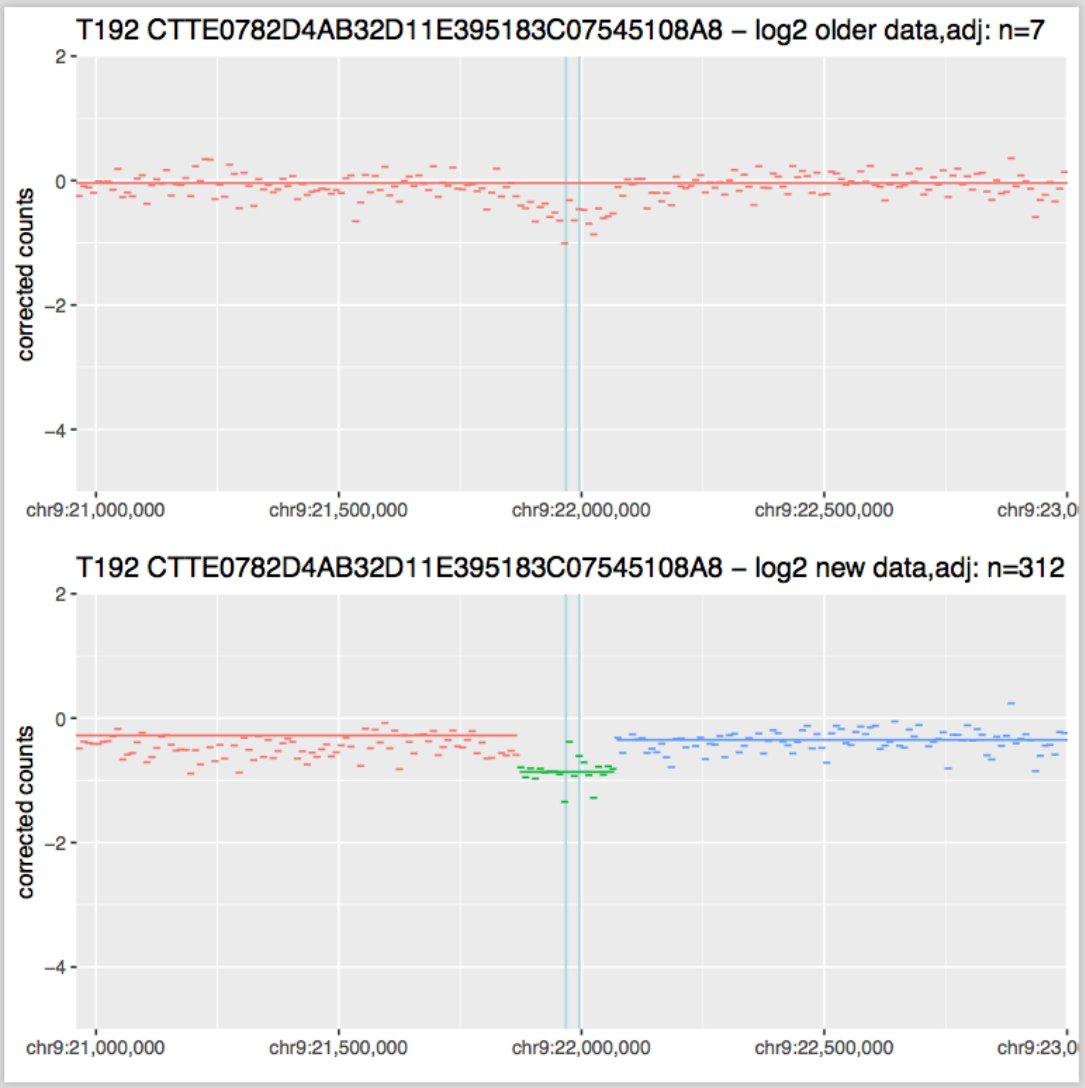


Figure 6.16. The *CDKN2A* region for Sample 192 showing improved resolution of *CDKN2A* region.

The top figure reflects the original analysis and the lower figure following more extensive QC.

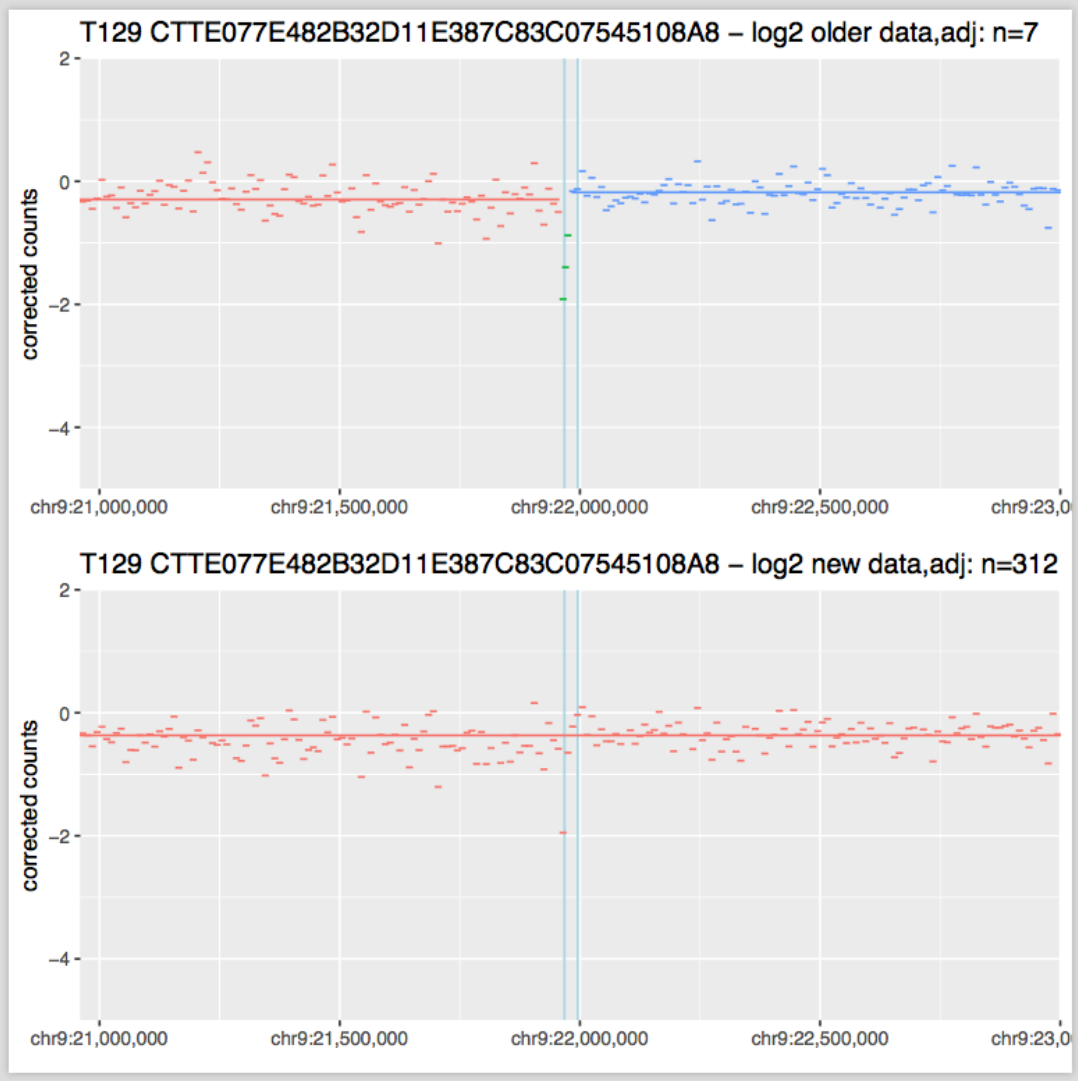


Figure 6.17. The *CDKN2A* region for Sample 129 showing improved resolution of *CDKN2A* region.

The top figure reflects the original analysis and the lower figure following more extensive QC.

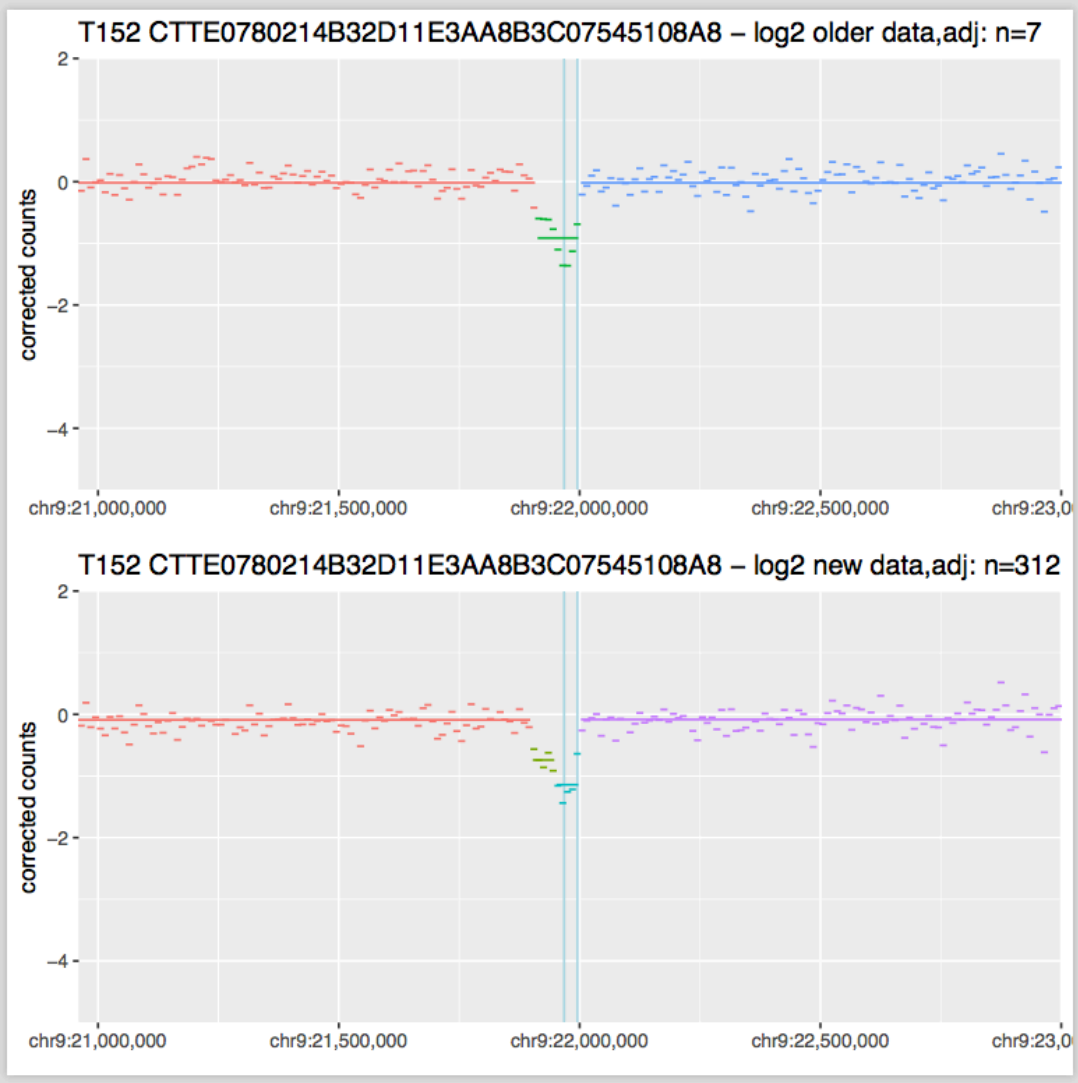


Figure 6.18. The *CDKN2A* region for Sample 152

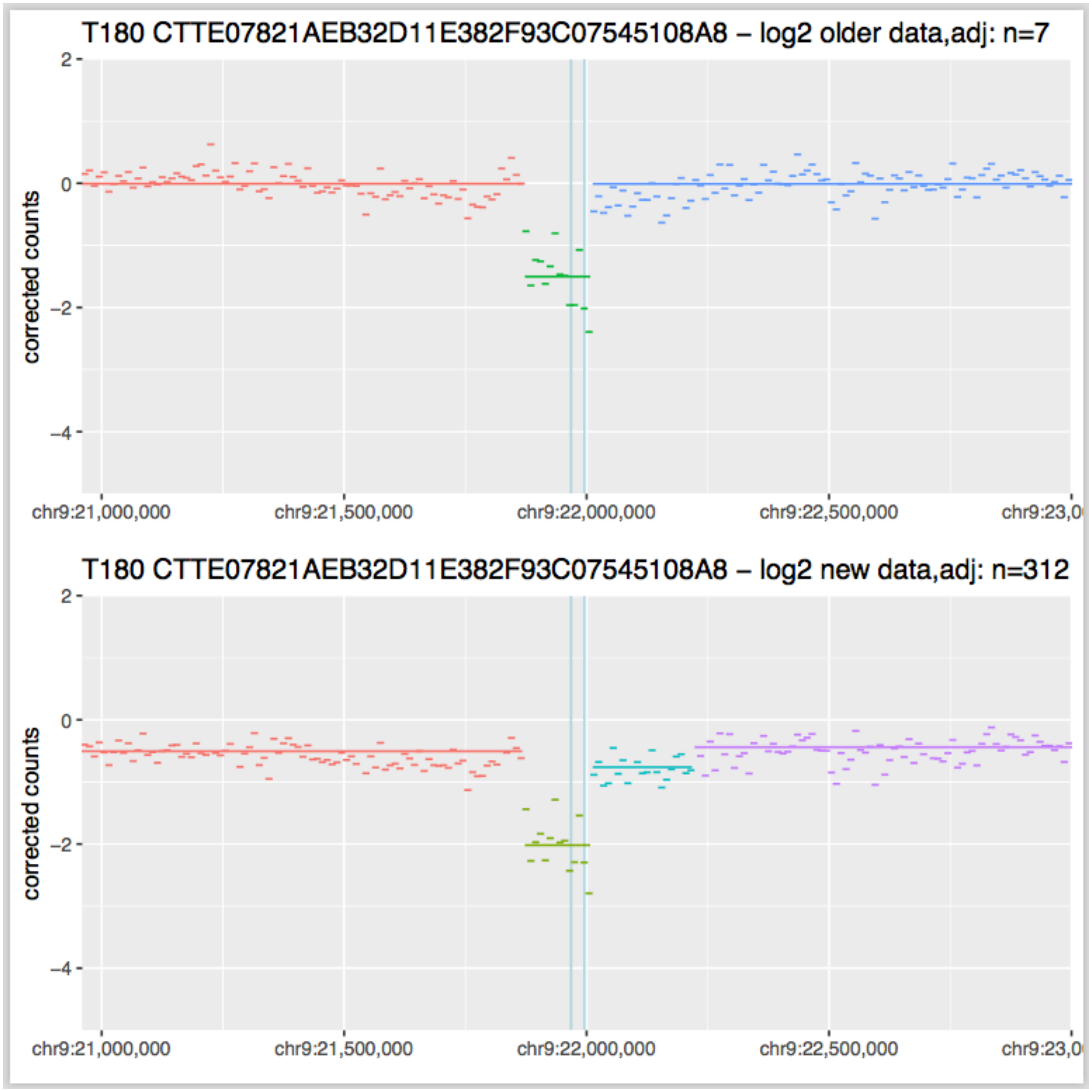


Figure 6.19. The *CDKN2A* region for Sample 180

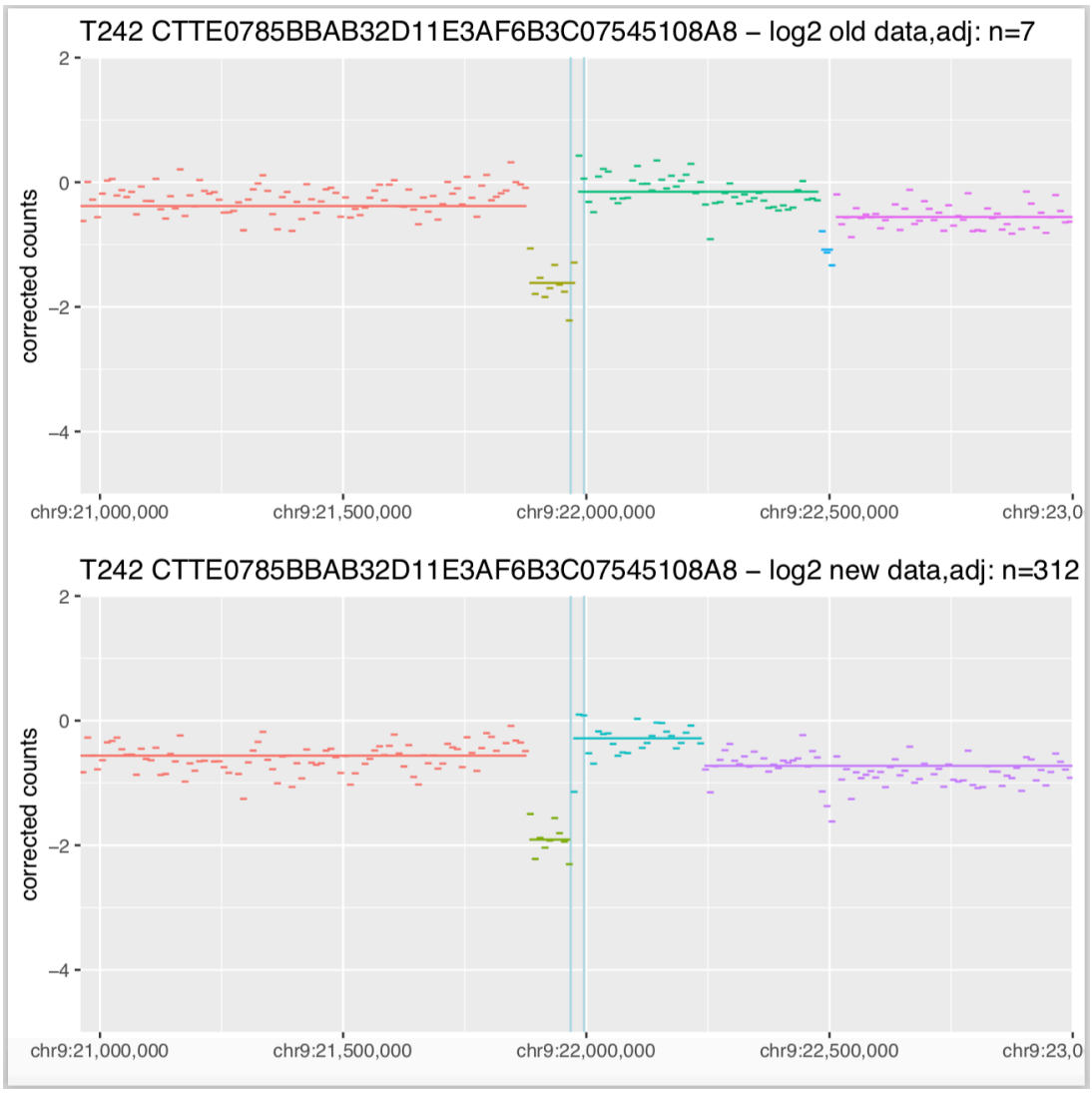


Figure 6.20. The *CDKN2A* region for Sample 242

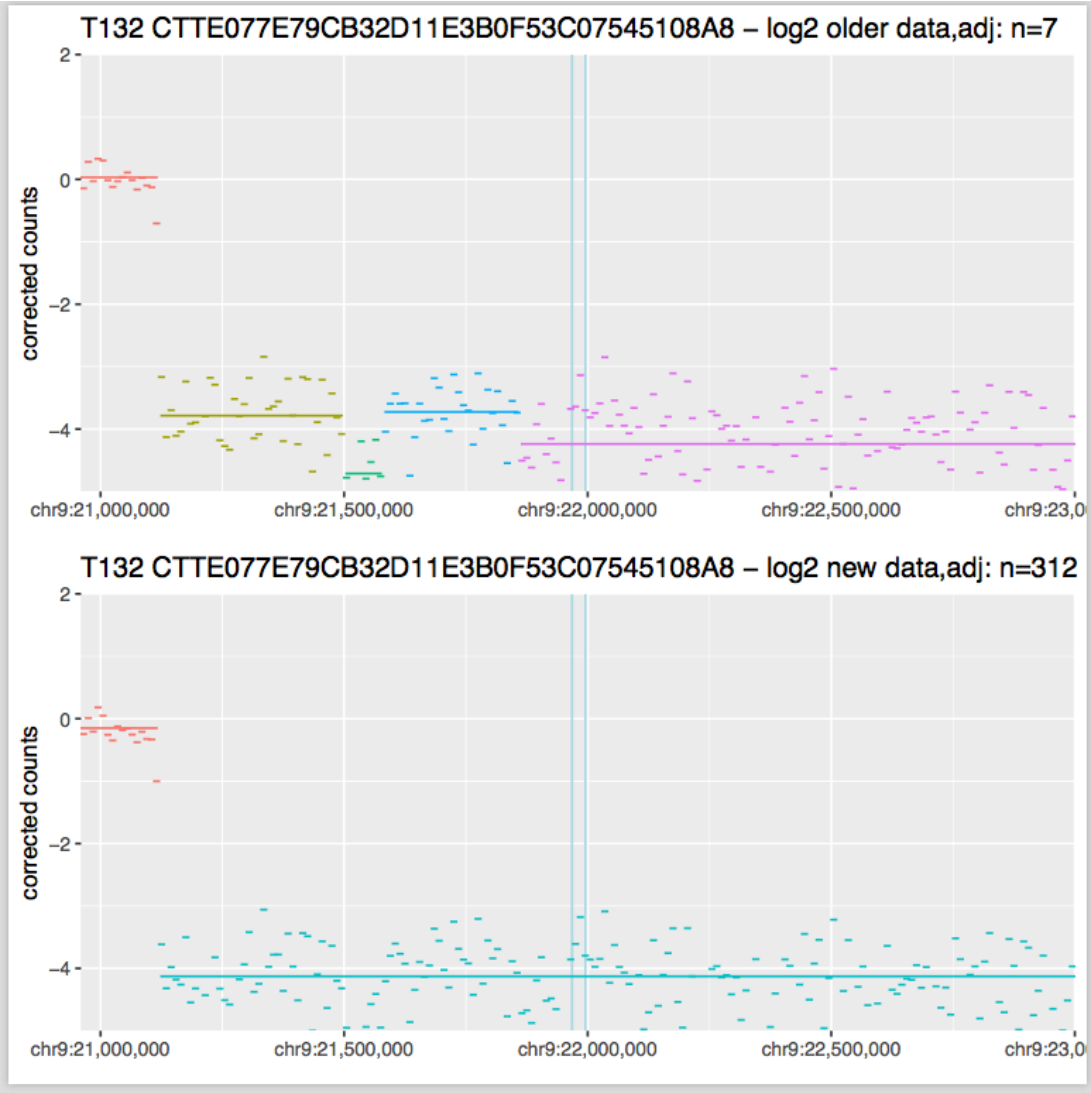


Figure 6.21. The *CDKN2A* region for Sample 132

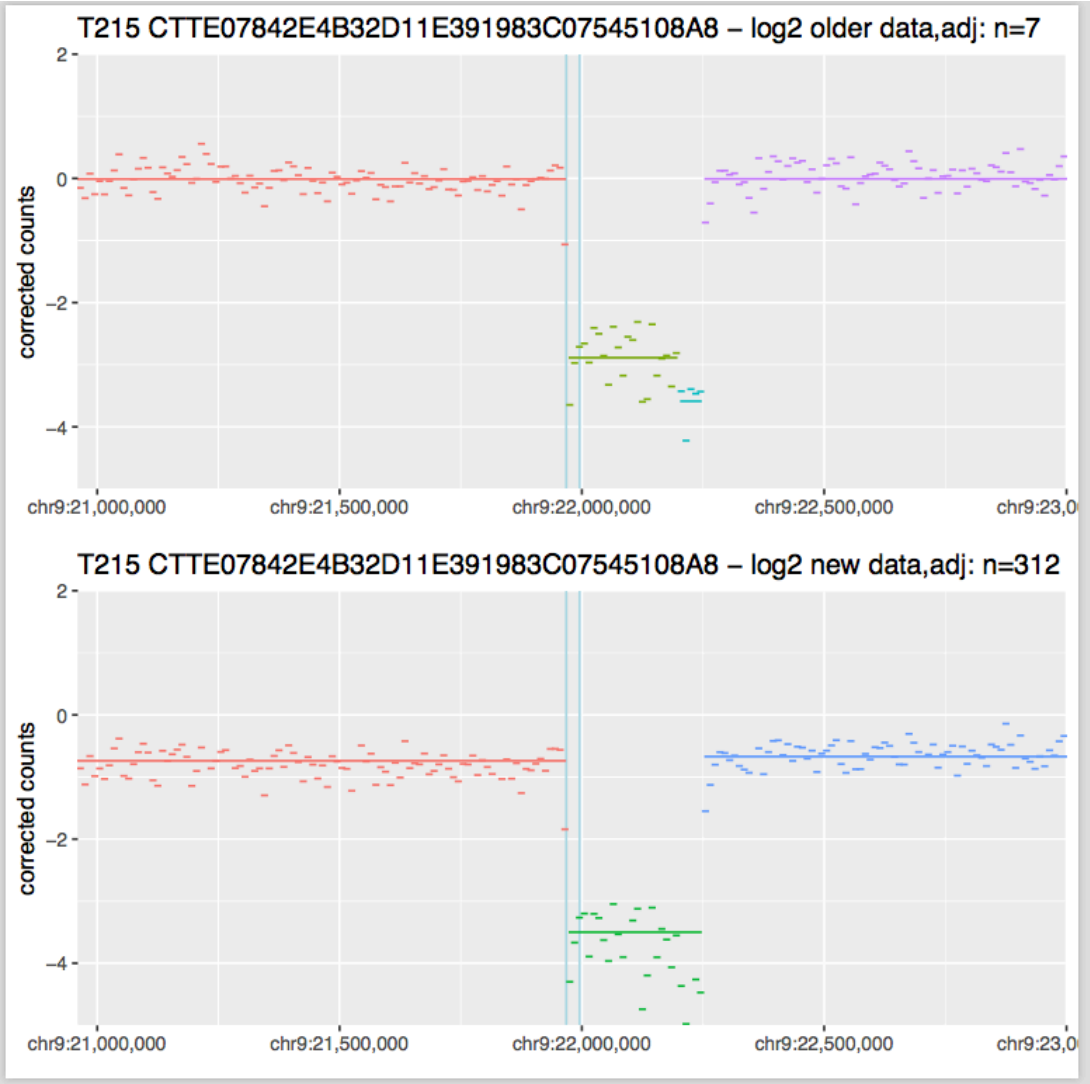


Figure 6.22. The *CDKN2A* region for Sample 215

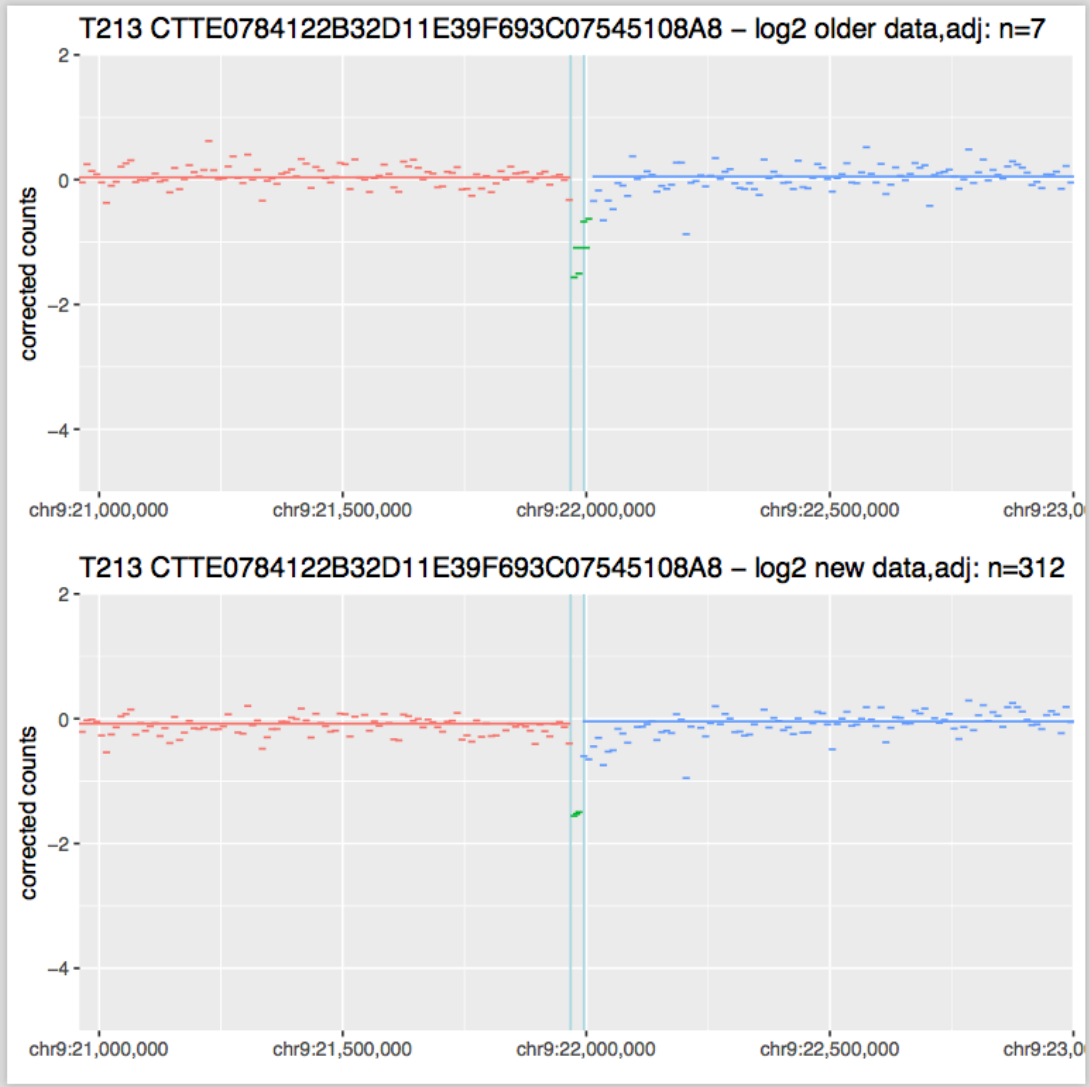


Figure 6.23. The *CDKN2A* region for Sample 213

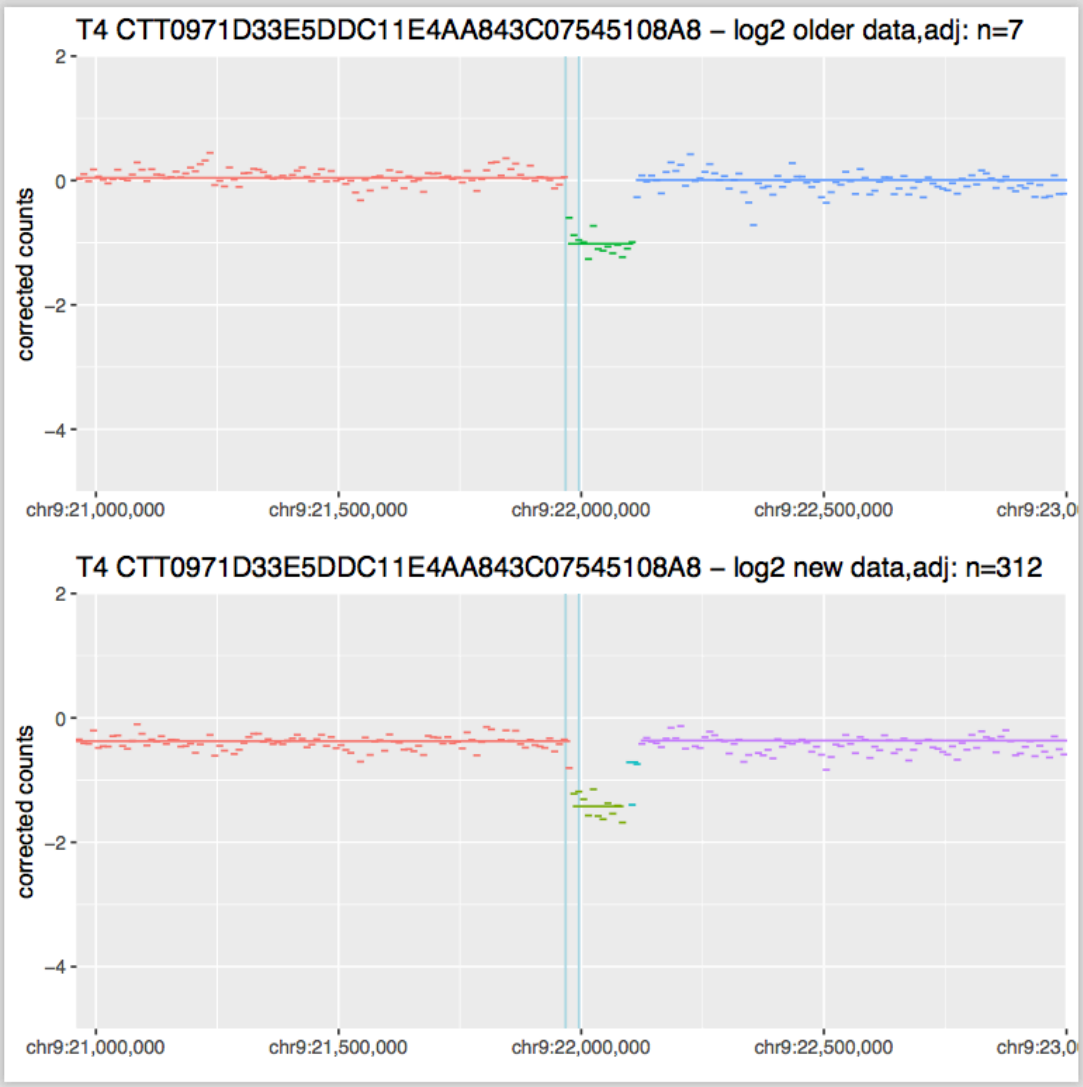


Figure 6.24. The *CDKN2A* region for Sample 4

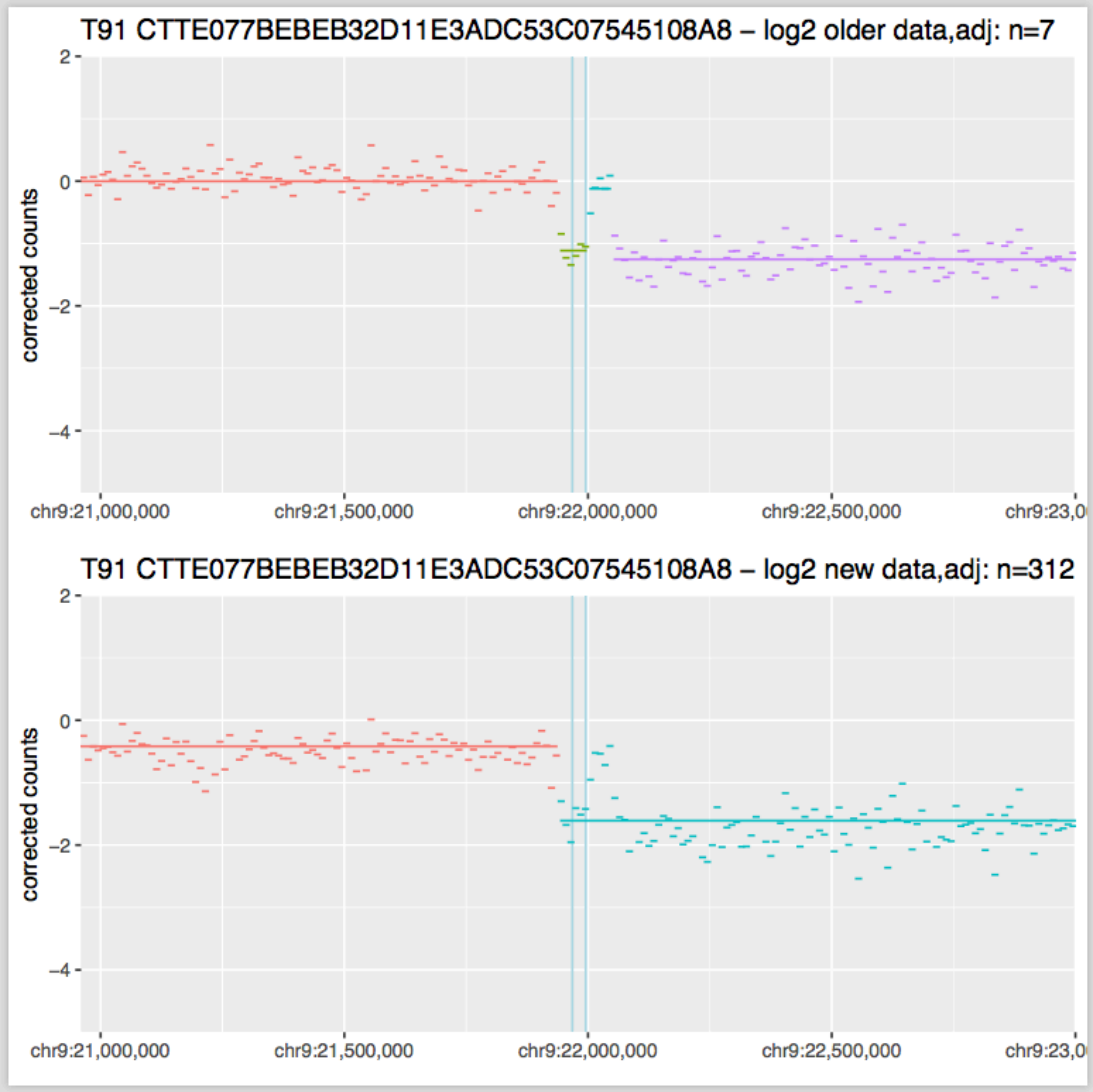


Figure 6.25. The *CDKN2A* region for Sample 91

6.3.6 The ESV Region

The analysis of the esv3620012 was previously performed in Chapter 4 [136]. This region is a direct example of common germline variation that could be potentially included in the blacklisted region identified in Chapter 5. Figure 6.26 and Figure 6.27 show the plot of two samples (T3 and T5) for the *CDKN2A* region with the ESV region next to it as highlighted by the second blue vertical line. For samples where deletion of this region was previously detected in the old data (shown by the drop of 3 windows in the second blue vertical line in the Figure), it can be seen that the data in this region has been successfully masked in the new data as resulted by additional data quality steps (blacklisting) that identifies poorly mappable regions, and common germline variations in the genome.

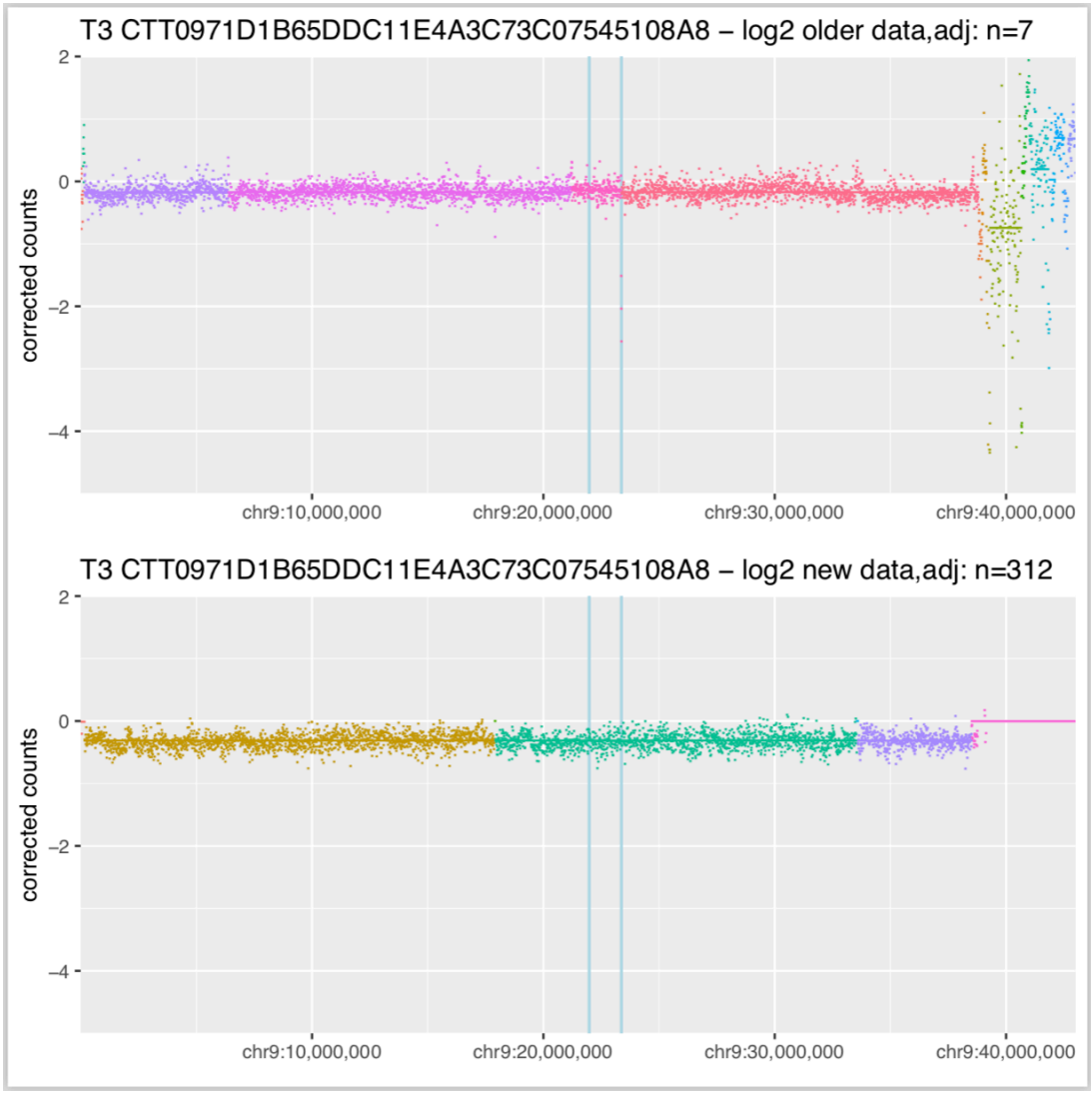


Figure 6.26. The ESV region for T3

Figure 6.27. The ESV region for T5

6.3.7 Comparison of NGS Data versus MLPA results

As done in Chapter 4, I replotted the sample previously used to visually demonstrate the comparison of copy number analysis between MLPA and NGS focusing on the *CDKN2A* region. Figure 6.28 below displays the plot of copy number profile of one sample displaying the copy number profile in the *CDKN2A* region with the light green region covering *MTAP*, light blue covering *CDKN2A*, and dark blue covering *CDKN2B*. The ends of the plot show normal copy number profile around zero and in the middle shows deletions of different regions. The 11 probes (blue dots) from MLPA analysis was plotted against the copy number segments with 95% confidence limits represented by red vertical lines. It can be observed that copy number results in all the MLPA probes matches those of the NGS data.

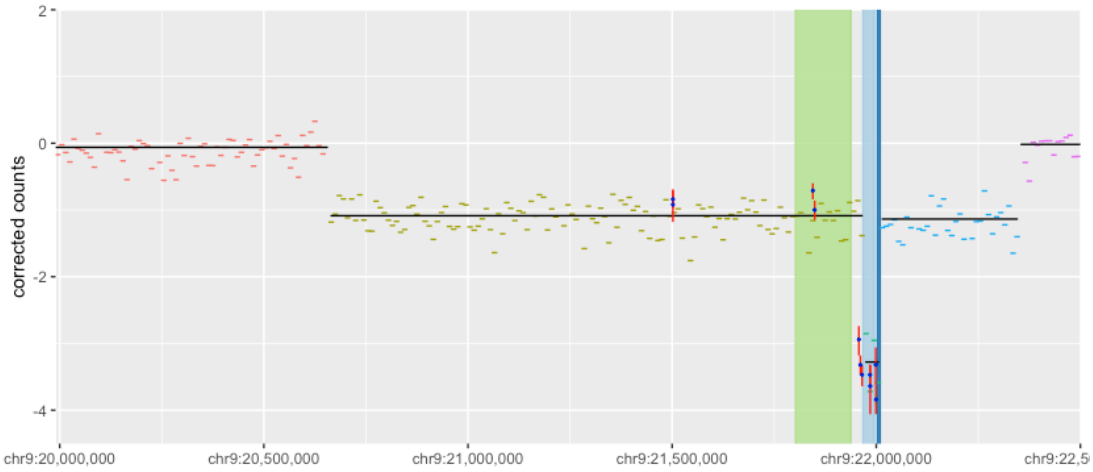


Figure 6.28. MLPA versus NGS CNA data

6.3.8 GISTIC Identified Significant Copy Number Peaks

Employing *GISTIC2.023* in the segmented LMC dataset identified regions of significant aberrations using the default cut-off of 0.25 residual q-value. For deletions, 44 autosomal regions were identified including with 9p21.3 ($8.3x10^{-107}$), 12q21.2 ($7.4x10^{-86}$), 7q34 ($1.5x10^{-36}$), 19q13.41 ($1.5x10^{-25}$), 15q11.2 ($1.7x10^{-25}$) as the top 5 most significant (Figure 6.29). The chromosome 9p21.3 or the *CDKN2A* region is known to be the most common form of deletion in melanoma and most cancers and consistently found to be most deleted in both LMC and TCGA. Aside from *CDKN2A*, other examples of genes that are commonly deleted in LMC are *CDKN2B*, *CDKN2B-AS1* (9p21.3), *SYT1*(12q21.2), *MGAM* (7q34), *SIGLEC5*, *SIGLEC14* (19q13.41), *PWRN3*(15q11.2), *UGT2B28*(4q13.2), *VPS53*(17p13.3), *ETS1* (11q24.3), *MIR3169*

(13q21.2), *MIR3668* (6q24.1), *ROPN1* (3q21.1), *HCG4B*, *HCG4B* (6p22.1), *GBP6* (1p22.2), *GOLGA8A* (15q14), and *PTEN* (10q23.31)

Figure 6.30 shows the significant peaks of amplification in the LMC. There were 36 autosomal regions identified as significantly amplified with chromosomes 3p11.1 (residual q-value=$1.6x10^{-115}$), 17q12($6.4x10^{-82}$), 15q31.1($1.3x10^{-40}$), 6p24.3($2.0x10^{-23}$), and 22q13.2 ($1.0x10^{-18}$) as the top 5 most significant.

There is no gene that exactly mapped to the most amplified region (3p11.1) identified but the closest gene to it is *EPHA3*. Other examples of genes identified to be highly amplified in LMC are *ARHGAP23* (17q12), *HERC2*, *GOLGA8G*, *GOLGA8F*, *MIR1268A*, *MIR4509*-1, *MIR4509*-2 (15q13.1), *TFAP2A*, *TFAP2A*-AS1 (6p24.3), *RRP7A*, *SERHL* (22q13.2), *ANKRD20A9P* (13q11), *ADSS*, *HNRNPU*, *CEP170*, *AKT3*, *ZBTB1*, *SDCCAG8*, *DESI2*, *EFCAB2*, *COX20*, *C1orf100*, *C1orf101*, *LOC339529*, *LINC01347*, *MIR4677*, *LOC101928068*, *SNORA100* (1q43), and *TPPP* (5p15.33). Comparison of the LMC and TCGA copy number data in terms of proportion of sample with amplification or deletion is shown later in this chapter.

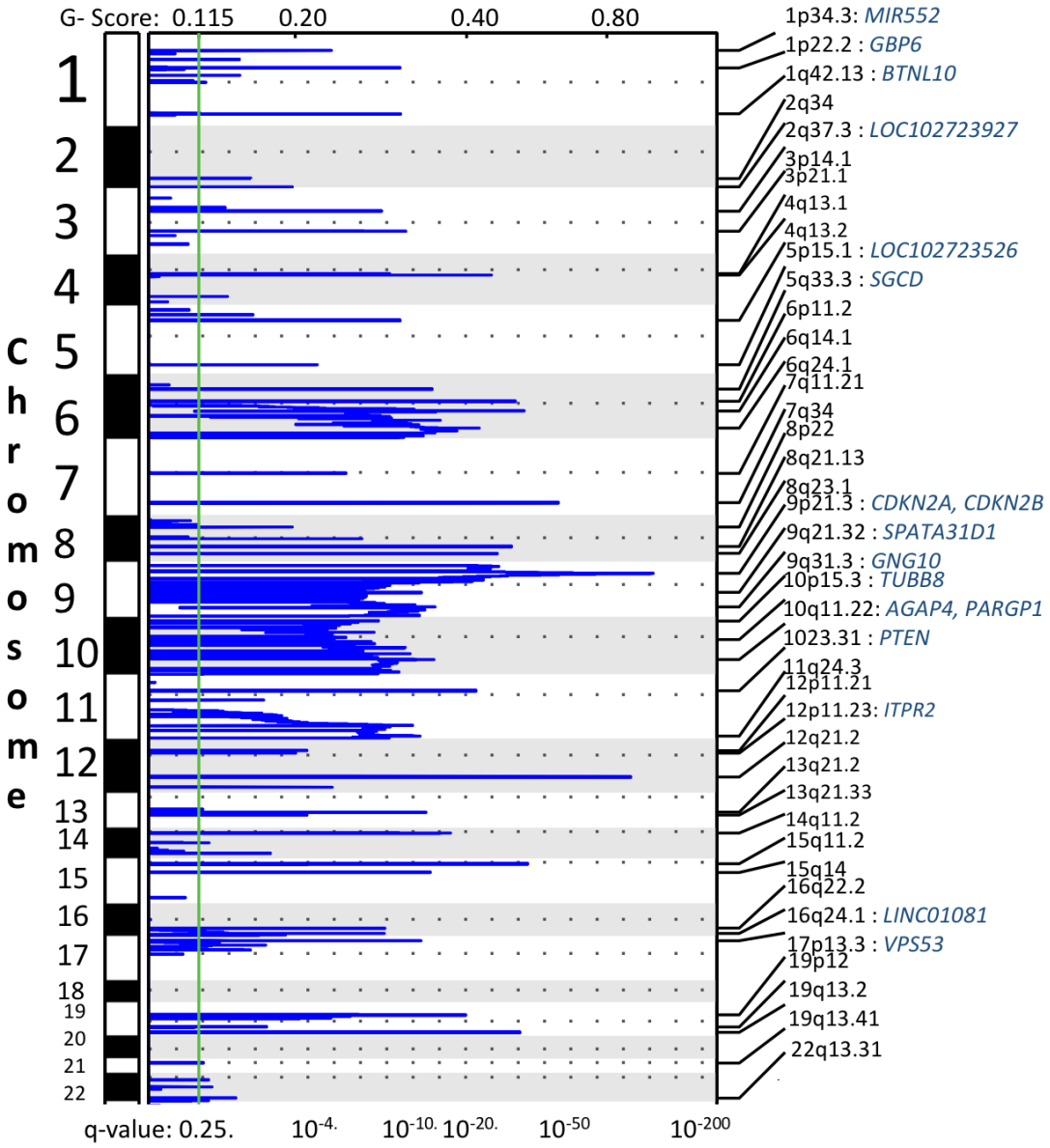


Figure 6.29. Visualisation of significantly deleted regions in LMC using GISTIC2.023.

Top X axis represents G-score where G = -log (Probability|Background) scores computed on markers or genes, p values computed by random permutation of markers or bins across genome. Lower X-axis represents the q-value where a green vertical is drawn to represent the default cutoff of 0.25.

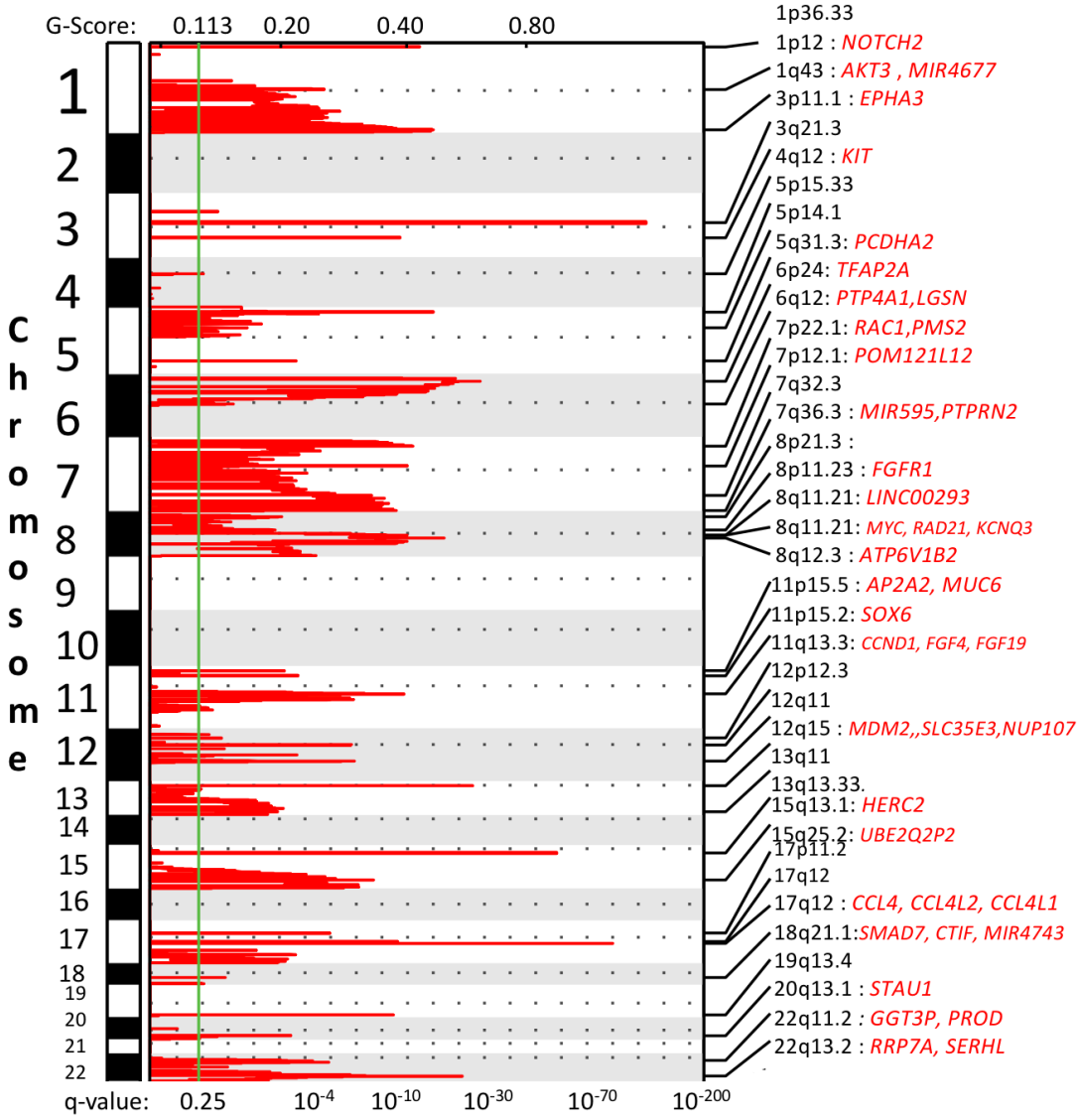


Figure 6.30. Visualisation of significantly amplified regions in LMC using GISTIC2.023.

Top X axis represents G-score where G = -log (Probability|Background) scores computed on markers or genes, p values computed by random permutation of markers or bins across genome. Lower X-axis represents the q-value where a green vertical is drawn to represent the default cutoff of 0.25.

6.3.9 Comparison with TCGA List of Genes with Deletion

The list of genes with significant rates of deletion in TCGA were obtained as in Chapter 4. The proportion of samples with deletion was plotted on the y axis while the gene labels and genomic band labels were located on the x axis as shown in Figure 6.31. For both datasets, *CDKN2A* gene had the most proportion of deletion

(73 % in LMC vs 76% in TCGA). Comparing the old and the new LMC CNA dataset, the new LMC dataset showed more similarity with the TCGA data in terms of distribution of samples with deletion in the given list of genes. An interesting observation is an obvious difference of proportion of samples with deletion in some genes in the list such as in the region of 17p13.3 (*ABR, DBIL5P, FAM57A, GEMIN4, GLOD4, MIR3183, NXN, TIMM22, VPS53*, and *BHLHA9* at 18% vs 36% in TCGA) , also in 17p13.3 (*C17orf97*, *DOC2B*, and *RPH3AL* at 10% vs 36% in TCGA), 16q24.3 (*FAM157C* at 14% vs 36% in TCGA), and 19p13.3 (*SBNO2* at 4% vs 26% in TCGA). This difference could potentially suggest markers of disease progression. Further analysis on this observation is discussed in the Chapter 7. For reference, a whole genome comparison of rates of deletion among the genes common to both TCGA and LMC lists are found in Appendix D.2.

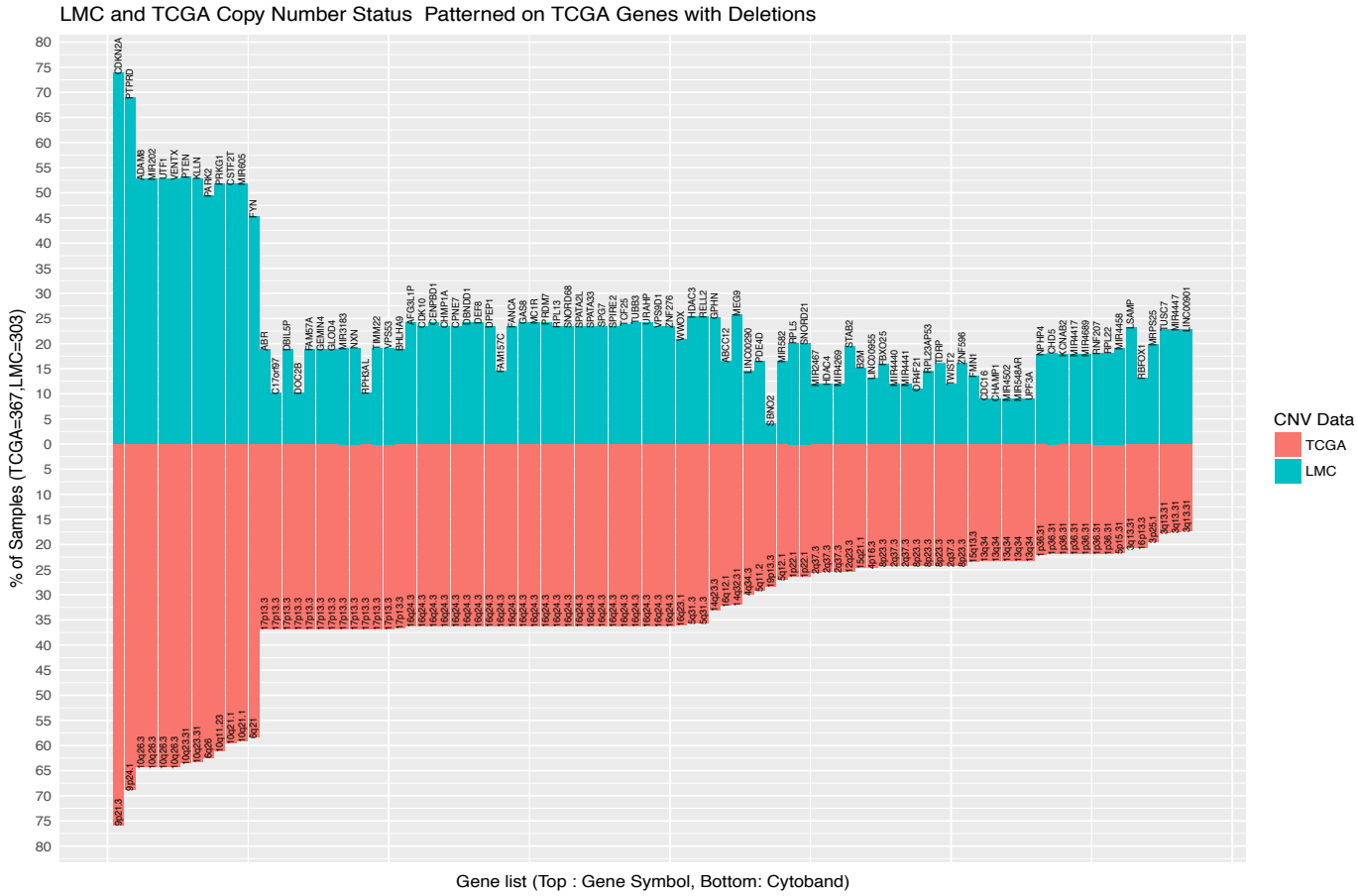


Figure 6.31. Proportion of Samples with Deletion in LMC and TCGA.

Proportion of samples in LMC and TCGA was calculated by using the GISTIC estimated copy number values and assigning values less than -0.1 as deletion.

6.3.10 Comparison with TCGA List of Genes with Amplification

Similarly, for amplification, the TCGA list of genes with amplifications was obtained and the same locations examined in LMC. The proportion of samples with amplification were estimated for both LMC and TCGA datasets and plotted in an upside-down bar plot. The proportion of samples with amplification was plotted on the y axis while the gene labels and genomic band labels were located on the x axis as shown in Figure 6.32. There is broadly similar distribution of proportion of samples with amplification between the new LMC and TCGA datasets. Differences in the proportion can be attributed to the fact the LMC dataset was derived from primary melanoma samples and TCGA datasets was derived from metastatic samples. A very similar proportion of amplification can be observed in the region 7p22.1 (*CYTH3* at 56% vs 58% in TCGA) and may indicate that the rate of amplification in this region does not change directly with disease progression. The region of 1p12 (*NOTCH2* at 20% vs 54% in TCGA) appeared to have twice the proportion of LMC in TCGA which may either be biologically relevant or due to the fact this region is close to the centromere which is blacklisted. Additionally, part of *NOTCH2* is included in the blacklisted region. For reference, a whole genome comparison of rates of amplification among the genes common to both TCGA and LMC lists are found in Appendix D.1.

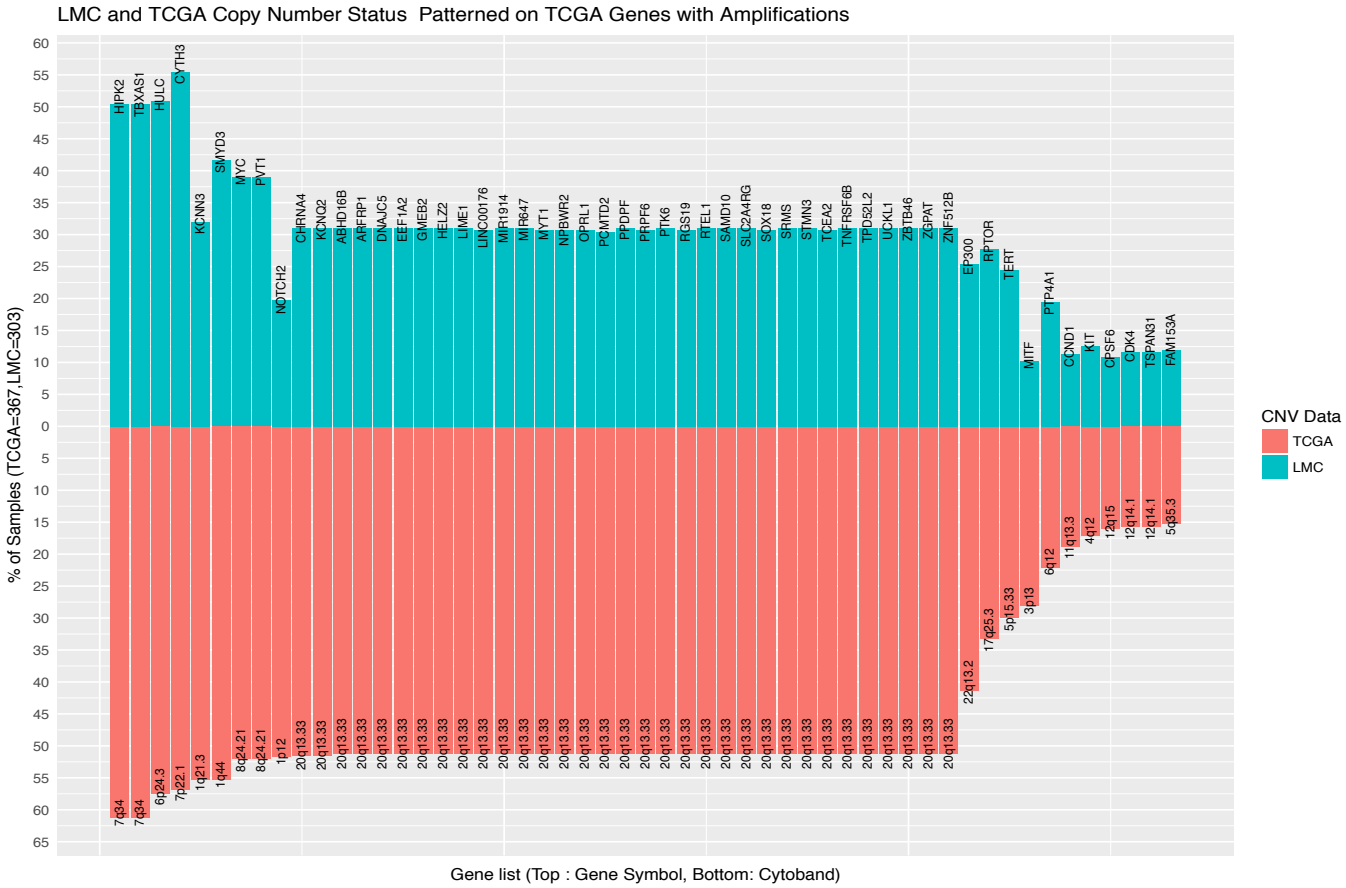


Figure 6.32. Proportion of Samples with Amplification in LMC and TCGA.

Proportion of samples in LMC and TCGA was calculated by using the GISTIC estimated copy number values and assigning values greater than 0.1 as amplification

6.4 Discussion

The success of the initial generation of the high-resolution copy number data from LMC samples was previously described by our team with analyses concentrating in the known melanoma genes especially the *CDKN2A* region [1, 168]. Extending the analysis of this data in the whole genome basis requires quality control assessment on the basis of the whole genome, particularly the autosomal genome.

This Chapter discussed the final assessment conducted on the new copy number data which had undergone further quality control measure as discussed in Chapter 5. To compare the old and the new data, several analyses in Chapter 4 were repeated starting with the analyses of 30 paired samples consisting of technical and biological replicates. For the biological replicates that consist of samples from different tumours, and samples from cores obtained from the same tumours, correlations are modest. This likely reflects the inherent heterogeneity of tumours [169] plus the limited sample size for this comparison. Technical replication was done by direct resequencing of the library and composed of processing the DNA in different methods of library preparation, or using different DNA concentration for library preparation. Correlation analysis using Pearson's r reveal a very high association (r=0.91) between the number of the genome fragments for the paired samples which are technically replicated. Copy number estimates from the LMC samples were generated in the whole genome basis as well as in terms of the *CDKN2A* region showing informative copy number profiles. These plots were used to validate computationally observed variable regions in the genome.

Another helpful way of identifying highly variable regions in the genome is by plotting the number of segments by segmented length. In this plot, we expect a longer chromosome to have more segments. Comparing the old and the new CNA data, both datasets revealed a linear relationship between number of segments per chromosome and the segmented length of the chromosome. The new CNA data exhibited better linear fit than the old data indicating improvement in the quality of the data ($P=1.9 \times 10^{-5}$ for new data, $P=5.2 \times 10^{-3}$ for old data). Similar trend with very slight decrease ($P=2.2 \times 10^{-5}$) in the significance of linear regression was observed when replicates were excluded in the analysis.

This improvement in the copy number data as measured on the segment level provides more confidence in proceeding with the next step of copy number analysis which is identification of significant copy number peaks in the genome which uses segmented copy number data as input to the R package *GISTIC2.023*. External comparison of the LMC copy number data with the published TCGA copy number data was repeated using the new data.

Broad similarity between the old LMC data and TCGA data was observed when checking for proportion of samples with deletion highlighting *CDKN2A* which is known to be of comparable proportion between metastatic and primary melanoma. The same trend with improved similarity on the pattern was observed when comparing the LMC data with the TCGA proportion of samples with deletions. On the other hand, a poor resemblance of the old LMC CNA proportion of samples with amplification was observed when comparing with the TCGA proportions. Using the new LMC data, a closer resemblance of the proportion of samples with amplification was observed when comparing with the TCGA data. This Chapter justifies the importance of the additional quality control steps performed in the LMC CNA data to improved quality on the whole genome level. On the focal sense, one of my colleague Dr. Joanna Pozniak on her PhD project on classifying primary melanoma patients based on immune groups using gene expression data found that increased expression of *MYC* and decreased expression of *NFKB* is associated with poor survival [170]. Partially explaining the variation in gene expression of the patients from different immune groups using copy number alteration was explored. Using the old CNA data, the association between copy number and gene expression of *MYC* was not captured whereas association was detected after performing additional quality control steps with the data. Similar association of 8q24 region copy number (where *MYC* is located) and *MYC* gene expression was reported in the study of Pouryazdanparast, Brenner (2012) which looked at the role of this association in amelanotic cutaneous melanoma [171].

This analysis also showed the high number of copy number changes on chromosome 7 in melanoma. At this time, the interpretation is unclear as the changes are not focal. Chromosome 7 has been shown to contain the highest content of segmental duplications in the human genome [172] indicating complex genomic sequence. This may indicate regions which are less stable and hence liable to copy number alterations. A number of interesting genes and diseases maps to this chromosome as summarized by Tsui (1988) and Scherer et al. (2003) [173] [151].

The overall conclusion of this chapter is that the QC steps have produced data which are consistent both internally and externally with the TCGA so in the next Chapter, I will focus on changes associated with the clinical characteristics of the tumours.

Chapter 7
Copy Number Alterations and Patient Clinical Characteristics Including Survival

This chapter discusses my work on estimating the overall measure of genome instability using current and proposed methods, followed with testing for association with clinical and tumour characteristics. Window level analysis and testing for association with clinical and tumour characteristics was also performed. This included estimation of study specific genome wide significance to address the concern on correlated neighbouring windows in the genome.

7.1 Introduction

Genomic instability is a hallmark of cancer that leads to an increase in genetic alterations that facilitates the acquisition of additional mechanisms required for the development and progression of a tumour [174]. After establishing the quality of the CNA data, I proceeded with the next steps of my project which generally aims to test for association of copy number alteration with patient and tumour characteristics and survival. This chapter discusses the methodologies performed to check and identify the magnitude and significance of association of copy number alterations in melanoma with the aforementioned attributes. An initial approach to checking for this is by identifying to what extent an overall measure of aberration varies with patient clinical and tumour characteristics and survival. I estimated several parameters calculated from the different measures of the genome profile and tested for association, initially with survival. I chose the parameter that varies most with survival and proceeded testing its association to the different patient clinical characteristics. A focal analysis using 10kb window CNA data was done in recognition that a focal aberration driving progression would be of major interest to melanoma biologists.

For precision, in this and the other chapters, I use the term survival to indicate “melanoma specific survival (MSS)” censored at 12 years past diagnosis chosen because the majority of participants are followed up at least 12 years from recruitment. The desire to impose a censoring date reflects that as described before the decision to conclude a death to be a “melanoma specific death” was made on careful assessment of clinical records but it is clear it is more likely to be accurate closer to the time of recruitment given that patients are more likely to be under regular medical surveillance earlier in the disease course. The patient clinical and tumour data were part of the

Leeds Melanoma Cohort as previously described in Section 3.1 [105, 106]. The information on deaths were obtained from the National Health Service (NHS) GP records linked to Office for National Statistics (ONS).

7.2 Methodology

7.2.1 Measuring the instability in the genome of cutaneous melanoma samples

Overall copy number aberration measures the amount of damage in the genome. Various measures are calculated to identify the parameter that best associates with patient clinical characteristics - initially examining survival. These are described in the subsections below. Other measures of genomic aberration such as the z-score based measures were examined in the study of Heitzer et al. (2013) but where not tested in this study as they were focused on plasma sample derived copy number estimates[175]. Calculations were done specifically for the 277 cutaneous melanoma samples after excluding the 26 melanoma samples which were acral tumours, subungual, genital and mucosal tumours.

7.2.1.1 Aneuploidy Score

A measure called aneuploidy score was adapted from the work of Taylor et al. (2018) which calculated the number of autosomal chromosome arms that have aberrations in the genome excluding the short arms of 13, 14, 15, 21, and 22; these regions were previously mentioned in Chapter 5 as part of the blacklisted regions [176]. This score ranges from 0 to 39 with zero indicating a genome without an arm level aberration and 39 indicating at least one aberration on all chromosome arms. Arm level copy number data were directly obtained from the *GISTIC2.023* output. A detailed guide on identifying output from the *GISTIC* results folder can be found at ftp://ftp.broadinstitute.org/pub/GISTIC2.0/GISTICDocumentation_standalone.htm as previously mentioned in Chapter 6.

7.2.1.2 Fraction of Genome Altered

The study of Taylor et al. (2018) described the calculation of the Fraction of Genome Altered (FGA) as a measure related to aneuploidy score. This was obtained by taking the product of each arm level copy number aberration to its length then divided by the length of the genome [176]. In this study, we followed the FGA calculation similar to Domcke et al. (2013) in evaluating cell lines as tumour models by comparison of

genomic profiles [177] and of Luebker (2017) which compared genomes of cutaneous melanoma tumours to commercially available cell lines [178]. The formula for FGA is:

Equation 7 $$FGA = \sum_{i=1}^{n} I[|CNi| > T]Li \,/\, \sum_{i=1}^{n} Li$$

where *i* is a counter over segments, *Li* denotes the length of each segment *i* and *CNi* denotes the segment copy number mean for each segment *i* that exceeds a chosen cutoff. The value of *T* is usually estimated based on the variation of the data. In this study , I adapted *T*= 0.1 from the study of the TCGA network [71], for *n* total segments [177].I also tried several values for T such as 0.1, 0.25, and 0.30 and observed that 0.1 best captures the aberration in the data. In summary, FGA is the sum of all segment lengths of segments with absolute segment mean greater than 0.1 divided by the sum of all segments lengths in the autosomal genome.

7.2.1.3 Mean Weighted Segment Mean

This is a genomic instability measure that I devised based on previous studies that utilized and segment lengths but I incorporated the segment means. The study of Knijnenburg et al. (2018) used total number of segments which was obtained by counting all the segments present in a sample [179]. While FGA in the study of Taylor et al. (2018) used purely the segment lengths in its calculation, I added segment mean as a weighting factor bearing in mind that both length and the magnitude of aberration contributes in the genomic instability. Figure 7.1 below shows how each given region in the genome is given a height and length by drawing squares dictated by the length and the magnitude of the aberration.

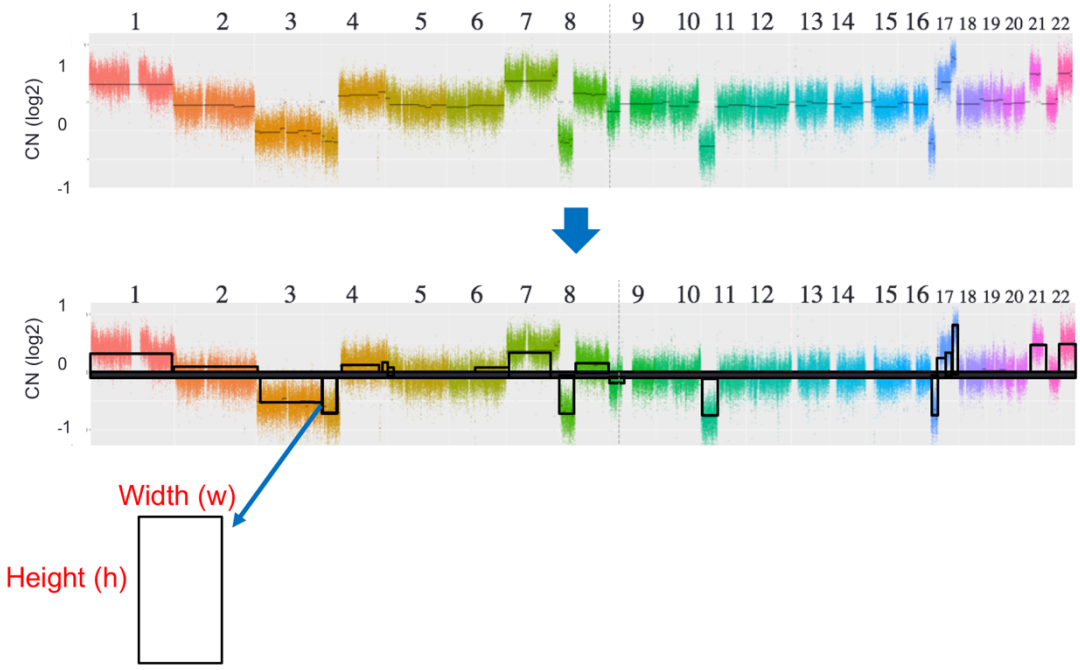


Figure 7.1. Calculating the mean weighted segment mean (MWSM)

The formula below is used to calculate mean weighted segment mean (MWSM):

$$MWSM = \sum_{i}^{n} LiWi \,/\, 22$$

Where *Li* is the length (alternatively labelled as *w* width in Figure 7.1) of a segment with absolute segment mean (alternatively labelled as *h* height in Figure 7.1) > 0.1 for segment *i*. *Wi* is the segment mean segment *i* with an absolute value greater than 0.1. The sum product of *Li* and *Wi* from *i* to *n* was taken and then divided by 22 to produce an autosomal chromosome average.

7.2.2 The Patient Clinical Characteristics

A description of patient clinical characteristics was provided in Chapter 3 (Methodology). In this chapter, the following measures are used and tested for association with the selected measure of overall genomic aberration: age at diagnosis (years), sex (male or female), stage (I, II, III), site of melanoma (Limb, Head and Neck, Trunk), mutation type (*NRAS* mutated, *BRAF* mutated, double wild type for *BRAF* and *NRAS*), ulceration status (ulcerated vs. non-ulcerated), mitotic rate (<1 mm^2, 1-2 mm^2, and >2 mm^2), tumour infiltrating lymphocytes or TILs (absent, brisk, non-brisk), Breslow thickness (<=2mm, 2-4 mm, >4 mm), and percentage of stroma (%) . Survival time for melanoma specific was censored at 12 years. Any survival time greater that 12 years was reset at 12 years and had the survival status set to “alive” if the survival status after 12 years was “died”. These data were part of the Leeds Melanoma Cohort as

previously described in Section 3.1 [105, 106]. The death data was obtained from the National Health Service (NHS) GP records linked to Office for National Statistics (ONS). Since these characteristics are commonly investigated in melanoma, their associations with copy number were tested[58, 67, 183-185].

7.2.3 Plot of Copy Number Profile by Patient Characteristics

To visually inspect whether there is an inherent pattern of the copy number aberration across the samples when categorised according to different patient characteristics, whole genome copy number plots were generated. Copy number information where presented in 10kb window level. There was a total of 248, 736 windows left for the whole genome analysis after blacklisting.

7.2.4 Testing for Association with Patient Characteristics

Testing for association of overall CNA load and 10k windows with clinical features was conducted using Spearman rank correlation for pairs of variables which are both continuous, Wilcoxon rank sum test for comparing two groups in terms of quantitative copy number, and Kruskal-Wallis test for comparing three or more groups in terms of quantitative copy number. Manhattan plot was used to visualise the significance of each region in the genome. This plots the $-\log_{10}$ of the P-value of the test on each10k window on the y-axis and the 10k window location on the x-axis.

7.2.5 Permutation Analysis on Clinical Characteristics

One challenge in statistical testing of copy number data is that neighbouring windows tend to be highly correlated. To address this, I carried out a permutation analyses on the selected clinical information (site and Breslow thickness) and survival (using quantitative copy number data). For each characteristic tested, the characteristic is randomly assigned to each tumour genome. Then, a 10k window level analysis was done obtaining the significance of the test employed. Each simulation yielded test results for 262, 827 windows. This underwent filtering to ensure excluding unaccounted blacklisted regions yielding the 248,736 windows. For each genome divided into 10k windows, the windows were ranked by calculated P-value. The 95^{th} percentile of this list was obtained and identified as basis as the critical study-specific genome-wide significance threshold. Each iteration analysing the whole autosomal genome lasted from 5 to 9 hours. The number of iterations submitted is at least 500 for the site of tumour and Breslow thickness, and 1,500 for survival analysis using the quantitative copy number profile. However, for logistical reasons, not all permutations produced

output (e.g. elapse time on high performance computer exceeds allowance or memory restrictions inhibit completion); 297 simulations were completed for the analysis on the site of the tumour, 138 for Breslow thickness, and 930 for survival analysis. Due to lack of time, permutation analysis on other tumour and clinical factors were not conducted. Identification of interesting regions for these factors were done by ranking the results by test significance.

7.2.6 Level of stroma and copy number profile

The data on proportion of stroma was based on the work of Dr. Sally O'Shea. Sally scored stroma using RandomSpot© to derive a measure called percentage of stroma (POS) [180] by examining whether randomly chosen "spots" on the H & E slide image were "stroma" or "non-stroma". The stromal levels are defined using tertiles created by the function *quant_groups* under the R package *dvmisc* [181]. The tertiles based on the three intervals that cover the proportion of stroma are presented in Table 7.1 below:

Table 7.1. Categorisation of percentage of stroma (POS)

Level of stroma	POS Interval	n=206
Low POS	[1.9%, 23.3%]	69
Medium POS	(23.3%, 42.5%]	68
High POS	(42.5%, 93.3%]	69

The derived data on level of stroma were then merged with the whole genome copy number data. A total of 206 tumours have matching data on level of stroma as categorised above. Whole genome profiles are then plotted by each level of stroma to produce three plots for comparison.

7.2.7 Survival analysis

The majority of the analysis started with Cox proportional hazard models predicting MSS using one window at a time adjusting for age, sex, and stage as these are the commonly known confounders of melanoma survival [182]. A Kaplan-Meier curve was used to plot and compare survival curves among sample groups. Log-rank test was used to test significant difference between or among survival curves of two or more groups of samples.

7.2.7.1 The Kaplan-Meier Curve

The Kaplan – Meier curve was devised by Edward L. Kaplan and Paul Meier to address the problem of dealing with survival data with incomplete observations [183-

185]. It is a nonparametric method used to estimate the survival function from lifetime data. The survival function $S(t)$ is estimated by the formula below:

Equation 8 $\widehat{S}(t) = \prod_{i:t_i \leq t}(1 - \frac{d_i}{n_i})$

Where $S(t)$ the probability that life is longer than time *t* (years)

t_i is a time when at least one event (e.g. death) happened

d_i is the number of events (deaths) that happened at time t_i

n_i is the number of individuals who survived (did not have the event or die or have not been censored) up to time t_i

This method is primary useful for visually comparing survival curves between or among groups of subjects in a study. It requires three pieces of information to perform the analysis namely the survival time, the status at the end of the survival period and the study group of origin. Analysis using Kaplan-Meier estimator has three assumptions. Firstly, it is assumed that the survival prospects between individuals who are censored and those who continue to be followed up are the same. Secondly, it is assumed that survival probabilities of individuals are the same regardless if they are recruited earlier or later in the study. Thirdly, it is assumed that the event (e.g. death) happened at the time indicated. In some cases a maximum discrepancy in recording the actual time of event is defined (e.g. maximum of 1 day)[186]

7.2.7.2 Log-rank test

Log-rank or logrank test is used to test the hypothesis that at least one of the survival curves of one or more samples (e.g. a group with a specific copy number amplification or deletion) are significantly different from that of a reference group (e.g. normal copy number). It is a nonparametric test that is appropriate to use when the data are skewed to the right and censored. This is also known as the Mantel-Cox test named after Nathan Mantel and David Cox. It was initially proposed by Nathan Mantel and was termed the logrank test by Richard Peto and Julian Peto [187-189]. Log-rank test statistic compares estimates of the hazard functions of the two groups being compared (one group versus the reference at a time) at each observed event time. This is constructed by calculating the observed and expected number of events in one of the groups at each event time and taking the sum of these resulting to a single value (χ^2) that summarises all time points where an event occurred. This is showed in the formula below:

Equation 9 $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

Where *k* is the number of groups compared

O_i is the number of death/s in the *kth* group when at least one death was observed

E_i is the expected number of deaths in the *kth* group when at least one death was observed

For each year when death occurred, E_i is calculated by the formula below:

$$E_i = R_i * D_i/N_i$$

Where R_i is the number of samples at risk at year *i*

D_i is the number of deaths observed at year *i* for a given group

R_i is the total number of samples at year *i*

For the case where the survival time is censored, that individual is considered to be at risk of dying in the year of the censoring but not in subsequent years. This way of handling censored observations is similar for calculating the Kaplan-Meier survival curve[190]. The significance of this test is approximated from a χ^2 distribution with *k-1* degrees of freedom. Note that this test only identifies the significant difference of the survival curves between groups and does not provide a measure of effect size. In this case, providing the hazard ratios calculated using the Cox proportional hazard model is useful.

7.2.7.3 Survival analysis using Cox Proportional Hazard Model

Survival analysis was performed on one window at a time using Cox proportional hazard model[183] . This model addresses the weakness of the logrank test which does not provide a measure of the effect size, by providing hazard ratio estimates. The key assumption for this model is the proportional hazards function assumption which means that the model assumes that each covariate has a multiplicative effect in the hazard function that is constant over time [188, 196]. In this step, the model takes the general form:

Equation 10 $h(t) = h_0(t) * e^{b_1x_1+b_2x_2+\ldots+b_px_p}$

where *t* represents the survival time

$h(t)$ is the hazard function determined by a set of covariates (x_1, x_2...x_p)

$b_1, b_2,\ldots b_p$ represents the effect size of the covariates in the model

h_0 is the baseline hazard (the value of the hazard all x's is zero)

The hazard ratio (HR) for a specific covariate is calculated by exponentiating the corresponding effect size and is generally interpreted as a good prognostic factor when HR is less than 1, no effect when HR =1 and a bad prognostic factor when HR is greater than 1 (because here MSS (event) = 1, means the patient died or did not survive).

7.2.8 Mapping the Window to the Gene

To understand better the results of tests for associations that identified genomic windows that are associated with tumour and clinical characteristics including survival, each window was assigned a numerical index corresponding to the numerical value attached at the end of its window label. As an example, the index for the window 10k000000079 is 79. This index was applied similarly to the list of all 10k windows mapping to the genome derived using both the *biomaRt [140]* and *mcnv [135]* packages. *biomaRt* was used to identify the genes in the human genome excluding those from the sex chromosomes[140]. This uses different search characteristics to find information from the genome. For the purpose of identifying genes from the somatic genome, search based on a list of chromosomes (1-22) were used. This then resulted to a list of chromosomes together with the genes and their genomic locations obtained from *Ensembl* [140, 191]. Each gene maps to at least one or more 10k windows depending on its size. In case a window of interest did not map to any gene, the nearest gene name before or after that window is taken as a proxy. If two genes are equidistant to the window of interest, both are taken as proxies as long as they are located in the same cytoband as that of the window of interest.

7.3 Results

7.3.1 Distribution of melanoma tumours by site and location

The distribution of the primary melanoma tumours by site and location is presented in Table 7.2 below. It can be observed that tumours are most commonly found on the limbs region such as those in the lower leg (n=48), upper arm (n=39), and thigh (n=26). Overall, a total of 139 out of 303 samples were found in the limbs comprising about half of the samples (46%). This is followed by the tumours in the trunk region such as those in the back (n=59), chest (n=17), and abdomen (n=13). A total of 92 samples

were located on the trunk comprising 30% of the samples. There was a total of 46 samples located on the Head comprising 15% of the samples. The least common location of samples is found in mostly hidden parts of the body as shown below with a total of 26 samples or 9% of all the samples in this study.

These 26 samples were excluded in the further analysis being classified as mucosal or non-cutaneous melanoma leaving a total of 277 cutaneous melanoma samples.

Table 7.2. Tumours by site and location

All the 26 samples located in mostly hidden parts of the body (classified as 'Other' in the table below) were excluded in all analyses.

	Site			
Location	**Head**	**Limbs**	**Trunk**	**Other (Excluded)**
abdomen	0	0	13	0
acral	0	0	0	7
anal	0	0	0	1
back	0	0	59	0
buttock	0	0	3	0
chest	0	0	17	0
elbow	0	1	0	0
ENT	0	0	0	3
foot	0	8	0	0
hand	0	1	0	0
head/neck	46	0	0	0
knee	0	3	0	0
lower arm	0	16	0	0
lower leg	0	45	0	0
penis	0	0	0	1
subungual	0	0	0	5
thigh	0	26	0	0
upper arm	0	39	0	0
vaginal	0	0	0	1
vulval	0	0	0	8
Total	**46**	**139**	**92**	**26**
%	**15%**	**46%**	**30%**	**9%**

7.3.2 Measures of Genomic Instability

While defining whether a genomic instability observed in all the sample is specific to or combination of changes in the nucleic acid sequences, chromosomal rearrangement, or aneuploidy was not part of this section, I have identified samples that exhibits known characteristic of chromothripsis. Figure 7.2 and Figure 7.3 show patterns of copy number alternation in at least two states - a characteristic of chromothripsis mentioned by Maher CA, Wilson RK (2012), Forment JV, Kaidi A, Jackson SP (2012), and Korbel JO, Campbell PJ (2013) [74, 75, 192]. The following sections discusses the three measures of genomic instability calculated for all the samples in this study.

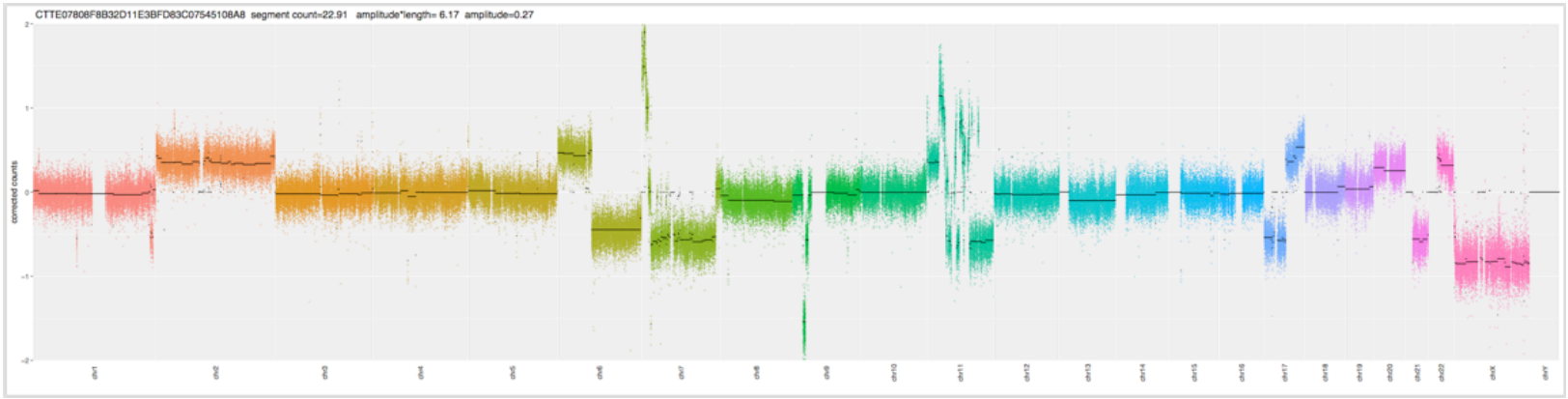


Figure 7.2. Sample 1 exhibiting chromothripsis

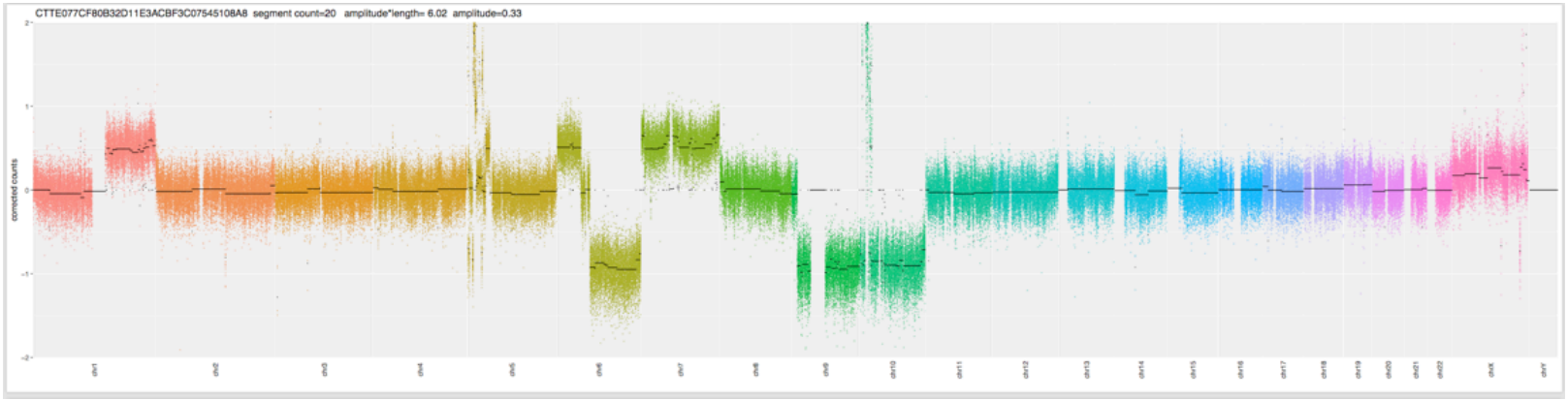


Figure 7.3. Sample 2 exhibiting chromothripsis

7.3.2.1 Aneuploidy Score

The aneuploidy score was calculated for all the patients showing a minimum value of 0 and maximum value of 35. The mean aneuploidy score is 10.3 arms with a standard deviation of 8.0 (Figure 6.9). This is similar to the study of Taylor et al. (2018) where the mean aneuploidy score of 88% of cancer samples examined had a mean of 10.0 [176].

To understand more the distribution of the aneuploidy score, a table containing the frequency and rate of deletion, amplification and the combination of both in each chromosome arm (overall arm level aberration) was calculated and is shown in Table 7.3. The most aberrated chromosome arm is 9p (52%) which is expected as it is the most common site of deletion in melanoma and, indeed, most cancers [193]. A mapping of different disease associated to chromosome 9 was done by Gilbert and Kauff (2001) [194]. This is followed by aberration of 7p (42%) and 7q (41%) which suggests a whole chromosome amplification of the chromosome 7. This chromosome was also previously observed to be the most variable chromosome when checking for the number of segments in the genome. Similar aberrations in chromosome 7 were reported by the studies of Tsui (1988), Stagni et al. (2018), Sherer et al. (2004), and Hellman et al. (2002) [151, 172, 195, 196]. As previously mentioned in Chapter 6, mapping of diseases associated with chromosome 7 was summarised by summarized by Tsui (1988) and Scherer et al. (2003) [173] [151]. The fourth and fifth most aberrated regions are chromosomes 10q (39%) and 10p (38%), also suggesting whole chromosome aberration. These chromosomes have already been associated with melanoma in the literature [197, 198].

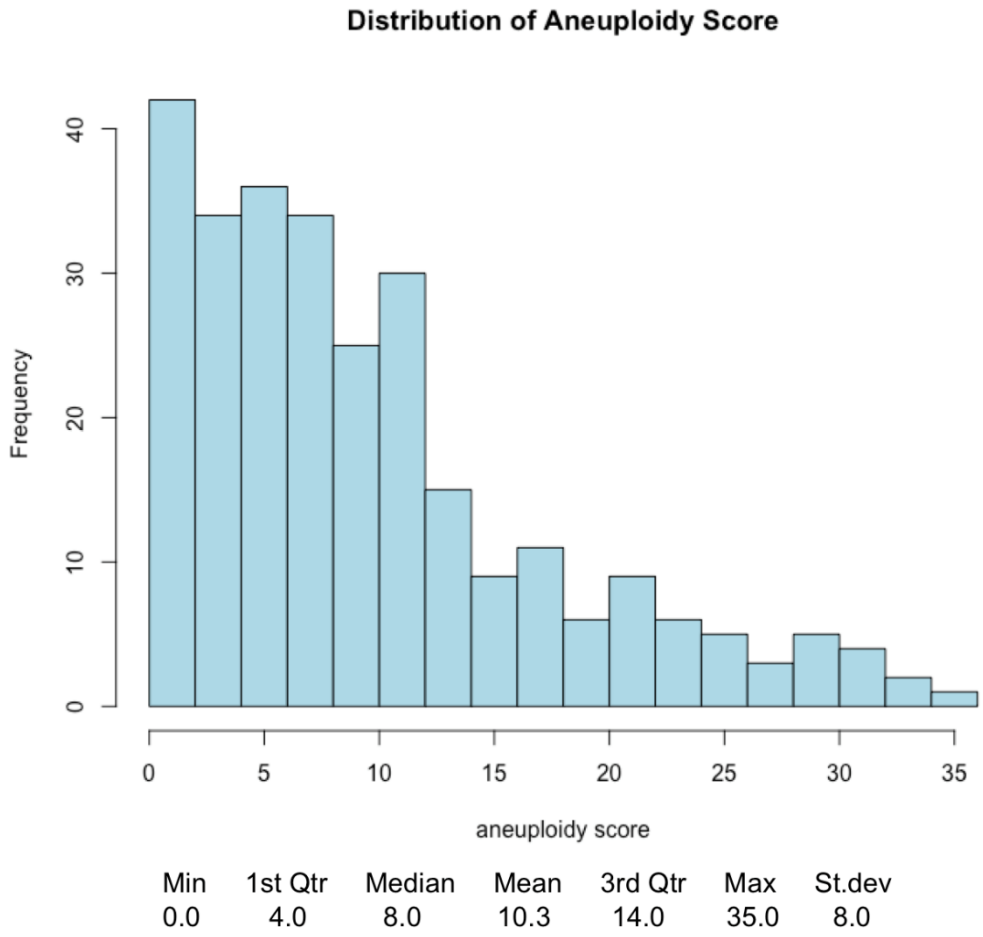


Figure 7.4. Distribution of aneuploidy score (n=277)

Table 7.3. Frequency of aberration by type of change

(n=277). The short arms of chromosomes 13,14,15, 21 and 22 are blacklisted.

Arm	Deletion	Normal	Amplification	% Deletion	% Amplification	% Aberration
1p	27	199	51	10%	18%	28%
1q	10	192	75	4%	27%	31%
2p	26	226	25	9%	9%	18%
2q	22	229	26	8%	9%	17%
3p	38	223	16	14%	6%	19%
3q	39	222	16	14%	6%	20%
4p	22	233	22	8%	8%	16%
4q	22	231	24	8%	9%	17%
5p	35	206	36	13%	13%	26%
5q	47	206	24	17%	9%	26%
6p	17	186	74	6%	27%	33%
6q	65	184	28	23%	10%	34%
7p	4	164	109	1%	39%	41%
7q	4	160	113	1%	41%	42%
8p	15	203	59	5%	21%	27%
8q	8	185	84	3%	30%	33%
9p	139	133	5	50%	2%	52%
9q	95	175	7	34%	3%	37%
10p	99	173	5	36%	2%	38%
10q	104	169	4	38%	1%	39%
11p	49	219	9	18%	3%	21%
11q	49	221	7	18%	3%	20%
12p	28	226	23	10%	8%	18%
12q	36	228	13	13%	5%	18%
13q	25	214	38	9%	14%	23%
14q	44	221	12	16%	4%	20%
15q	16	222	39	6%	14%	20%
16p	21	233	23	8%	8%	16%
16q	26	231	20	9%	7%	17%
17p	31	221	25	11%	9%	20%
17q	18	217	42	6%	15%	22%
18p	23	226	28	8%	10%	18%
18q	26	223	28	9%	10%	19%
19p	9	190	78	3%	28%	31%
19q	10	194	73	4%	26%	30%
20p	9	202	66	3%	24%	27%
20q	4	199	74	1%	27%	28%
21q	17	231	29	6%	10%	17%
22q	20	215	42	7%	15%	22%

7.3.2.2 Fraction of Genome Altered (%)

The distribution of fraction of genome altered (FGA) is shown in Figure 7.5 below. The minimum FGA is almost zero at 0.01% and the maximum is 97.35%. The mean FGA is 28.29 % with a standard deviation of 20.32%. These figures are close to the study of Luebker et al. (2017) where the mean FGA for primary melanoma samples (n=101) is 33% with an standard deviation of 19% [178].

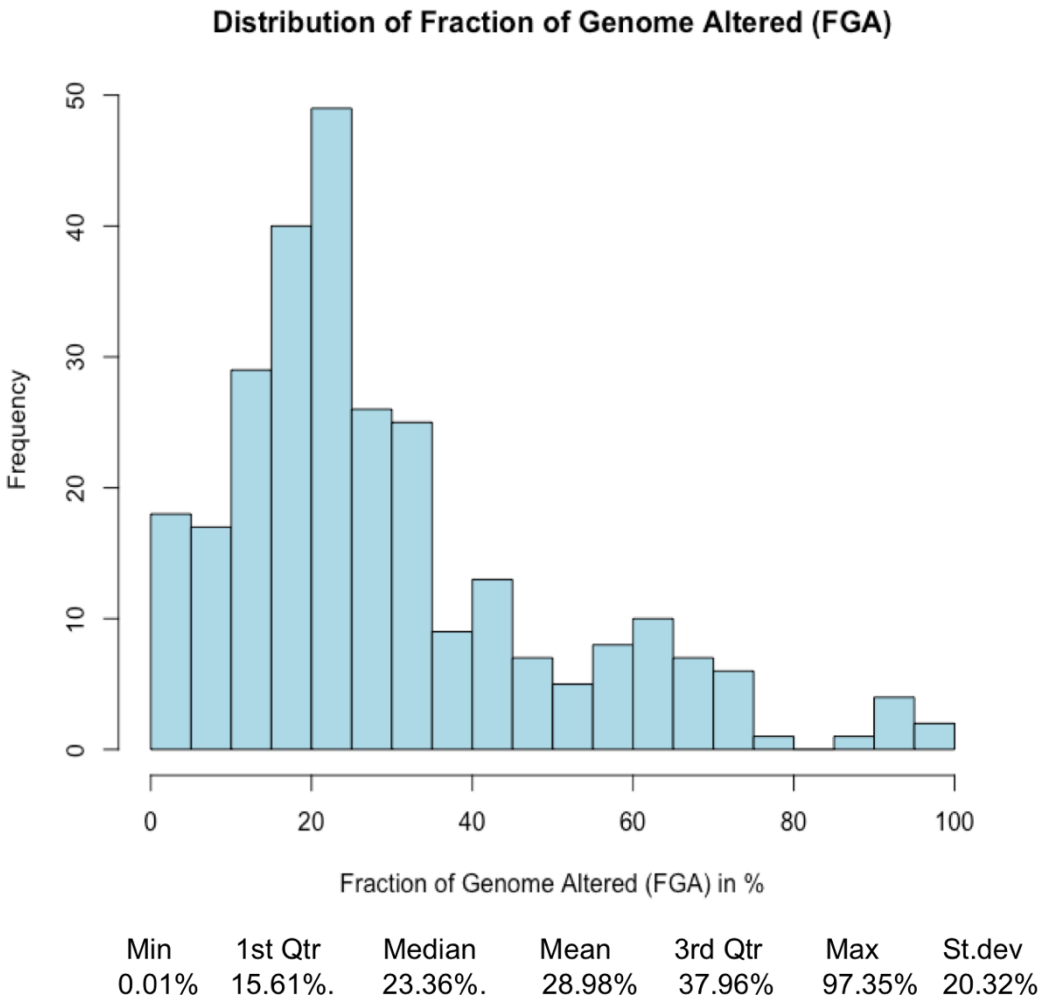


Figure 7.5. Distribution of Fraction of Genome Altered

7.3.2.3 Mean Weighted Segment Means (MWSM)

The distribution of mean weighted segment means (MWSM) is presented in Figure 7.6 below. The mean weighted segment mean for the samples has a lowest value of 1.12 and a highest value of 9.47. This has a mean of 3.91 and a standard deviation of 1.47. Of the three measures of genomic instability considered in this study, MWSM appears to have a distribution that is closest to the statistical normal distribution.

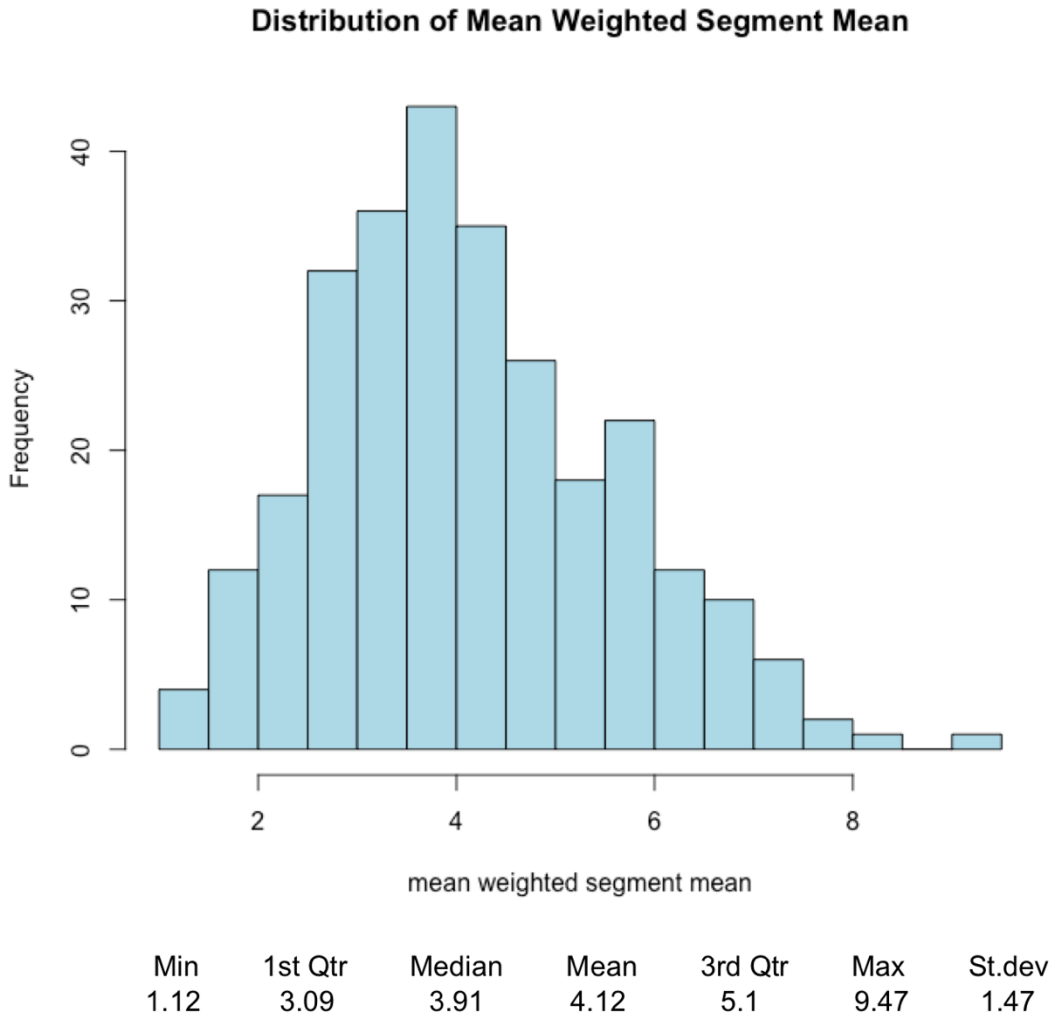


Figure 7.6. Mean Weighted Segment Mean

7.3.3 Association of Three Measures of Genomic Instability with Clinical Characteristics

A summary of test for association of different patient clinical characteristics with three different measures of genomic instability such as aneuploidy score, fraction of genome altered (FGA), and mean weighted segment mean (MWSM) is presented in Table 7.4 below. The p-values in these results were not adjusted for multiplicity. Overall, the measures showed the same patterns with each other, with some modest differences comparisons e.g. by sex, although formal significance differences varied by measure. For each measure, expected patterns were seen with increasing tumour severity e.g. for Breslow thickness.

Almost all clinical characteristics except mutation status ($P=3.8 \times 10^{-3}$) showed no significant association with aneuploidy score which was the least informative of the measures. Figure 7.7A shows a boxplot of aneuploidy score by mutation status. *BRAF*

mutant tumours tend to have higher aneuploidy score than both *NRAS* mutant tumours and double wild type (non-mutant for both *BRAF* and *NRAS*) tumours. There was not a significant difference in the aneuploidy score between *NRAS* mutant tumours and Double wild type tumours.

Fraction of genome altered (FGA) showed significant associations with Breslow thickness ($P= 2.2 \times 10^{-4}$), AJCC stage ($P=4.3 \times 10^{-3}$), ulceration ($P= 3.8 \times 10^{-3}$), mitotic rate (P=0.03), mutation status ($P=7.5 \times 10^{-3}$), and percentage of stroma (P=0.02). Ulcerated tumours tend to have higher fraction of genome altered (FGA %) as compared to those which are non-ulcerated (Figure 7.7B). Fraction of genome of altered (FGA %) tend to increase with the increase in mitotic rate (Figure 7.7C).

Among the three measures of genomic instability used in this study, only mean weighted segment mean (MWSM) showed statistical evidence of significant correlation with patient age at diagnosis (P=0.03). It is also the measure that is most associated with Breslow thickness ($P=2.8 \times 10^{-5}$), AJCC stage ($P=1.3 \times 10^{-3}$), and percentage of stroma (POS) ($P=4.4 \times 10^{-4}$). Similarly, to FGA, MWSM is also associated with ulceration ($P=4.7 \times 10^{-3}$), and mitotic rate (P=0.05) but of lesser significance. Figure 7.7D shows that mean weighted segment mean (MWSM) increases with the increase in the Breslow thickness. This also increase with the melanoma progression (Figure 7.7E). For tumour purity, a purer tumour is associated with lower mean weighted segment mean (MWSM) as shown in Figure 7.7F

Table 7.4.Testing for association of clinical characteristics with three measures of genomic instability

Clinical characteristics		Measure of Overall Genome Instability					
		Aneuploidy Score	P	Fraction of Genome Altered (FGA %)	P	Mean Weighted Segment Mean (MWSM)	P
Sex m(r)	Male (n=132)	7 (0,31)	0.52	23.8 (0.02, 91.5)	0.58	4.08 (1.31, 9.47)	0.26
	Female (n=145)	8 (0,35)		22.1 (0.01,97.4)		3.77 (1.12, 8.01)	
Site n=277 m(r)	Head (n=46)	7.5 (0,30)	0.13	23.0 (0.02,91.5)	0.23	3.87 (1.1,1.5)	0.32
	Limbs (n=139)	9.0 (0,35)		25.4 (0.01,97.4)		4.12 (1.4,8.0)	
	Trunk (n=92)	7.0 (0,33)		21.4 (0.02, 95.2)		3.8 (1.3,7.7)	
Age at diagnosis, rho (median=57.09 years)		0.07	0.26	0.11	0.07	0.13	0.03
Breslow thickness , rho (median=2.4 mm)		0.11	0.08	0.22	2.2E-04	0.25	2.8E-05
Breslow thickness m(r)	<= 2mm (n=109)	7 (0,34)	0.25	21.1 (0.01,97.4)	0.01	3.74 (1.1,8)	2.3E-03
	2 - 4 mm (n=100)	8 (1,33)		23.7 (3.8,95.2)		3.9 (1.5,7.7)	
	> 4 mm (n=65)	10 (0,35)		27.4 (0.02,92.8)		4.45 (1.3,9.5)	
AJCC Stage n=275 m(r)	I (n=88)	7 (0,34)	0.12	20.4 (0.13, 97.4)	4.3E-03	3.6 (1.1,8)	1.3E-03
	II (n=147)	8 (0,35)		23.9 (0.17, 95.2)		4.1 (1.3,9.5)	
	III (n=40)	10.5 (0,31)		32.6 (0.64,69.64)		4.4 (1.8,7.7)	
Ulceration, m(r)	Ulcerated (n=92)	9 (0,35)	0.39	27.5 (0.02,92.8)	3.8E-03	4.2 (1.3,9.5)	4.7E-03
	Non ulcerated (n=185)	8 (0,34)		21.8 (0.01,97.4)		3.8 (1.1,8.0)	
Mitotic rate, m(r)	>1 (n=39)	6 (0,31)	0.19	20.3 (0.6,75.6)	0.03	3.6 (1.6,7.4)	0.05
	1-2 (n=29)	8 (0,34)		21.5 (0.02,73.1)		3.9 (1.6,6.5)	
	>2 (n=163)	9 (0,35)		25.2 (0.01,97.4)		4.2 (1.1,9.5)	
TILs, m(r)	Absent (n=33)	6 (1,33)	0.16	23.7 (4.6,95.2)	0.21	3.9 (2.3,9.5)	0.13
	Brisk (n=18)	7.5 (0,21)		20.7 (0.02,49.7)		3.5 (1.6,6.2)	
	Non-Brisk (n=126)	9 (0,34)		25.1 (0.01,97.4)		4.1 (1.1,8.0)	
Mutation, m(r)	BRAF (n=126)	10 (0,35)	3.8E-03	25.2 (0.02,97.4)	7.5E-03	3.87 (1.1,8)	0.95
	NRAS (n=62)	6.5 (0,34)		20.9 (2.3,73.1)		3.88 (1.5,7.7)	
	Double wild type (n=89)	7 (0,30)		21.9 (0.01,91.5)		4.17 (1.4,9.5)	
Percentage of Stroma (POS), rho (n=206)		-0.09	0.22	-0.17	0.02	-0.24	4.4E-04

** Where n is the number of samples, m is the median , r is the range, rho is spearman rank correlation coefficient, and P is the test significance*

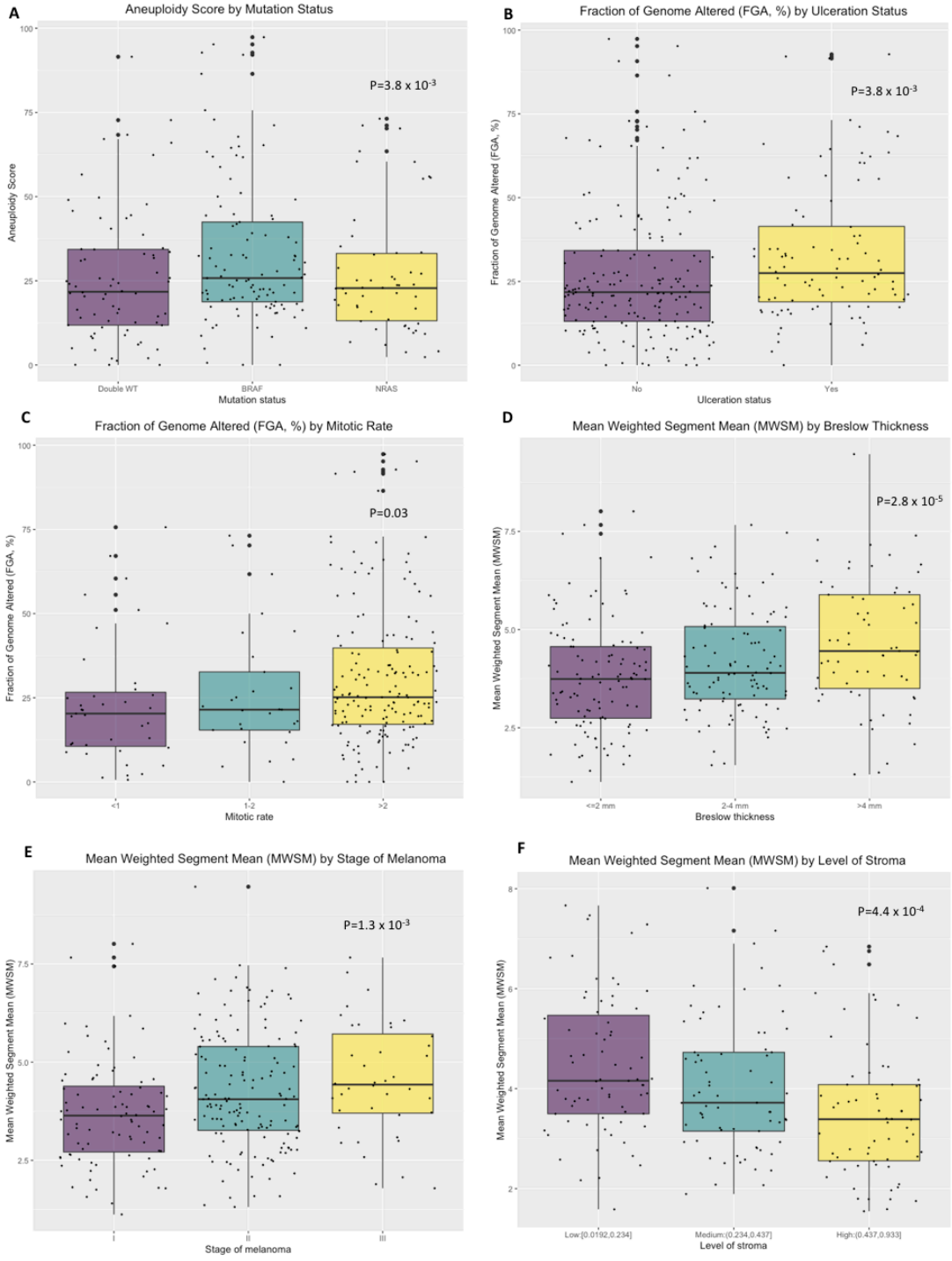


Figure 7.7. Clinical Characteristics Associated with Genome Instability for the most significant comparisons

7.3.4 Prognostic Value of Genomic Instability

7.3.4.1 Prognostic Value of Aneuploidy Score

Survival analysis shows that aneuploidy score predicts melanoma-specific survival (MSS) in univariable testing (HR=1.02, P=0.042) but not in multivariable (HR=1.01, P=0.22) testing using continuous measure after adjusting for age, sex, and stage (Table 7.5). When using the two-quantile measure, both univariable and multivariable analysis did not show significant association of aneuploidy score with MSS (Table 7.6) but showed hazard ratios that are indicative of a trend that increases with the quantile corresponding to higher aneuploidy score. This is graphically depicted by the Kaplan-Meier curve presented in Figure 7.8 below. The proportion of MSS deaths for the first quantile is 36% while it is 43 % for the second quantile.

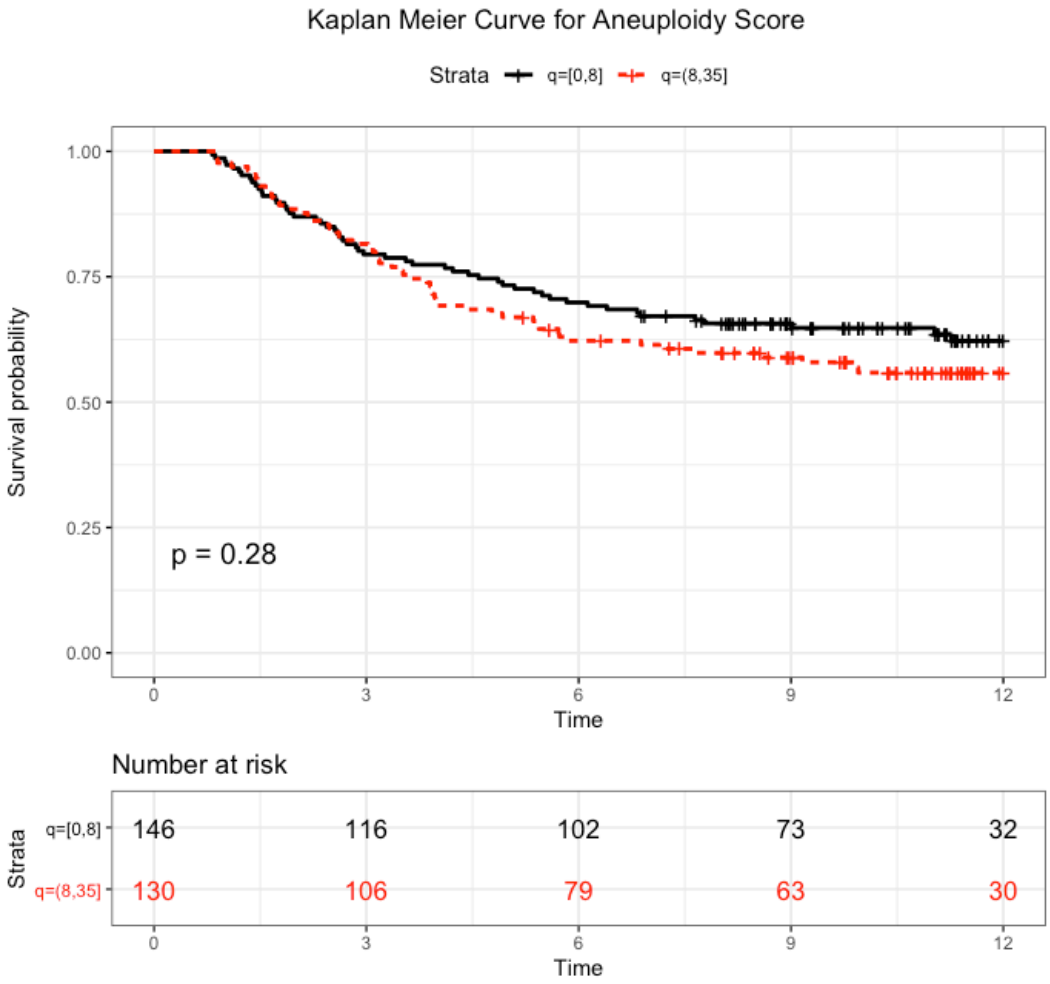


Figure 7.8. Kaplan Meier Curves for Aneuploidy Score

Table 7.5. Cox Hazard Model for Aneuploidy Score as a Continuous Variable

The variable n is the total number of samples, deaths are the number of melanoma specific (MSS) deaths, CI is the confidence interval, HR is the hazard ratio, and P is the significance of the test.

	Univariable					Multivariable (n=274, deaths =107)			
	Covariate(n)	**HR**	**95% CI**	**P**		**Covariate(n)**	**HR**	**95% CI**	**P**
Aneuploidy Score (n=276,deaths=109)	-	1.02	1.00-1.04	0.042	**Aneuploidy Score**	-	1.01	0.99-1.04	0.22
Age (n=276,deaths=109)	-	1.03	1.01-1.04	0.002	**Age**	-	1.03	1.01-1.04	0.005
Sex (n=276,deaths=109)	**F (144)**	1.00	-	-	**Sex**	**F (143)**	1.00	-	-
	M (132)	1.39	0.96-2.02	0.085		**M (131)**	1.32	0.90-1.93	0.16
Stage (n=274, deaths=107)	**I (88)**	1.00	-	-	**Stage**	**I (88)**	1.00	-	-
	II (146)	1.36	0.85-2.17	0.20		**II (146)**	1.27	0.79-2.04	0.32
	III (40)	3.81	2.22-6.54	1.26E-06		**III (40)**	3.55	2.05-6.14	6.23E-06

Table 7.6. Cox Hazard Model for Aneuploidy Score in Two-Quantile

The variable n is the total number of samples, deaths are the number of melanoma specific (MSS) deaths, CI is the confidence interval, HR is the hazard ratio, and P is the significance of the test.

	Univariable					Multivariable (n=274, deaths =107)			
	Covariate(n)	**HR**	**95% CI**	**P**		**Covariate(n)**	**HR**	**95% CI**	**P**
Aneuploidy Score (n=276,deaths=109)	**q(0,8] (146)**	1.00	-	-	**Aneuploidy Score**	**q(0,8] (146)**	1.00	-	-
	q(8,35] (130)	1.23	0.84-1.79	0.28		**q(8,35] (130)**	1.07	0.73-1.58	0.72
Age (n=276,deaths=109)	-	1.03	1.01-1.04	0.002	**Age**	-	1.03	1.01-1.04	0.005
Sex (n=276,deaths=109)	**F (144)**	1.00	-	-	**Sex**	**F (143)**	1.00	-	-
	M (132)	1.39	0.96-2.02	0.08		**M (131)**	1.31	0.90-1.93	0.16
Stage (n=274, deaths=107)	**I (88)**	1.00	-	-	**Stage**	**I (88)**	1.00	-	-
	II (146)	1.36	0.85-2.17	0.20		**II (146)**	1.30	0.81-2.08	0.28
	III (40)	3.81	2.22-6.53	1.26E-06		**III (40)**	3.71	2.15-6.40	2.62E-06

7.3.4.2 Prognostic Value of Fraction of Genome Altered (FGA%)

Fraction of genome altered (FGA, %) predicts MSS in univariable testing (HR=1.02, P=1.95 x 10^{-4}) in continuous measure and in multivariable testing (HR=1.02, P=0.001) after adjusting for age, sex, and stage (Table 7.7). When using the two-quantile measure, the univariable analysis significantly predicts MSS (HR= 1.72 compared with the first quantile, P=0.006) and show borderline significance (HR=1.47, P=0.056) in the multivariable analysis after adjusting for age, sex, and stage (Table 7.8) The difference in survival distribution between the first quantile (FGA= 0 - 23.3%) and second quantile (FGA = 23.3 - 97.4%) is graphically depicted by the Kaplan-Meier curve presented in Figure 7.9 below. Nonparametric test using log-rank test reveals that there is a significant difference between the survival curves of the two groups (P=0.005). The proportion of MSS deaths for the first quantile is 31% while it is 48 % for the second quantile.

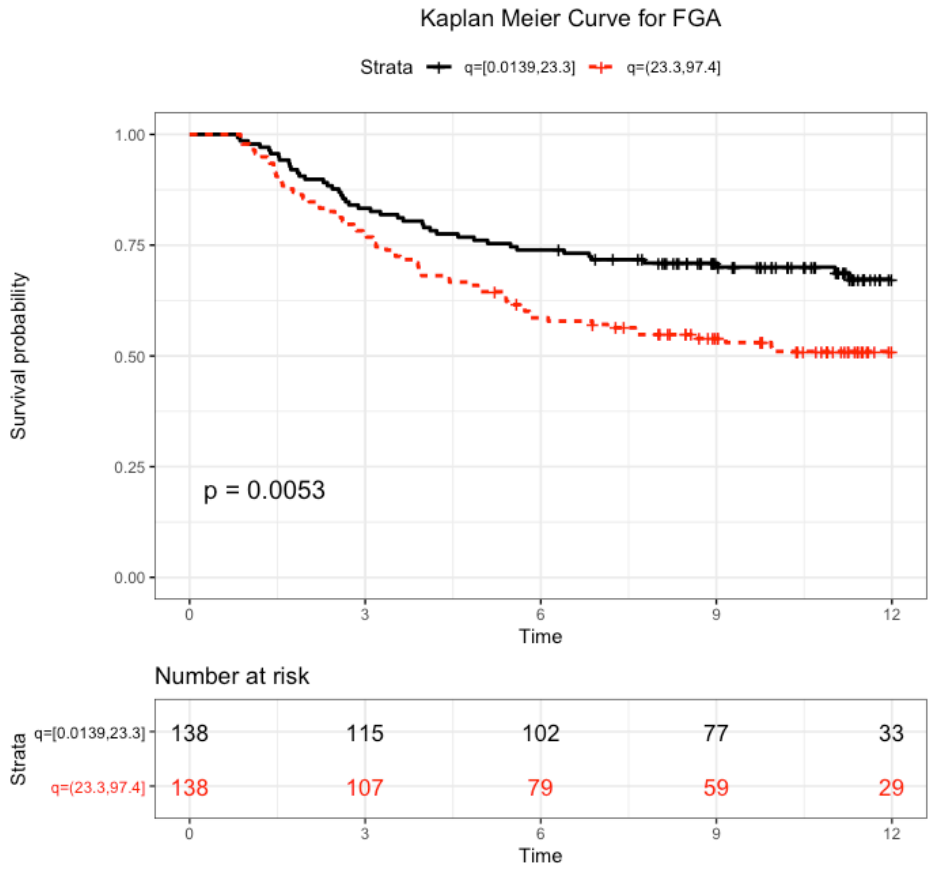


Figure 7.9. Kaplan Meier Curves for Fraction of Genome Altered (FGA %)

Table 7.7. Cox Hazard Model for Fraction of Genome Altered (FGA %) as a Continuous Variable

The variable n is the total number of samples, deaths are the number of melanoma specific (MSS) deaths, CI is the confidence interval, HR is the hazard ratio, and P is the significance of the test.

	Univariable					Multivariable (n=274, deaths =107)			
	Covariate(n)	**HR**	**95% CI**	**P**		**Covariate(n)**	**HR**	**95% CI**	**P**
Fraction of Genome Altered (FGA %) (n=276,deaths=109)	-	1.02	1.01-1.02	1.95E-04	**Fraction of Genome Altered (FGA %)**	-	1.02	1.01-1.02	0.001
Age (n=276,deaths=109)	-	1.03	1.01-1.04	0.002	**Age**	-	1.02	1.01-1.04	0.008
Sex (n=276,deaths=109)	**F (144)**	1.00	-	-	**Sex**	**F (143)**	1.00	-	-
	M (132)	1.39	0.96-2.02	0.085		**M (131)**	1.36	0.93-2.00	0.11
Stage (n=274, deaths=107)	**I (88)**	1.00	-	-	**Stage**	**I (88)**	1.00	-	-
	II (146)	1.36	0.85-2.17	0.20		**II (146)**	1.20	0.75-1.93	0.44
	III (40)	3.81	2.22-6.53	1.26E-06		**III (40)**	3.41	1.98-5.87	1.00E-05

Table 7.8. Cox Hazard Model for Fraction of Genome Altered (FGA %) in Two-Quantile

The variable n is the total number of samples, deaths are the number of melanoma specific (MSS) deaths, CI is the confidence interval, HR is the hazard ratio, and P is the significance of the test.

	Univariable					Multivariable (n=274, deaths =107)			
	Covariate(n)	**HR**	**95% CI**	**P**		**Covariate(n)**	**HR**	**95% CI**	**P**
Fraction of Genome Altered (FGA %) (n=276,deaths=109)	**q(0,23.3] (138)**	1.00	-	-	**Fraction of Genome Altered (FGA %)**	**q(0,23.3] (138)**	1.00	-	-
	q(23.3,97.4] (138)	1.72	1.17-2.52	0.006		**q(23.3,97.4] (138)**	1.47	0.99-2.18	0.056
Age (n=276,deaths=109)	-	1.03	1.01-1.04	0.002	**Age**	-	1.02	1.01-1.04	0.011
Sex (n=276,deaths=109)	**F (144)**	1.00	-	-	**Sex**	**F (143)**	1.00	-	-
	M (132)	1.39	0.96-2.02	0.085		**M (131)**	1.30	0.89-1.91	0.17
Stage (n=274, deaths=107)	**I (88)**	1.00	-	-	**Stage**	**I (88)**	1.00	-	-
	II (146)	1.36	0.85-2.17	0.20		**II (146)**	1.25	0.78-2.01	0.35
	III (40)	3.81	2.22-6.53	1.26E-06		**III (40)**	3.52	2.04-6.07	6.12E-06

7.3.4.3 Prognostic Value of Mean Weighted Segment Mean (MWSM)

Table 7.9 summarises the survival analysis performed on mean weighted segment mean in the continuous scale. In univariable testing, MWSM predicts MSS (HR=1.29, P=4.89 x 10^{-5}). Multivariable testing reveals that MWSM remains significant (HR=1.23, P=0.002) after adjusting for age, sex, and stage. When using the two-quantile MWSM, MWSM predicts MSS (HR=2.21, P=8.38 x 10^{-5}) in the univariate testing. This remained statistically significant (HR=1.80, P=0.005) after adjusting for age, sex, and stage (Table 7.10). The proportion of MSS deaths for the first quantile is 28 % while it is 51 % for the second quantile.

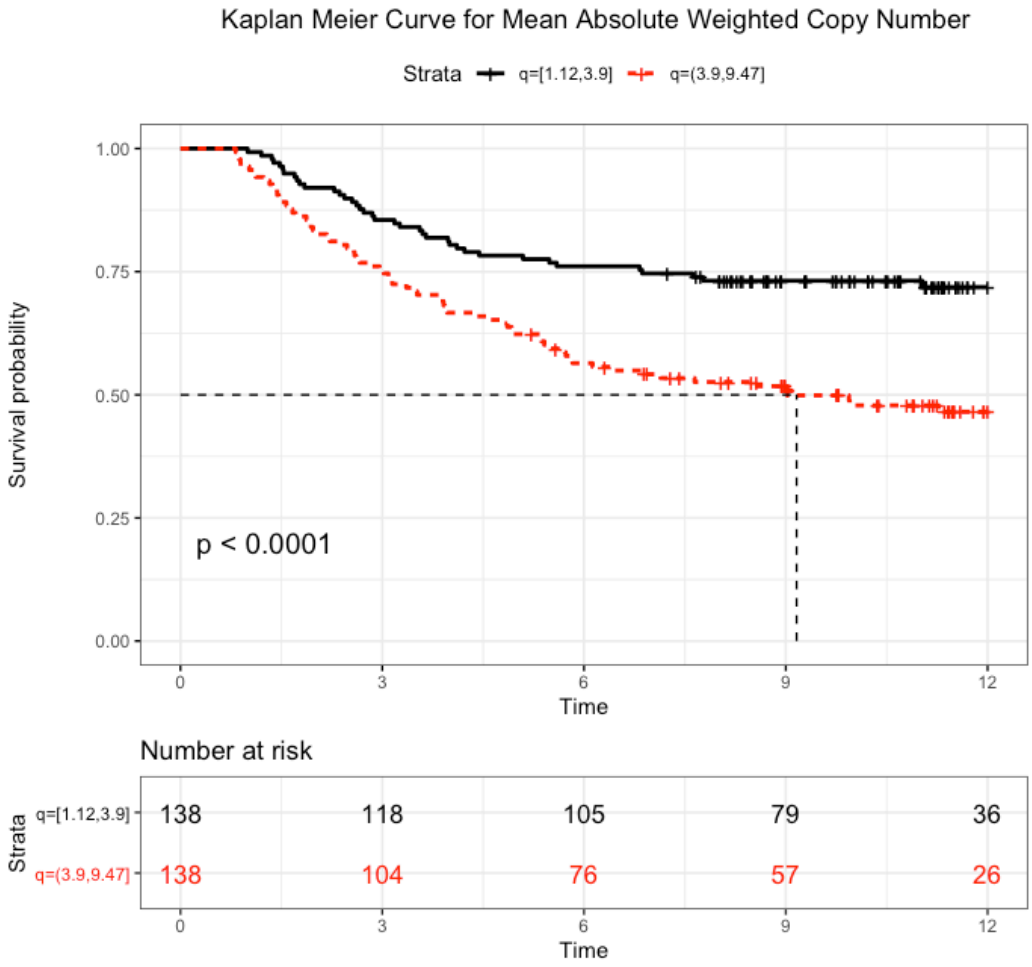


Figure 7.10. Kaplan Meier Curves for Mean Weighted Segment Mean (MWSM)

Table 7.9. Cox Hazard Model for Mean Weighted Segment Mean (MWSM) as a Continuous Variable

The variable n is the total number of samples, deaths are the number of melanoma specific (MSS) deaths, CI is the confidence interval, HR is the hazard ratio, and P is the significance of the test.

	Univariable					Multivariable (n=274, deaths =107)			
	Covariate(n)	**HR**	**95% CI**	**P**		**Covariate(n)**	**HR**	**95% CI**	**P**
Mean Weighted Segment Mean(MWSM) (n=276,deaths=109)	-	1.29	1.14-1.46	4.89E-05	**Mean Weighted Segment Mean(MWSM)**	-	1.23	1.08-1.39	0.002
Age (n=276,deaths=109)	-	1.03	1.01-1.04	0.002	**Age**	-	1.02	1.00-1.04	0.015
Sex (n=276,deaths=109)	**F (144)**	1.00	-	-	**Sex**	**F (143)**	1.00	-	-
	M (132)	1.39	0.96-2.03	0.085		**M (131)**	1.30	0.89-1.90	0.18
Stage (n=274, deaths=107)	**I (88)**	1.00	-	-	**Stage**	**I (88)**	1.00	-	-
	II (146)	1.36	0.85-2.17	0.20		**II (146)**	1.21	0.75-1.94	0.43
	III (40)	3.81	2.22-6.54	1.26E-06		**III (40)**	3.27	1.89-5.66	2.29E-05

Table 7.10. Cox Hazard Model for Mean Weighted Segment Mean (MWSM) in Two-Quantile

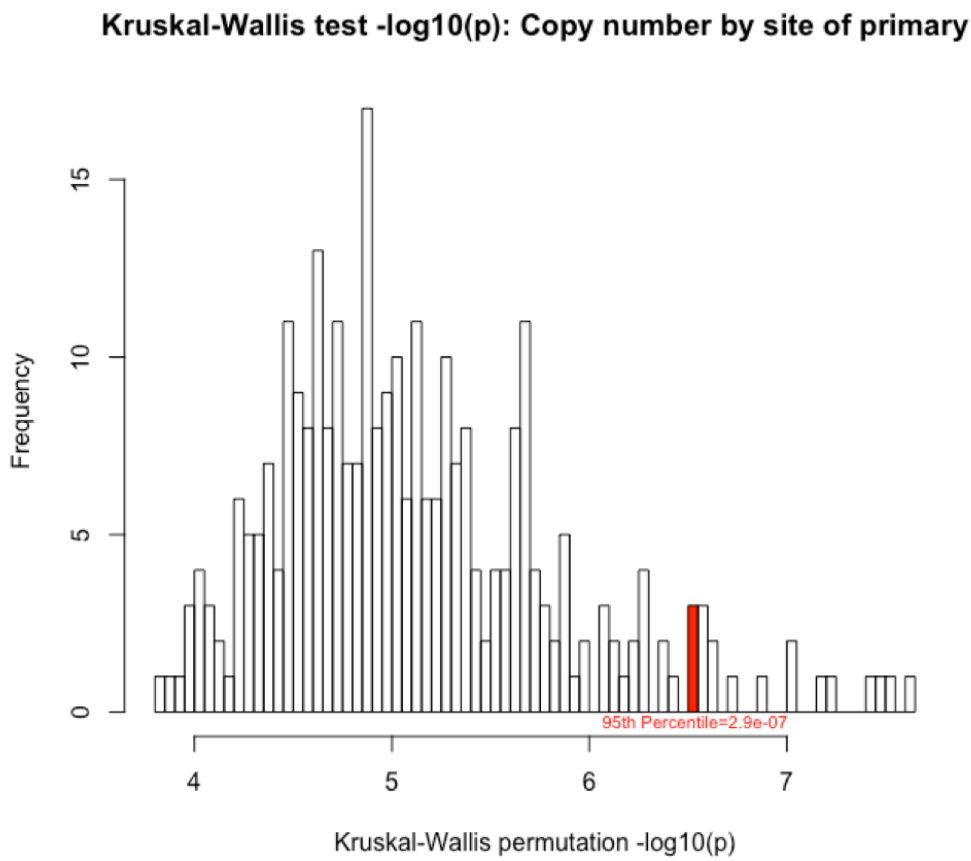
The variable n is the total number of samples, deaths are the number of melanoma specific (MSS) deaths, CI is the confidence interval, HR is the hazard ratio, and P is the significance of the test.

	Univariable					Multivariable (n=274, deaths =107)			
	Covariate(n)	**HR**	**95% CI**	**P**		**Covariate(n)**	**HR**	**95% CI**	**P**
Mean Weighted Segment Mean(MWSM) (n=276,deaths=109)	**q(0,3.9] (138)**	1.00	-	-	**Mean Weighted Segment Mean(MWSM)**	**q(0,3.9] (138)**	1.00	-	-
	q(3.9,9.47] (138)	2.21	1.49-3.27	8.38E-05		**q(3.9,9.47] (138)**	1.80	1.19-2.72	0.005
Age (n=276,deaths=109)	-	1.03	1.01-1.04	0.002	**Age**	-	1.02	1.00-1.04	0.014
Sex (n=276,deaths=109)	**F (144)**	1.00	-	-	**Sex**	**F (143)**	1.00	-	-
	M (132)	1.39	0.96-2.03	0.085		**M (131)**	1.23	0.83-1.80	0.30
Stage (n=274, deaths=107)	**I (88)**	1.00	-	-	**Stage**	**I (88)**	1.00	-	-
	II (146)	1.36	0.85-2.17	0.20		**II (146)**	1.22	0.76-1.96	0.41
	III (40)	3.81	2.22-6.54	1.26E-06		**III (40)**	3.21	1.85-5.57	3.44E-05

7.3.5 Permutation Analysis on Clinical Characteristics

7.3.5.1 Site of tumour

For the permutation analysis on the site of tumour based on a Kruskal-Wallis test, a total of 297 simulations were retrieved. The distribution of minimum P-values per genome were shown in Figure 7.11 below. The smallest P-value from the permutations is 2.4 x 10^{-8} while the highest is 1.5 x 10^{-4} with a median P-value of 1.03 x 10^{-5}. The 95th percentile of the distribution of the obtained P-values is 2.9 x10^{-7} (~0.00000029153414) and is marked as a red bar in the figure below.



Summary

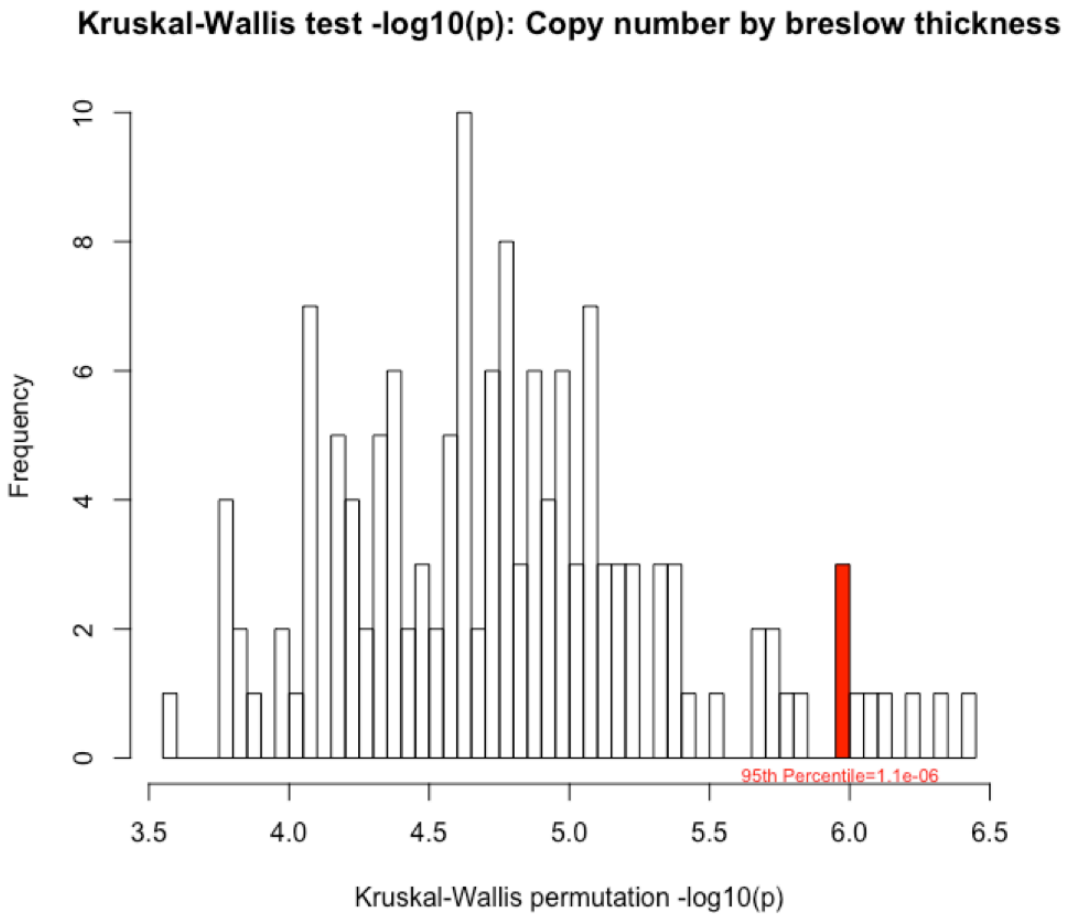
Min.	1st Qtr	Median	3rd Qtr	Max.	95th %
2.35E-08	3.08E-06	1.03E-05	2.37E-05	0.000154	2.92E-07

Figure 7.11. Distribution of P-values from Permutation Analysis for Site

7.3.5.2 Breslow thickness

Shown below in Figure 7.12 is the distribution of minimum P-values per genome resulting from the permutation analysis on the Breslow thickness of the primary melanoma using Kruskal-Wallis test. There were 138 iterations retrieved with the smallest P-value of 3.7 x 10^{-7}, a highest P-value of 2.6 x 10^{-4} and a median P-value of

1.8 x 10^{-5}. The 95th percentile of the distribution of the obtained P-values is 1.1 x10^{-6} (*~0.00000107880738*) and is marked as a red bar in the figure below.



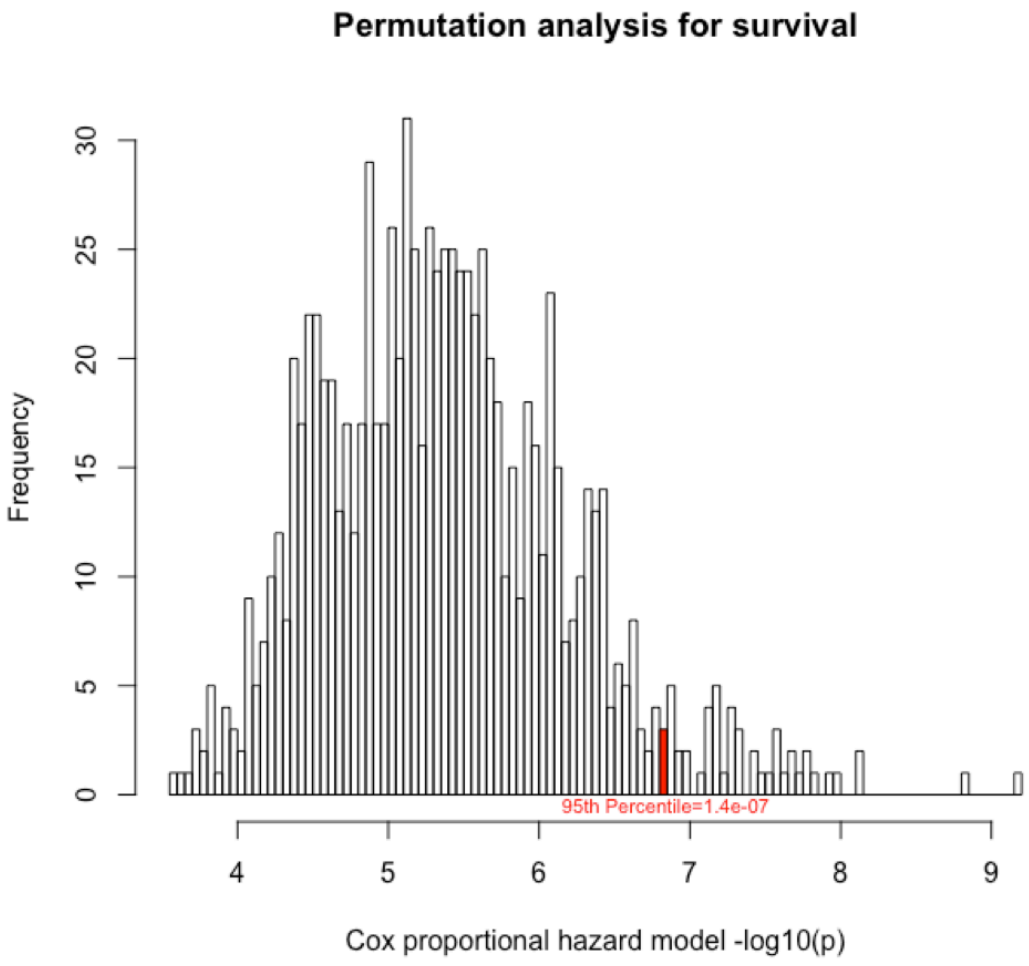
Summary

Min.	1st Qtr	Median	3rd Qtr	Max.	95th %
3.74E-07	8.51E-06	1.82E-05	4.36E-05	0.000261	1.08E-06

Figure 7.12. Distribution of P-values from Permutation Analysis for Breslow Thickness

7.3.5.3 Quantitative 10k windows

Figure 7.13 summarises the distribution of the most significant P-values for each of the 930 iterations retrieved from the permutation analysis. The minimum P-value is 6.3 x 10^{-10} while the highest is 2.6 x 10^{-4} with a median P-value of 4.8 x 10^{-6}. The 95th percentile of the distribution of the obtained P-values is 1.4 x10^{-7} (~0.00000014438647) and is marked as a red bar in the figure below.



Summary

Min.	1st Qtr	Median	3rd Qtr	Max.	95th %
6.38E-10	1.32E-06	4.82E-06	1.63E-05	0.000264	1.44E-07

Figure 7.13. Distribution of P-values from Permutation Analysis for Survival

7.3.6 Association of 10K windows with Clinical Characteristics

7.3.6.1 Sex and 10k window copy number

Figure 7.18 shows the whole autosomal genome copy number profile for cutaneous melanoma categorised by sex: Male (n=132) and Female (n=145). Visual inspection of chromosome arm by chromosome arm comparison of two groups suggest higher rate of copy number gain on the regions of chromosomes 1p, 6p, 7p, 7q, 8p, 8q and a higher rate of copy number loss for male group as compared to female group. A Manhattan plot of Wilcoxon rank sum test p-values testing for significant difference between copy number of male and female groups in each 10k window is presented in Figure 7.28. While a generally flat pattern is observed, eight 10k windows show statistical significance at 0.0001 level of significance, noting that a permutation-based study-

specific genome-wide significance threshold was not performed for the analysis of sex. Table 7.11 below shows the table describing each window with the corresponding mapping genes which are predominantly broadly expressed in the testis (*LINC01917 [199], LINC01919 [200], MDC1-AS1 [201]*), highly expressed in the prostate (*LINC01090 [202], MIR99AHG [203]*) or associated with prostate cancer (MIR5692B [204], *ANK3 [205]*). One window also maps with *MDC1* which is known to be associated with the regulation of *p53* [206, 207], and controls the formation of damage-induced 53BP1, *BRCA1* and *MRN* foci partly by promoting efficient *H2AX* phosphorylation [208].

The distribution of copy number of the most significant window (*LINC01917/ LINC01919*) is displayed in Figure 7.14 showing that female copy number for this window falls around normal (approximately zero) while male presents higher copy number.

Table 7.11. 10k copy number windows that are most significantly different between male and female

10k Window	P	FC	Mean (F)	Mean (M)	Location	Gene/s	Note
10k000262770	1.8E-44	0.17	0.004	1.76	18q21.2	*LINC01917, LINC01919*	Broadly Expressed in Testis
10k000109193	1.5E-06	0.85	0.16	0.32	6p21.33	*MDC1,* *MDC1-AS1*	Associated with regulation of *p53, and* formation of damage-induced 53BP1, BRCA1 and MRN foci Broadly Expressed in Testis
10k000282052	4.2E-06	0.86	-0.22	-0.07	21q22.3	*MIR5692B*	Associated with Metastatic Prostate Cancer
10k000173511	5.2E-06	0.84	-0.15	0.03	10q21.2	*ANK3*	Associated with poorer survival in Prostate Cancer
10k000279386	1.6E-05	1.13	0.09	-0.02	21q21.1	*MIR99AHG*	Highly Expressed in Prostate
10k000041643	5.5E-05	0.87	0.10	0.25	2q24.3	-	-
10k000063833	7.1E-05	0.89	0.02	0.13	3q24	-	-
10k000043720	7.3E-05	1.11	0.03	-0.07	2q32.1	*LINC01090*	Highly Expressed in Prostate

Fold Change ,exp(Mean_F)/exp(Mean_M) , F= Female , M=Male

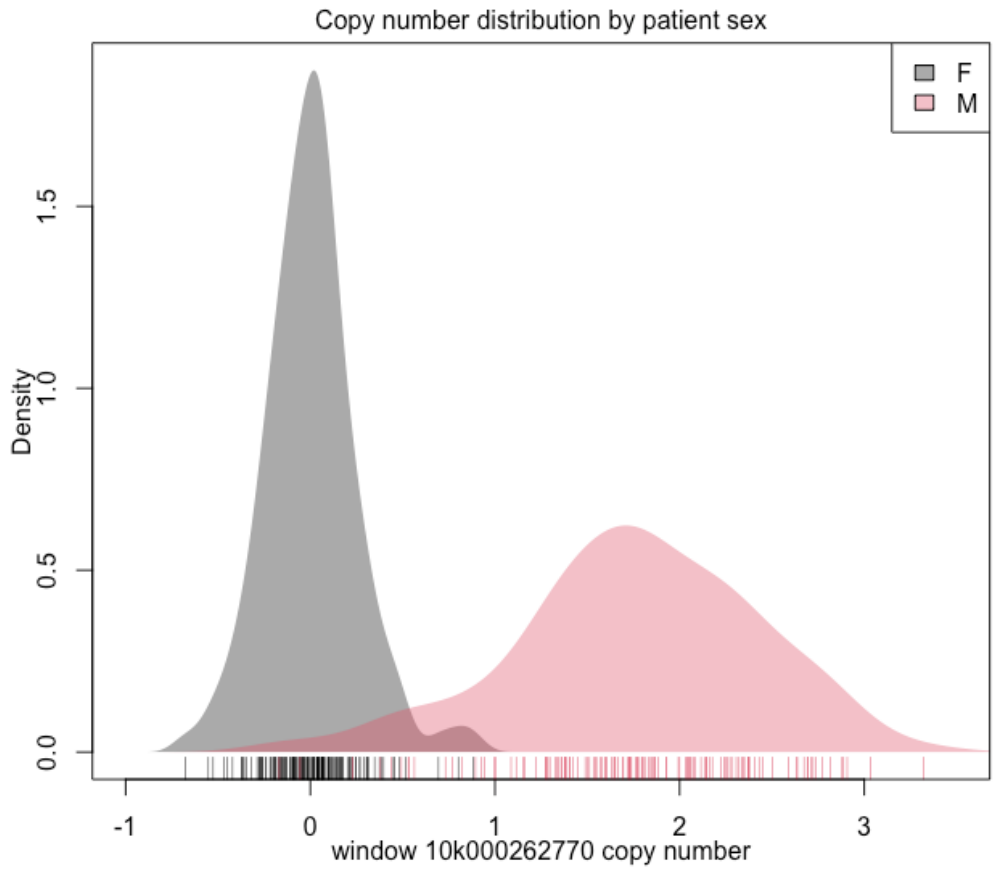


Figure 7.14. Distribution of LINC01917/LINC01919 10k window copy number by sex

7.3.6.2 Site and 10k window copy number

The 10k window copy number profile categorised by the site of primary as to Head (n=46), Limbs (n=139), and Trunk (92) did not present any clear pattern of difference. It suggests slightly higher rate of increased copy number in the chromosome 1q region for the Limbs group as compared with Head and Trunk groups and for chromosome 6p in the Head group as compared with Limbs and Trunk groups (Figure 7.19). Based on permutation analysis, any window with P-value less than 2.9 $x10^{-7}$ is deemed study-specific genome-wide significant. The Manhattan plot presenting the P-values testing for difference of copy number among sites of primary is presented in Figure 7.29. Checking the summaries of ranked windows according to test significance, only one window is lower than this cutoff. This window maps to *LINC01917* and *LINC01919* (Table 7.12). This is the same window that is found to be most significantly different in terms of copy number when comparing male and female groups. As previously mentioned, this gene is known to be broadly expressed in testis [199, 200]. Figure 7.15 shows the distribution of this 10k window in terms of site of primary melanoma.

This clearly shows that Limbs copy number tend be close to normal (approximately zero) while the copy number for Head and Trunk tend to be higher than zero.

Table 7.12. 10k copy number windows that are most significantly different among sites of primary melanoma

10k Window	P	Median (Head)	Median (Limbs)	Median (Trunk)	Location	Gene/s	Note
10k000262770	*3.82E-10*	*1.55*	*0.10*	*1.52*	*18q21.2*	*LINC01917, LINC01919*	*Broadly Expressed in Testis*
10k000019159	6.00E-06	-0.05	0.12	-0.02	1q31.2	LINC01680	Higher Expression in Testis
10k000024793	8.47E-06	-0.07	0.11	-0.05	1q44	OR2AJ1, OR2T8	Expressed in spleen, bone marrow, and Thyroid gland
10k000016763	1.62E-05	0.07	0.23	0.11	1q24.2	RCSD1	Higher expression in lymph node, spleen, appendix, and bone marrow
					1q24.2	CREG1	both activates and represses gene expression to promote cellular proliferation and inhibit differentiation
10k000006594	1.86E-05	-0.12	-0.11	-0.24	1p31.3	PDE4B	Higher expression in Bone marrow, brain, and appendix
10k000016887	2.11E-05	-0.13	0.03	-0.16	1q24.2	LINC00626	Associated with Head and neck squamous cell carcinoma tumors harboring human papillomavirus
10k000236545	2.32E-05	-0.09	0.02	-0.01	15q22.32	LINC02206	Higher expression in prostate, splee, fat, kidney, lymph node, and skin
10k000041345	2.71E-05	-0.10	0.04	0.05	2q24.3	GRB14	Higher expression in Kidney, liver, testis, ovary, and adrenal
10k000166290	3.55E-05	0.07	-0.12	0.11	9q33.3	MVB12B	Higher expression in brain, lung, spleen , and placenta
10k000017118	3.71E-05	0.09	0.19	0.04	1q24.3	FMO6P	Higher expression in salivary gland, stomach, and lung

**Row/s in blue are study-specific genome-wide significant. Genes in red are nearest approximate when no gene maps to the window*

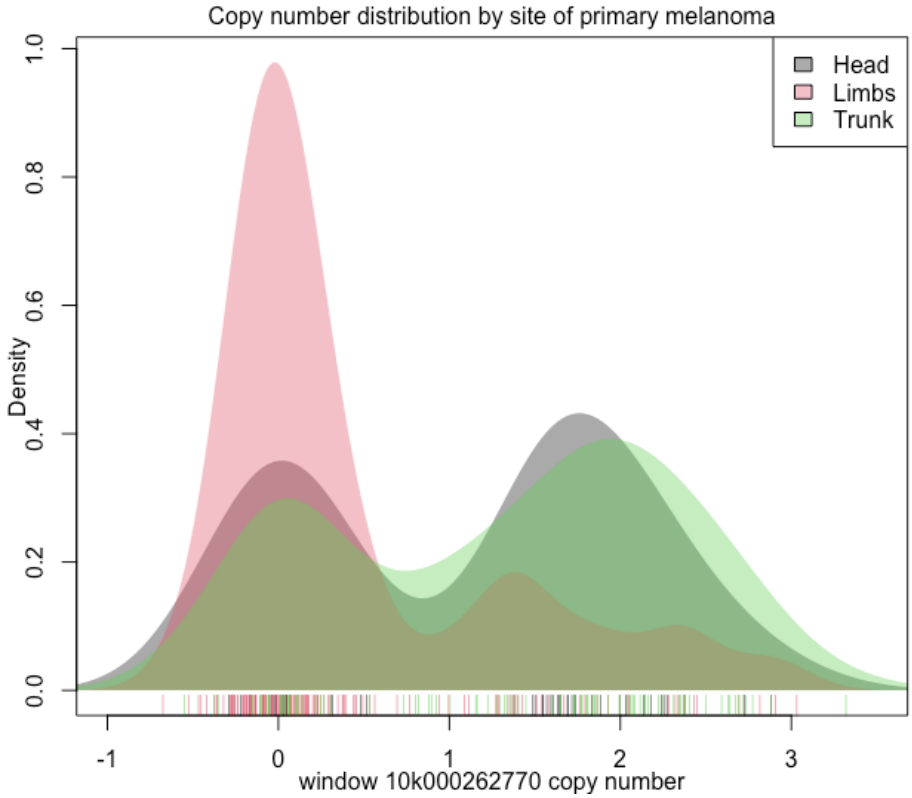


Figure 7.15. Distribution of LINC01917/LINC01919 10k window copy number by site

Other genes that show some evidence of copy number difference but did not reached the permutation based significance in terms of site of primary melanoma are: *LINC01680* which is also known to present higher expression in the testis [209], *LINC00626* which is known to be associated with head and neck squamous cell carcinoma tumours harbouring human papillomavirus [210, 211], *OR2AJ1, OR2T8 , RCSD1,PDE4B, LINC02206, GRB14, MVB12B , FMO6P*,which are known to be expressed in one or some of the following organs/tissues: spleen, bone marrow, lymph nodes, appendix, brain, prostate, kidney, fat, skin, liver, ovary, adrenal, lung, placenta, salivary gland, stomach, and thyroid gland [40, 212-218], and *CREG1* which is known to be play a role in both activation and repression of gene expression to promote cellular proliferation and inhibit differentiation [219].

A further investigation on the *LINC01917* / *LINC01919* copy number was done visually by plotting this region using both 10k window data and segment level data. Since the segment mean values and are affected by the values of the neighbouring windows with which it was segmented with, other samples that contributed to the copy number difference in Table 7.12 and Figure 7.15 and were segmented together with the neighbouring windows and reduced their magnitudes i.e. absolute segment means less than 0.10 were not included. The segment level plot is shown in Figure 7.16 below showing 24 samples with absolute segment means > 0.1. Two of these samples have segment means focused on the *LINC01919* and *LINC01917* specific regions. Checking for these two samples, both are Males and the tumour are both from the Head area. The remaining 22 samples have segments that are longer. These 22 samples are located in the regions (Head and Trunk) known to have higher copies of this *LINC01917*/ *LINC01919* 10k window are from Male samples. This is tested in the 277 samples and revealed significant association between site of the primary tumour and patient sex ($P=9.078 \times 10^{-10}$) as expected (Table 7.13).

Secondly, the window level plots are checked for window 10k000262770 to see how it compares around the neighbouring windows. Sample plots of clear copy number difference of this window over around the neighbouring windows is shown in Figure 7.17. This was observed in 128/277 (46%) of the samples.

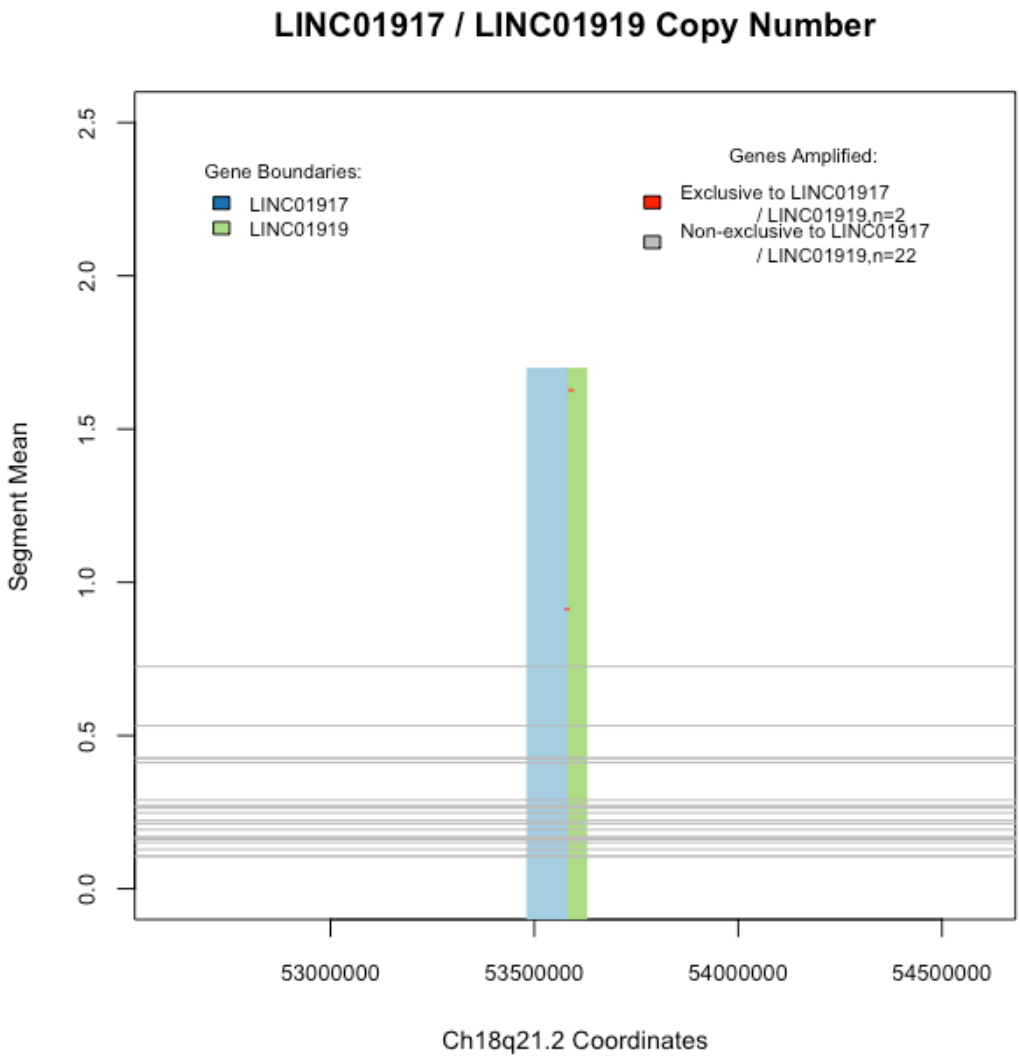


Figure 7.16. Plot of copy number segments around LINC01917 and LINC01919 genes

Table 7.13. Test for association between patient sex and site of primary melanoma

Site	Sex	
	Female	Male
Head	18(39%)	28(61%)
Limbs	99(71%)	40(29%)
Trunk	28(30%)	64(30%)
Total	145(52%)	132(48%)
Fisher test P-value:		9.08E-10

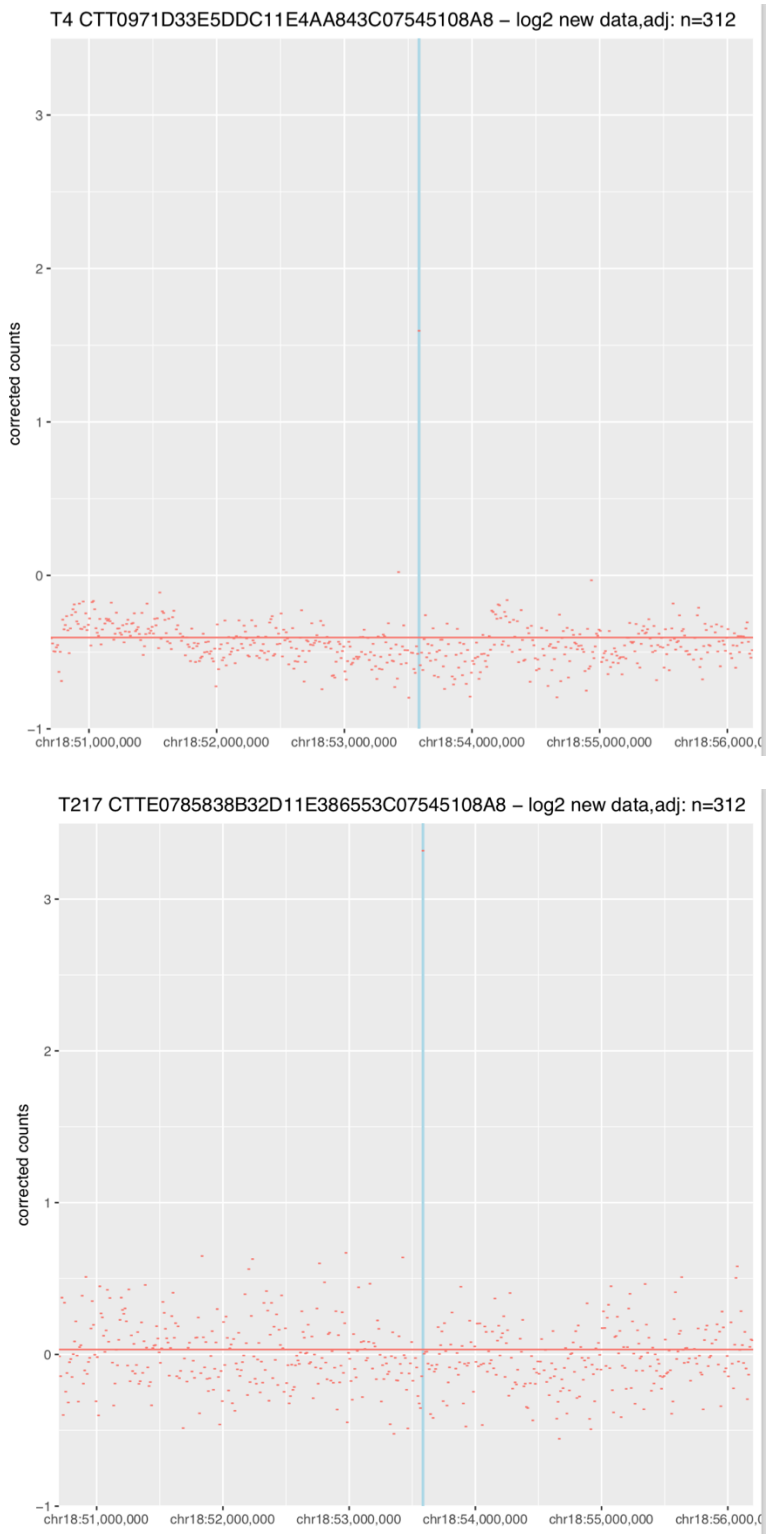


Figure 7.17. Plot of two samples which have extreme amplification for window 10k000262770

7.3.6.3 Age and 10k window copy number

In terms of age, Figure 7.20 shows the whole copy genome copy number plot of the samples grouped by two age categories with reference to the median age: 57.09 years and below (n= 139), and higher than 57.09 years (n=138). Visual comparison suggests that except for chromosome 6p where older patients tend to have higher rate of copy number aberrations, there does not seem to be clear visual difference between the copy number profiles of the two age groups. Association between copy number and age was tested using Spearman correlation as displayed in Figure 7.30. High correlation can be visually observed in the regions of chromosome 1, chromosome 4, chromosome 6, chromosome 8, chromosome 15, and chromosome 17. Table 7.14 below shows the 10k windows that tops the list of those which are most associated with age when ranked according to increasing test significance. First in the list is *MAPK10* which is known to be involved in cell proliferation, differentiation, transcription, regulation, and development [220]. This is followed by genes such as *FAM50B* which is related to a plant protein that plays role in the circadian clock [221], *ZNF184* (associated with higher expression in testis, brain, and endometrium [222]), *KRBA2* (which has higher expression in heart, kidney, and thyroid[223]), *NOTCH4* (which regulates interaction between physically adjacent cells [224]), *LNC-LBCS* (long-coding RNA bladder and prostate cancer suppressor), *RPP40* (mentioned as prognostic marker in renal cancer, endometrial cancer, and liver cancer [225]), *PPP1R3G* (higher expression in liver cancer), *RN7SKP293* (which regulates the activity of positive transcription elongation factor b [226]) , *RIPK1* (plays a role in inflammation and cell death in response to tissue damage, pathogen recognition, and as part of developmental regulation [227]), and *PSMG4* (influence the age at onset of clinical symptoms of multiple sclerosis [228]).

Table 7.14. 10k copy number windows that are most significantly associated with age

10k Window	P	Spearman	Location	Gene/s	Note
10k000077553	1.57E-07	-0.31	4q21.3	MAPK10	involved in cell proliferation, differentiation, transcription regulation and development
10k000106504	4.09E-07	0.30	6p25.2	FAM50B	related to a plant protein that plays a role in the circadian clock
10k000108866	1.62E-06	0.28	6p22.1	ZNF184	higher expression in testis, brain, and endometrium
10k000249922	1.63E-06	-0.28	17p13.1	KRBA2	higher expression in heart, kidney, and thyroid
10k000109343	2.04E-06	0.28	6p21.32	NOTCH4	regulates interactions between physically adjacent cells
10k000108102	2.44E-06	0.28	6p22.3	LNC-LBCS	lncRNA bladder and prostate cancer suppressor,
10k000106624	2.88E-06	0.28	6p25.1	RPP40	prognostic marker (unfavourable) in renal cancer, endometrial cancer, and liver cancer
				PPP1R3G	higher expression in liver cancer
10k000107356	3.26E-06	0.28	6p24.1	RN7SKP293	regulates the activity of positive transcription elongation factor b (P-TEFb)
10k000106433	3.36E-06	0.28	6p25.2	RIPK1	plays a role in inflammation and cell death in response to tissue damage, pathogen recognition, and as part of developmental regulation
10k000106449	3.65E-06	0.27	6p25.2	PSMG4	influence the age at onset of clinical symptoms of multiple sclerosis

**Genes in red are nearest approximate when no gene maps to the window*

7.3.6.4 Breslow thickness and 10k window copy number

For Breslow thickness, the 10k window whole genome copy number profile samples grouped according to three Breslow thickness categories: Breslow thickness <= 2mm (n=109), Breslow thickness 2-4 mm (n=100), and Breslow thickness >4mm (n=65) are shown in Figure 7.22. Similar to previous observation about patient age and copy number, there seems to be no clear pattern of difference among the three groups of samples according to Breslow thickness in terms of copy number except for chromosome 6 and chromosome 7. Figure 7.31 displays the P-values of the Kruskal-Wallis test performed between each 10k window of the genome and Breslow thickness. Notable differences can be observed in the regions of chromosomes 1q, 2, 3p, 6p, 7, and 20. A list of top windows and their corresponding mapping gene is presented in Table 7.15 below. Recalling the significance threshold identified in the permutation analysis for Breslow thickness which is 1.1 $x10^{-6}$ (*~0.00000107880738*), there were three windows which are considered genome-wide significant which map to *CXCR4* which is associated with cancer progression and cell survival [229-231], *SDK1* for which its silencing leads to cell rounding and blunted CaP cell migration [232], and *ITGA4* which was reported to be upregulated in melanoma[233].

Notable genes but did not reached study-specific genome-wide significance include *DDX1* which is known to be involved in transcription, viral replication, mRNA/miRNA processing, and transfer ribonucleic acid (tRNA) splicing and plays important role in the regulation of gene alternative splicing and insulin secretion in pancreatic β cells

[234] ; *CHST12* which is associated with tumour regrowth in non-functioning pituitary adenoma (NFPA) [235]; *GPR39* for which its overexpression contributes to malignant development of human oesophageal squamous cell carcinoma[236]; *SUN1* for which silencing of this gene inhibits cell growth through G0/G1 phase arrest in lung adenocarcinoma [237]; and *VEGFA* which activates an epigenetic pathway upregulating ovarian cancer-initiating cells [238]

Table 7.15. Top 10k copy number windows that are most significantly associated with Breslow thickness

10k Window	P	Spearman	Spearman	Spearman	Location	Gene/s	Note
10k000038511	*5.32E-08*	*-0.14*	*-0.23*	*-0.29*	*2q22.1*	*CXCR4*	*the binding of CXCL12 to CXCR4 induces intracellular signaling through several divergent pathways initiating signals related to chemotaxis, cell survival and/or proliferation, increase in intracellular calcium, and gene transcription*
10k000123539	6.22E-07	0.20	0.34	0.50	7p22.2	SDK1	silencing leads to cell rounding and blunted CaP cell migration
10k000043040	7.51E-07	-0.11	-0.19	-0.33	2q31.3	ITGA4	reported to be upregulted in melanoma
10k000123448	1.34E-06	0.07	0.17	0.29	7p22.3	CHST12	associated with tumor regrowth in non-functioning pituitary adenoma (NFPA)
10k000026455	1.59E-06	-0.03	-0.07	-0.18	2p24.3	DDX1	involved in transcription, viral replication, mRNA/miRNA processing, and tRNA splicing; important role in the regulation of gene alternative splicing and insulin secretion in pancreatic β cells
10k000111014	1.69E-06	0.09	0.17	0.39	6p12.3	RNU7-65P	RNA, U7 Small Nuclear 65 Pseudogene
10k000038151	1.91E-06	0.03	-0.04	-0.10	2q21.2	GPR39	overexpression contributes to malignant development of human esophageal squamous cell
10k000025748	2.53E-06	0.03	-0.02	-0.14	2p25.1	LINC01814	Long Intergenic Non-Protein Coding RNA 1814
10k000123284	3.31E-06	0.39	0.52	0.62	7p22.3	SUN1	silencing of this gene inhibits cell growth through G0/G1 phase arrest in lung adenocarcinoma
10k000110498	5.34E-06	0.12	0.15	0.32	6p21.1	VEGFA	activates an epigenetic pathway upregulating ovarian cancer-initiating cells

**Row/s in blue are study-specific genome-wide significant. Genes in red are nearest approximate when no gene maps to the window*

7.3.6.5 AJCC stage and 10k window copy number

In terms of AJCC stage, Figure 7.23 shows the 10k whole genome copy number grouped by AJCC stage. The chromosomes 6p and 7 resembles regions are clearly associated with higher rate of copy number amplifications for more advanced stage of melanoma. Difference among the three AJCC stages was tested and the resulting significance of the tests were plotted on the Manhattan plot in Figure 7.32. It suggests significant copy number differences in the regions of chromosomes 2, 7, and 11. In terms of the top 10k windows that have highest significance when testing for copy number difference in terms of stage (Table 7.16), genes associated with at least one form of cancer (e.g. liver, colorectal, prostate, pancreatic, thyroid, renal cell carcinoma, biliary tract, non-functioning pituitary adenoma, gastric carcinoma, acute myeloid leukemia, intrahepatic (ICC) , extrahepatic (ECC) cholangiocarcinoma, lung cancer and B-Cell Acute Lymphoblastic Leukemia) or its progression and cell survival are observed such as *CXCR4* which was previously identified in this study as genome-wide significant in terms of association with Breslow thickness [229-231], *UVRAG* [239,

240], *CYP3A54P* [241] ,*ACER3* [242], *ELFN1* [243], *ANO5* [244-246], *PRKAR1B* [247, 248], *CHST12* [235], *C7orf26* [249], *FAM20C* [250], and *PPME1* [251].

Table 7.16. Top 10k copy number windows that are most significantly associated with AJCC stage

10k Window	P	Stage 1*	Stage 2*	Stage 3*	Location	Gene/s	Note
10k000038511	2.45E-06	-0.13	-0.24	-0.30	2q22.1	CXCR4	multiple essential functions include homing of stem cells and metastasis of cancer cells; trafficking and homeostasis of immune cells such as T lymphocytes and plays important role in cancer progression
10k000188464	1.17E-05	0.09	0.04	0.21	11q13.5	UVRAG	frameshift mutation of UVRAG is associated with switching a tumor suppressor to an oncogene in colorectal cancer, and in gastric carcinomas with microsatellite instability
10k000123648	1.71E-05	-0.05	0.15	0.16	7p22.1	CYP3A54P	pseudogene of CYP3A - which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids
10k000188571	2.25E-05	0.13	0.01	0.20	11q13.5	ACER3	supports development of acute myeloid leukemia
10k000123372	2.29E-05	0.14	0.31	0.46	7p22.3	ELFN1	location intersects with ELFN1-AS1 which accelerates the proliferation and migration of colorectal cancer
10k000183047	3.65E-05	-0.01	-0.03	0.17	11p15.1	ANO5	associated in in prostate cancer progression ; regulates cell proliferation and migration in pancreatic cancer; regulates cell migration and invasion in thyroid cancer
10k000123274	3.77E-05	0.24	0.38	0.52	7p22.3	PRKAR1B	associated with poorer survival in renal cell carcinoma; significantly altered in intrahepatic (ICC) and extrahepatic (ECC) cholangiocarcinoma
10k000123448	4.43E-05	0.06	0.19	0.18	7p22.3	CHST12	associated with tumor regrowth in non-functioning pituitary adenoma (NFPA)
10k000123864	4.56E-05	0.20	0.30	0.38	7p22.1	C7orf26	higher expression in in B-Cell Acute Lymphoblastic Leukemia
10k000123222	4.61E-05	0.13	0.27	0.35	7p22.3	FAM20C	part of hypoxia-related key genes in Lung adenocarcinoma progression, which were regulated by DNA methylation
10k000188291	5.46E-05	0.05	0.04	0.24	11q13.4	PPME1	amplification is associated with gastric and lung cancer and its potential as a novel therapeutic target

** Median copy number* ***Genes in red are nearest approximate when no gene maps to the window*

7.3.6.6 Ulceration and 10k window copy number

Ulcerated tumours tend to have higher rate of aberration in the regions of chromosomes 6p and 9p (Figure 7.21). Testing this formally reveals that aside from these two regions, differences in copy number can also be observed in the regions of chromosomes 1p, 3p, 10p, 11q, 12q, 14p, and 14q as shown in Figure 7.33. The top 10 windows in terms of test significance for copy number by ulceration status are shown in Table 7.17 below. This list comprises *NEBL* which is enriched in the heart muscle and reported to be a favourable prognostic marker in renal cancer and unfavourable prognostic marker in urothelial cancer [252], *LINC01623* (long intergenic non-protein coding RNA 1623), *GGNBP1* which is expressed higher in testis [253] , *RN7SL26P* (RNA, 7SL, Cytoplasmic 26, Pseudogene), *RNF182* which has higher expression in brain, bone marrow, testis, and lung [254], *ADAMTSL1* which was reported to be differentially methylated between paired tumour and normal tissues from breast cancer patients [255], *ELOCP20* (Elongin C Pseudogene 20), *ITPR3* which encodes a receptor for inositol 1,4,5-trisphosphate, a second messenger that mediates the release of intracellular calcium, *RNMTL1P1* (RNA methyltransferase like 1

pseudogene 1), *SLC22A23* which has high expression in stomach, colon, small intestine, duodenum, prostate, and oesophagus, *ATP6V1G1P4* (ATPase H+ Transporting V1 Subunit G1 Pseudogene 4), and *CREM* which was reported to be progressively and significantly upregulated from controls to nondysplastic UC (ulcerative colitis) to UC with neoplasia.

Table 7.17. Top 10k copy number windows that are most significantly associated with ulceration status

10k Window	P	FC	Non-ulcerated*	Ulcerated*	Location	Gene/s	Note
10k000169597	5.5E-06	1.19	-0.154	-0.33	10p12.31	*NEBL*	Tissue enriched (heart muscle), prognostic marker in renal cancer (favourable) and urothelial cancer (unfavourable)
10k000109008	6.1E-06	0.85	0.44	0.61	6p22.1	*LINC01623*	long intergenic non-protein coding RNA 1623
10k000109477	6.4E-06	0.87	0.15	0.29	6p21.31	*GGNBP1*	higher expression in testis
						RN7SL26P	RNA, 7SL, Cytoplasmic 26, Pseudogene
10k000107512	6.9E-06	0.86	0.09	0.24	6p23	*RNF182*	higher expression in brain, bone marrow, testis, and lung
10k000155490	7.3E-06	1.25	-0.24	-0.46	9p22.2	*ADAMTSL1*	differentially methylated between paired tumor and normal tissues from breast cancer patients
10k000011549	9.0E-06	0.84	0.00	0.17	1p13.2	*ELOCP20*	Elongin C Pseudogene 20
10k000109491	9.7E-06	0.86	0.23	0.38	6p21.31	*ITPR3*	encodes a receptor for inositol 1,4,5-trisphosphate, a second messenger that mediates the release of intracellular calcium
10k000169633	1.1E-05	1.20	-0.16	-0.34	10p12.31	*RNMTL1P1*	RNA methyltransferase like 1 pseudogene 1
10k000106471	1.1E-05	0.85	0.12	0.29	6p25.2	*SLC22A23*	high expression in stomach, colon, small intestine, duodenum, prostate, and esophagus
10k000171008	1.1E-05	1.24	0.02	-0.19	10p11.21	*ATP6V1G1P4*	ATPase H+ Transporting V1 Subunit G1 Pseudogene 4
	1.1E-05	1.24	0.02	-0.19	10p11.21	*CREM*	progressively and significantly upregulated from controls to nondysplastic UC (ulcerative colitis) to UC with neoplasia

** Mean copy number **Genes in red are nearest approximate when no gene maps to the window*

7.3.6.7 Mitotic rate and 10k window copy number

Tumours with higher mitotic rate tend to be more aberrated in terms of chromosomes 6p (amplification), 9p (deletion), 10q (deletion), and 11q (deletion) as shown in Figure 7.24. Checking the Manhattan plot in Figure 7.34, it can be observed that the regions with significant differences across the three groups in terms of mitotic rate (>1 , 1-2 , >2) are in the regions of chromosome 6p, 7p, 7q, 10p, 10q, 11q, 15q, 18q, and 20q. Table 7.18 below summarizes the top 10 windows when testing for copy number difference by with mitosis which include *MIR100HG* that promotes colorectal cancer metastasis and is associated with poor prognosis [256], *NRSN1* which is reported to be significantly related to grade and prognosis of Glioma [257], *ST3GAL4* which is associated with Hyperglycemia and Inflammatory Breast Carcinoma and biological pathways such as Metabolism of proteins and Pre-*NOTCH* Expression and

Processing [258], *GSEC* which is associated with Colorectal Cancer and Al-Raqad Syndrome[259], *LINCMD1* [260], MIR133B [261], *MIR206* [262], *SLC25A20P1* (Solute Carrier Family 25 Member 20 Pseudogene 1), IL17F which is expressed by activated T cells, and has been shown to stimulate the production of several other cytokines, including *IL6*, *IL8*, and CSF2/GM_CSF [263], *ARHGEF12* [264], *DCPS* [265], *ALDH5A1* [266], *KIAA0319* [267], and *F13A1* which is reported as a new biomarker for the screening of colorectal cancer [268].

Table 7.18. Top 10k copy number windows that are most significantly associated with mitosis

10k Window	P	< 1*	1 to 2*	> 2*	Location	Gene/s	Note
10k000193092	3.53E-07	-0.16	-0.24	-0.41	11q24.1	MIR100HG	promotes colorectal cancer metastasis and is associated with poor prognosis
10k000108537	5.44E-07	0.11	0.29	0.40	6p22.3	NRSN1	significantly related to grade and prognosis of Glioma
10k000106734	5.52E-07	0.01	-0.01	0.23	6p25.1	F13A1	reported as a new biomarker for the screening of colorectal cancer
10k000111337	5.75E-07	-0.12	-0.01	0.08	6p12.2	LINCMD1	associated with Increased Risk to Hip and Knee Osteoarthritis
	5.75E-07	-0.12	-0.01	0.08	6p12.2	MIR133B	related pathways are MicroRNAs in cardiomyocyte hypertrophy and Cell Differentiation - Index
	5.75E-07	-0.12	-0.01	0.08	6p12.2	MIR206	related pathways are miRs in Muscle Cell Differentiation and Glial Cell Differentiation
	5.89E-07	-0.21	-0.01	0.01	6p12.2	SLC25A20P1	Solute Carrier Family 25 Member 20 Pseudogene 1
	5.89E-07	-0.21	-0.01	0.01	6p12.2	IL17F	expressed by activated T cells, and has been shown to stimulate the production of several other cytokines, including IL6, IL8, and CSF2/GM_CSF; also found to inhibit the angiogenesis of endothelial cells and induce endothelial cells to produce IL2, TGFB1/TGFB, and monocyte chemoattractant protein-1
10k000193513	6.32E-07	0.10	0.03	-0.08	11q24.2	ST3GAL4	associated with Hyperglycemia and Inflammatory Breast Carcinoma; related pathways are Metabolism of proteins and Pre-NOTCH Expression and Processing
10k000192918	6.37E-07	0.11	0.09	-0.06	11q23.3	ARHGEF12	encodes a protein that may form a complex with G proteins and stimulate Rho-dependent signals and has been observed to form a myeloid/lymphoid fusion partner in acute myeloid leukemia
10k000109880	6.78E-07	0.04	0.14	0.25	6p21.2	MIR4462	MicroRNA 4462
10k000193508	8.42E-07	0.04	-0.03	-0.15	11q24.2	GSEC	associated with Colorectal Cancer and Al-Raqad Syndrome
	8.42E-07	0.04	-0.03	-0.15	11q24.2	DCPS	associated with DCPS include Al-Raqad Syndrome and Autosomal Recessive Non-Syndromic Intellectual Disability; related pathways are Gene Expression and Deadenylation-dependent mRNA decay
10k000108575	1.01E-06	0.18	0.24	0.42	6p22.3	ALDH5A1	associated with ALDH5A1 include Succinic Semialdehyde Dehydrogenase Deficiency and Gamma-Amino Butyric Acid Metabolism Disorder; s related pathways are Valproic Acid Pathway, Pharmacodynamics and Neurotransmitter Release Cycle.
10k000108583	1.02E-06	0.11	0.18	0.37	6p22.3	KIAA0319	Involved in neuronal migration during development of the cerebral neocortex; may function in a cell autonomous and a non-cell autonomous manner and play a role in appropriate adhesion between migrating neurons and radial glial fibers; may also regulate growth and differentiation of dendrites.

** Mean copy number **Genes in red are nearest approximate when no gene maps to the window*

7.3.6.8 Tumour Infiltrating Lymphocytes (TILs) and 10k window copy number

Higher rate of aberration for tumours with no TILs and tumours with Brisk TILs as compared with tumours with non-brisk TILs is observed in the regions of chromosome 6p (amplification) as depicted in Figure 7.25. When checking for significant difference of copy number among the three groups based on TILs, the window that shows the highest significance is in chromosome 4, followed by chromosome 11 and chromosome 3 (Figure 7.35). Table 7.19 summarises the top 10k windows which show copy number difference by TILs including the description of the genes mapping to them. An interesting gene on the list is *LRRIQ3* which was previously reported to be associated with response to chemotherapy in rectal cancer [269]. Other genes on the list are *ZSWIM5P3* (Zinc Finger SWIM-Type Containing 5 Pseudogene 3), *LRRC4C* which is a specific binding partner for netrin G1 [270], LINC02053 which has higher expression in testis [271], and *RNU1-89P* which was reported to have differential expression of small nuclear RNA (snRNA) in oesophageal adenocarcinoma[272].

Table 7.19. Top 10k copy number windows that are most significantly associated with TILS

10k Window	P	Absent*	Brisk*	Non-brisk*	Location	Gene/s	Note
10k000081873	1.29E-05	-0.34	-0.10	-0.08	4q28.2	ZSWIM5P3	Zinc Finger SWIM-Type Containing 5 Pseudogene 3
10k000184907	2.43E-05	-0.29	0.01	-0.03	11p12	LRRC4C	a specific binding partner for netrin G1
10k000067167	5.54E-05	-0.22	0.03	-0.13	3 q26.33	LINC02053	higher expression in testis
10k000082560	6.08E-05	-0.29	0.03	-0.05	4q28.3	RNU1-89P	differential expression of snRNA in esophageal adenocarcinoma
10k000007403	9.87E-05	-0.28	0.04	-0.09	1p31.1	LRRIQ3	associated with response to responses to chemoradiotherapy in rectal cancer

** Median copy number* ***Genes in red are nearest approximate when no gene maps to the window*

7.3.6.9 Mutation status and 10k window copy number

In terms of mutation status, *NRAS* group tend to have higher rate of copy number amplification in chromosomes 1q (*NRAS* is located in chromosome 1p) and 6p as compared to *BRAF* group and double wild type group. *BRAF* group has higher rate of copy number amplification in chromosome 7 (where *BRAF* is located) as compared to *NRAS* group and double wild type group (Figure 7.26). Checking the manhattan plot for mutation status in Figure 7.36, it can be seen that the *NRAS* (chromosome 1p near centromere) and *BRAF* (chromosome 7q) regions shows highly significant copy number difference across the sample groups in terms of mutation status. Similar observation is found in chromosomes 10q - a region which is reported to be aberrated in melanoma [146]. The summary of the top significant 10k windows is presented in Table 7.20 below. The list is topped by *ADAM12* which has been reported to contribute to increased tumour proliferation, metastasis and endocrine resistance[273]. Other

cancer related genes on the list are *ALDH18A1* which is associated with associated with luminal B breast cancer [274], *VTI1A* which is associated with susceptibility to colorectal and lung cancers [275], *SORBS1* which has been reported to suppress tumour metastasis and improves the sensitivity of cancer to chemotherapy drug [276], *LGI1* which has been reported to be differentially expressed in early and late-stage oral squamous cell carcinoma [277], and *ZNF777* which inhibits proliferation at low cell density through down-regulation of *FAM129A* [278].

Table 7.20. Top 10k copy number windows that are most significantly associated with mutation status

10k Window	P	BRAF*	NRAS*	DWT*	Location	Gene/s	Note
10k000180124	2.68E-13	-0.35	-0.07	-0.05	10q26.2	ADAM12	contributes to increased tumor proliferation, metastasis and endocrine resistance
10k000176807	5.45E-13	-0.47	-0.15	-0.11	10q23.33	XRCC6P1	associated with Open-angle glaucoma risk
10k000137098	5.47E-13	0.32	0.03	0.11	7q34	KIAA1549	associated with pilocytic astrocytoma (PA)
10k000177057	7.88E-13	-0.40	-0.06	-0.10	10q24.1	ALDH18A1	associated with luminal B breast cancer
10k000137887	7.96E-13	0.45	0.21	0.21	7q35	CNTNAP2	prominent susceptibility gene implicated in multiple complex neurodevelopmental disorders, including autism spectrum disorders , intellectual disability , and schizophrenia
10k000178743	8.39E-13	-0.34	0.04	0.02	10q25.2	VTI1A	Polymorphisms in this gene have been associated with binocular function, and also with susceptibility to colorectal and lung cancers
10k000176582	9.32E-13	-0.49	-0.11	-0.17	10q23.31	RPP30	higher expression in testis, ad lymph node
10k000177037	9.39E-13	-0.46	-0.04	-0.15	10q24.1	SORBS1	suppresses tumor metastasis and improves the sensitivity of cancer to chemotherapy drug
10k000176877	9.40E-13	-0.50	-0.15	-0.15	10q23.33	LGI1	differentially expressed in early- and late-stage oral squamous cell carcinoma
10k000138144	1.16E-12	0.31	0.04	0.10	7q36.1	ZNF777	inhibits proliferation at low cell density through down-regulation of FAM129A

** Median copy number* ***Genes in red are nearest approximate when no gene maps to the window*
DWT = Double wild type

7.3.6.10 Percentage of stroma and 10k window copy number

The rate of copy number amplification in chromosome 6 tend to increase with the decrease in percentage of stroma while the rate of copy number deletion in chromosome 9p tend to decrease with the decrease in percentage of stroma as shown in Figure 7.27. Test for correlation using Spearman's rho identifies the region of

chromosome 1q as the most correlated with percentage of stroma as shown in Figure 7.37.

Table 7.21 summarizes the list of top 10k windows that shows top correlation significance with percentage of stroma and shows *KCNT2* as top on the list. The expression of this gene is correlated with the prognosis of skin cutaneous melanoma (SCM) and significantly different between normal skin and SCM [279]. This followed by *OR2G6* which is associated in breast cancer [280]. Other genes in the list include *FAM163A* which is reported to be positive regulator of *ERK* signalling pathway, interacts with 14-3-3β and promotes cell proliferation in squamous cell lung carcinoma [281], *AXDND1* which is reported to be downregulated in gastric cancer [282], *INTS7* which is associated with Gastric Cancer and Ivic Syndrome, and its increased levels is associated with the aggressiveness of prostate cancer [283, 284], *RALGPS2* which is reported to be essential for survival and cell cycle progression of lung cancer cells [285], *PLD5* which is associated with survival of thyroid cancer patients [286], *LELP1* which is significantly increased in atopic dermatitis skin [287], *RSL24D1P4* (Ribosomal L24 Domain Containing 1 Pseudogene 4), *EFCAB2* which is considered as one of the most potential targets for four breast cancer subtypes, a specific therapeutic targets for luminal A, and identified as a susceptibility gene for Colorectal Cancer in East Asian populations [288, 289], and *PKP1* where its phosphorylation by *RIPK4* regulates epidermal differentiation and skin tumorigenesis [290].

Table 7.21. Top 10k copy number windows that are most significantly associated with percentage of stroma

10k Window	P	Rho	Location	Gene/s	Note
10k000019628	3.43E-09	-0.40	1q31.3	KCNT2	expression was correlated with the prognosis of SCM and significantly different between normal skin and SCM
10k000024853	5.96E-09	-0.39	1q44	OR2G6	associated in human breast cancer
10k000017980	7.13E-08	-0.36	1q25.2	FAM163A	positive regulator of ERK signaling pathway, interacts with 14-3-3β and promotes cell proliferation in squamous cell lung carcinoma
10k000017940	7.23E-08	-0.36	1q25.2	AXDND1	downregulated in gastric cancer
10k000021188	7.85E-08	-0.36	1q32.3	INTS7	associated with Gastric Cancer and Ivic Syndrome; increased levels is associated with the aggressiveness of prostate cancer
10k000017896	8.70E-08	-0.36	1q25.2	RALGPS2	essential for survival and cell cycle progression of lung cancer cells
	9.52E-08	-0.36	1q43	PLD5	associated with survival of thyroid cancer patients
10k000015321	9.85E-08	-0.36	1q21.3	LELP1	significantly increased in atopic dermatitis skin
10k000024288	1.08E-07	-0.36	1q43	RSL24D1P4	Ribosomal L24 Domain Containing 1 Pseudogene 4
10k000024513	1.29E-07	-0.36	1q44	EFCAB2	considered one of the most potential targets for four breast cancer subtypes ; specific therapeutic targets for luminal A; identified as a susceptibility gene for Colorectal Cancer in east asian populations
10k000020133	1.53E-07	-0.36	1q32.1	PKP1	Phosphorylation by RIPK4 regulates epidermal differentiation and skin tumorigenesis

***Genes in red are nearest approximate when no gene maps to the window*

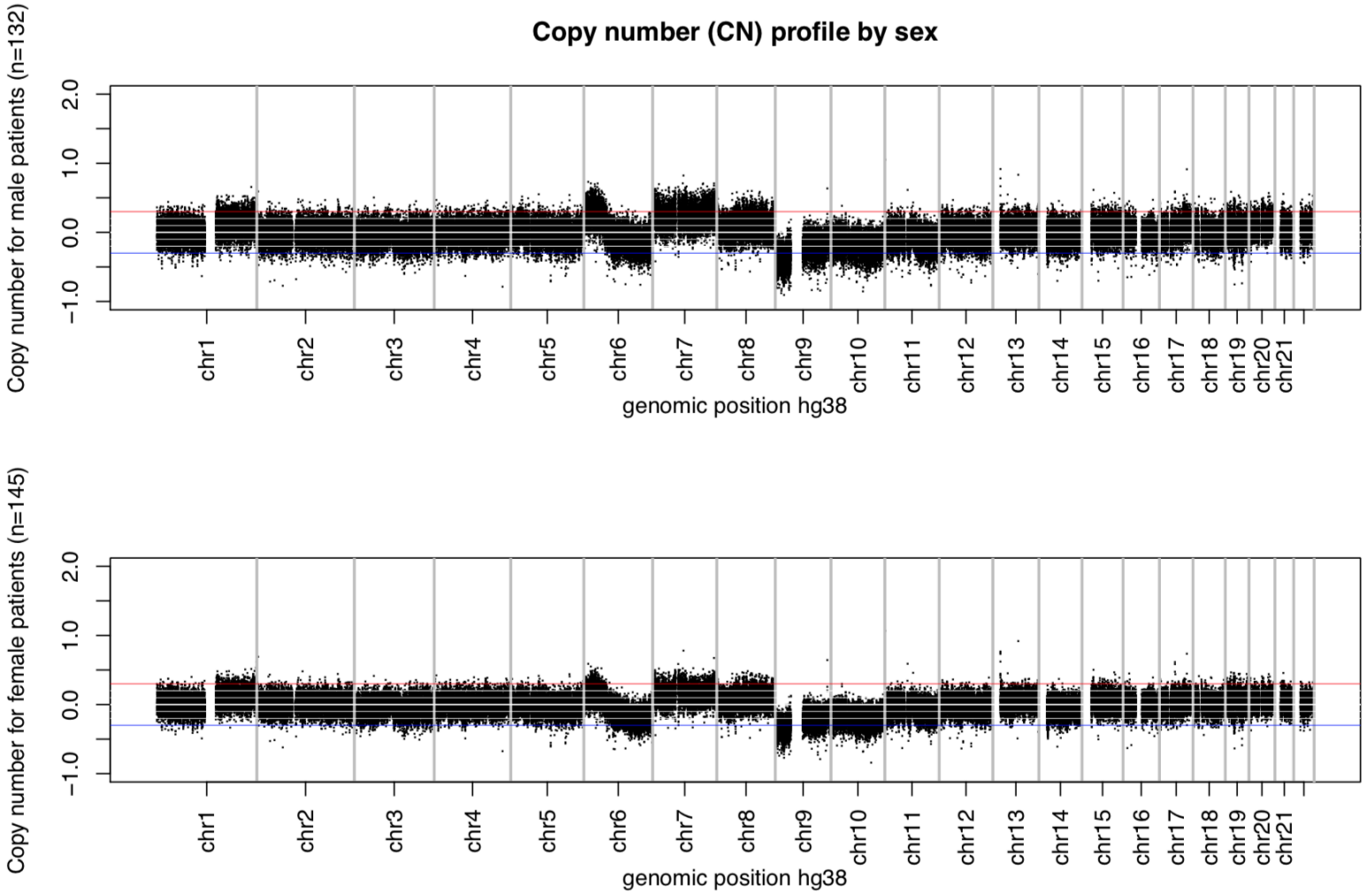


Figure 7.18. 10k Window copy number profile by sex

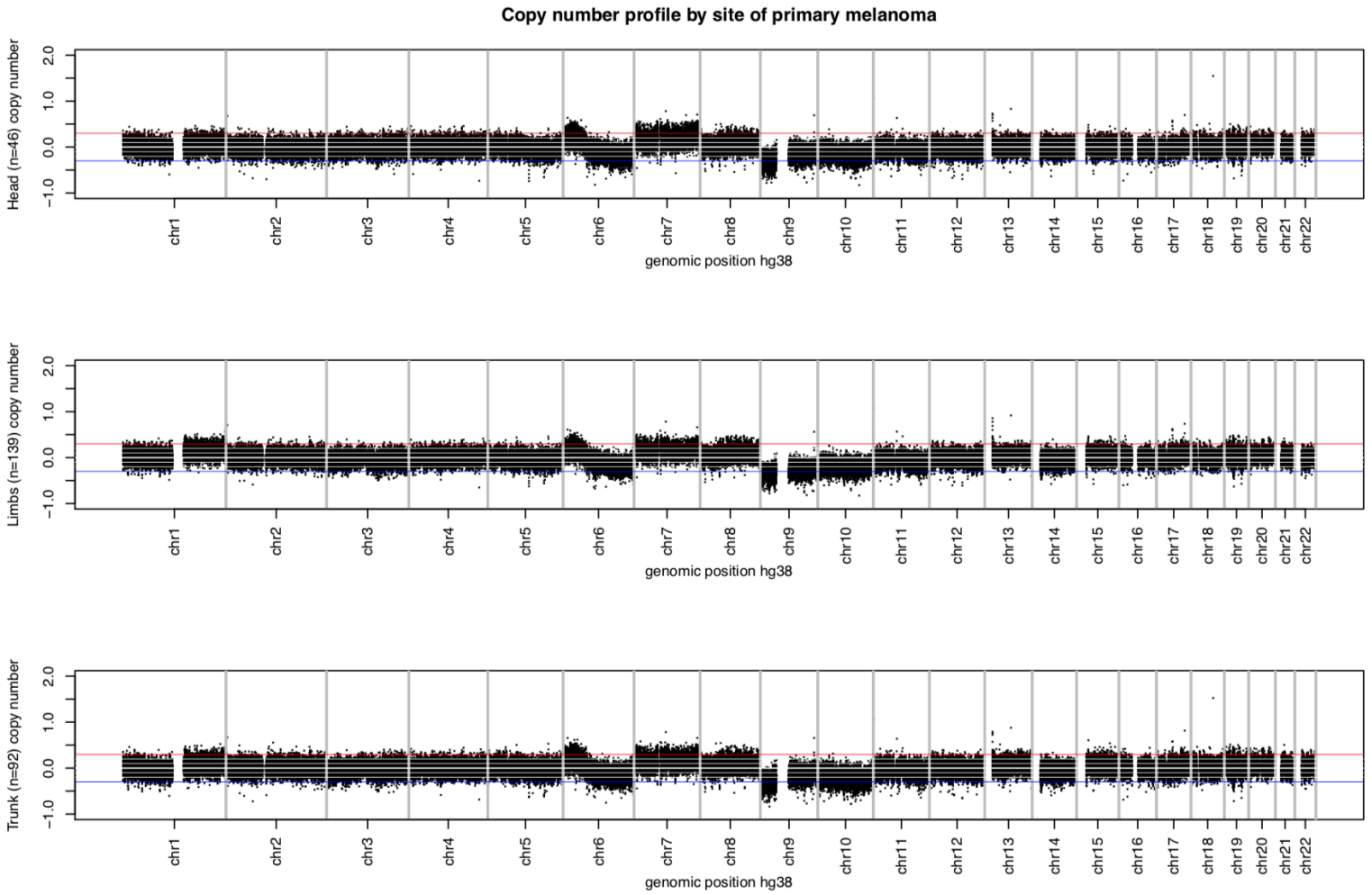


Figure 7.19. 10k Window copy number profile by site of primary melanoma

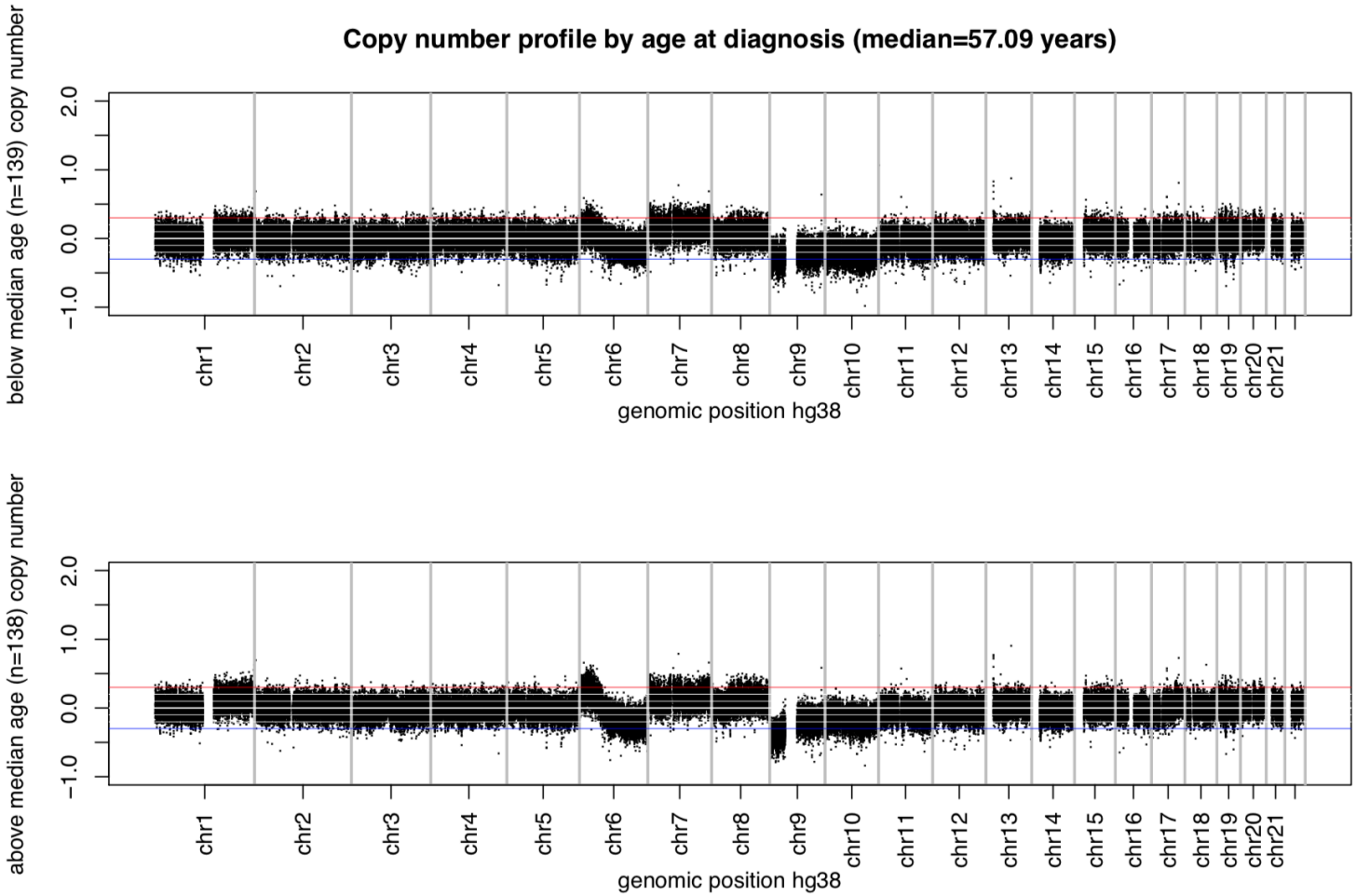


Figure 7.20. 10k Window copy number profile by age at diagnosis

Median age at diagnosis = 57.09 years

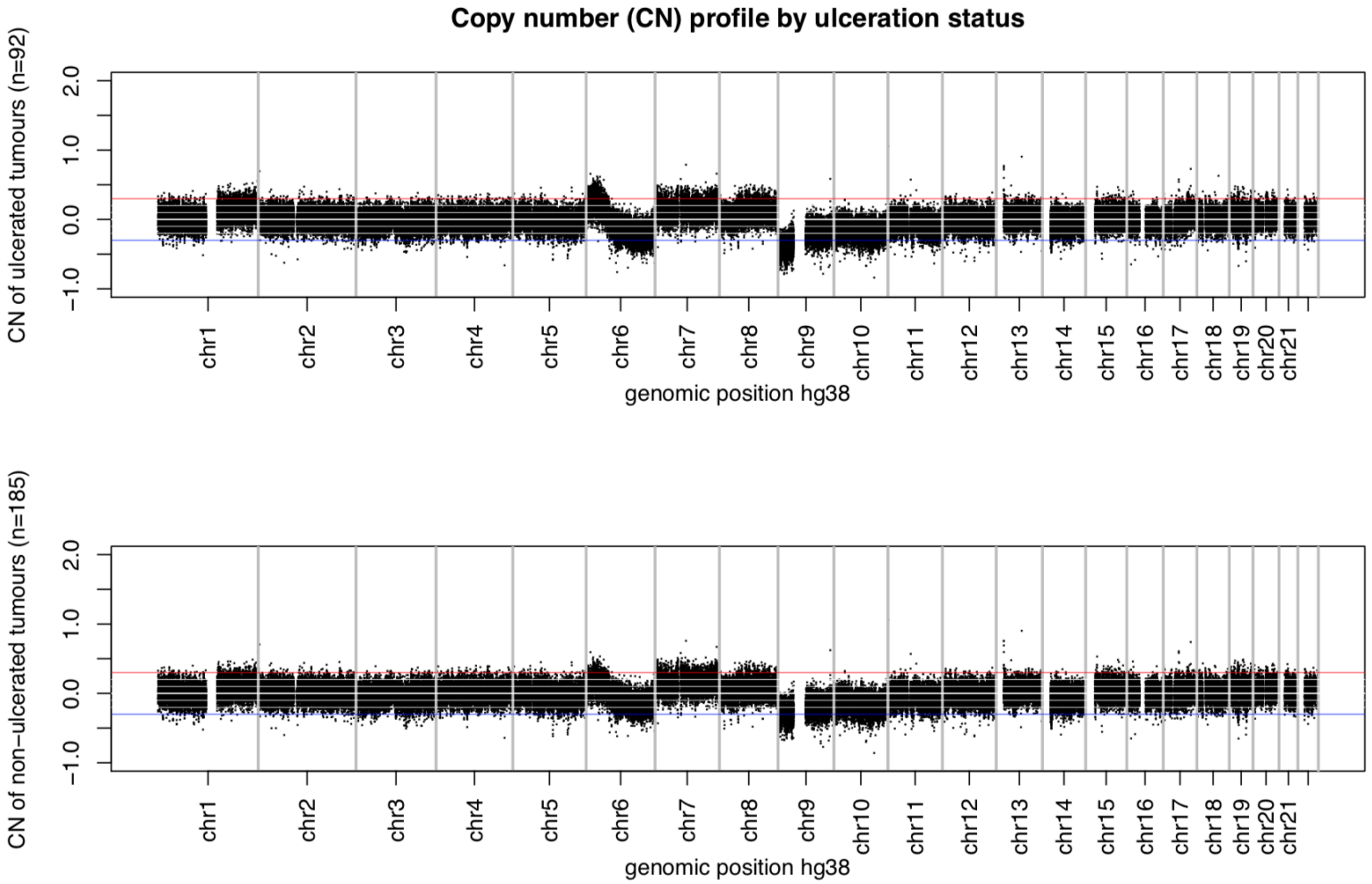


Figure 7.21. 10k Window copy number profile by ulceration status

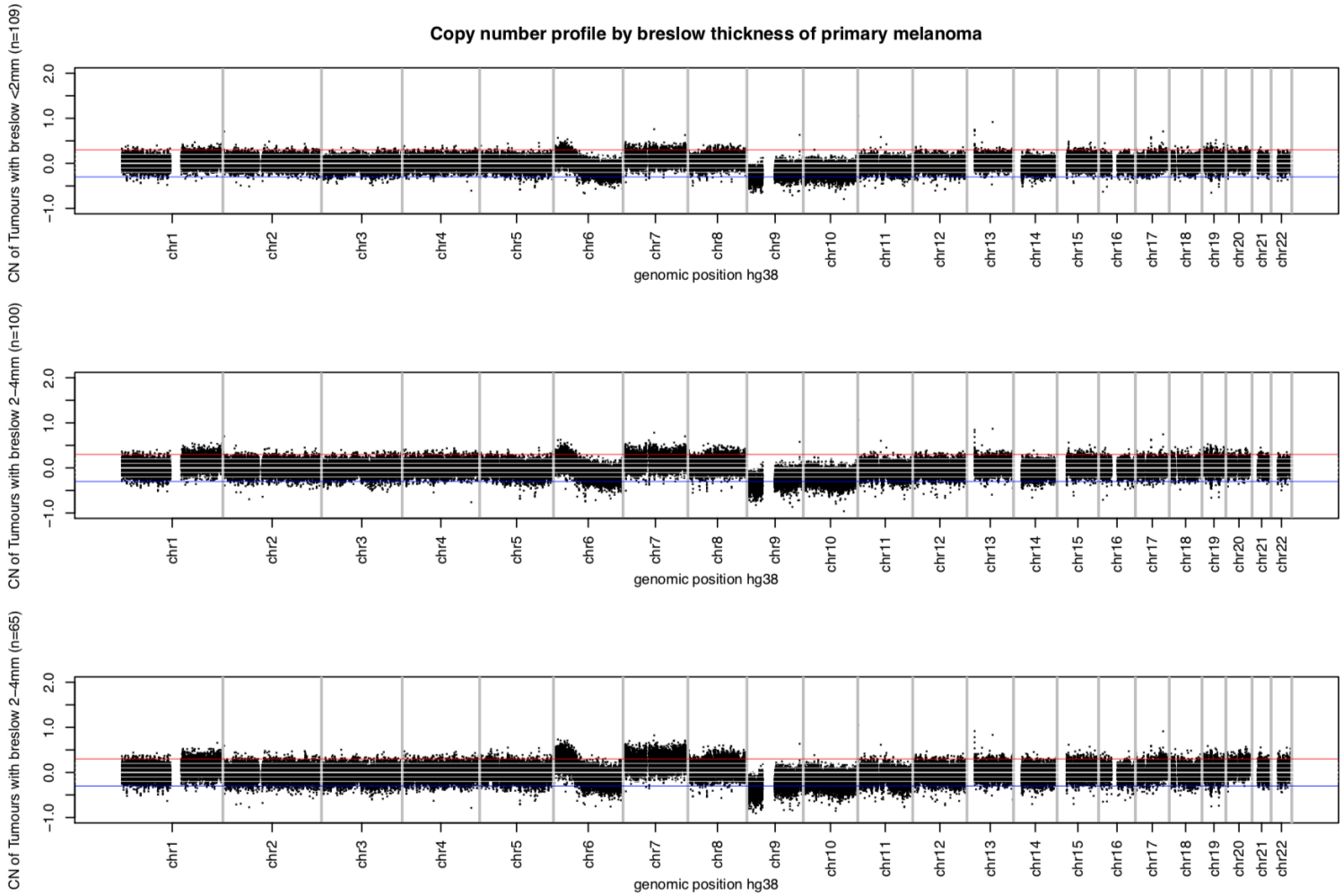


Figure 7.22. 10k Window copy number profile by Breslow thickness

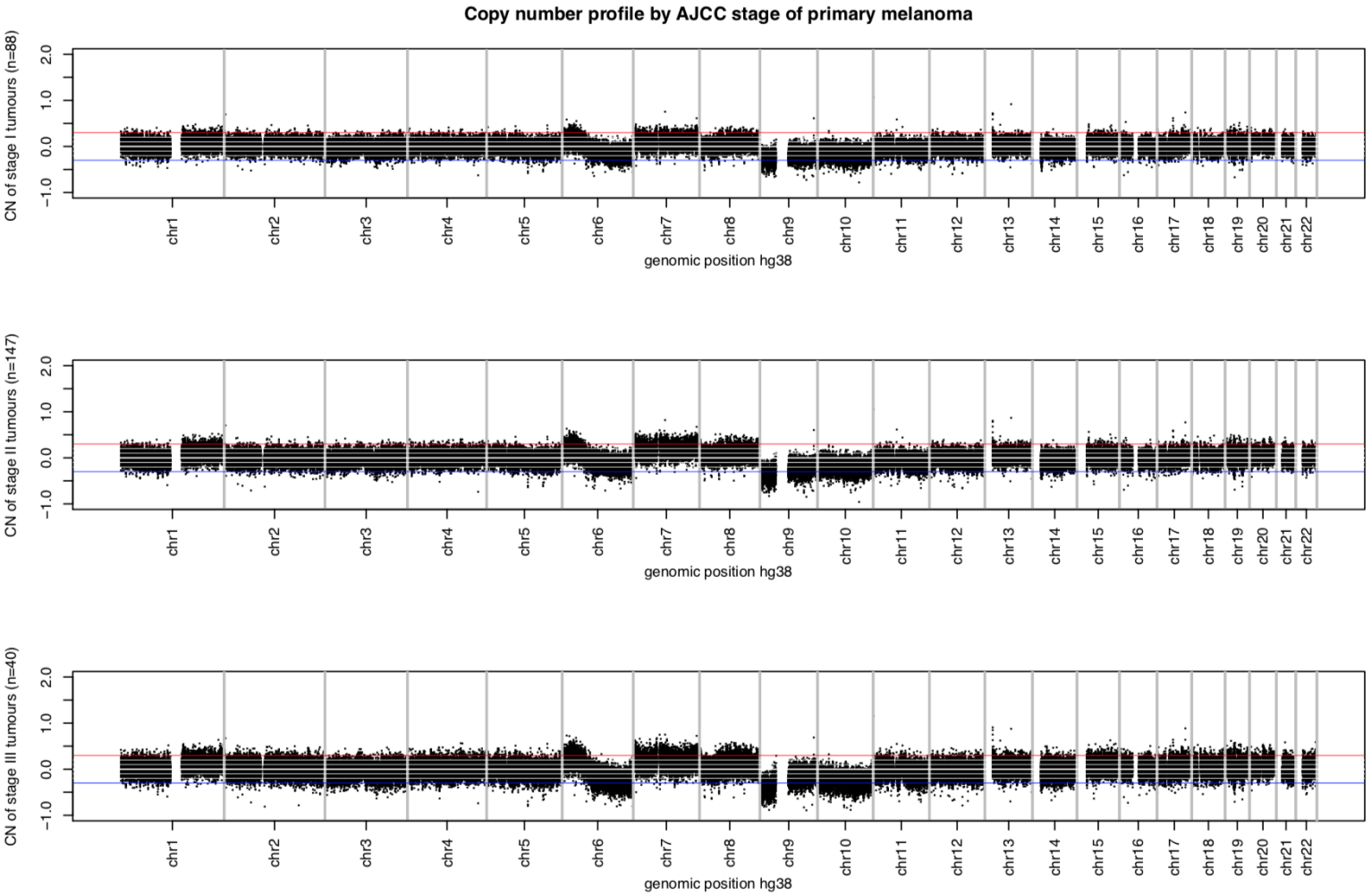


Figure 7.23. 10k Window copy number profile by AJCC stage

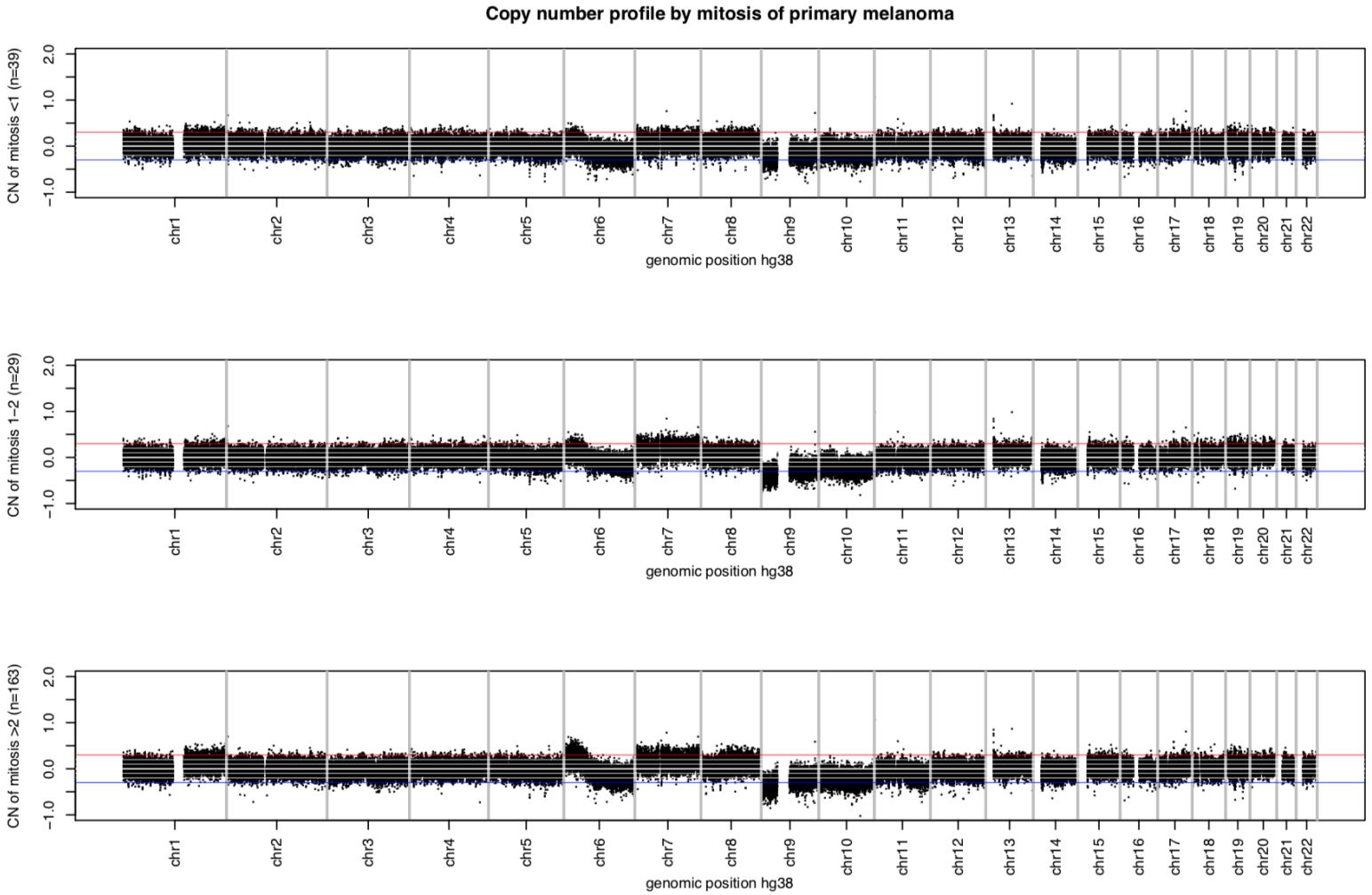


Figure 7.24. 10k Window copy number profile by mitosis

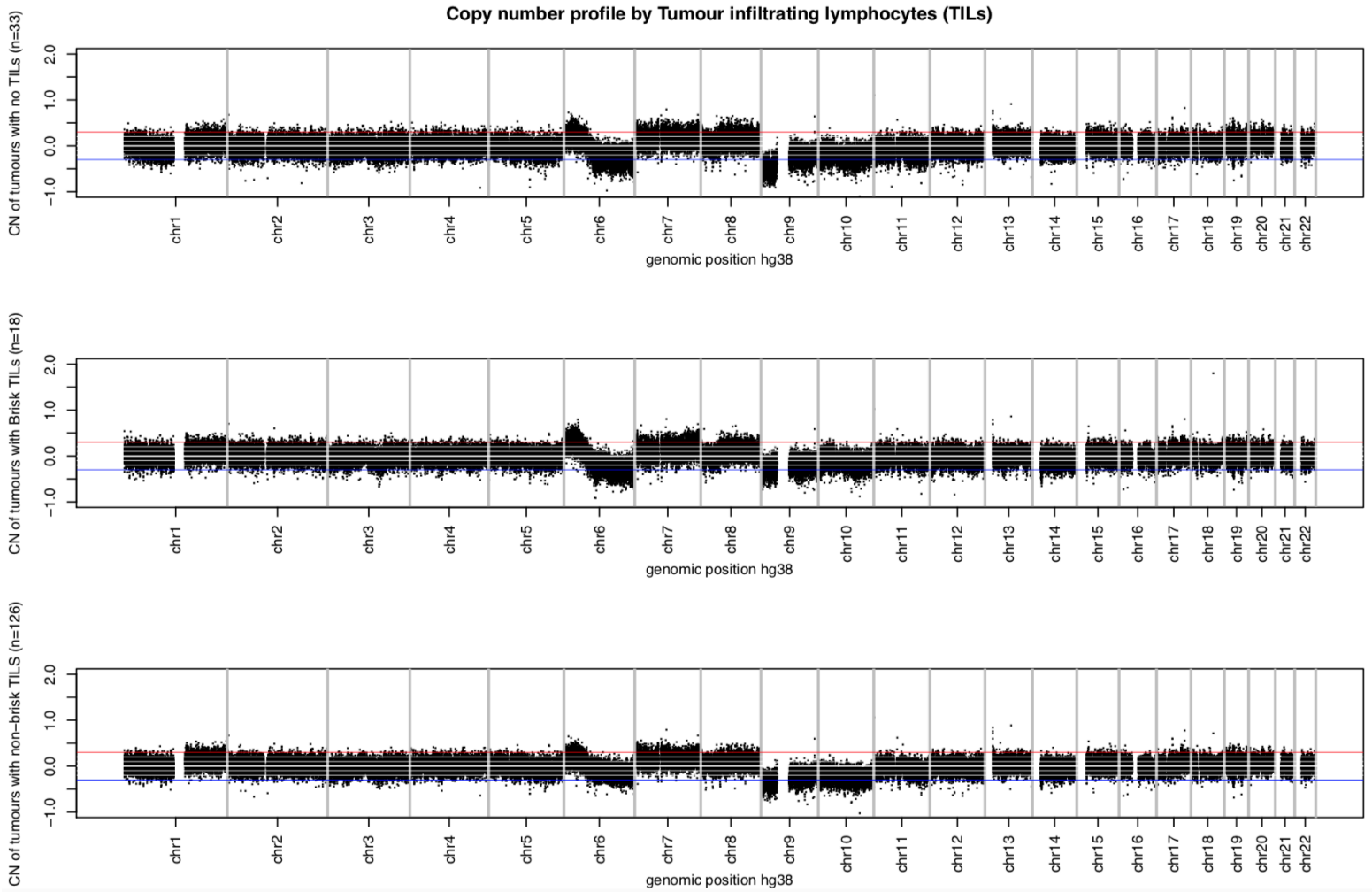


Figure 7.25. 10k Window copy number profile by tumour infiltrating lymphocytes (TILs)

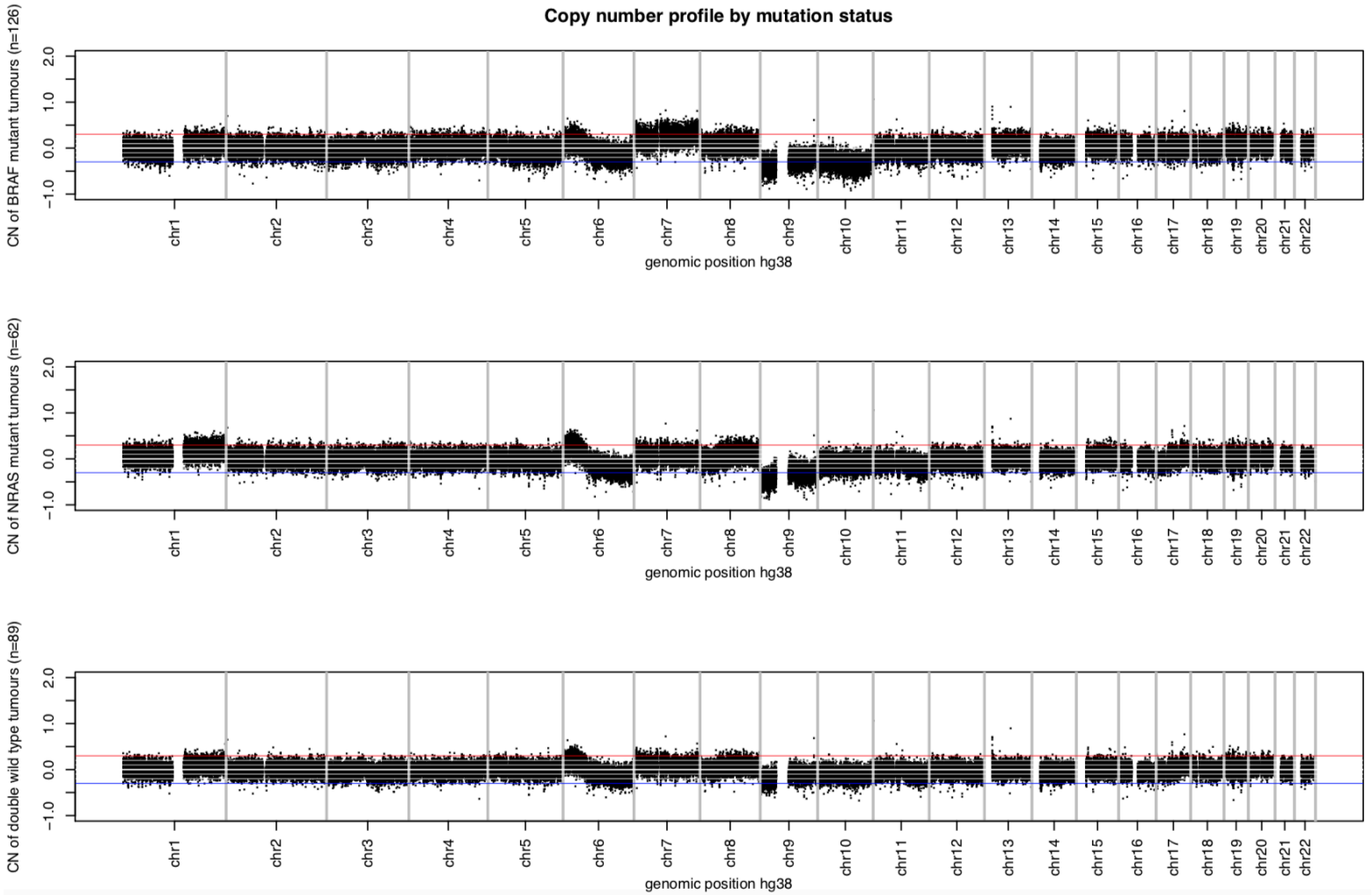


Figure 7.26. 10k Window copy number profile by mutation status

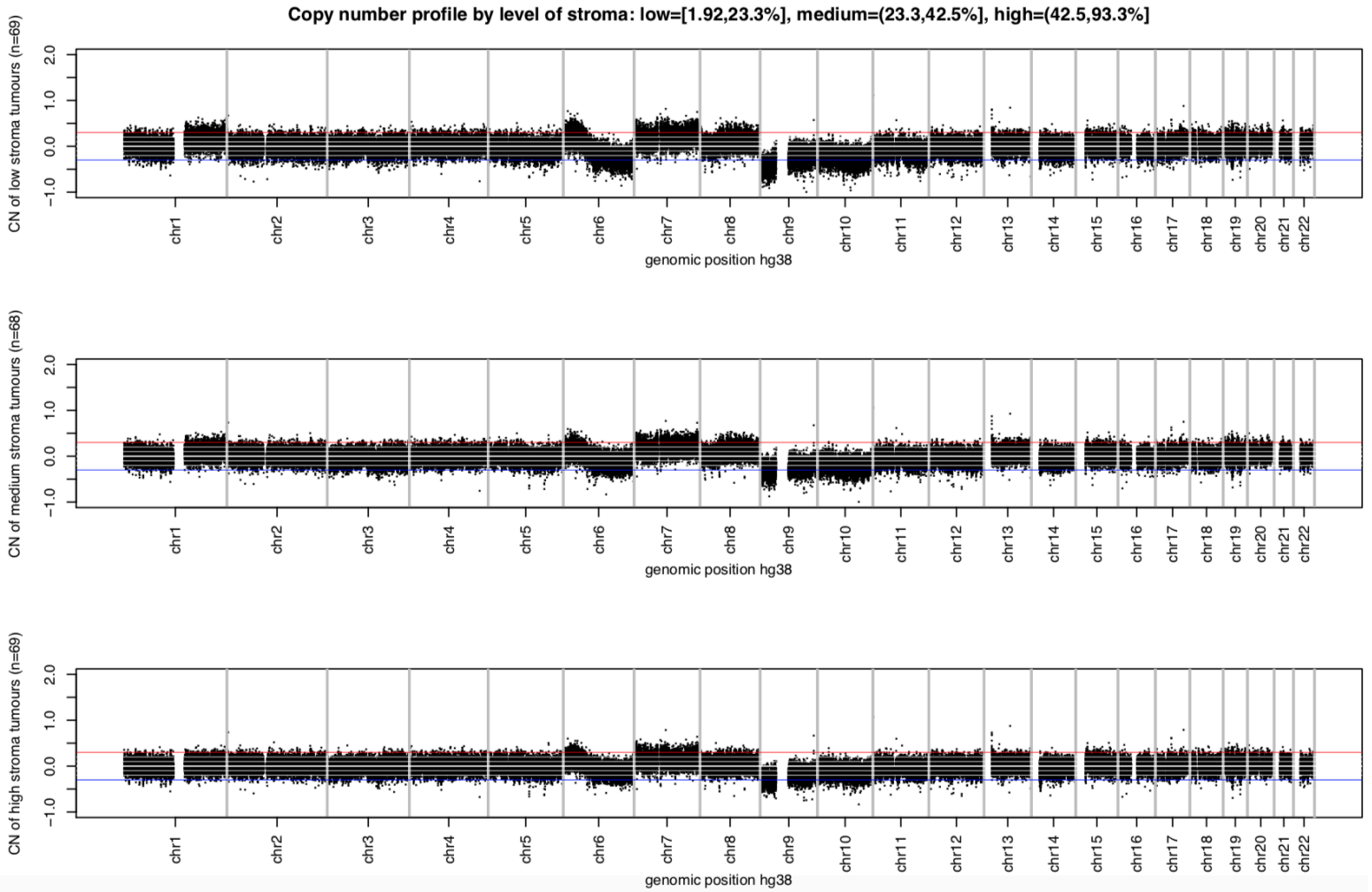


Figure 7.27. 10k Window copy number profile by level of stroma.

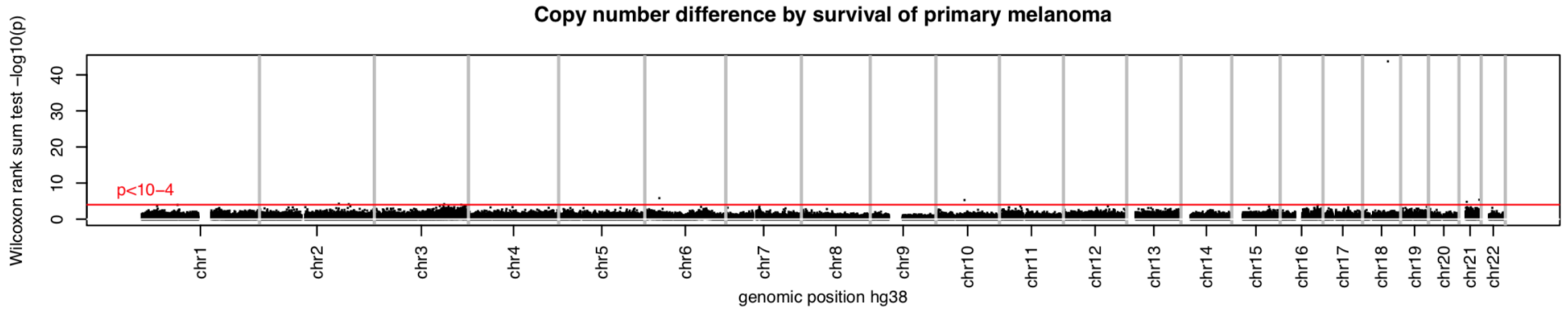


Figure 7.28. Manhattan plot for sex

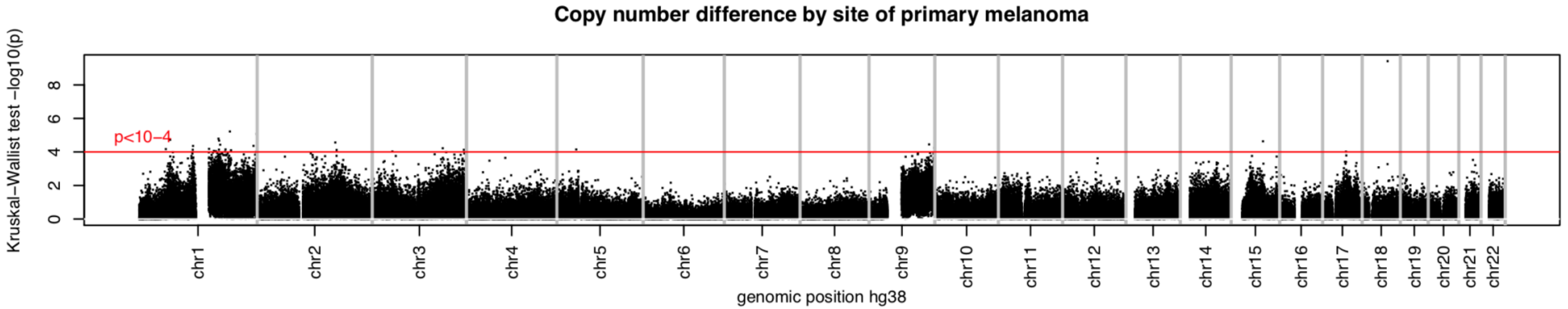


Figure 7.29. Manhattan plot for site

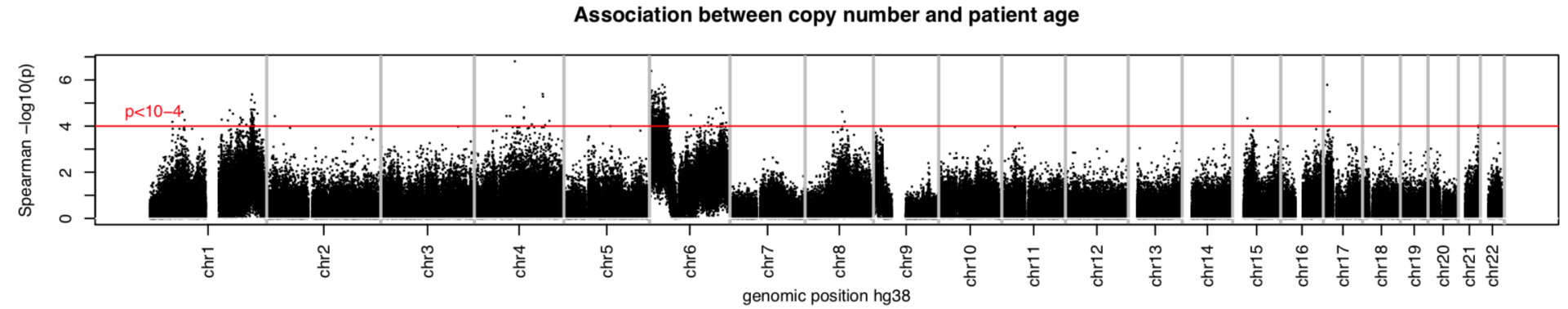


Figure 7.30. Manhattan plot for age

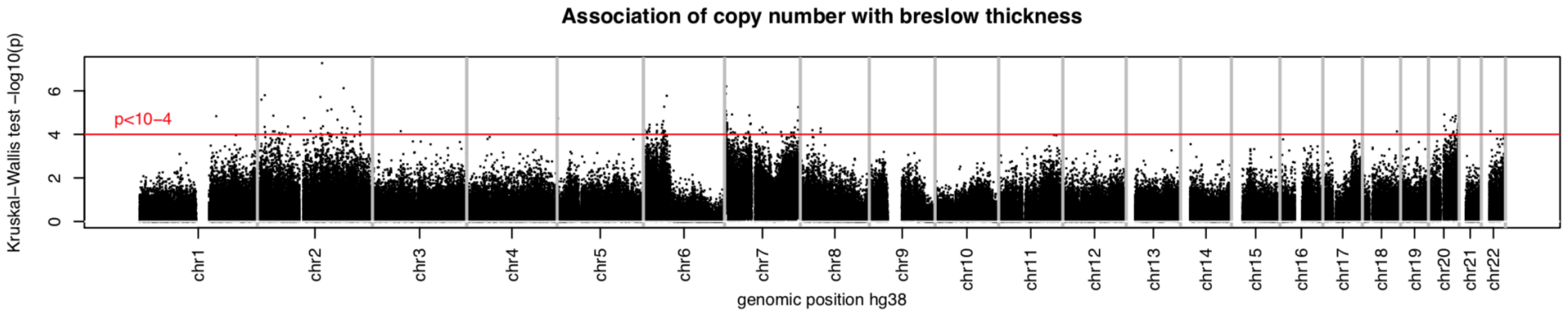


Figure 7.31. Manhattan plot for Breslow thickness

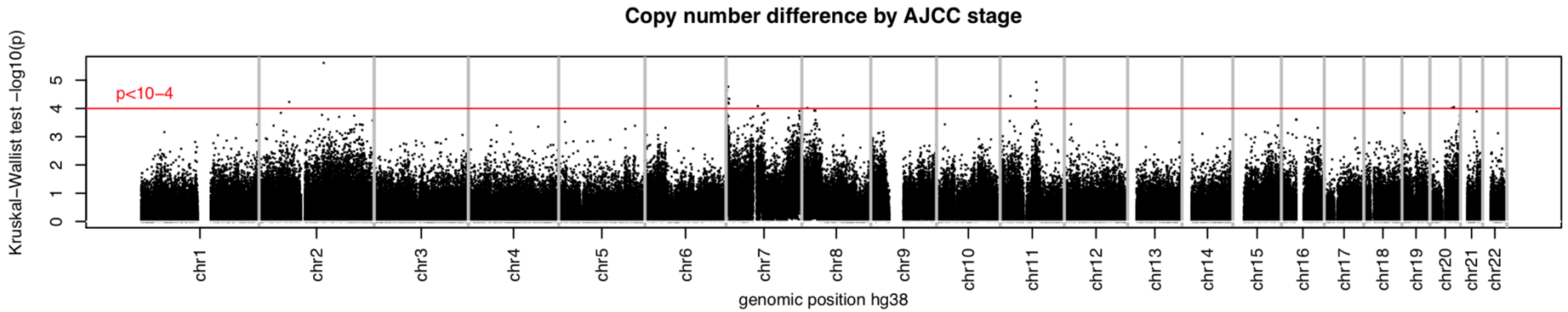


Figure 7.32. Manhattan plot for AJCC stage

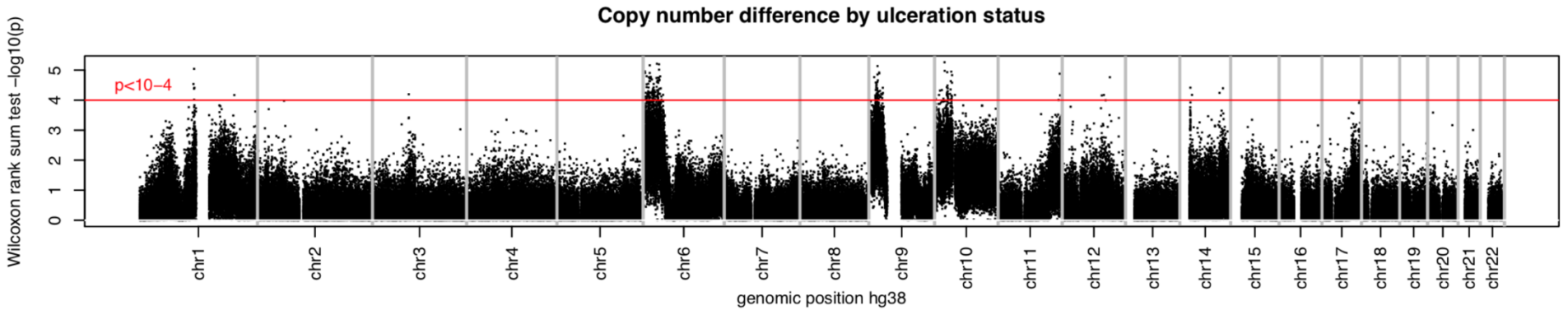


Figure 7.33. Manhattan plot for ulceration status

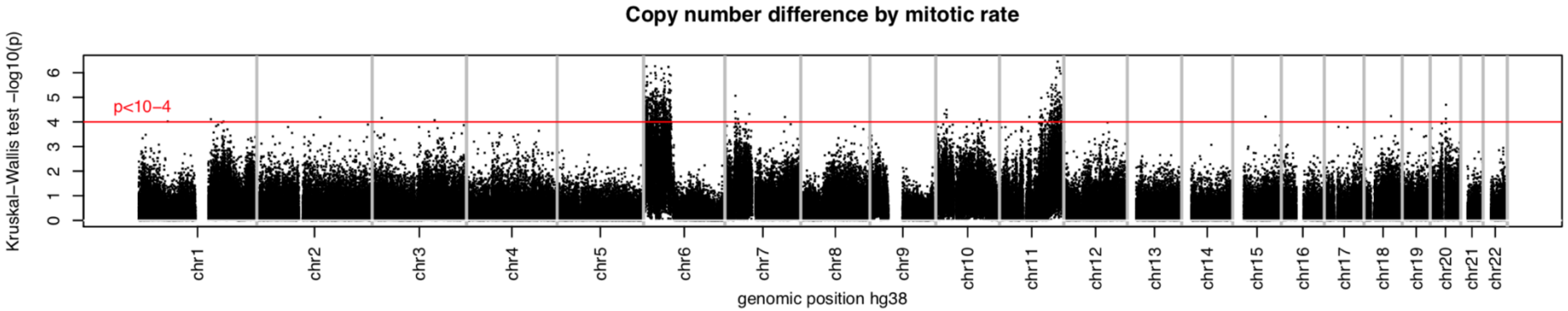


Figure 7.34. Manhattan plot for mitotic rate

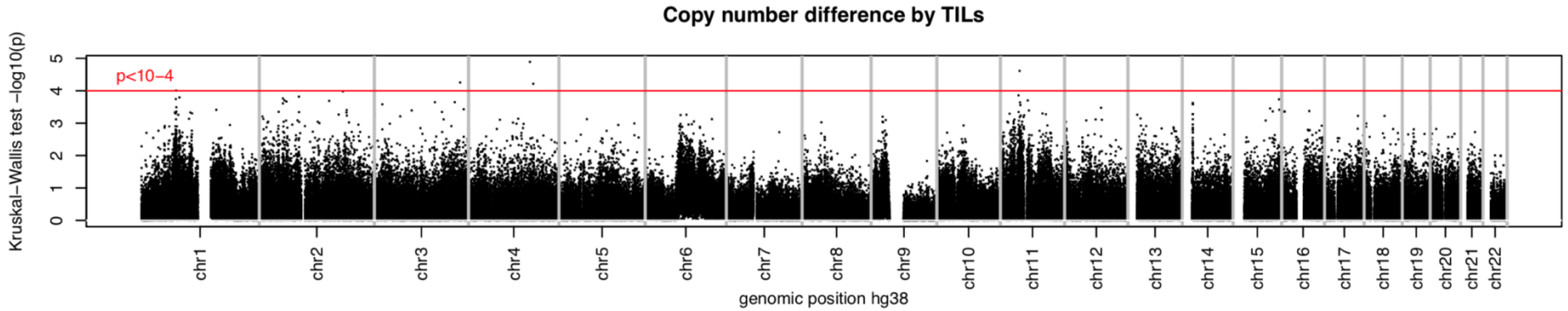


Figure 7.35. Manhattan plot for TILs

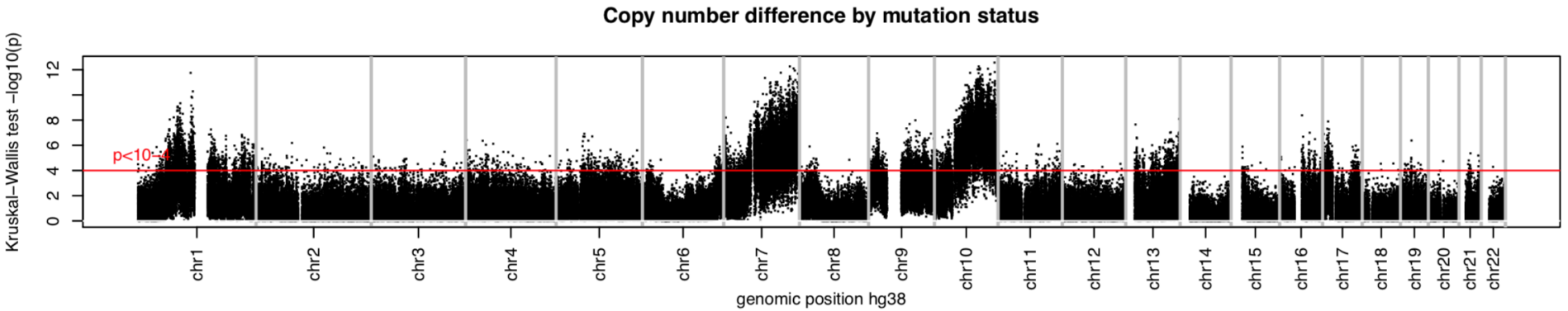


Figure 7.36. Manhattan plot for mutation status

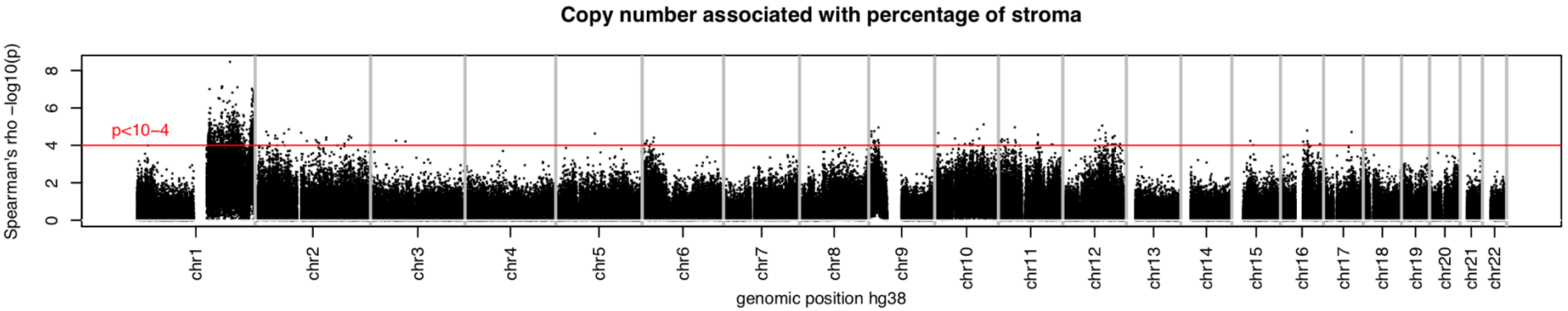


Figure 7.37. Manhattan plot for percentage of stroma

7.3.7 Association of 10K windows with Survival

Survival analysis of the 10k windows across the genome was performed in two ways. Firstly, using the quantitative copy number data which uses the adjusted read count on each window; and secondly, using the qualitative copy number which assigns a label to a copy number value as to normal [-0.10, 0.10], deletion (< -0.10), and amplification or gain (>0.10). The cutoff used is consistent with what was used by the TCGA study in assigning a qualitative value to a given quantitative copy number.

7.3.7.1 Survival analysis on quantitative copy number data

The results for whole autosomal genome window level copy number (continuous scale) survival analysis is presented in Figure 7.40. It can be observed that the region of chromosome 10 has the highest significance in terms of survival. This is closely followed by chromosome 13, chromosome 2, chromosome 18, chromosome 8, chromosome 9, chromosome 4, chromosome 6, and chromosome 3 which were previously reported to contain melanoma genes.

Table 7.22 displays the top 10k windows in terms of association with survival. The observed highest significance is 1.8 $x10^{-7}$ and is greater than the study-specific genome-wide significance threshold for this analysis which is 1.4 $x10^{-7}$ (~0.00000014438647). This window corresponds to *BMPR1A* which inhibit the tumorigenic potential of human brain tumour-initiating cells [291]. Based on the simulation results, the observed significance for this window translates to an actual study-specific significance of 0.05698925 which is close to the 5% level of significance. The Kaplan-Meier plot in Figure 7.38 shows survival curves for *BMPR1A* using the median cutoff. This indicates that above median copy number of this window corresponds to better survival (HR=0.46, logrank P=7.3 x 10^{-5}).

Table 7.22. Top 10k quantitative copy number windows that are significantly associated with survival

10k Window	P	HR*	Location	Gene/s	Note
10k000176179	1.80E-07	0.23	10q23.2	BMPR1A	inhibit the tumorigenic potential of human brain tumour-initiating cells
10k000170306	6.40E-07	0.24	10p12.1	MPP7	activates YAP1 (a transcriptional coactivator in the Hippo pathway), which in turn promoted autophagy in in Pancreatic Ductal Adenocarcinoma
10k000179025	9.69E-07	0.26	10q25.3	ATRNL1	enhances radiosensitivity of oral squamous cell carcinoma cells
10k000169476	1.59E-06	0.29	10p12.31	PLXDC2	identified as cell-surface receptors for Pigment Epithelium Derived Factor (PEDF). PEDF is known for promoting cell survival and proliferation, as well as its antiangiogenic, antitumor, and anti-metastatic properties
10k000177164	1.66E-06	0.23	10q24.1	PIK3AP1	essential to miR-567-mediated suppression of gastric cancer cell behaviour and oncogenic signalling
10k000175114	1.69E-06	0.26	10q22.2	LRMDA	mutation in this gene might lead to poor prognosis for pancreatic ductal adenocarcinoma patients
10k000211174	1.93E-06	4.88	13q13.2	LINC00457	associated with Thyroid Cancer
10k000167754	2.52E-06	0.26	10p15.3	LINC02645	Restricted expression toward testis
10k000175726	2.54E-06	0.25	10q23.1	NRG3	potential regulator of normal and malignant breast epithelial cells in vivo; associated with significantly longer overall survival in an ovarian cancer IV stage subgroup.
10k000176866	3.24E-06	0.32	10q23.33	LGI1	differentially expressed in early- and late-stage oral squamous cell carcinoma

**HR=Hazard ratio **Genes in red are nearest approximate when no gene maps to the window*

** adjusted for age, sex, and stage*

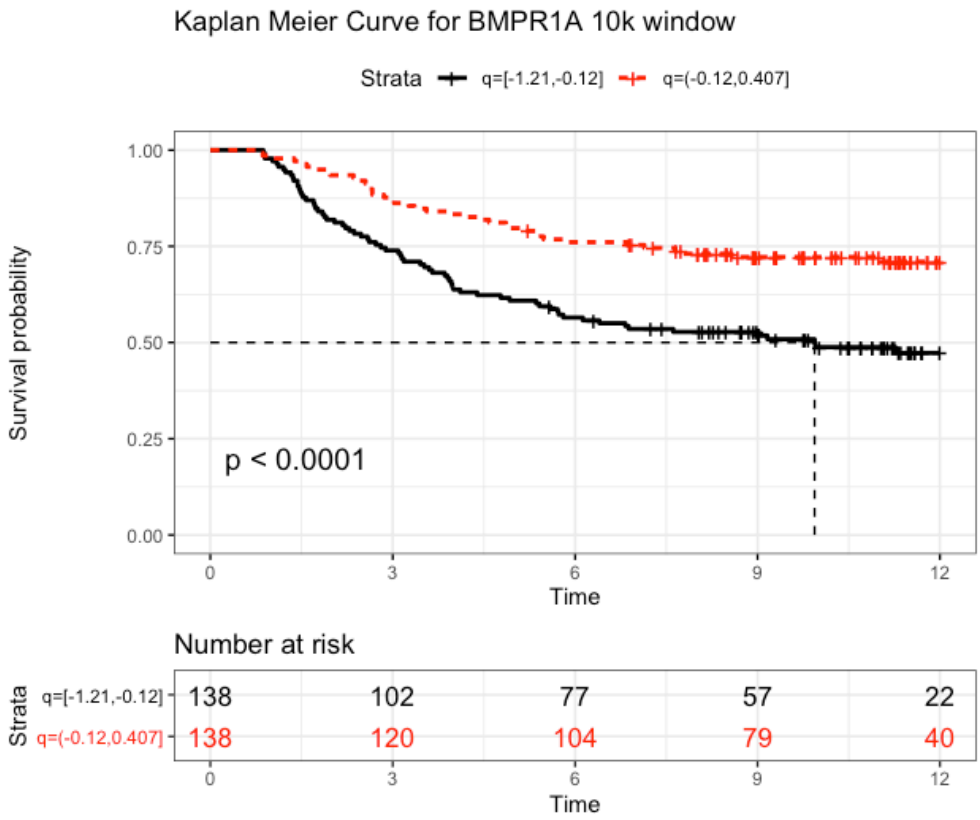


Figure 7.38. Kaplan-Meier curve for *BMPR1A* using median cutoff

Next on the list were *MPP7* which activates *YAP1* (a transcriptional coactivator in the Hippo pathway), which in turn promoted autophagy in in Pancreatic Ductal Adenocarcinoma [292, 293], *ATRNL1* which enhances the radiosensitivity of oral squamous cell carcinoma cells [294], *PLXDC2* which is identified as cell-surface receptors for Pigment Epithelium Derived Factor (*PEDF*) [295], *PIK3AP1* which is essential to miR-567-mediated suppression of gastric cancer cell behaviour and oncogenic signalling [296], *LRMDA* in which mutation in this gene might lead to poor prognosis for pancreatic ductal adenocarcinoma patients [297], *LINC00457* which is associated with Thyroid Cancer [298], *LINC02645* (restricted expression toward testis) [299], *NRG3* which is reported to be potential regulator of normal and malignant breast epithelial cells in vivo, and associated with significantly longer overall survival in an ovarian cancer stage IV subgroup [300, 301], and *LGI1* which is differentially expressed in early and late-stage oral squamous cell carcinoma [277].

7.3.7.2 Survival analysis on qualitative copy number data (cutoff=0.1)

Figure 7.41 (qualitative copy number) captures significant 10k windows that are mostly not captured in the analysis using the quantitative copy number data in Figure 7.40 using Cox proportional model which used the parametric method. While the

previous figure shows copy number windows that are associated with survival and primarily concentrated in chromosome 10q, analysis of qualitative data using logrank test shows regions that have top significance level when testing for association with survival from different chromosomes in the genome is located in chromosome 7 and reaches a conventional genome wide significance level (> 1 $x10^{-8}$), noting that permutation based study-specific genome wide significance threshold was not identified for this analysis.

Table 7.23 displays the top qualitative 10k windows when testing for associated withsurvival. The hazard ratios are estimated using Cox proportional hazard model while the significance was tested using logrank test. The top window in the list reaches a genome wide significance (P=6.5 $x10^{-9}$) and maps to *BRAF*. *BRAF* is an oncogene linked to melanoma and some carcinomas, and functions to upregulate the *RAS/RAR/MEK* pathway and a common target in melanoma therapy [302, 303]. Checking for mutation status as shown in Figure 7.39, there is no significant difference between the survival curves of *BRAF* mutated and non - *BRAF* mutated tumours (HR=1.01, logrank P=0.9). Other genes in the list are *LINC01339* which has restricted expression toward testis[304], *SDCCAG8* which is reported to be regulated by *SOX11* to promote head and neck cancer progression[305]; *SNRPD1* which for mutation of this gene is associated with response of autologous T cells to a human melanoma [306], *MPP7* (also appeared in the analysis of quantitative copy number data in Table 7.22) which activates *YAP1* (a transcriptional coactivator in the Hippo pathway), which in turn promoted autophagy in in Pancreatic Ductal Adenocarcinoma [292, 293]; *RBMS3* where its low expression is associated with poor prognosis in patients with gastric cancer [307]; *ARL2BPP4* (ADP ribosylation factor like GTPase 2 binding protein pseudogene 4) was identified as novel loci for non-high-density lipoprotein cholesterol and its postprandial lipemic response [308]; and *KRT19P6* (Keratin 19 Pseudogene 6).

Table 7.23. Top 10k qualitative copy number windows that are significantly associated with survival

HR compares normal with deletion, and gain with deletion.

10k Window	HR*(normal)	P	HR*(gain)	P	logrank.P	Location	Gene	Note
10k000137277	0.26	1.21E-03	0.26	5.99E-05	6.50E-09	7q34	BRAF	a common target for melanoma therapy
10k000096993	0.89	6.34E-01	2.85	1.85E-05	1.42E-08	5q14.3	LINC01339	Restricted expression toward testis
10k000024336	0.24	9.20E-05	0.25	1.07E-05	1.26E-07	1q43	SDCCAG8	regulation of SDCCAG8 by Sox11 promotes head and neck cancer progression
10k000259574	0.43	4.09E-02	0.18	3.70E-06	1.57E-07	18q11.2	SNRPD1	mutation of this gene is associated with response of autologous T cells to a human melanoma
10k000170325	0.31	6.52E-06	0.40	1.54E-04	3.04E-07	10p12.1	MPP7	activates YAP1 (a transcriptional coactivator in the Hippo pathway), which in turn promoted autophagy in in Pancreatic Ductal Adenocarcinoma
10k000052065	0.21	3.25E-07	0.50	9.77E-04	6.59E-07	3p24.1	RBMS3	Low expression ois associated with poor prognosis in patients with gastric cancer
10k000101035	2.59	1.17E-05	1.38	2.60E-01	6.62E-07	5q23.3	ARL2BPP4	ADP ribosylation factor like GTPase 2 binding protein pseudogene 4
10k000165382	0.63	1.03E-01	2.10	7.06E-04	8.12E-07	9q33.1	ASTN2	associated with breast cancer
10k000044301	0.73	2.08E-01	2.60	2.31E-04	8.73E-07	2q32.3	GLULP6	identified as novel loci for non-high-density lipoprotein cholesterol and its postprandial lipemic response
10k000078128	0.34	2.78E-05	0.53	5.69E-03	1.00E-06	4q22.1	KRT19P6	Keratin 19 Pseudogene 6

**HR=Hazard ratio* ***Genes in red are nearest approximate when no gene maps to the window*

** adjusted for age, sex, and stage*

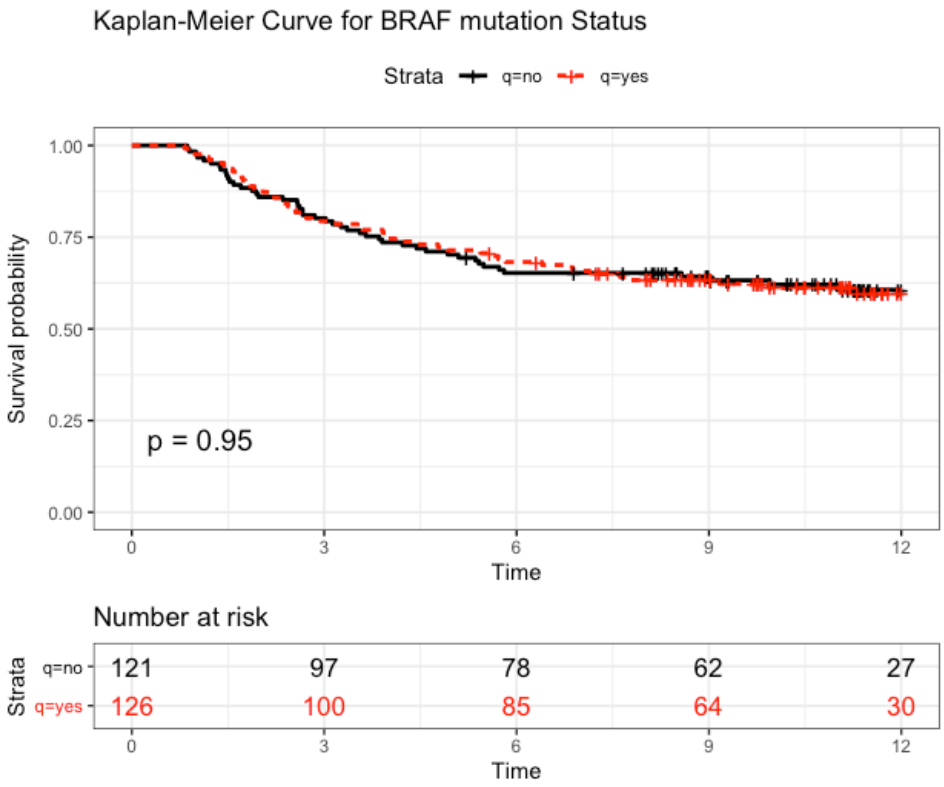


Figure 7.39. Kaplan-Meier curve for *BRAF* mutation status

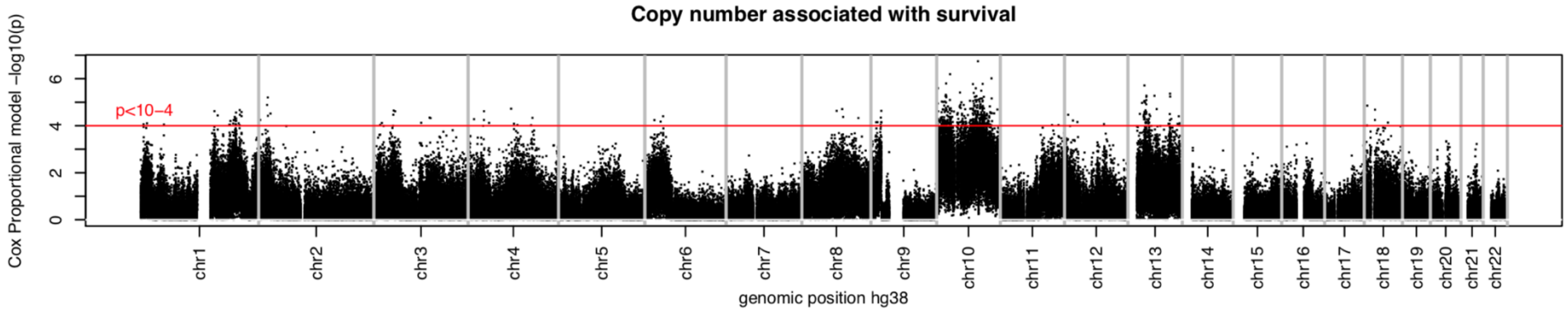


Figure 7.40. Manhattan plot for survival analysis

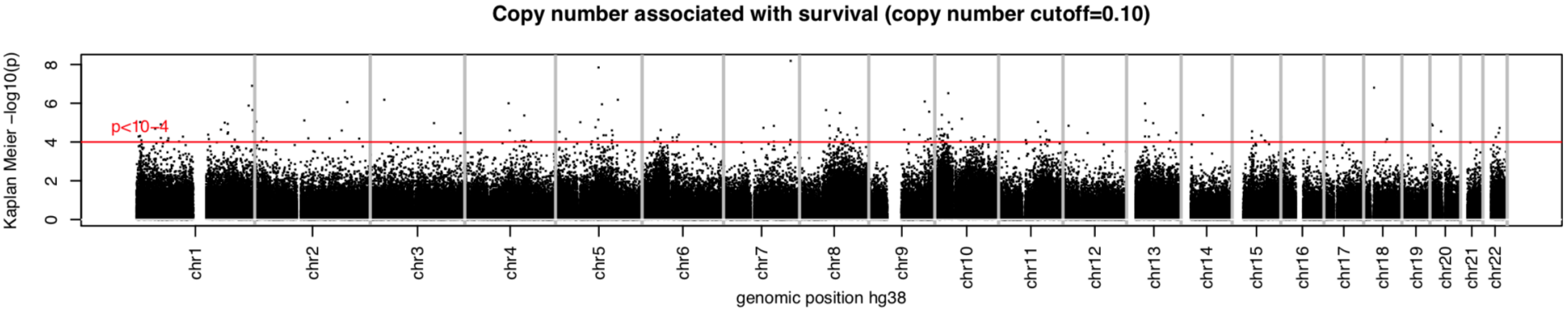


Figure 7.41. Manhattan plot for survival analysis (cutoff= 0.10)

7.3.8 Identification and analysis based on FAM190A window

I previously conducted an analysis that includes all the 303 samples (i.e. includes mucosal and non-cutaneous lesions) and without limiting the survival time to 12 years. The summary of the top significant 10k window is shown in Table 7.24 below. The most significant window maps to *FAM190A* (or *CCSER1*) which will be the focus of this section. This window did not crop up in the analysis of 277 samples because the 26 samples excluded had both poor prognosis and were commonly deleted for that *FAM190A* 10k window. Deficiency of *FAM190A* gene was reported to create a cell division effect [309] and was demonstrated in vivo to have oncogenic properties by in-frame deletions within the region of this gene. There was also an indication that these transcript variants are potential therapeutic targets in patients with cancer [310]. Further investigation on this window is shown in Figure 7.42 below. The upper portion of the figure (Figure A) shows that samples with deletion (n=37, cutoff= -0.10) on this window tend to have higher overall copy number aberration load as measured using mean weighted segment mean (MWSM) compared with samples having normal copy number in this window (Fold change=1.23, P=0.002). The lower part of the figure (Figure B) plots the survival curve comparison between the two samples groups. Samples with deletion (n=37) on the window being investigated tend to have poorer survival when compared with samples with normal copy number in this window (HR=0.40, P=3 x10^{-10}).

Table 7.24. Top 10k qualitative copy number windows that are significantly associated with survival using the 303 samples

10k Window	HR	P	Location	Gene
10k000077981	0.08	2.10E-07	4q22.1	CCSER1 / FAM190A
10k000021294	2.90	3.15E-07	1 q32.3	VASH2
10k000211585	6.31	3.37E-07	13q13.3	FREM2
10k000022031	2.82	5.01E-07	1q41	XRCC6P3
10k000176179	0.09	6.06E-07	10q23.2	BMPR1A
10k000022022	2.79	6.97E-07	1q41	RAB3GAP2
10k000022024	2.42	7.75E-07	1q41	RAB3GAP2
10k000211127	4.95	8.59E-07	13q13.2	VDAC1P12
10k000021311	2.11	8.72E-07	1q32.3	RPS6KC1
10k000211338	4.94	9.16E-07	13q13.3	CCDC169-SOHLH2
10k000211338	4.94	9.16E-07	13q13.3	CCDC169
10k000211376	5.09	9.28E-07	13q13.3	C13orf36/SERTM1

**HR=Hazard ratio* ***Genes in red are nearest approximate when no gene maps to the window*

* adjusted for age, sex, and stage

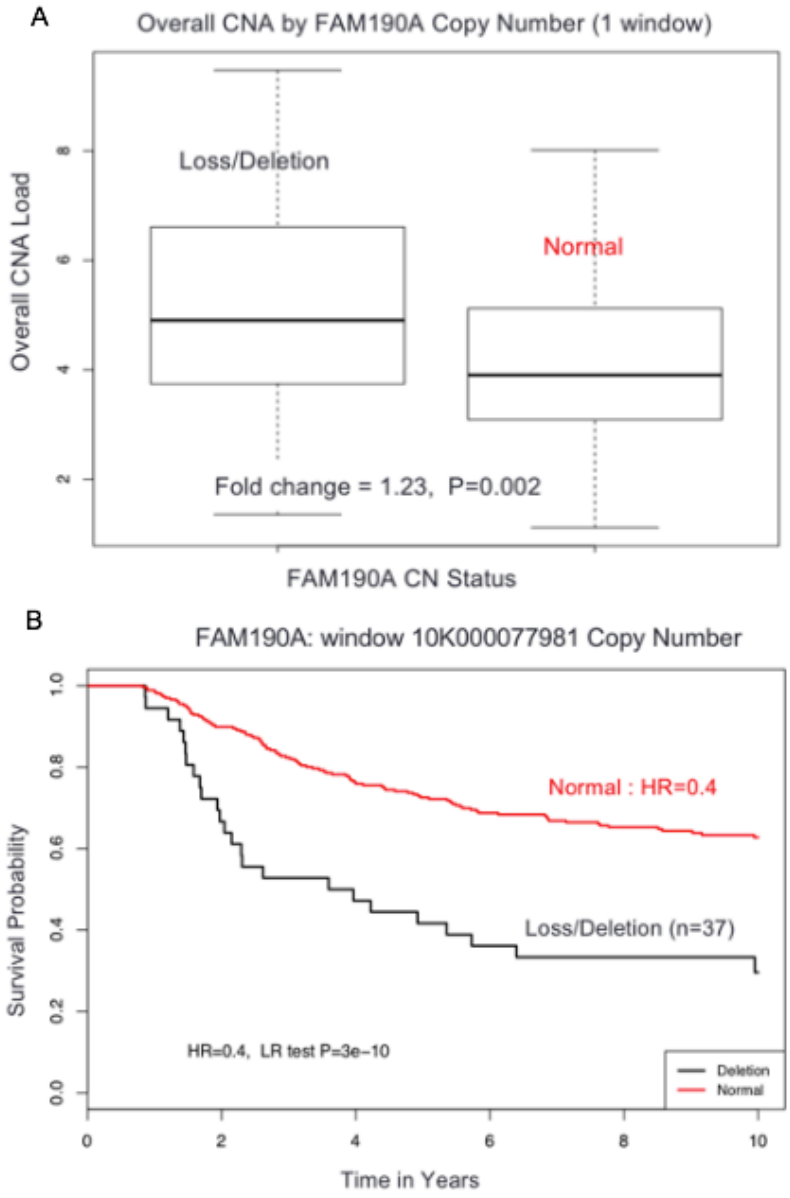


Figure 7.42. FAM190A deletion in LMC.

A, compares the distribution of 37 samples with deletion on *FAM190A* (window 10K000077981) and the rest of the LMC samples having normal copy number on this window in terms of the estimated overall CNA load using Mean Weighted Segment Mean (MWSM). B, shows the survival curves of samples with *FAM190A* (window 10K000077981) deletion (black curve) compared with the rest of the samples having normal copy number (red curve) in this window.

In Figure 7.43, the whole genome profile of LMC samples categorised as having deletion or not in the window under investigation is plotted. A more conservative cutoff of -0.30 was used this time to clearly capture the comparison between the two sample groups trimming the samples with deletion to 19. Obvious difference between the copy number profiles of the two sample groups is observed featuring more aberrations for the group with deletion while samples with no deletion can be generally seen as normal except in the regions chromosome 6, 7, 8, 9, and 10.

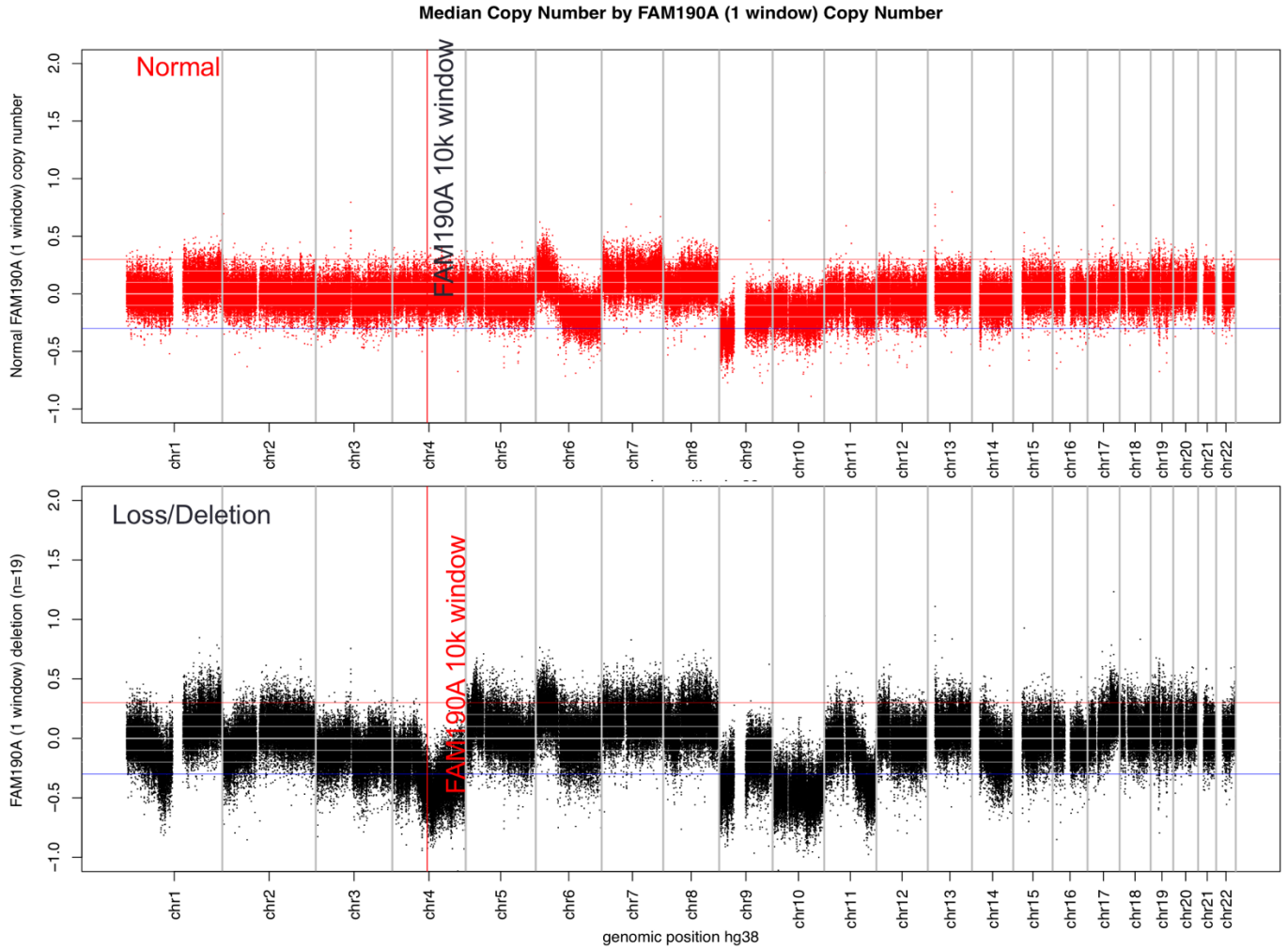


Figure 7.43. Whole genome profile by FAM190A (window 10K000077981) deletion in LMC

7.4 Discussion

This chapter tested for association of copy number with clinical and tumour characteristics and survival. This is done in two ways: (1) By using three different measures of overall CNA load as measures of genomic instability such as (FGA: fraction of genome altered, MWSM: mean weighted segment, and AS: aneuploidy score) and (2) by using 10k window level copy number data. The three measures of overall CNA load were assessed in terms of distribution and test for association with tumour and clinical characteristics and survival. While these all correlate with most of the tumour and clinical characteristics including survival, MWSM shows higher level of association as compared with FGA and AS indicating that MWSM retains most information and thus is more likely to produce more reliable results.

Checking for the association of MWSM with the patient tumour and clinical characteristics, sex, site, tumour infiltrating lymphocytes (TILs), and mutation status did not show statistically significant results. A follow up is recommended to verify these results. In terms of sex, the study of Lopes-Ramos et al. (2020) discussed the genome-wide sex and gender differences in cancer which may be attributed to a combination of environmental, genetic, and epigenetic factors, as well as differences in gene regulation, and expression [311]. The study of Li et al. (2018) provides a comprehensive catalogue of sex differences in somatic alterations which includes in cancer driver genes influencing prognostic biomarkers that predict patient outcome after therapy [312].

I went forward with analysing the 10k window data recognising that a focal aberration driving progression would be of major interest to melanoma biologists. One obvious challenge is identifying the threshold for significance when testing for associations across the whole genome since neighbouring windows are correlated as well as there is an inherent multiple testing problem as I analysed the 248,736 10k windows. Permutation tests were conducted to provide a guide in identifying significance threshold. Because this experiment requires significant computational time and memory (5-9 hours per iteration), this was only done for selected clinical characteristics such as site of tumour, Breslow thickness, and survival.

7.4.1 Overall CNA load

Aneuploidy score is not associated with survival when categorised into two levels by median cutoff. It is significant in the univariable analysis when using the continuous data, but loses significance when adjusting for age, sex, and stage (multivariable analysis). FGA is associated with survival when categorised into two levels by median

cutoff. It is significant in the univariable analysis when using the continuous data, but unlike Aneuploidy score, it remained significantly associated with survival after adjusting for age, sex, and stage.

Of the three measured of overall CNA load, MWSM provides the highest significance when testing for association with survival using the data categorised into two levels by median cutoff. It is also significantly associated with survival in the univariable analysis and even after adjusting for age, sex, and stage. The significance of MWSM in univariable analysis is higher than that of FGA but slightly lower (P=0.001 for FGA versus P=0.002 for MWSM, using Cox proportional hazard model) in the multivariable analysis. In terms of effect size, though FGA is more statistically significant, MWSW has higher effect size (HR = 1.23 for MWSG versus HR = 1.02 for FGA) thus capturing more information in terms of representing the overall aberration in the genome. While both FGA and MWSM both provide additional information over and above the standard clinical variables (i.e. age, sex, stage), MWSM provides better distribution of estimate for overall genomic aberration and is therefore easier to use for potential clinical decisioning.

7.4.2 Window level analysis for clinical characteristics

In terms of clinical features, sex is most significantly associated with a 10k window that maps to a genes (*LINC01919* [200] and *LINC01917* [199]) that is broadly expressed in testis. This is the same window that is most associated with site of tumour and is study-specific genome-wide significant. Further investigation showed that sex and site of tumour is significantly associated in the sample (n=277) of this study as indeed within the population. While some literatures mention about the sex differences in cancer and is more likely to drive the observed association than site, this results still require follow up [311, 312]. In terms of age, the most significantly associated window maps to *MAPK10* which is involved in cell proliferation, differentiation, transcription, regulation and development[220] while it is *RNU7-65P* (eukaryotic translation elongation factor 1 alpha 1 pseudogene 42) for Breslow thickness. Analysis of melanoma stage reveals that the window mapping to *CXCR4* is the most associated to it. This gene has multiple essential functions include homing of stem cells and metastasis of cancer cells; trafficking and homeostasis of immune cells such as T lymphocytes and plays important role in cancer progression[229, 230]. For ulceration, the window mapping to *NEBL* was shown to have the highest level of significance. This is enriched in the heart muscle tissue and reported to be a prognostic marker in renal cancer (favourable) and urothelial cancer (unfavourable) [252]. The window with highest level of association to mitotic rate maps to *MIR100HG* which was reported to

promote colorectal cancer metastasis and is associated with poor prognosis [256]. For tumour infiltrating lymphocytes (TILs), the top significantly associated windows map to *ZSWIM5P3* (Zinc Finger SWIM-Type Containing 5 Pseudogene 3) and *LRRC4C*. *LRRC4C* is a specific binding partner for netrin G1[270]. High level of associations of window level copy number are found when correlated with mutation status. The top window on the list maps to *ADAM12* which contributes to increased tumour proliferation, metastasis and endocrine resistance [273]. For percentage of stroma, the top on the list is *KCNT2* in which its expression was correlated with the prognosis of skin cutaneous melanoma (SCM) and significantly different between normal skin and SCM.

While several recent studies have mentioned the association of these clinical characteristics with melanoma and other clinical characteristics such as that of Li et.al (2018) and Ramos-Lopes et al. (2020) for sex differences, Enninga et.al (2017) for age, sex, and stage, and Jewel et al (2015) for ulceration , age, site of tumour, survival, sex, tumour infiltrating lymphocytes (TILs), mitotic rate, and Breslow thickness , these results on copy number association still require follow up [58, 311, 313].

7.4.3 Window level analysis for survival

Each valid 10k window of the LMC genome was then tested for association with survival. This was done in two ways: method 1. Using quantitative copy number data, and method 2. Using the qualitative copy number data (deletion, normal, amplification). The top hit for the first method is *BMPR1A* which inhibits the tumorigenic potential of human brain tumour-initiating cells [291] while it is *BRAF* for method 2. It is known that *BRAF* mutation is a common target for melanoma therapy [302, 303]. While the lists from the results of these two methods differ, *MPP7* is consistent on both. This activates YAP1 (a transcriptional coactivator in the Hippo pathway), which in turn promoted autophagy in in Pancreatic Ductal Adenocarcinoma [292, 293].

7.4.4 Analysis of FAM190A

Previous analysis using the 303 (277 cutaneous melanoma samples plus 26 samples from other sites) samples identified a window mapping to *FAM190A* (or *CCSER1*) gene as the most associated with survival. It was previously reported that deficiency of this gene was reported to create a cell division effect and was demonstrated in vivo to have oncogenic properties by in-frame deletions within the region of this gene [309]. There was also an indication that these transcript variants are potential therapeutic targets in patients with cancer [310]. Deletion in this window is associated with poorer survival and high CNA load as measured by mean weighted

segment mean (MWSM). A further graphical analysis using a stricter cutoff to define deletion i.e. -0.30 instead of -0.10 grouping the samples into whether having the deletion in this window or not shows difference in genomic profiles. Sample groups with deletion in this window shows generally more aberration compared with that of the group without the deletion.

This window did not crop up in the analysis of 277 samples because the 26 samples excluded had both poor prognosis and were commonly deleted for that *FAM190A* 10k window. Because of the limited size, I could not do more detailed analysis of these samples.

Chapter 8
Discussion and Conclusion

8.1 Summary of the aims of this study

This study was conducted to identify copy number alterations/aberrations (CNAs) in primary melanoma using NGS data derived from the formalin-fixed paraffin embedded (FFPE) primary tumour samples taken from participants in the Leeds Melanoma Cohort (LMC) and to test for associations with patient tumour and clinical characteristics including survival. This was divided into two main aims: the first one is providing data quality control measures and performing additional steps to increase data quality (Chapters 4, 5, and 6), and the second one is to test for association of the copy number data with patient clinical characteristics including survival (Chapter 7).

The first aim was achieved by firstly identifying the 10k window as the data resolution to be used consistently across all the analysis. Replicates were analysed in terms of visual comparison of the whole genome profile plots and showed that biological replicates were more variable and the technical replicates were more homogenous as expected; importantly, technical replicates showed consistency implying the approach provided reproducible data. Parameters such as mean number of segments and segmented lengths were defined and calculated and compared across replicates and showed similar findings with that of the graphical analysis. Methods of validating data quality includes comparison with MLPA copy number analysis and with the publicly available TCGA copy number data. Results led me to decide to perform additional steps to improve data quality. At that time, new resources were known which includes expanding the blacklisted regions in the genome based on ENCODE [159, 165] , Genome Reference Consortium [155], and computationally derived list of regions of common germline variation in Caucasian population using healthy Caucasian control samples (n=312) from the 1000 Genomes Project [2, 114]. A method of simultaneously adjusting for GC and mappability was applied instead of the previously used method which is sequential adjustment. Instead of using the 7 control samples in the LMC, I learned that using the 312 normal Caucasian samples from the 1000 Genomes Project was able to account for more germline variability in the genome and used it as the new reference control sample. This new resources led to a higher quality of the data as evidenced by visual comparison of copy number profiles from old and new copy number data, improvements in the previously done data quality control assessments, and the ability to uncover associations which were previously undetected in the old data as in the case of the study of a previous PhD

student in our lab Dr. Joanna Pozniak who, using the new data, identified *MYC* (8q24.21) copy number gain to be associated with increased expression of this gene[170].

The second and final aim of this study was to associate the copy number data with patient clinical features (sex, site of the tumour, age at diagnosis, Breslow thickness, AJCC stage, ulceration status, mitotic rate, tumour infiltrating lymphocytes, mutation status, and percentage of stroma) and survival. It started with identifying measures to estimate the overall genome stability which identified a new metric called mean weighted segment mean (MWSM) to have stronger association with most of the tumour and clinical characteristics including survival over the known fraction of genome altered (FGA) [176-178] and aneuploidy score (AS)[176]. Recognising that biologically interesting copy number aberrations may occur in smaller sizes (focal aberrations), association analyses with clinical characteristics and survival were repeated using the 10k window copy number data. With the challenge of computational complexity and time constraints, permutation experiments to identify study-specific genome-wide significance were prioritised for site of the tumour, Breslow thickness, and quantitative 10k window copy number. Study-specific genome-wide significance levels for the rest of the tumour and clinical factors were not calculated but the analyses conducted indicate that genome-wide significance of 0.05 is likely around 10^{-6} to 10^{-7}. Instead, they are ranked based on the decreasing level of calculated significance (ignoring multiple testing) to identify potential copy number windows of biological interest.

8.2 Establishing data quality

Our previous work showed the quality of the data focused on the *CDKN2A* region [1]. Analysing the whole genome requires data quality assessment on the basis of the whole genome which after being performed, led me to conclude the need of further steps to improve data quality. Several steps were applied to the LMC data based on additional new and updated information about the blacklisted regions including those which are highly variable regions in the human genome and methods used in estimating somatic copy number from FFPE samples without matched normal control. Reassessment of the new data derived from these steps revealed significant improvement of data quality and allowed detection of association (i.e. *MYC* copy number and gene expression in the study of Joanna Pozniak [170]) which was not previously detected in the older version of the data. The association of 8q24 region copy number (where *MYC* is located) and *MYC* gene expression had been previously

reported in the study of Pouryazdanparast, Brenner (2012) which looked at the role of this association in amelanotic cutaneous melanoma [171].

8.2.5 Assessment of data quality

Initial assessment of the whole genome data involved several steps such as: (1) Visually checking the whole genome profiles of all the samples in the study (2) Analysis of the replicates (3) Testing for the association of the mean number of segments per chromosome and the mean segmented length (4) Examination of the ESV region (*esv3620012* [136]) which is a known common germline variation in the human genome, and is located very close to the *CDKN2A* region (5) comparison of NGS derived copy number to results from MLPA experiments (6) and validating the LMC data against the published TCGA data. These steps showed the need for improving the data quality to improve interpretability as conclusions are restricted by the presence of extensive noise as shown in the plots of genome profiles. The steps performed were discussed in Chapter 5 and resulted to a new version of the LMC copy number data. Reassessment of this data using the initial steps performed revealed significant improvement in terms of data quality. Firstly, the copy number profiles were cleaner because significant amount of noise had been removed. Secondly, analysis of paired samples used as technical replicates showed increased correlation for the new data (Pearson's r=0.91, $P=4.4x10^{-8}$) when compared with the old one (Pearson's r=0.78, $P=7.8x10^{-5}$). Although the correlation for tumour cores from the old data is higher (Pearson's r =0.68, P=0.021) than that of the new data (Pearson's r =0.21, P=0.533), I strongly suspect that the correlations detected in the old data is primarily contributed by noise (spurious correlation in regions of poor sequence quality) as shown by the whole genome profile plots of these replicates. Thirdly, linear relationship between the average number of segments per chromosome and the segmented length for the new data is higher ($P=2.2 \times 10^{-5}$) than that of the old data (P=0.0052). Fourthly, presence of common germline variations present in the old data were mostly removed in the new data. This is evident when checking for the presence of the ESV region (*esv3620012* [136]) which was no longer detected in the new data indicating that the additional steps done to remove the highly variable regions including common germline variations in the genome was successful. Sixthly, comparison of the *CDKN2A* region from NGS derived copy in the new LMC data showed similar results with that MLPA, as was previously shown in the old LMC data. Finally, validation of the new LMC data with the TCGA data showed more similar distribution of samples with deletion or amplification between the two datasets as compared with that of the older LMC data

8.2.6 Additional steps to increase data quality

The need to perform additional steps to improve the quality of the LMC copy number data was emphasised after the initial data quality assessment. Additional steps such as: (1) incorporating updated and additional blacklisted regions (2) utilising information derived from the sequence data from the 1000 Genomes Projects (1KGP) to identify highly variable regions in the human genome such as regions of common germline variations and (3) accounting for the interaction effect between GC count and mappability in adjusting for their effects on the read counts were performed. This step also utilised the median of the whole genome read counts in contrary to the old data which used chromosome level read counts to adjust each 10k window read counts in the genome.

For the first step, I gathered different new and updated sources of blacklisted regions in the genome and added to the list generated by Dr. Alistair Droop. This includes list of gaps including centromeres, telomeres, and heterochromatins obtained from the UCSC Browser and Genome Reference Consortium Website [154, 155]. Secondly, I learned about the *QDNAseq* (Quantitative DNA sequencing for chromosomal aberrations) pipeline which is available as an R package[2]. It was used to simultaneously correct for GC content and mappability bias using a two-dimensional LOESS model and empirically identified highly variable regions (blacklist) in the genome of a given set of samples, in this case, the 312 Caucasian samples from the 1000 Genomes Project [114]. These steps facilitated the identification of more highly variable regions in the genome that were not accounted for in our previous work

8.3 Association of genomic instability with clinical characteristics and survival

The three measures of overall CNA load were assessed in terms of distribution and tested for associated with tumour and clinical characteristics including survival. These were Aneuploidy Score (AS), Fraction of genome altered (FGA), and Mean weighted segment mean (MWSM).

AS is not associated with survival when categorised into two levels by median cutoff but it is significant in the univariable analysis when using the continuous data. AS loses significance when adjusting for age, sex, and stage (multivariable analysis). FGA is associated with survival when categorised into two levels by median cutoff and is significant in the univariable analysis when using the continuous data. Unlike AS, FGA remained significantly associated with survival after adjusting for age, sex, and stage.

Of the three measured of overall CNA load, MWSM provides the highest significance when testing for association with survival using the data categorised into two levels by median cutoff. It is also significantly associated with survival in the univariable analysis and even after adjusting for age, sex, and stage. The significance of MWSM in both univariable and multivariable analysis is higher than that of FGA, indicating that MWSM captures more information in terms of representing the overall aberration in the genome

8.4 Association of 10k window copy number with clinical characteristics

In recognition that a focal aberration driving progression would be of major interest to melanoma biologists, I performed association analysis of the 10k windows in genome with the patient tumour and clinical characteristics including survival. Due to time and memory requirements of the calculation using high performance computing, permutation analysis to identify study-specific genome wide significance of a test was prioritised for site of melanoma, Breslow thickness, and survival. For other clinical and tumour characteristics, genes were ranked according to decreasing significance and those in the top of the list were reported and checked for the corresponding gene/s.

For the analyses of both sex and site, a 10k window mapping to *LINC01917 [199],* and *LINC01919 [200]* was the most significant. It is lower than the study-specific genome wide significance threshold for site. These genes are predominantly broadly expressed in the testis. Analysis of this window was repeated in 1000 Genomes Project (1KGP) samples (n=289) and showed significant overlap between the histograms of male and female groups which supports that that the difference observed in the LMC data is indeed a somatic variation. For age, it is the window mapping to *MAPK10* which is known to be involved in cell proliferation, differentiation, transcription, regulation, and development [220] that topped the list. Another interesting genes that popped up in this analysis are *RPP40* (a prognostic marker (unfavourable) in renal cancer, endometrial cancer, and liver cancer [225]) and *RIPK1* (plays a role in inflammation and cell death in response to tissue damage, pathogen recognition, and as part of developmental regulation [227]). Three windows were considered genome-wide significant when testing for association with Breslow thickness in this study. This maps to *CXCR4* which is associated with cancer progression and cell survival [229-231], *SDK1* for which its silencing leads to cell rounding and blunted CaP cell migration [232], and *ITGA4* which was reported to be upregulated in melanoma[233]. Notable genes but did not reached

study-specific genome-wide significance include *DDX1* which is known to be involved in transcription, viral replication, mRNA/miRNA processing, and transfer ribonucleic acid (tRNA) splicing and plays important role in the regulation of gene alternative splicing and insulin secretion in pancreatic β cells [234] ; *CHST12* which is associated with tumour regrowth in non-functioning pituitary adenoma (NFPA) [235]; *GPR39* for which its overexpression contributes to malignant development of human oesophageal squamous cell carcinoma[236]; *SUN1* for which silencing of this gene inhibits cell growth through G0/G1 phase arrest in lung adenocarcinoma [237]; and *VEGFA* which activates an epigenetic pathway upregulating ovarian cancer-initiating cells [238]

CXCR4 also popped up as the most associated with AJCC stage. Other genes associated with at least one forms of cancer (e.g. liver, colorectal, prostate, pancreatic, thyroid, renal cell carcinoma, biliary tract, non-functioning pituitary adenoma, gastric carcinoma, acute myeloid leukemia, intrahepatic (ICC) , extrahepatic (ECC) cholangiocarcinoma, lung cancer and B-Cell Acute Lymphoblastic Leukemia) are *UVRAG* [239, 240], *CYP3A54P* [241] ,*ACER3* [242], *ELFN1* [243], *ANO5* [244-246], *PRKAR1B* [247, 248], *CHST12* [235], *C7orf26* [249], *FAM20C* [250], and *PPME1* [251].

For ulceration, the top window in the list is *NEBL* which is enriched in the heart muscle and reported to be a favourable prognostic marker in renal cancer and unfavourable prognostic marker in urothelial cancer [252]. For mitotic rate, the top window maps to *MIR100HG* which promotes colorectal cancer metastasis and is associated with poor prognosis [256]. Another interesting gene in this analysis is *ADAMTSL1* which was reported to be differentially methylated between paired tumour and normal tissues from breast cancer patients [255]. Analysis of TILs did not show very high level of association with copy number. Interesting gene on the list is LRRIQ3 which was previously reported to be associated with response to chemotherapy in rectal cancer [269]. Other genes on the list are ZSWIM5P3 (Zinc Finger SWIM-Type Containing 5 Pseudogene 3), LRRC4C which is a specific binding partner for netrin G1 [270], LINC02053 which has higher expression in testis [271], and RNU1-89P which was reported to have differential expression of small nuclear RNA (snRNA) in oesophageal adenocarcinoma[272]. In terms of mutation status, the top window maps to *ADAM12* which has been reported to contribute to increased tumour proliferation, metastasis and endocrine resistance[273]. Other cancer related genes on the list are ALDH18A1 which is associated with associated with luminal B breast cancer [274], VTI1A which is associated with susceptibility to colorectal and lung cancers [275], SORBS1 which has been reported to suppress tumour metastasis and improves the sensitivity of cancer to chemotherapy drug [276], LGI1 which has been reported to be

differentially expressed in early- and late-stage oral squamous cell carcinoma [277], and ZNF777 which inhibits proliferation at low cell density through down-regulation of FAM129A [278].

Finally, for the analysis of percentage of stroma , the top window maps to *KCNT2*. The expression of this gene is correlated with the prognosis of skin cutaneous melanoma (SCM) and significantly different between normal skin and SCM [279]. This followed by *OR2G6* which is associated in breast cancer [280]. Other genes in the list include *FAM163A* which is reported to be positive regulator of ERK signalling pathway, interacts with 14-3-3β and promotes cell proliferation in squamous cell lung carcinoma[281], *AXDND1* which is reported to be downregulated in gastric cancer [282], *INTS7* which is associated with Gastric Cancer and Ivic Syndrome, and its increased levels is associated with the aggressiveness of prostate cancer [283, 284], *RALGPS2* which is reported to be essential for survival and cell cycle progression of lung cancer cells [285], *PLD5* which is associated with survival of thyroid cancer patients [286], *LELP1* which is significantly increased in atopic dermatitis skin [287], *RSL24D1P4* (Ribosomal L24 Domain Containing 1 Pseudogene 4), *EFCAB2* which is considered as one of the most potential targets for four breast cancer subtypes, a specific therapeutic targets for luminal A, and identified as a susceptibility gene for Colorectal Cancer in east asian populations [288, 289], and PKP1 where its phosphorylation by *RIPK4* regulates epidermal differentiation and skin tumorigenesis[290].

In terms of survival analysis using the quantitative 10k window copy number, no test yielded a study-specific genome-wide significance. The first in the list of results with highest significance has an equivalent study-specific significance of 0.05698925 which is close to the 5% level of significance. This window maps to *BMPR1A* which inhibit the tumorigenic potential of human brain tumour-initiating cells [291]. Above median copy number of in window corresponds to better survival (HR=0.46, logrank P=7.3 x 10^{-5}).

Each 10k window copy number was categorised as to normal, deletion, or amplification. In this analysis, the window that reached the highest significance corresponds to *BRAF* which is an oncogene linked to melanoma and some carcinomas, and functions to upregulate the *RAS/RAR/MEK* pathway and a common target in melanoma therapy [302, 303]. Lower (possibly deletion and normal, as compared to copy number gain) copy number of this gene corresponds to poorer survival (P=6.5 x 10^{-9}). This is coherent to the study of Stagni et.al (2018) which found out that a significantly higher risk of progression was observed in patients with normal (diploid) *BRAF* status versus those with *BRAF* gains (HR= 2.86; 95% confidence

interval (CI): 1.29 - 6.35, P=0.01) [201]. Checking the survival of patients for *BRAF* mutation status, there is no significant difference between the survival curves of *BRAF* mutated and non - *BRAF* mutated tumours (HR=1.01, logrank P=0.9). The study of Stagni et.al (2018) observed that patients with low percentage versus those with balanced BRAF mutant allele percentage have a significantly higher risk of progression (HR, 4.54; 95% CI, 1.33-15.53; P = 0.016) while the study of Carlino et. Al (2014) concluded that *BRAF* mutation status does not influence survival in metastatic melanoma [201, 318].

Analysis using the 303 (277 cutaneous melanoma samples plus 26 samples from other sites) samples identified a window mapping to *FAM190A* (or *CCSER1*) gene as the most associated with survival. Deletion in this window is associated with poorer survival and high CNA load as measured by mean weighted segment mean (MWSM). The study of Patel et al. (2013) reported that deficiency of this gene was reported to create a cell division effect and was demonstrated in vivo to have oncogenic properties by in-frame deletions within the region of this gene [309]. There was also an indication that these transcript variants are potential therapeutic targets in patients with cancer based on the study of Kang et al. (2019) [310]. This window did not crop up in the analysis of 277 samples because the 26 samples excluded had both poor prognosis and were commonly deleted for that *FAM190A* 10k window. Because of the limited size, I could not do more detailed analysis of these samples.

8.5 Strength and limitations

8.5.1 Strengths of this study

The following are the strengths of this study:

1.) This study is currently the biggest population-based somatic copy number study in primary melanoma. There maybe a few similar studies about primary melanoma but they are significantly smaller and are not population based (e.g. ~100 samples for the TCGA primary melanoma copy number data). This also was sequenced with median coverage where most FFPE studies used shallow sequencing, providing us more potential to uncover biologically and clinically relevant somatic aberrations.

2.) Our copy number study includes information on careful follow up and detailed clinical and tumour information (extensive phenotypic information, biological samples and information describing patterns of UV exposure, the primary risk

factor for melanoma as well as measures of pigmentation and naevi which are also associated with risk) which are very useful in association studies.

3.) The version of the data in this study was given careful consideration in terms of data quality and utilised newly published information such blacklisted regions, sequence data from normal control samples of the 1000 Genomes Projects which are realigned to GRCh 38 genome build and readily available for analysis in identifying regions of common germline variation in the human genome. Some of which are potentially specific to Caucasian population which is an important consideration as our samples are from British Caucasian populations.

4.) Utilising normal control samples from the 1000 Genomes Projects (1KGP) also addressed the challenge of absence of matched normal samples for each tumour in this study by using the median for each 10k window of the genome profiles of the 312 1KGP normal control sample. We have available 7 normal samples from our study but by visual comparison of copy number profiles, we identified that using the 312 1KGP sample provides better adjustment of tumour copy number than our 7 normal sample. This could be due to the fact that because the 1KGP normal control sample is significantly larger, it was able to capture more germline variations (potentially including those which are specific to the ethnicity of the samples used i.e. Caucasian samples) and provided better adjustment to estimate a somatic copy number.

8.5.2 Limitations of this study

The following are the limitations of this study:

1.) Matched normal DNA was not available for this study.

2.) Although this study is currently the biggest population-based somatic copy number study in primary melanoma, the statistical power may still be not enough when testing for differences that normally exist in smaller quantity.

3.) Although there are publicly available data for validation like that of the TCGA, lack of datasets derived from similarly systematic population-based study for validation of the interesting my findings, especially for the measures of overall copy number aberration and focal changes (10k window). Our lab tried to addressed this by performing MLPA experiments but this is limited to the *CDKN2A* region as this experiment requires significant amount of DNA.

8.6 Conclusions and Recommendations of the study

This study showed the success of generating high quality somatic copy number data from FFPE tissue. It also demonstrated the success of the implementation of additional steps and a different read count adjustment method (using interaction of GC and mappability) in addressing the variations inherent in the human genome, degradation-prone FFPE samples, and sequencing methods to increase the data quality. This is shown by the improvement in both the qualitative and quantitative measures used in the data quality assessment as well as in terms of comparison with the TCGA copy number data.

In terms of the analysis of the copy number data, mean weighted segment mean (MWSM) showed the highest association with patient tumour and clinical characteristics as compared with fraction of genome altered (FGA) and Aneuploidy Score (AS).

8.7 Future work

Lack of matched normal samples was a primary challenge in the estimation and analysis of copy number data in this study. If funding and DNA samples are not a restriction, extending this study to include matched normal samples provide direct comparison of copy number data derived with and without matched normal samples. Analysis of the 10k window copy number data revealed copy number windows that are most associated to patient tumour and clinical characteristics. Validation of these results once similar population-based study cohort becomes available.These results also serve as target for future studies. Computational time required to estimate study specific genome wide significance also posed difficulty as one iteration takes about 4 to 8 hours. Future study increasing the number of iterations to further assess how the assigned genome-wide threshold behaves is recommended.

Appendix A

A.1 Rejected Samples

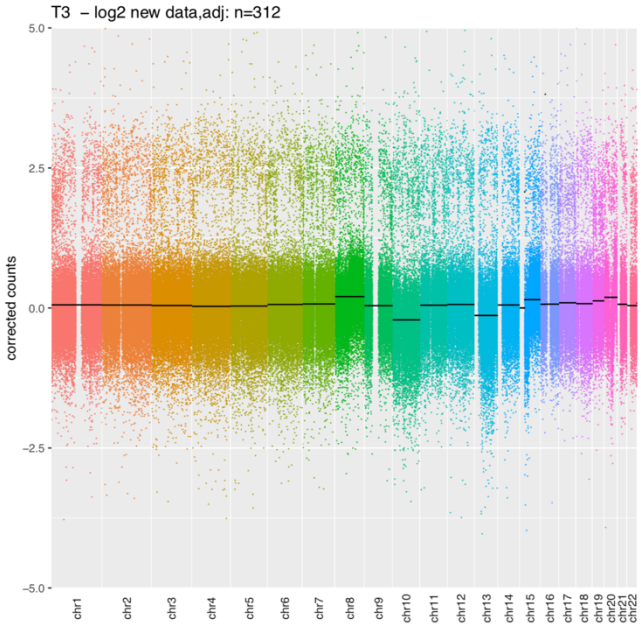


Figure A.1. Rejected Sample 3

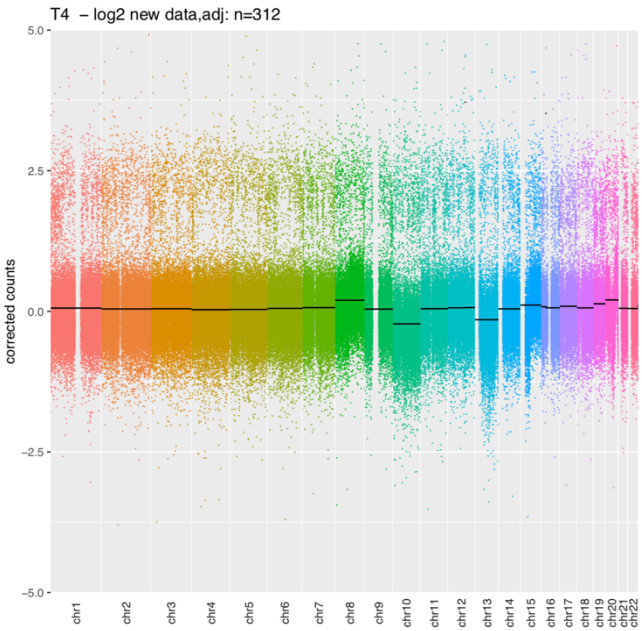


Figure A.2. Rejected Sample 4

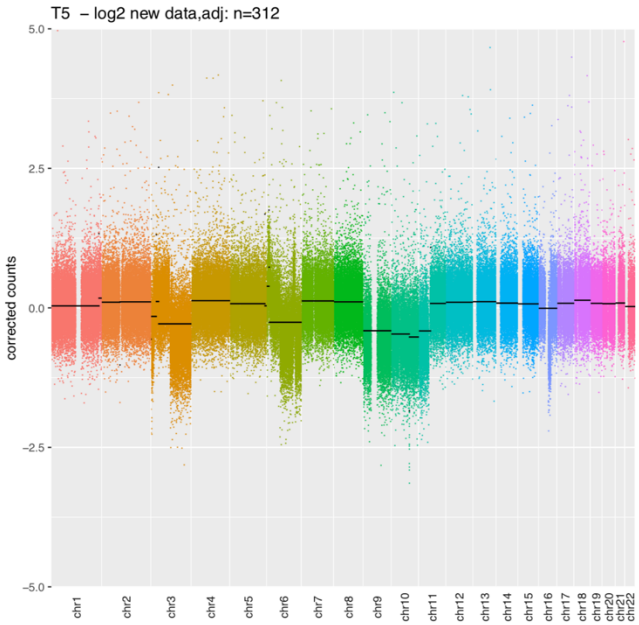


Figure A.3. Rejected Sample 5

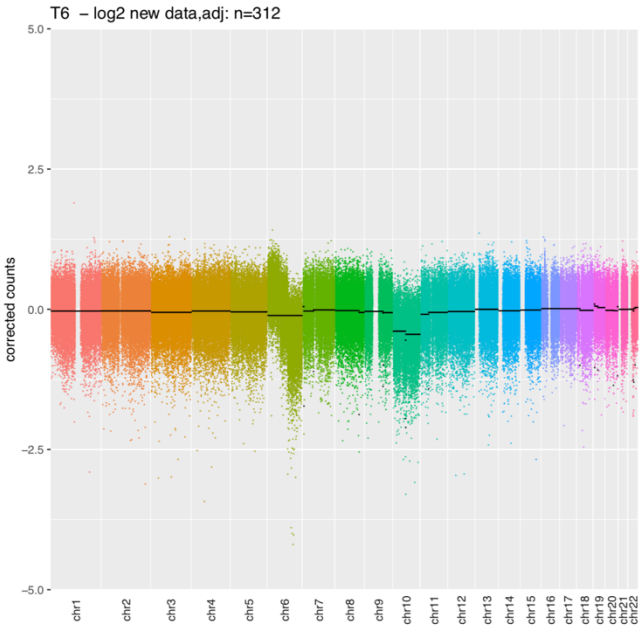


Figure A.4. Rejected Sample 6

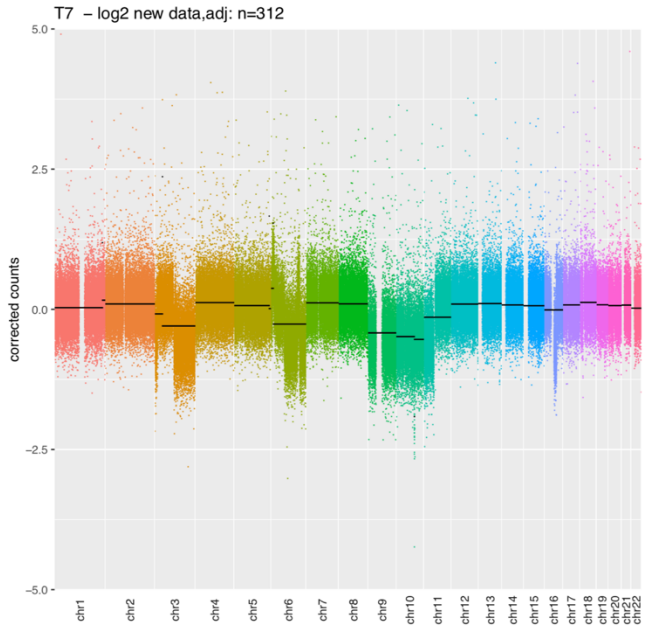


Figure A.5. Rejected Sample 7

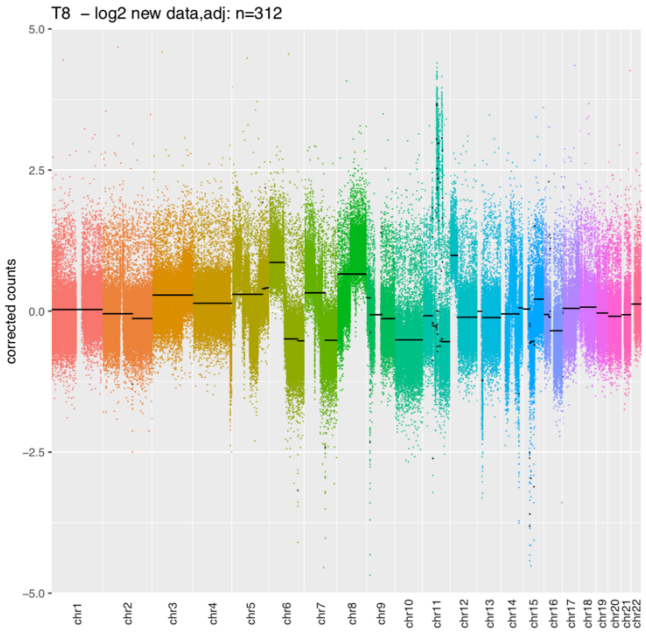


Figure A.6. Rejected Sample 8

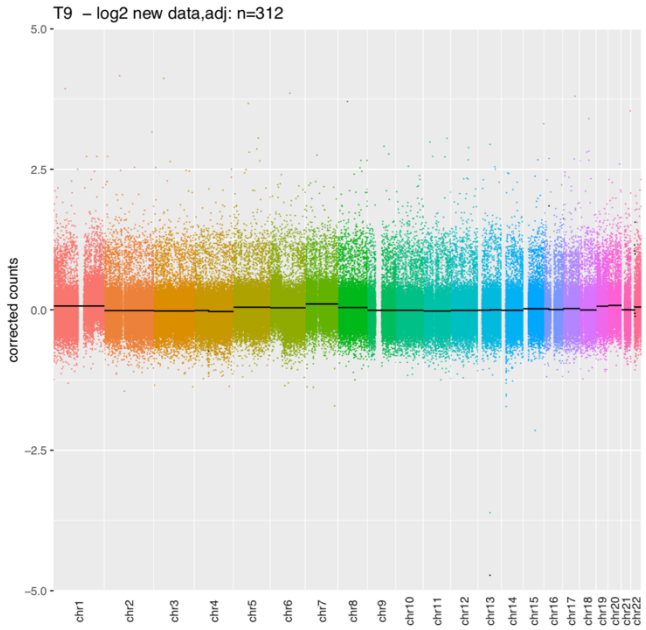


Figure A.7. Rejected Sample 9

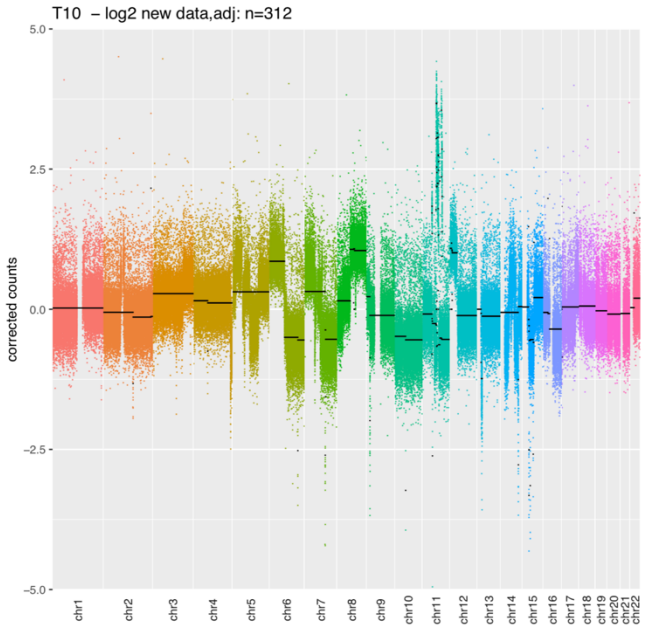


Figure A.8. Rejected Sample 10

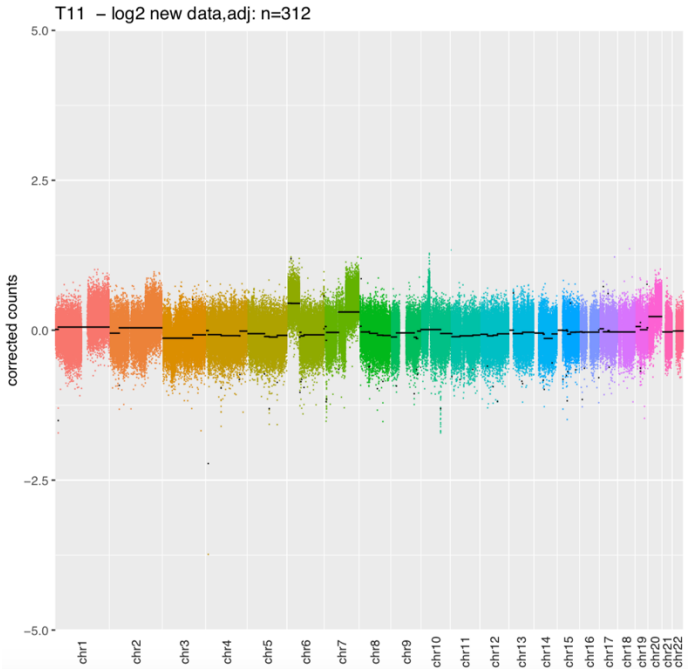


Figure A.9. Rejected Sample 11

Appendix B

B.1 Replicates

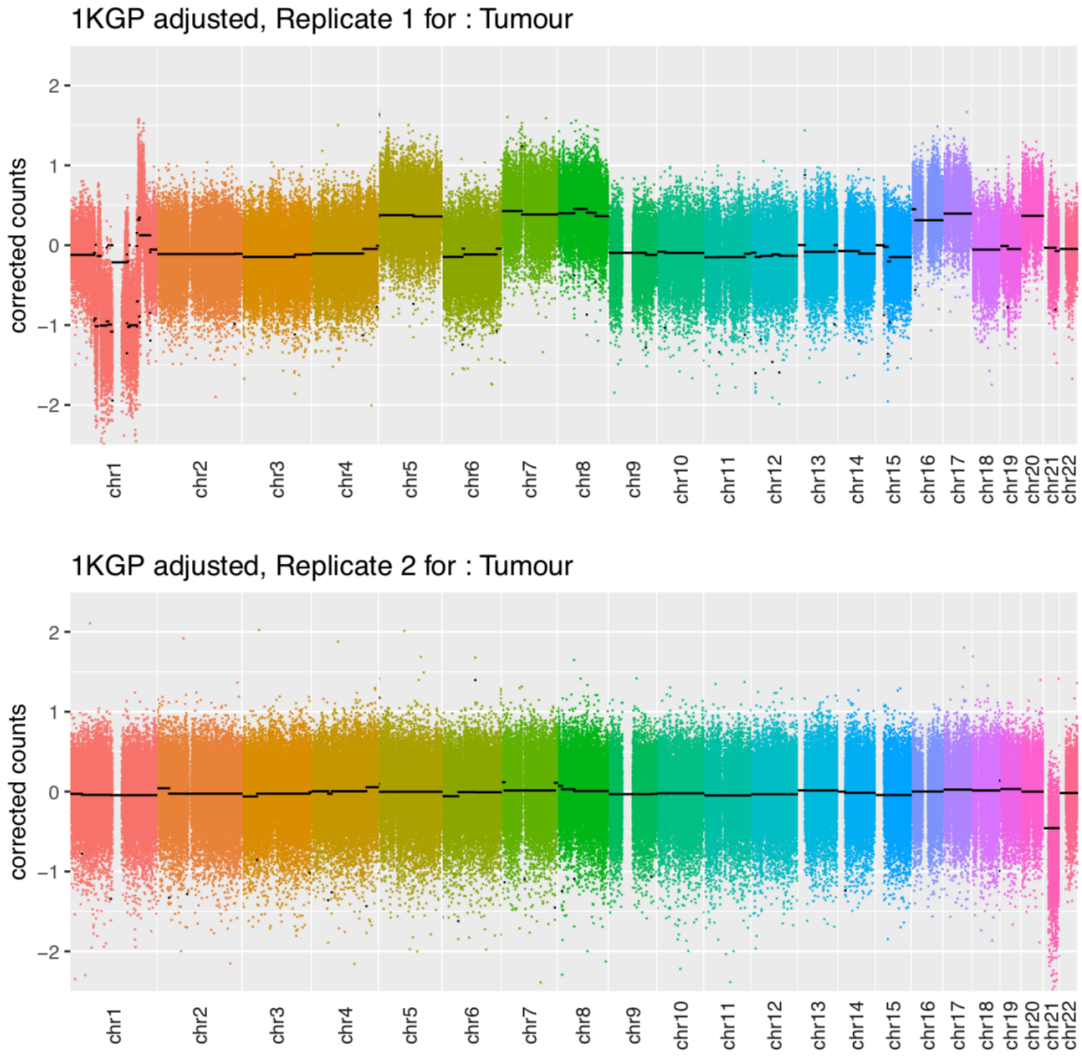


Figure B.1. Replicates for Tumour: Pair 2

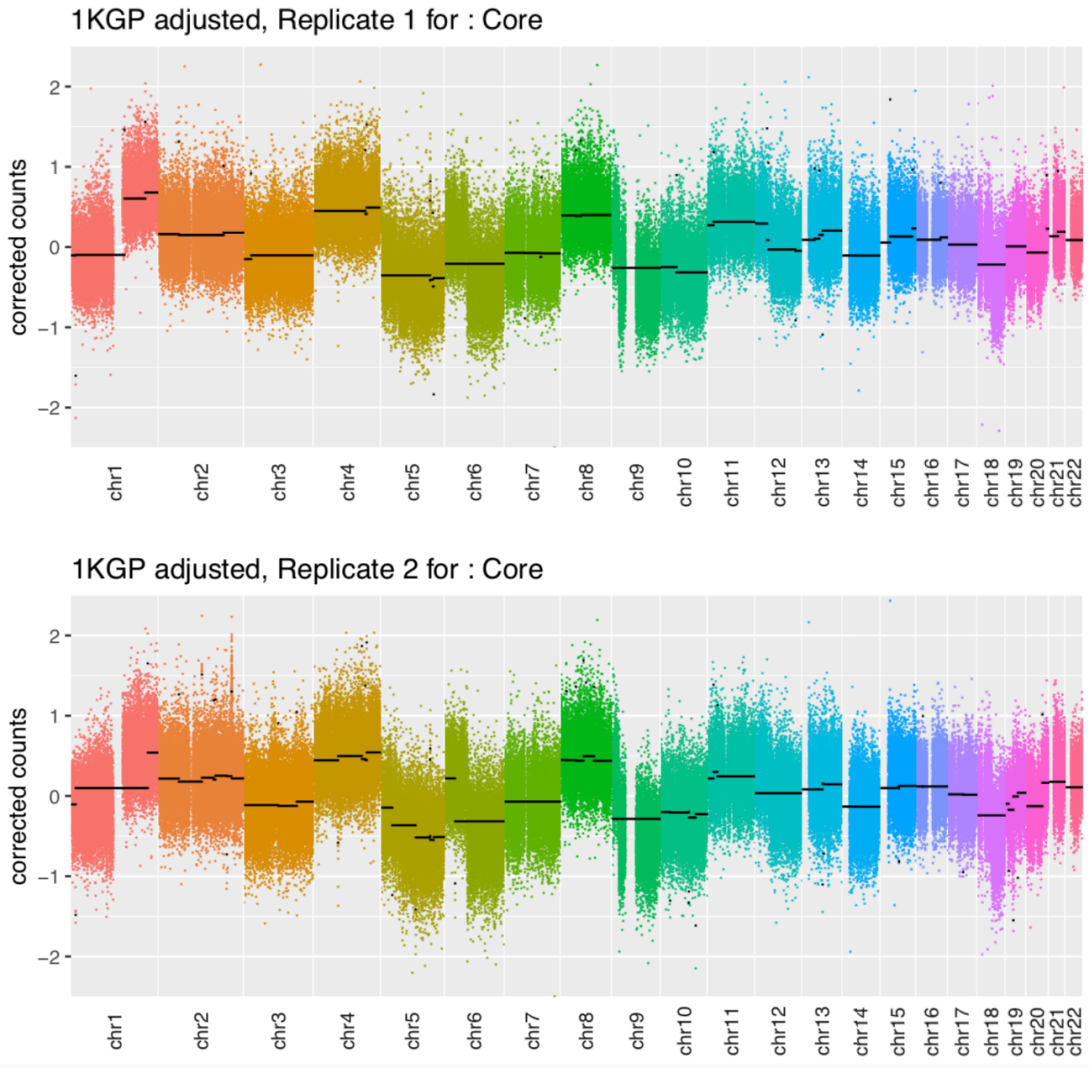


Figure B.2. Replicates for Core: Pair 2

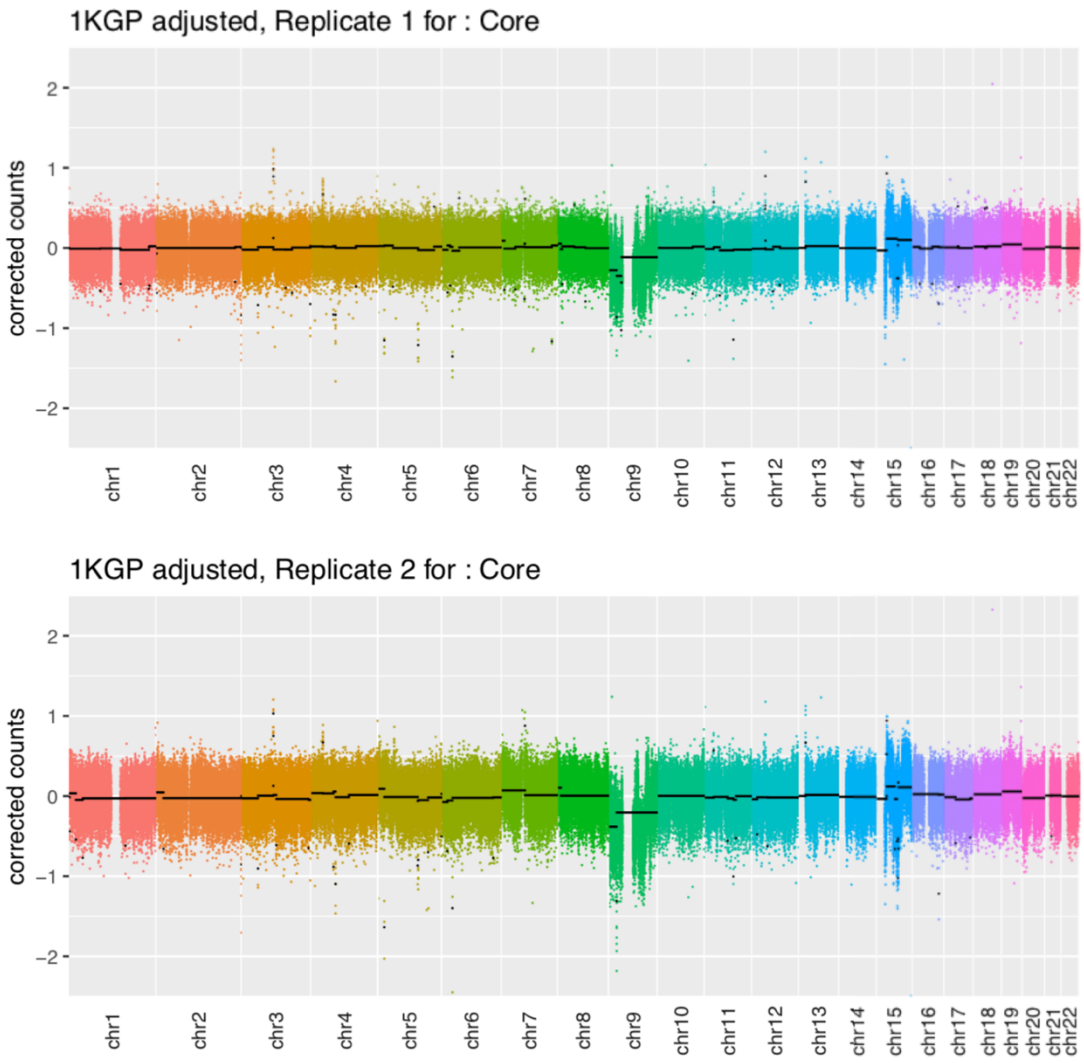


Figure B.3. Replicates for Core: Pair 3

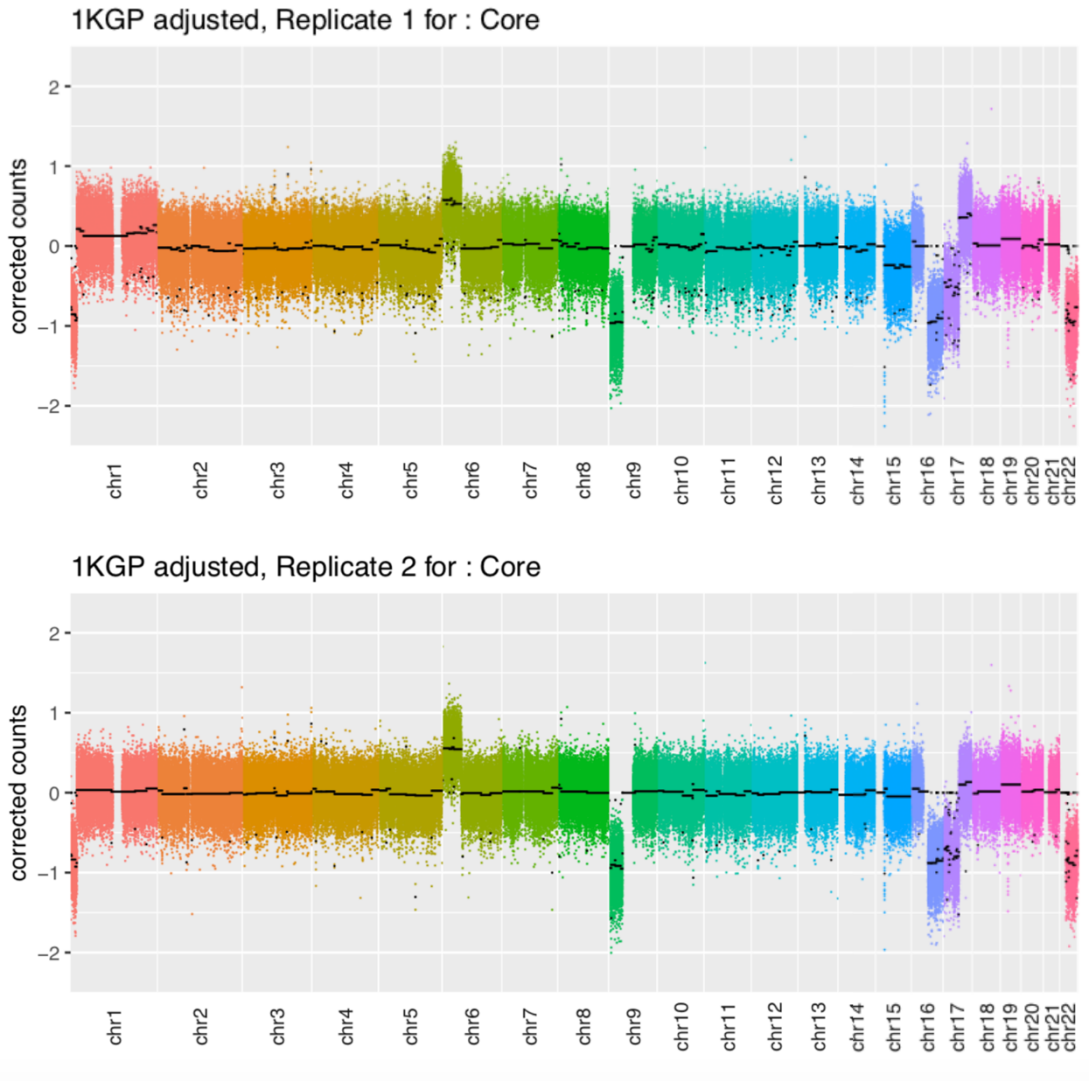


Figure B.4. Replicates for Core: Pair 4

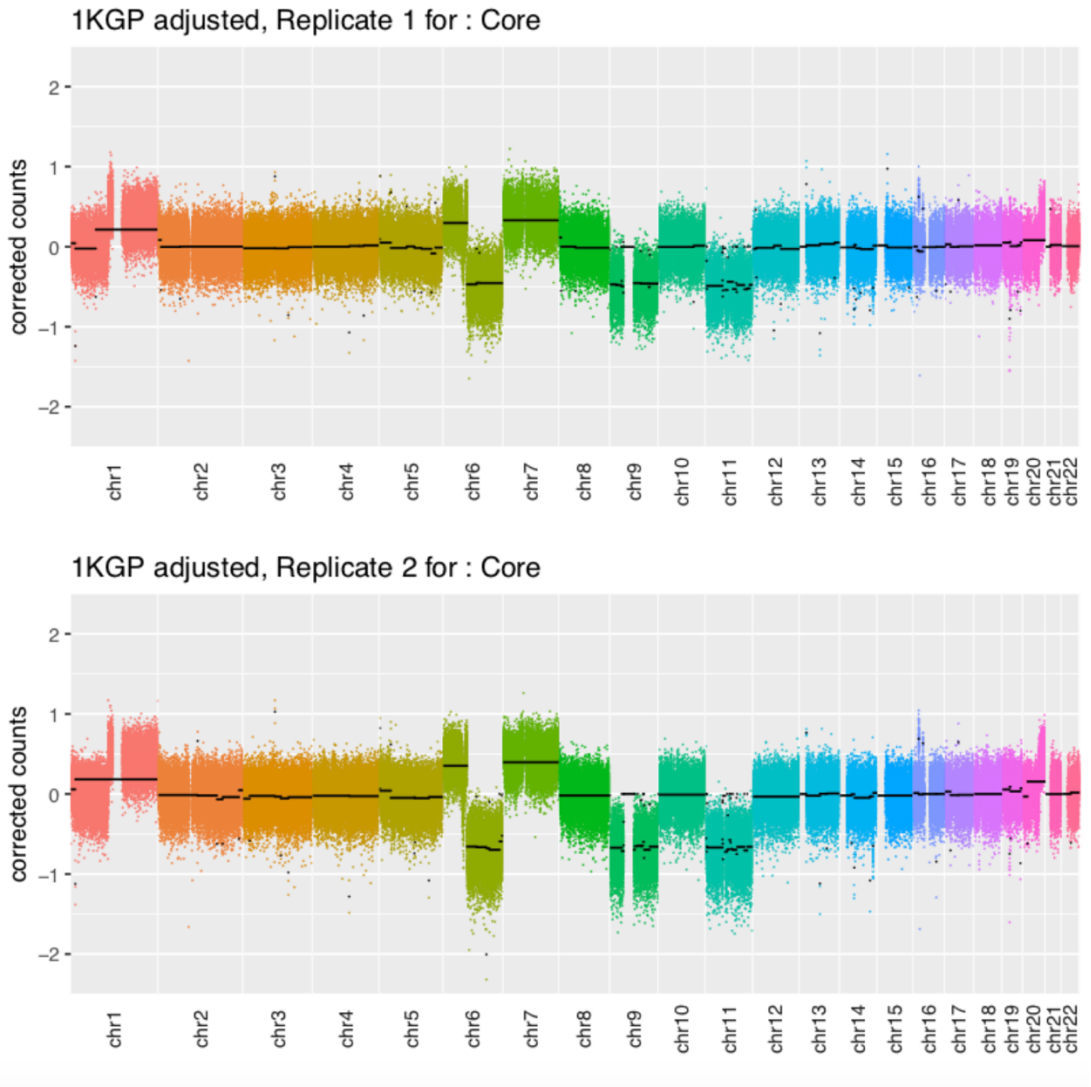


Figure B.5. Replicates for Core: Pair 5

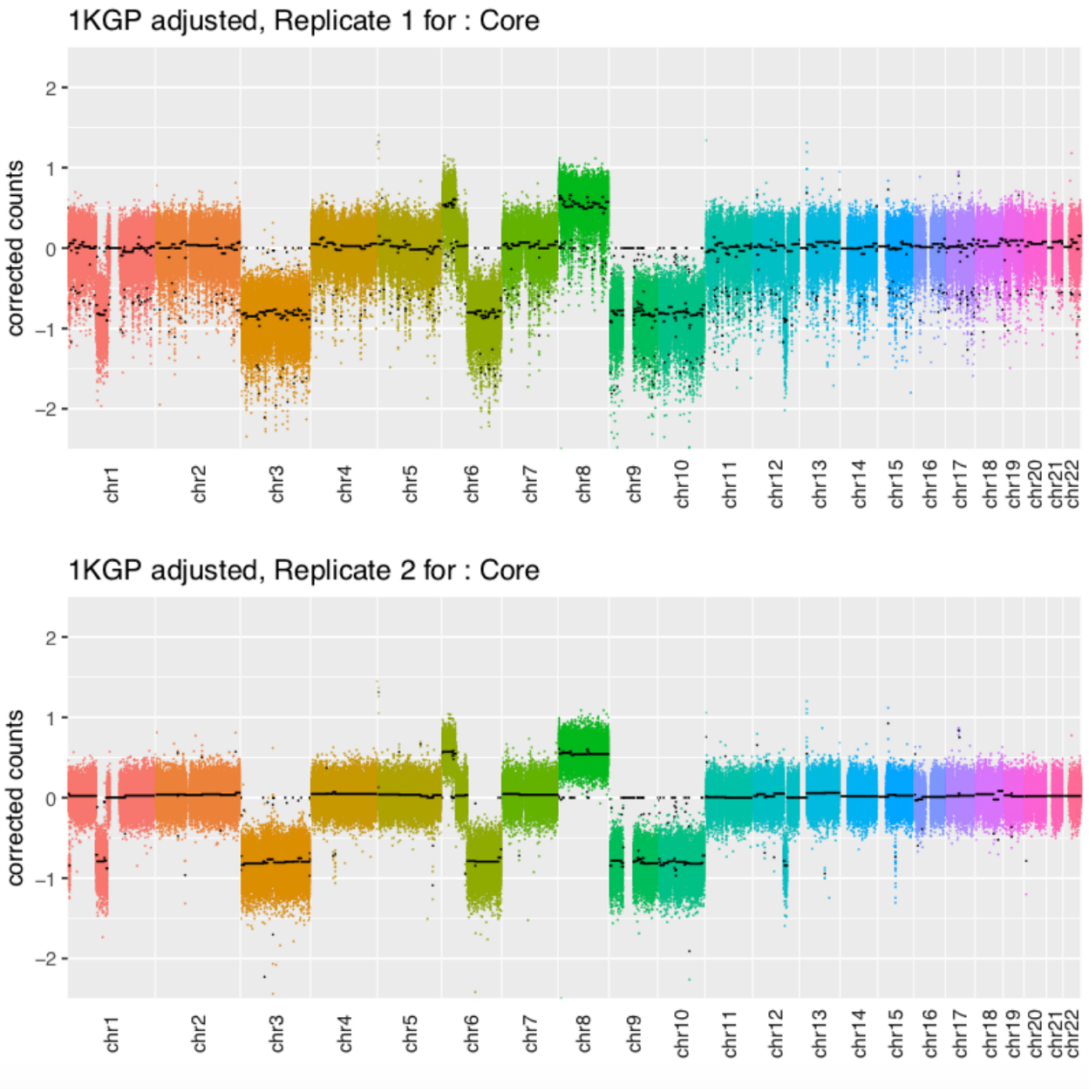


Figure B.6. Replicates for Core: Pair 6

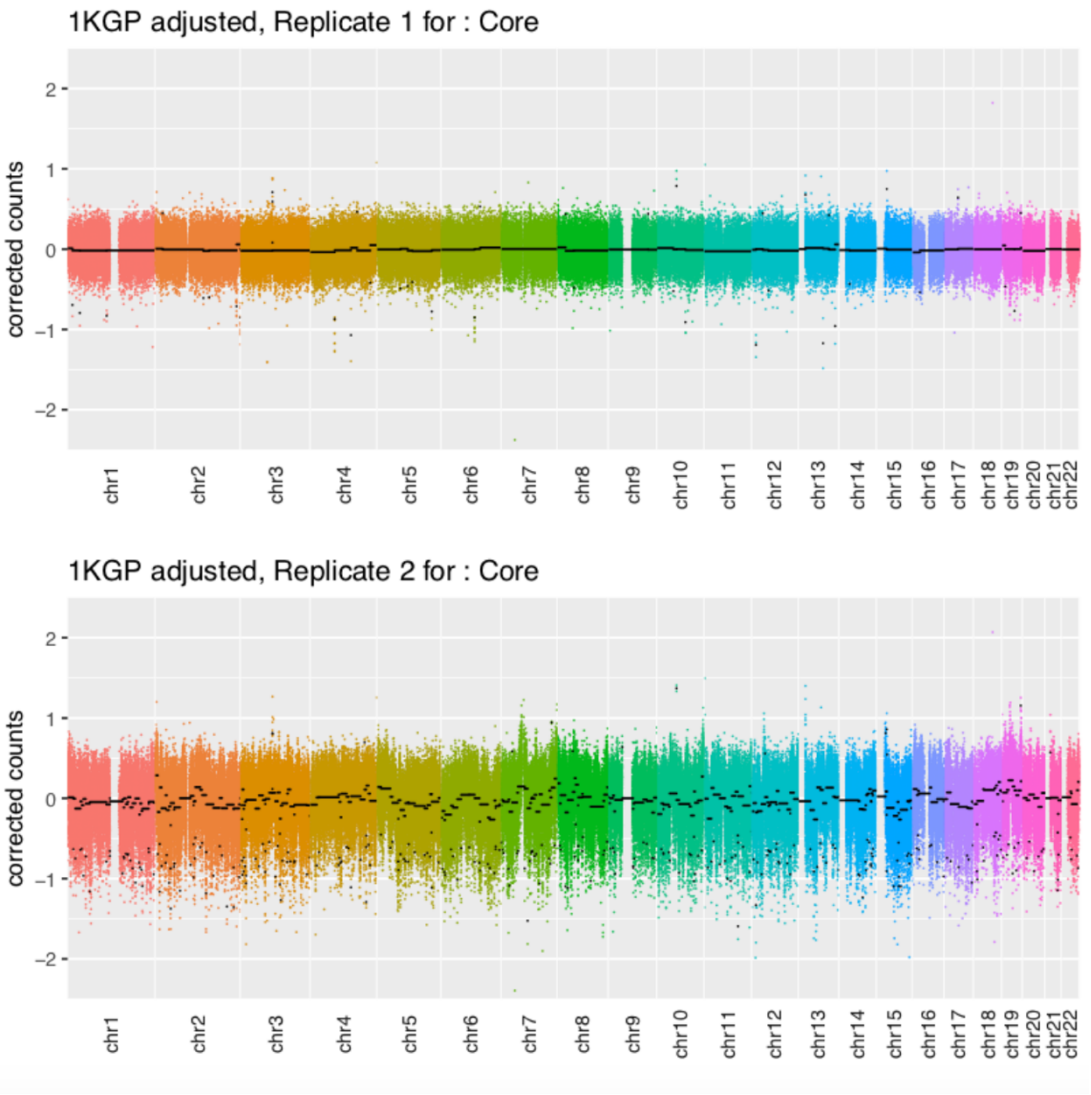


Figure B.7. Replicates for Core: Pair 7

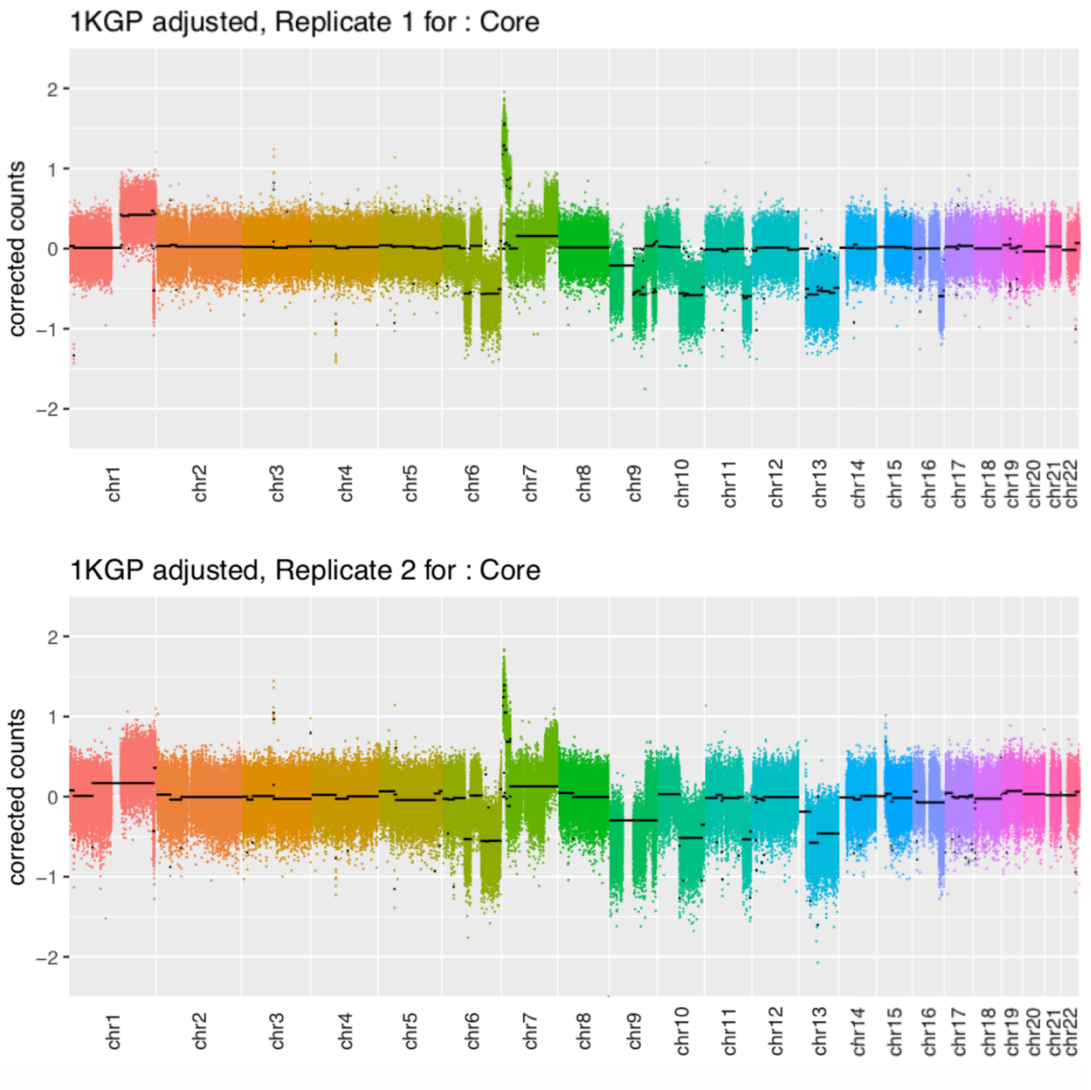


Figure B.8. Replicates for Core: Pair 8

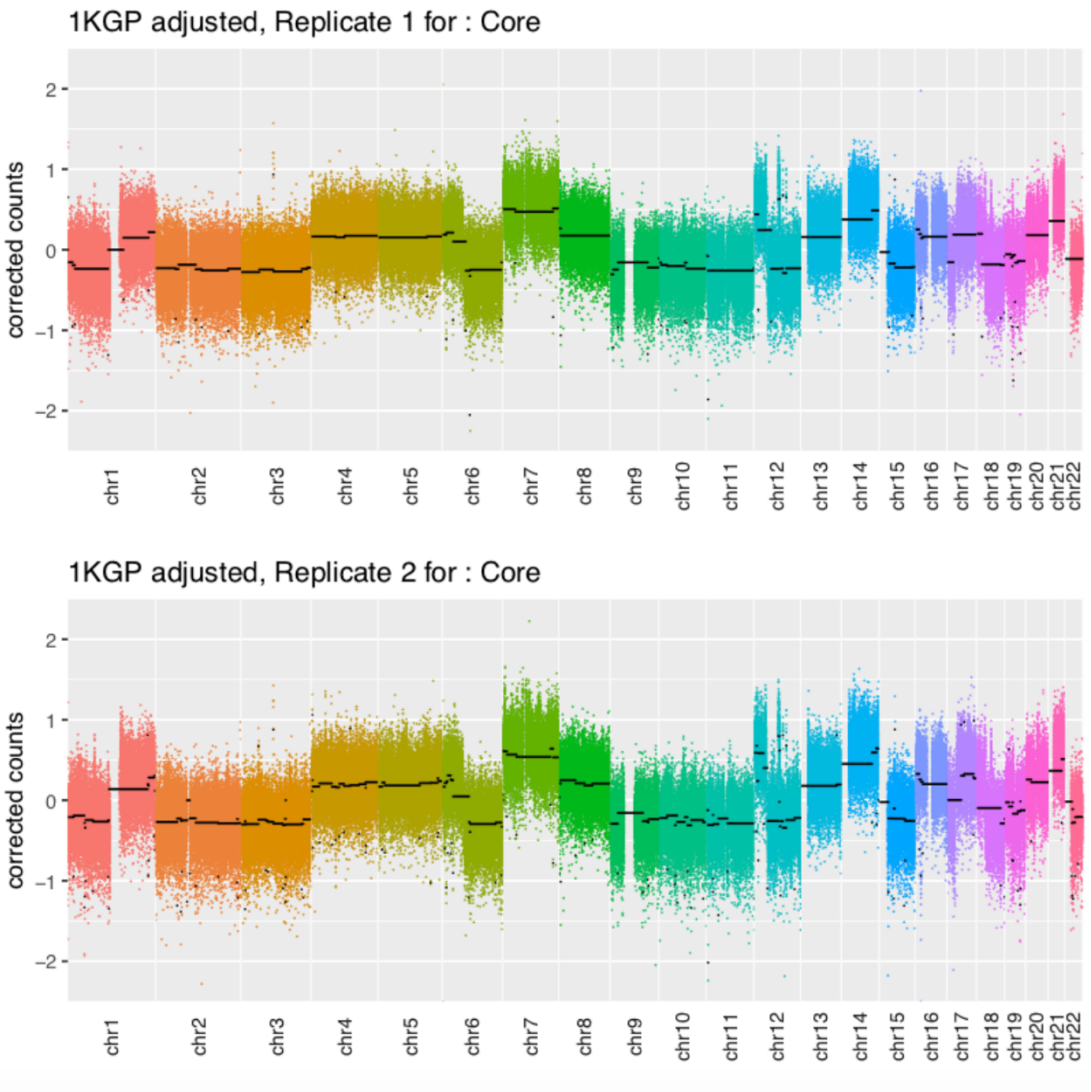


Figure B.9. Replicates for Core: Pair 9

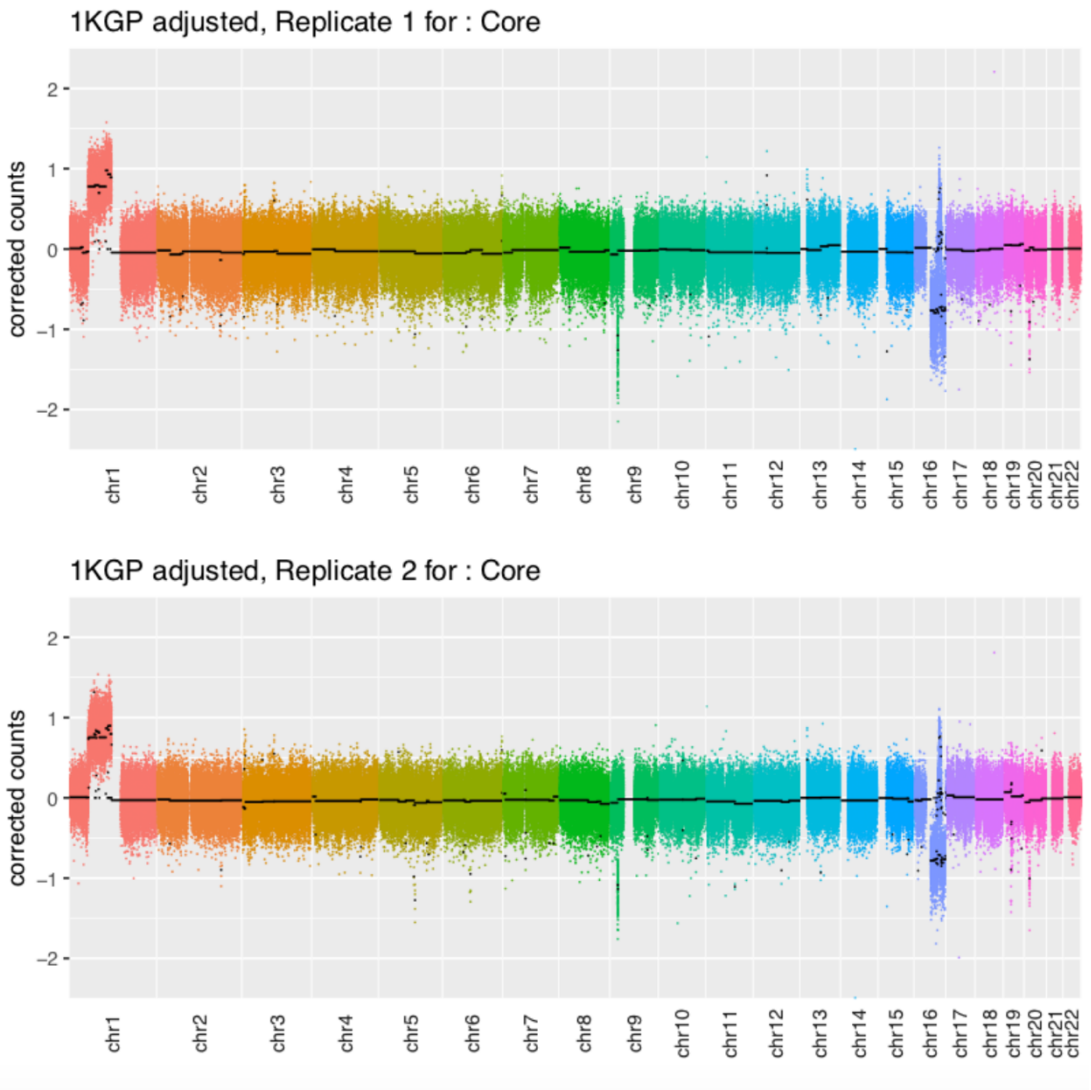


Figure B.10. Replicates for Core: Pair 10

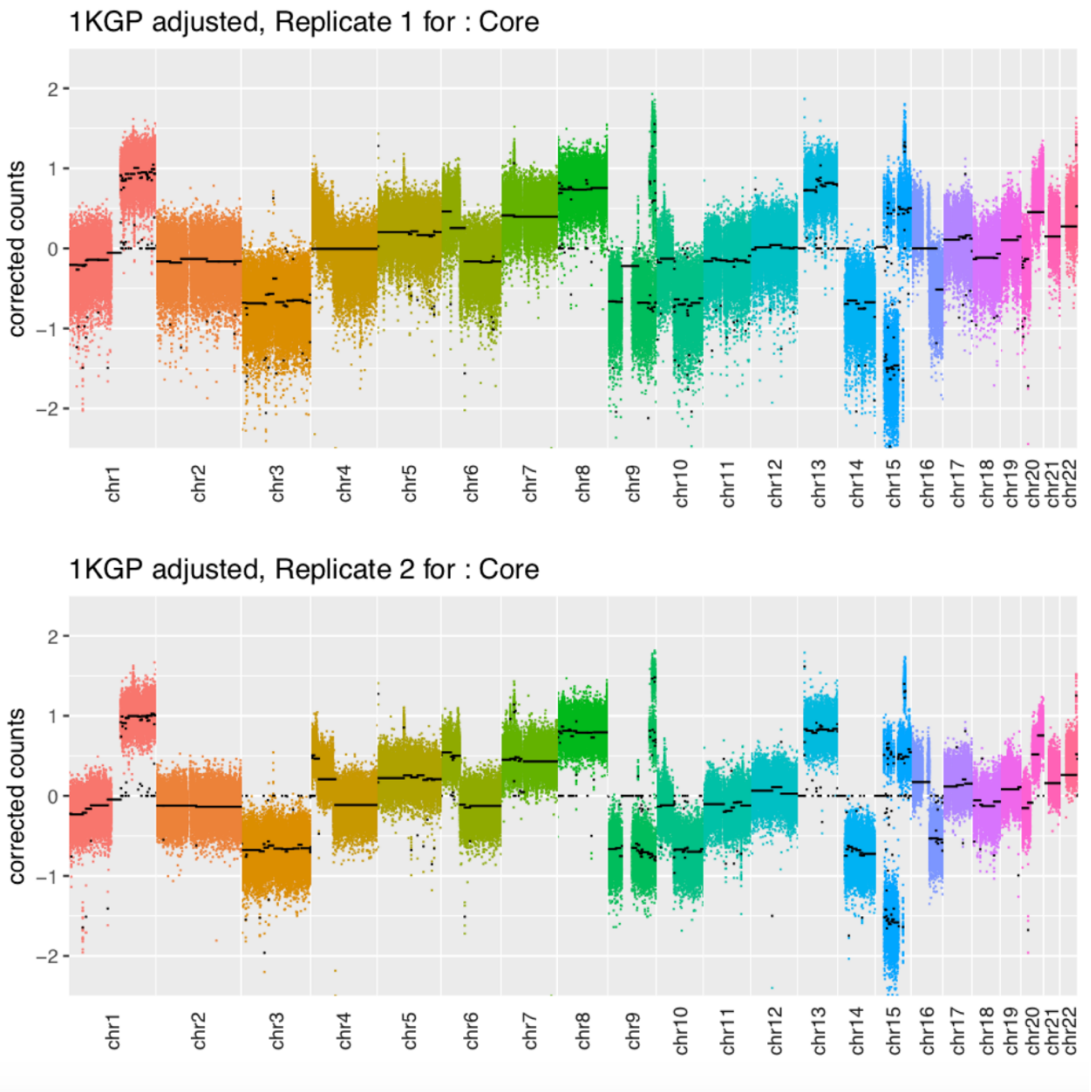


Figure B.11. Replicates for Core: Pair 11

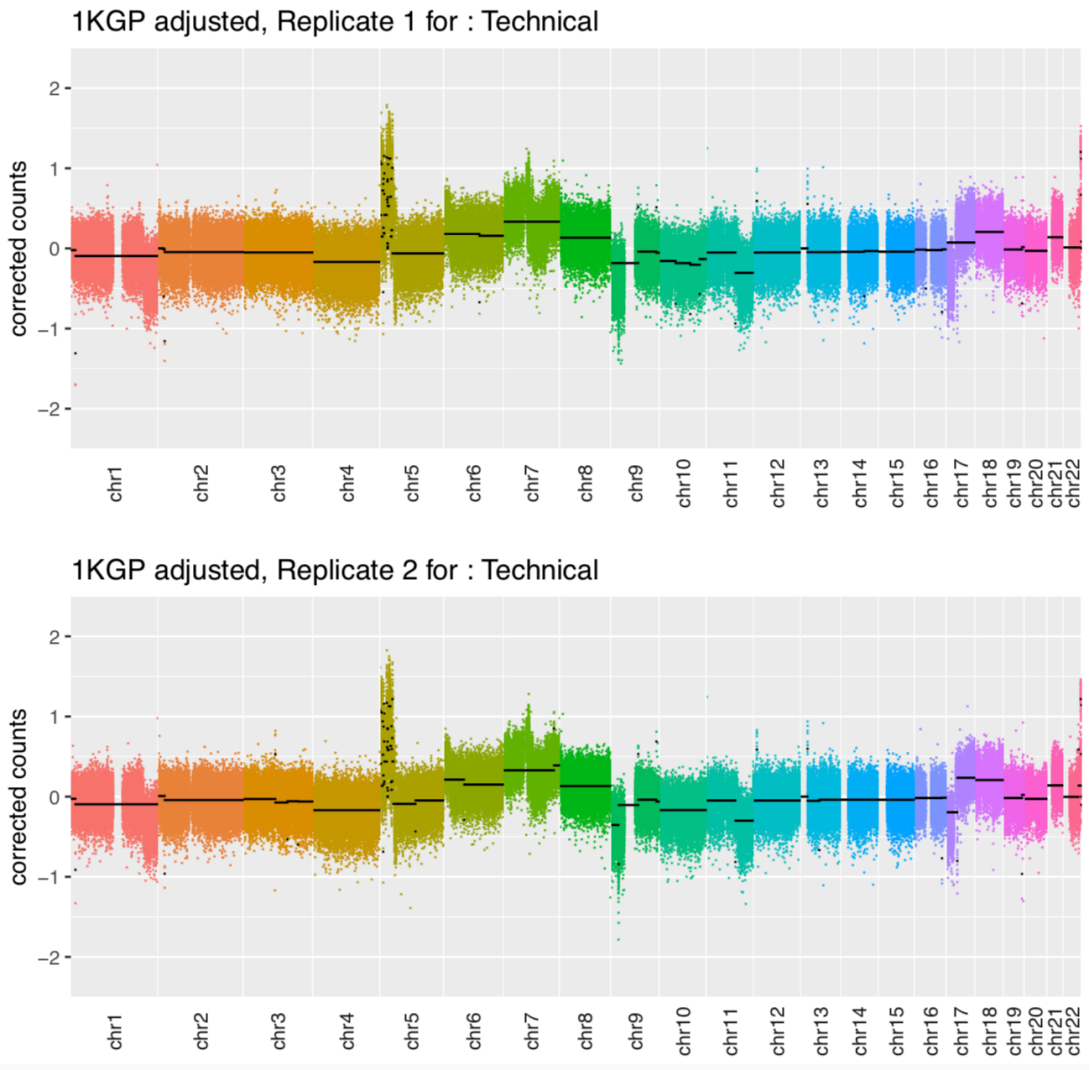


Figure B.12. Replicates for Technical: Pair 2

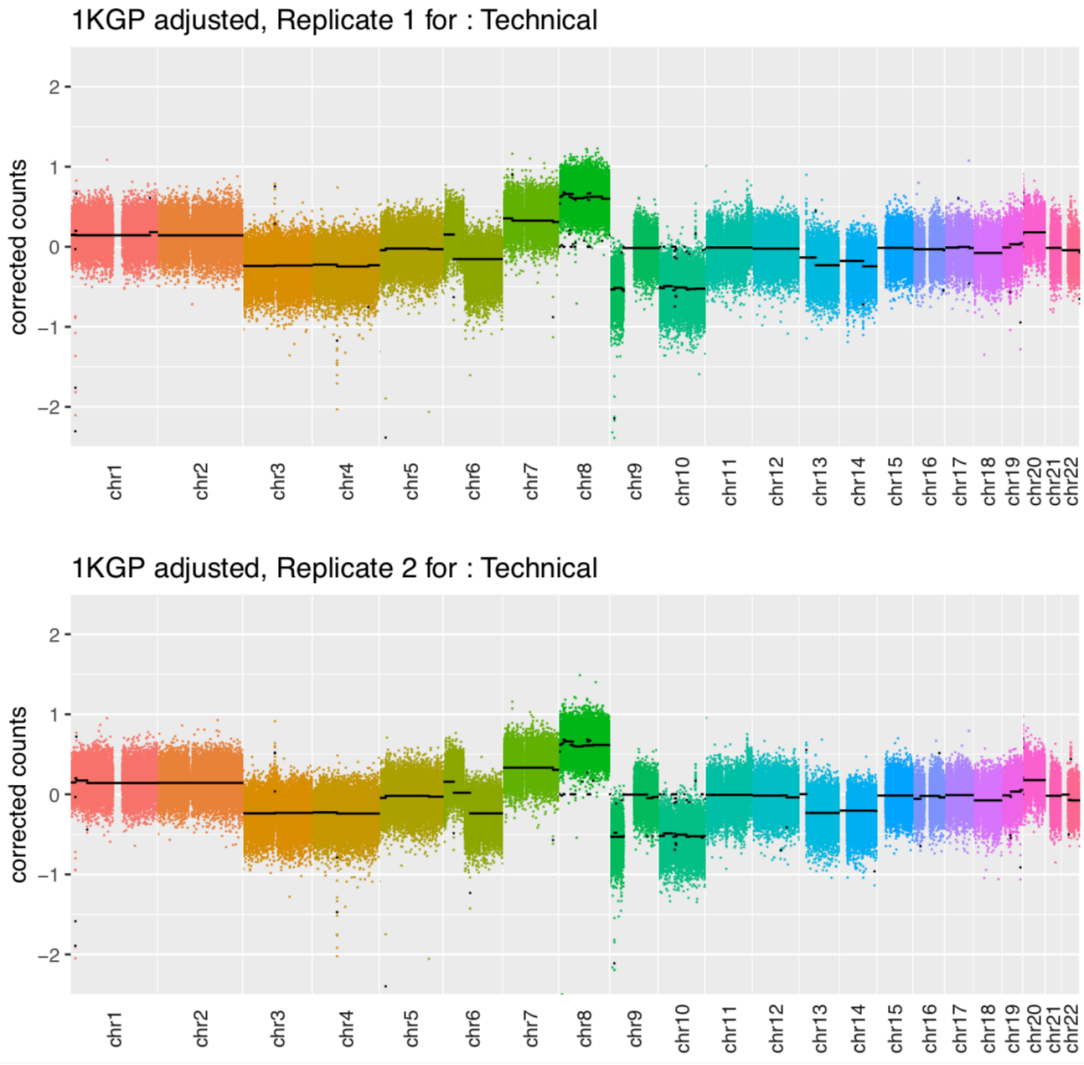


Figure B.13. Replicates for Technical: Pair 3

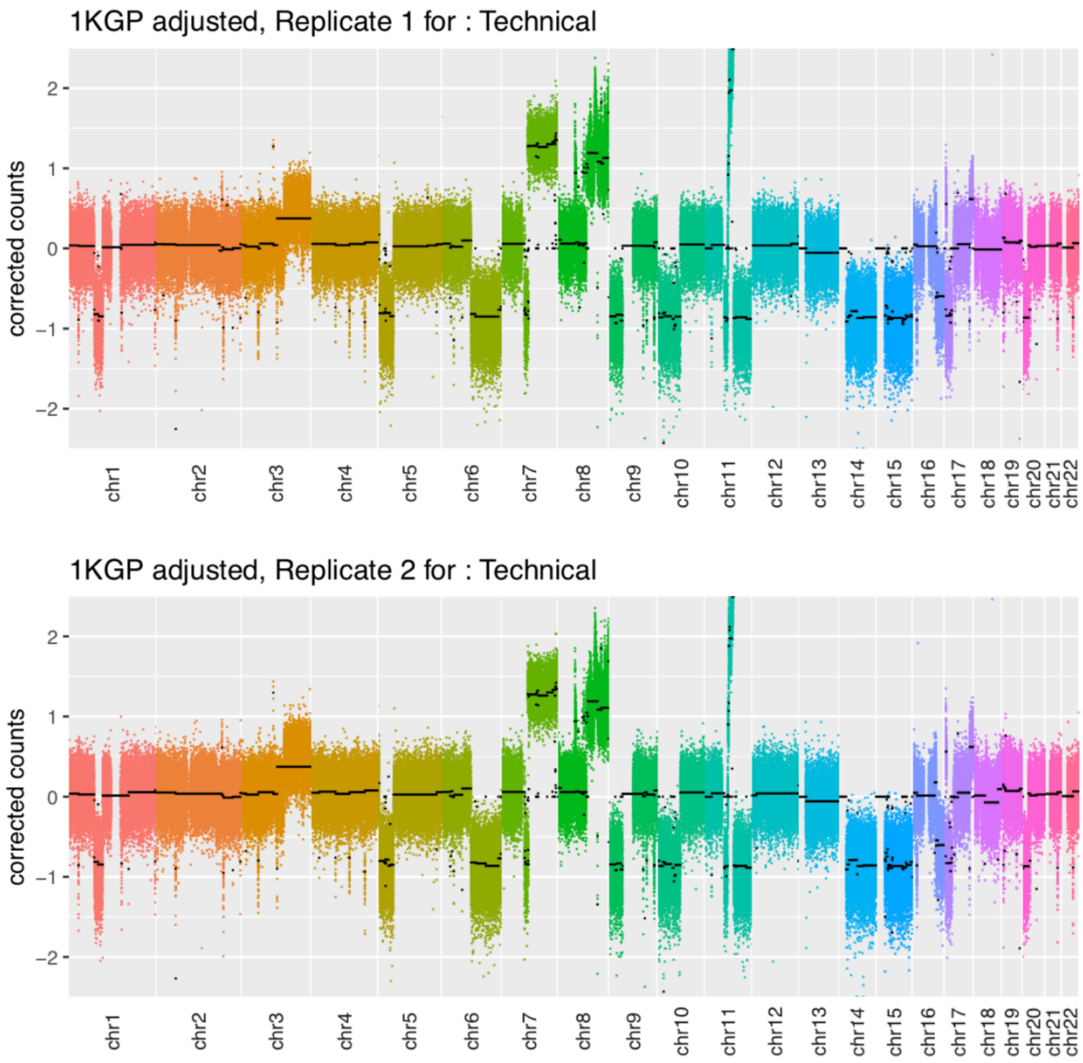


Figure B.14. Replicates for Technical: Pair 4

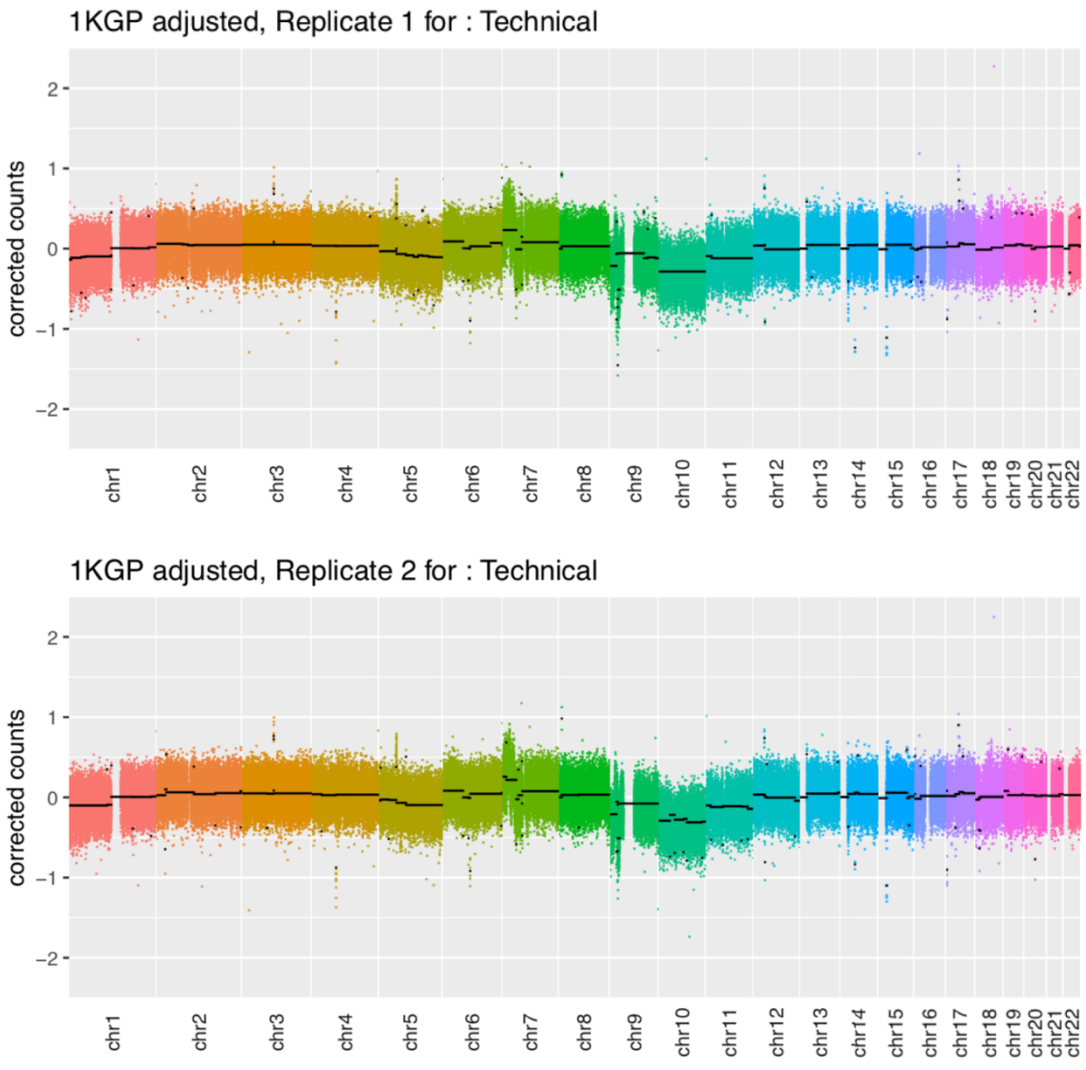


Figure B.15. Replicates for Technical: Pair 5

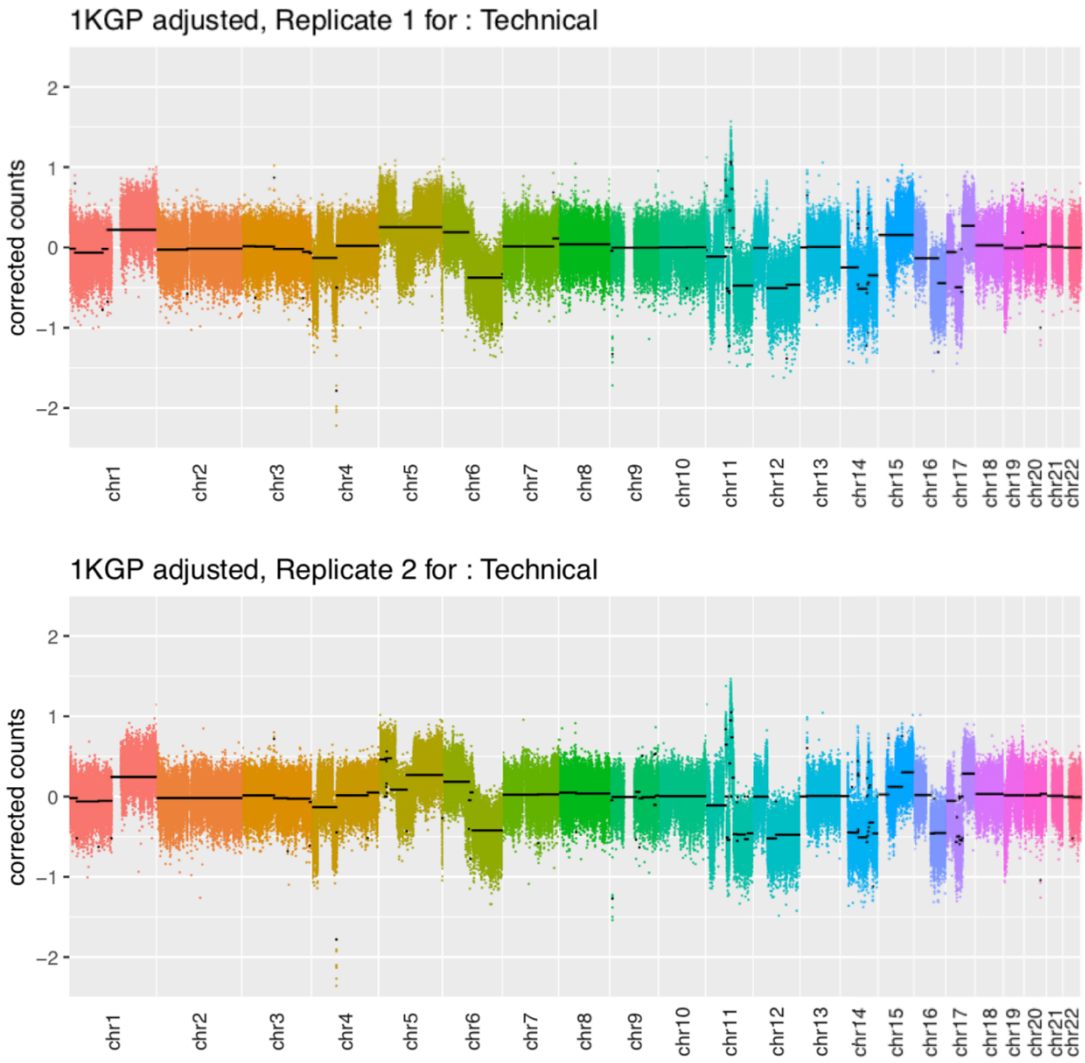


Figure B.16. Replicates for Technical: Pair 6

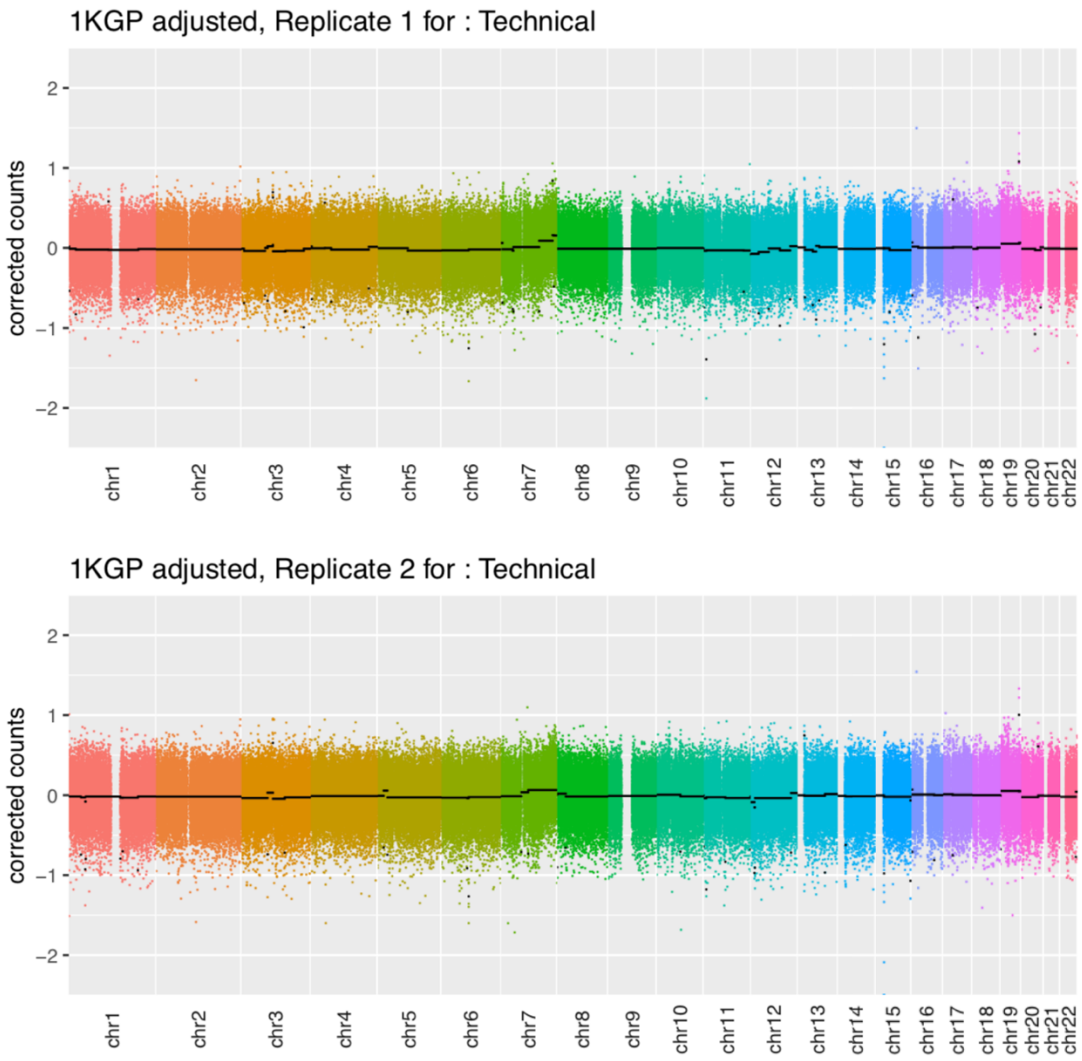


Figure B.17. Replicates for Technical: Pair 7

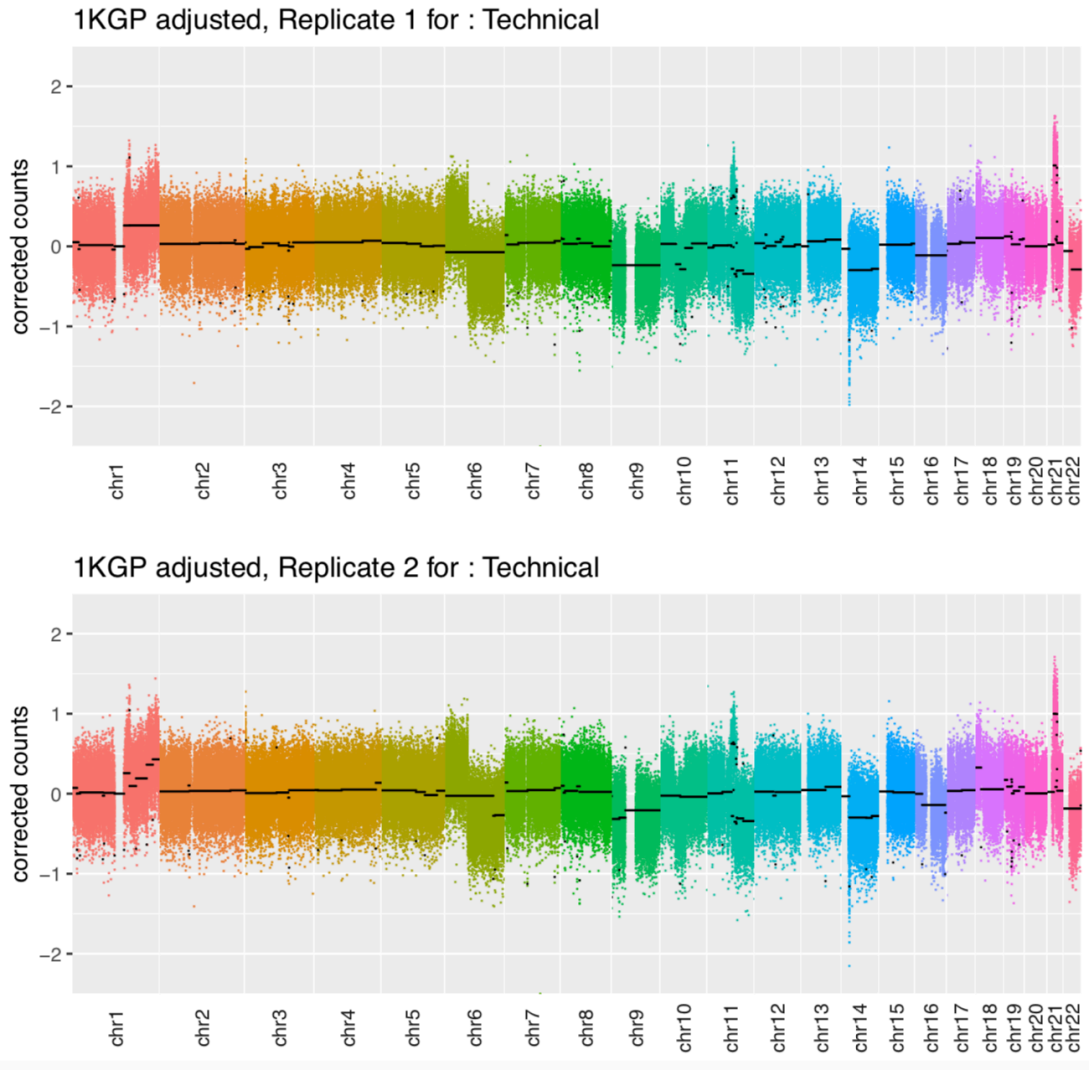


Figure B.18. Replicates for Technical: Pair 8

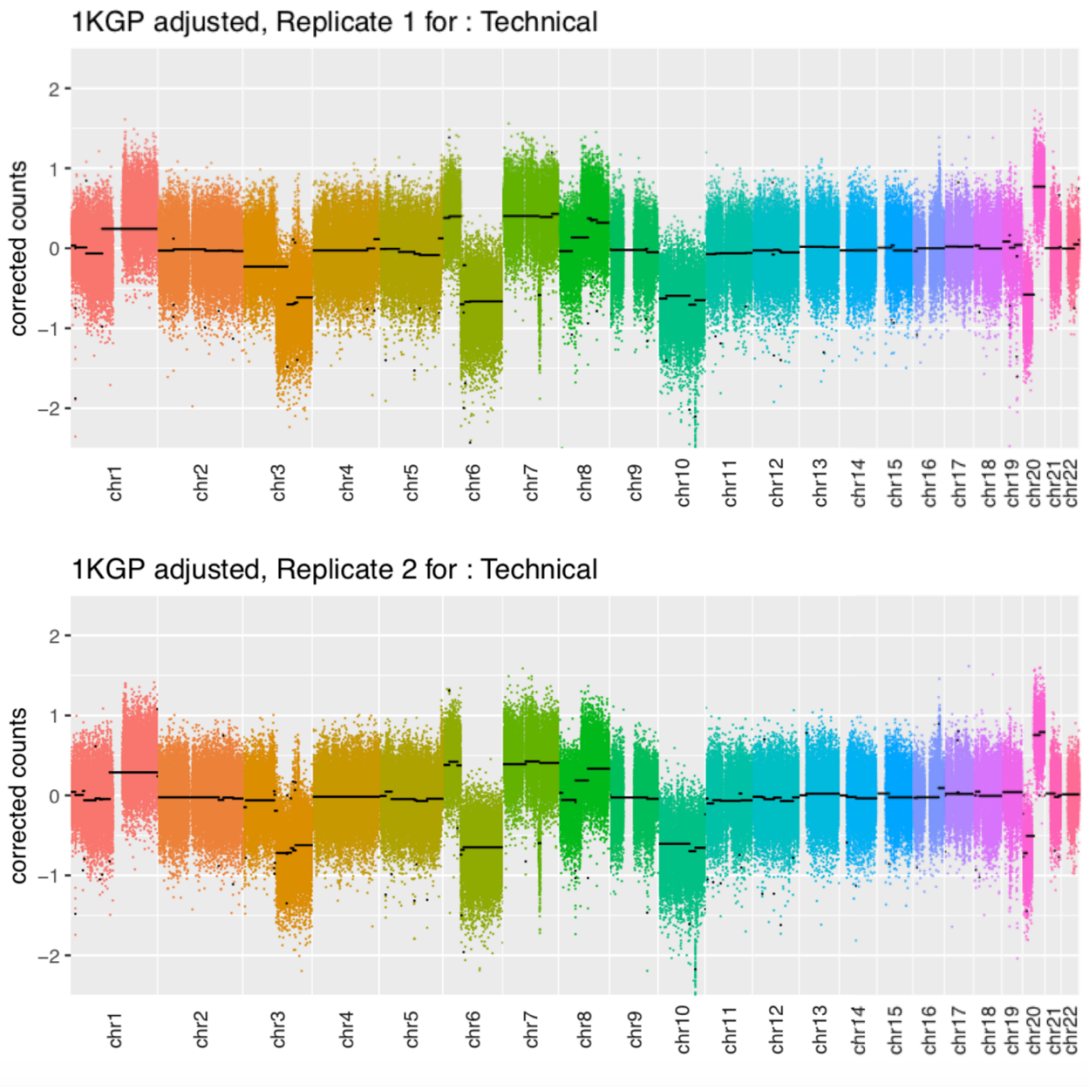


Figure B.19. Replicates for Technical: Pair 9

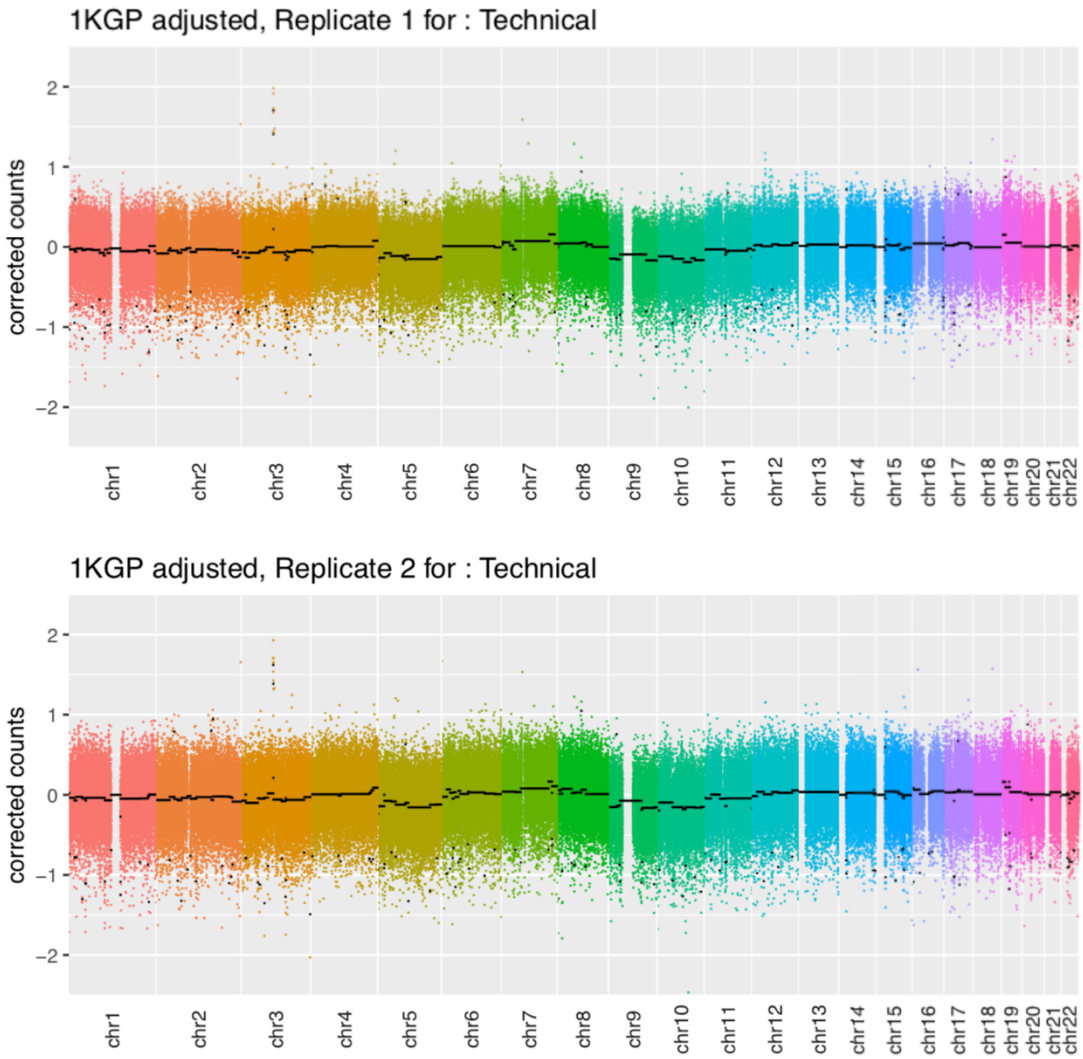


Figure B.20. Replicates for Technical: Pair 10

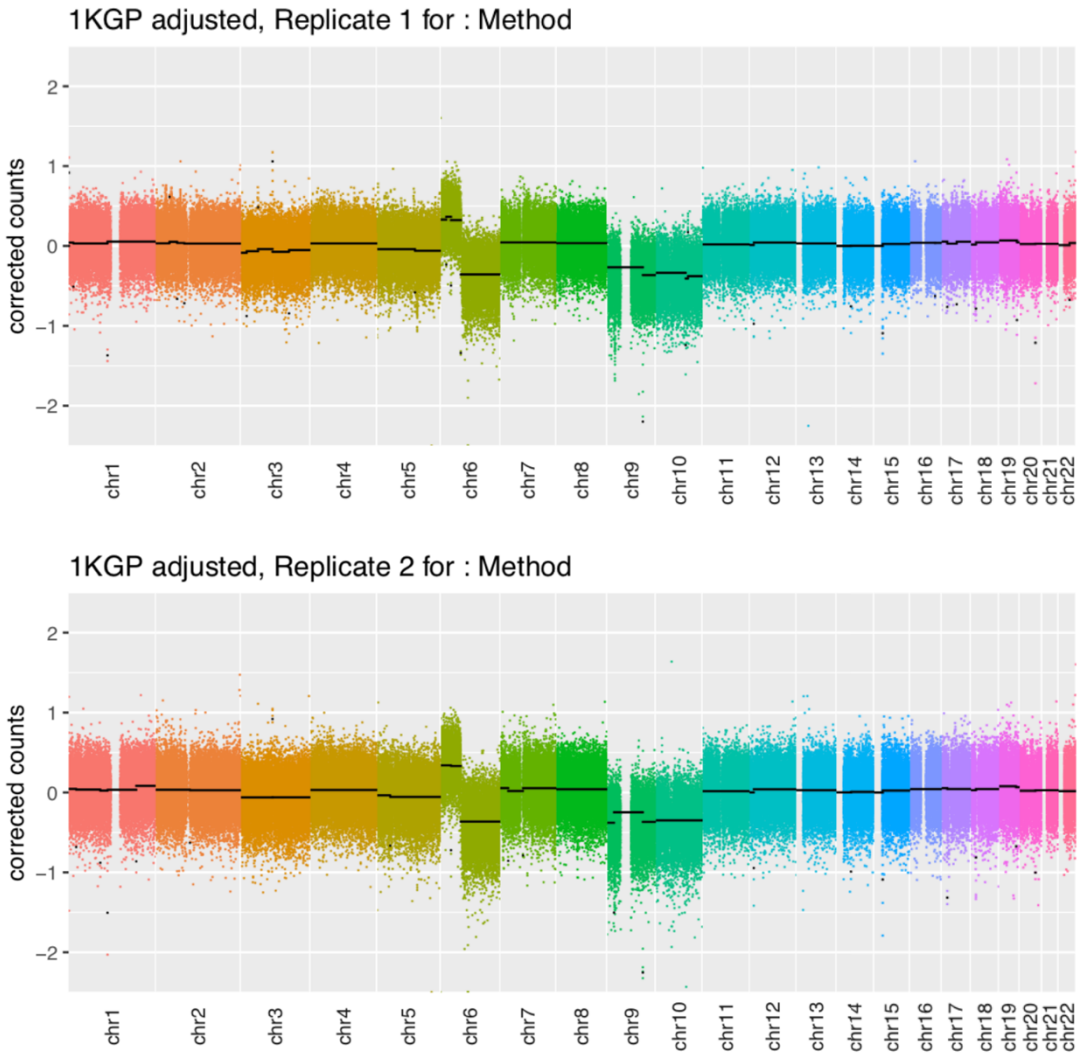


Figure B.21. Replicates for Method: Pair 2

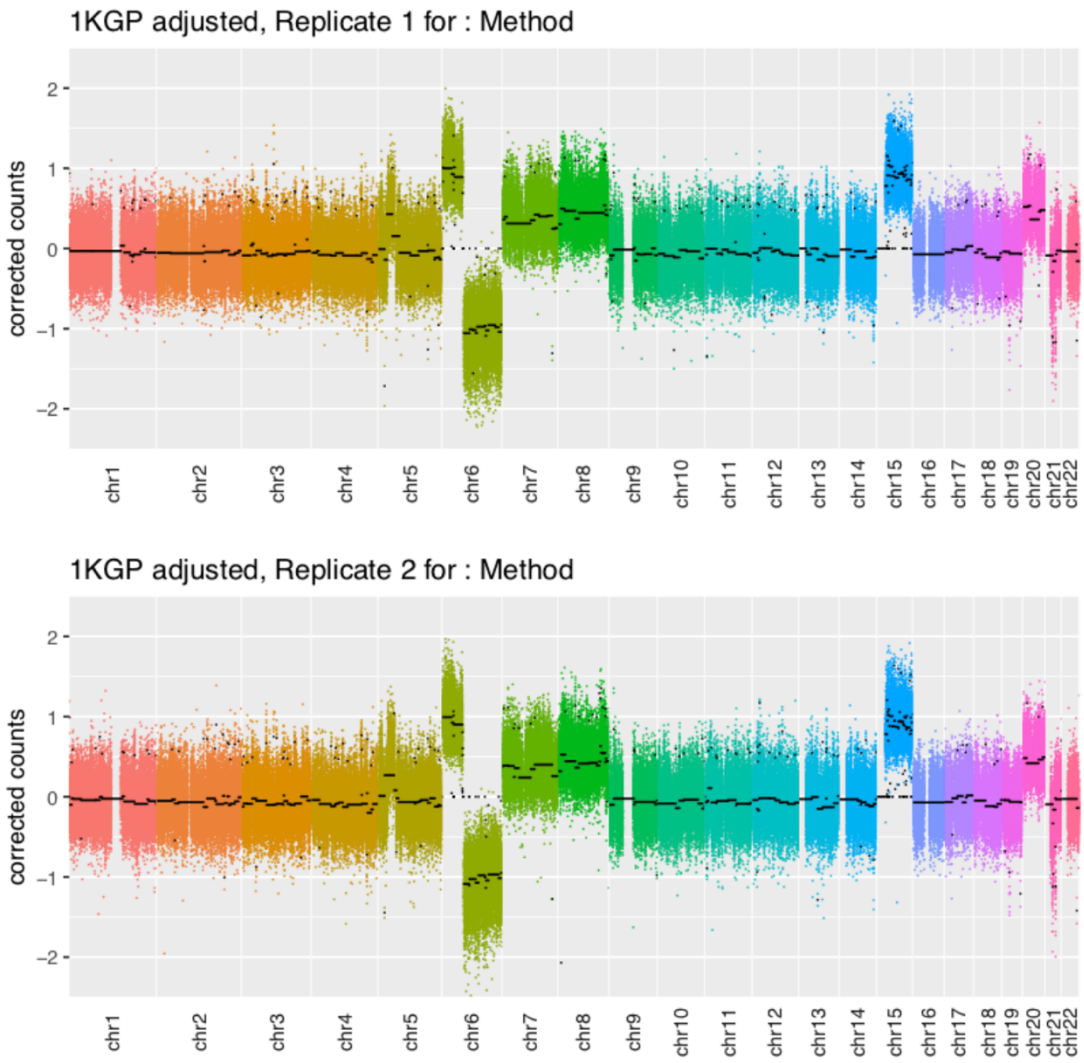


Figure B.22. Replicates for Method: Pair 3

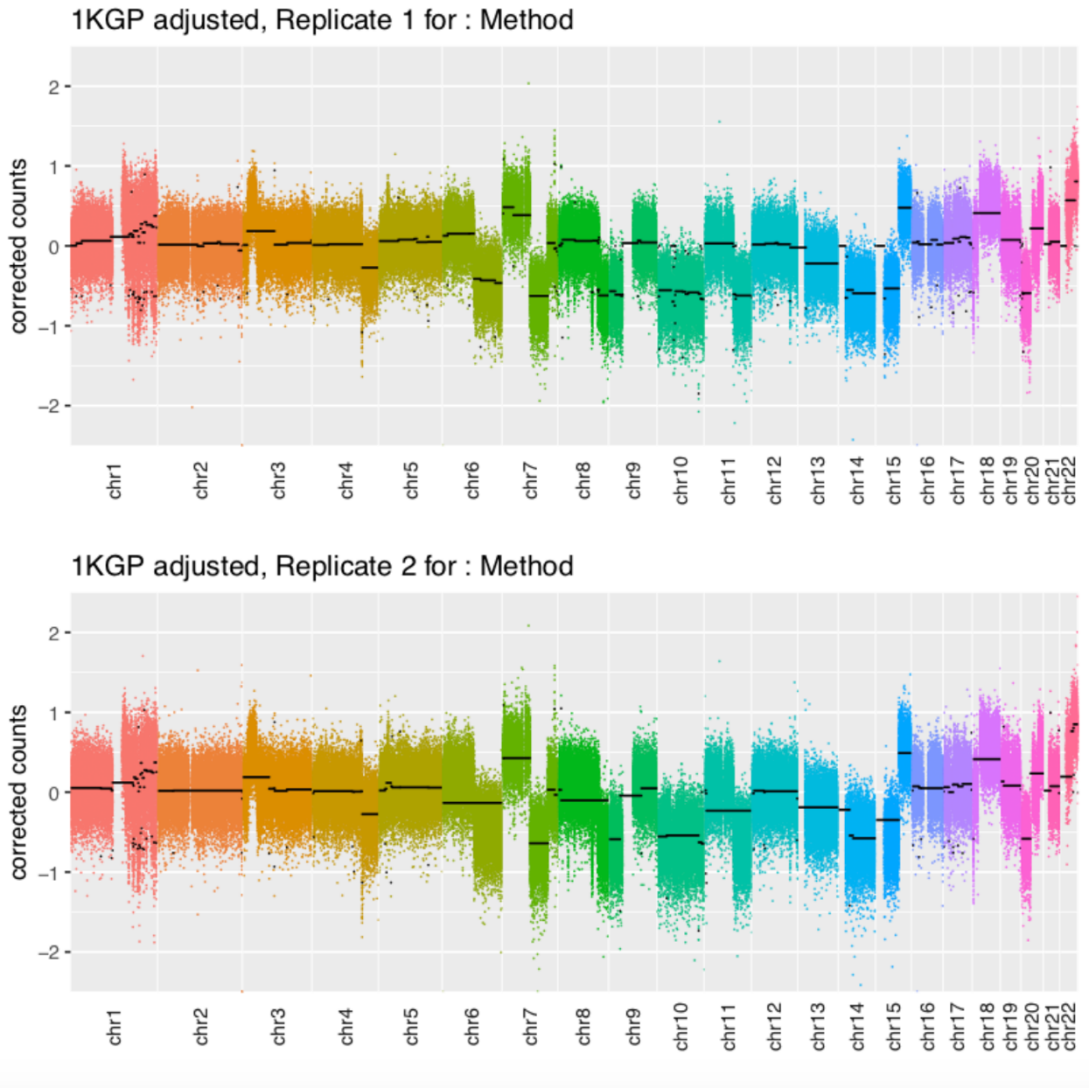


Figure B.23. Replicates for Method: Pair 4

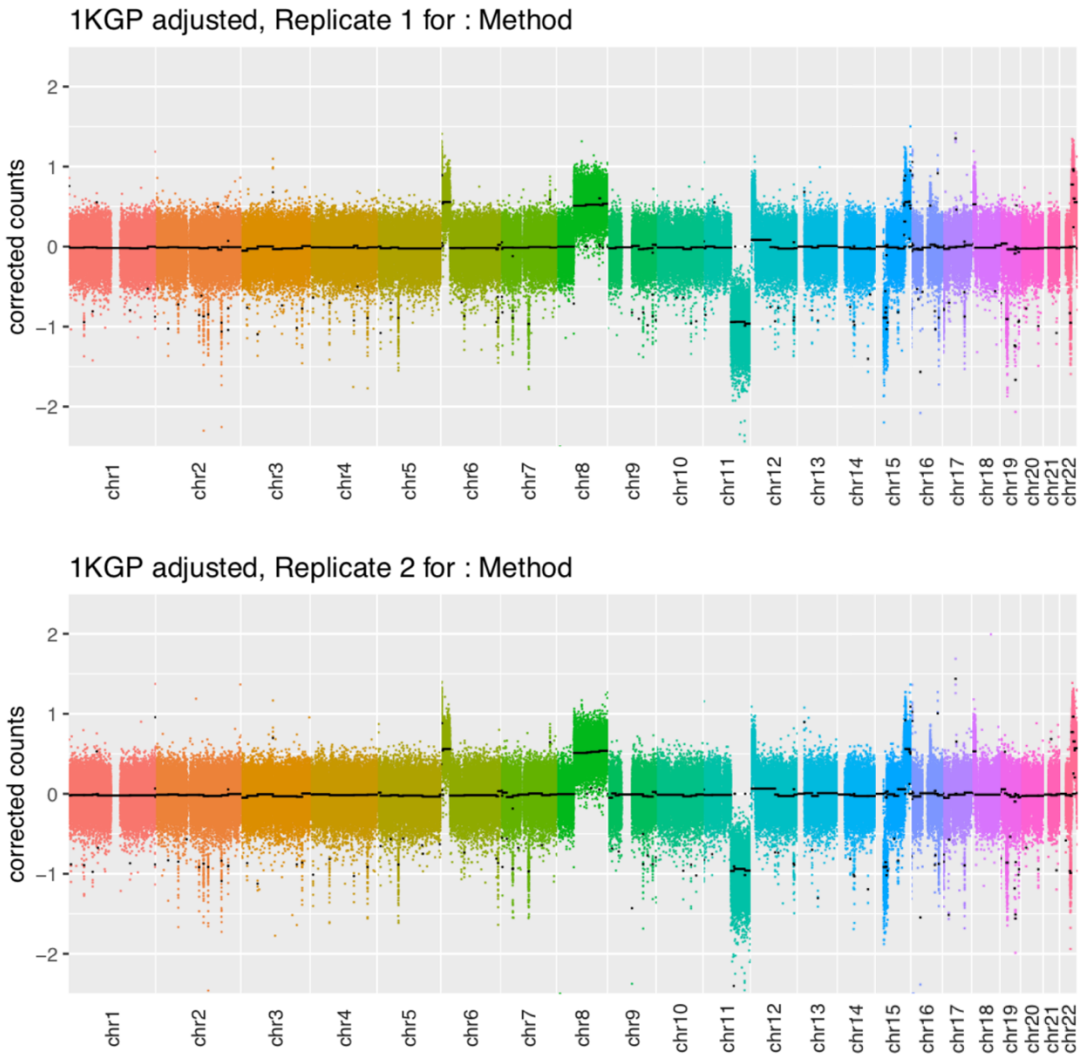


Figure B.24. Replicates for Method: Pair 5

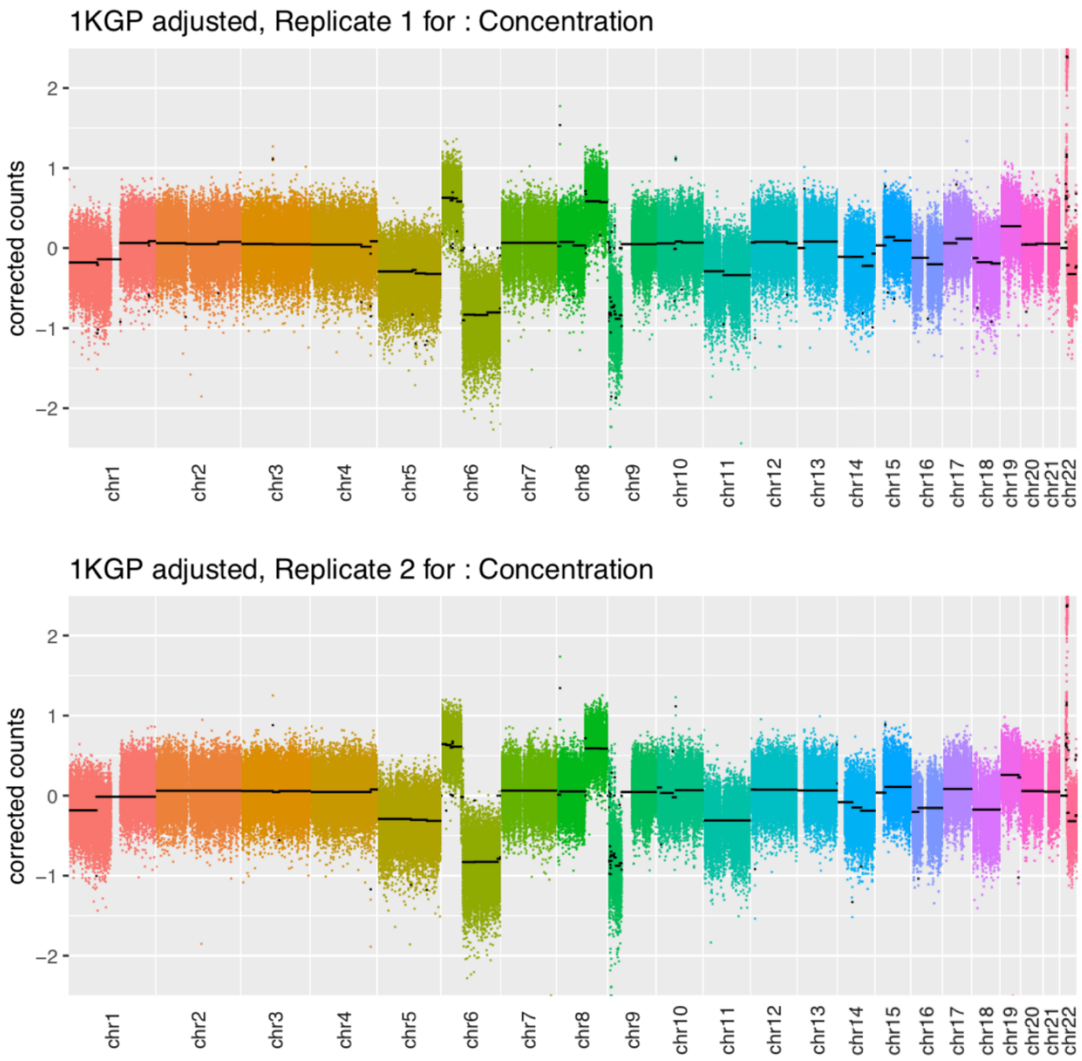


Figure B.25. Replicates for Concentration: Pair 2

Appendix C

C.1 Deletion in the *CDKN2A* region not identified in both the old and new LMC CNA data

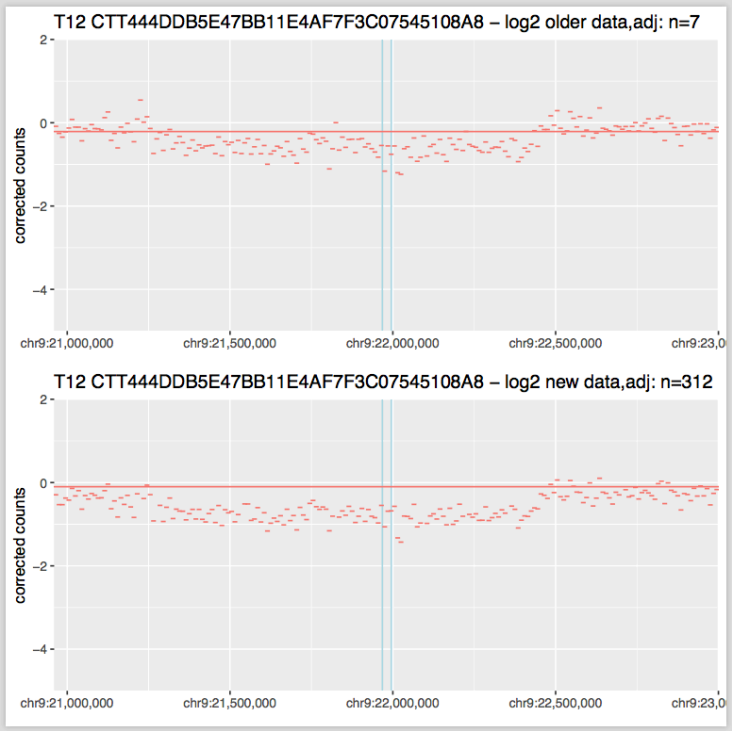


Figure C.1. The *CDKN2A* region for Sample 12

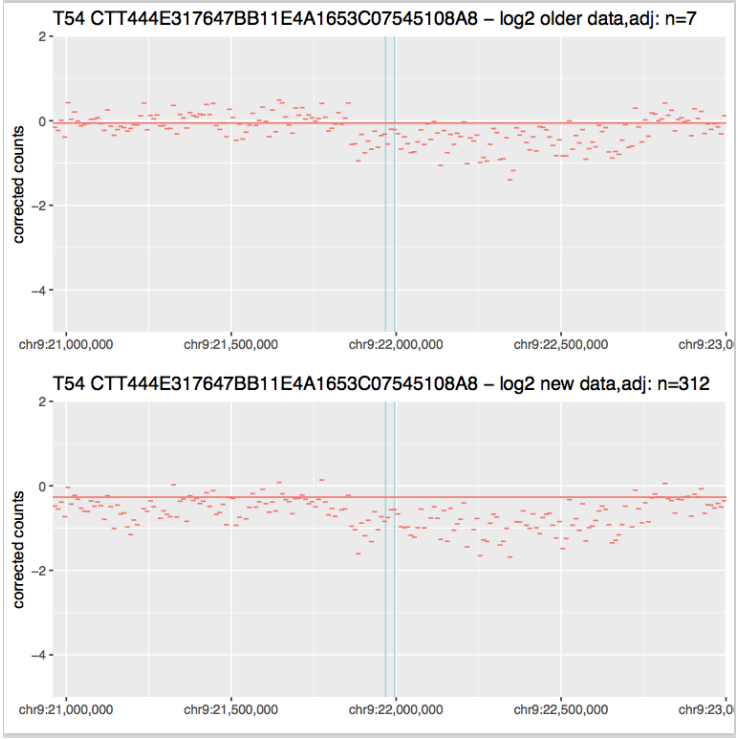


Figure C.2. The *CDKN2A* region for Sample 54

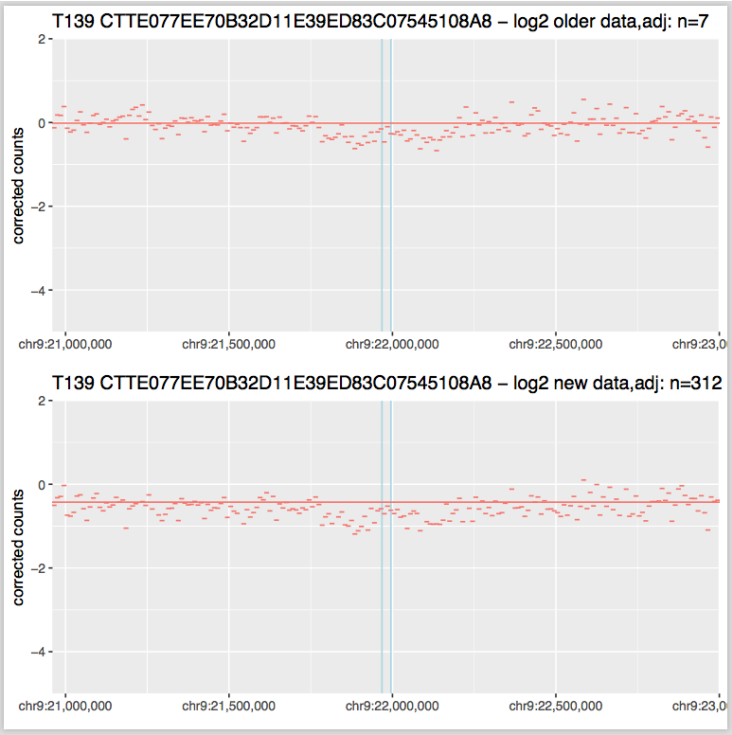


Figure C.3. The *CDKN2A* region for Sample 139

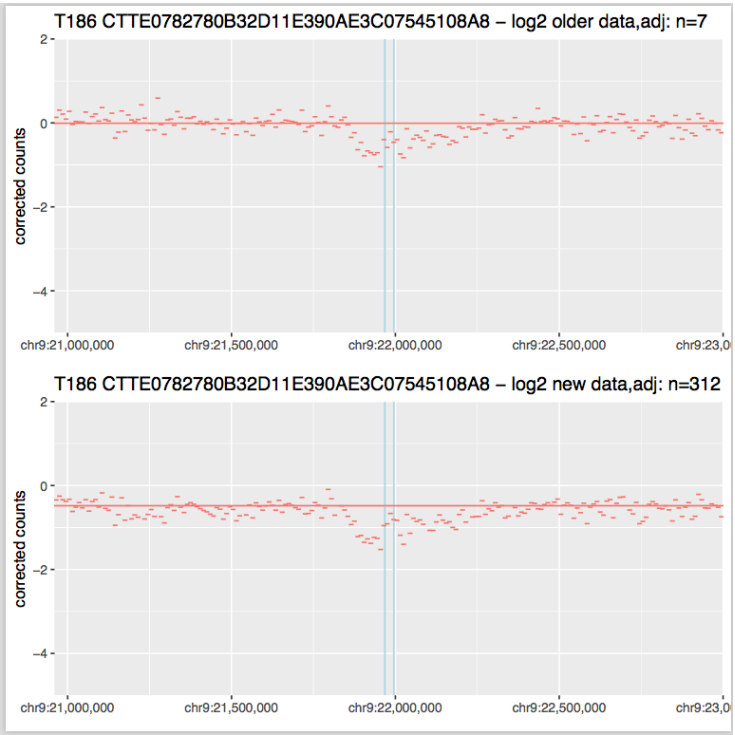


Figure C.4. The *CDKN2A* region for Sample 186

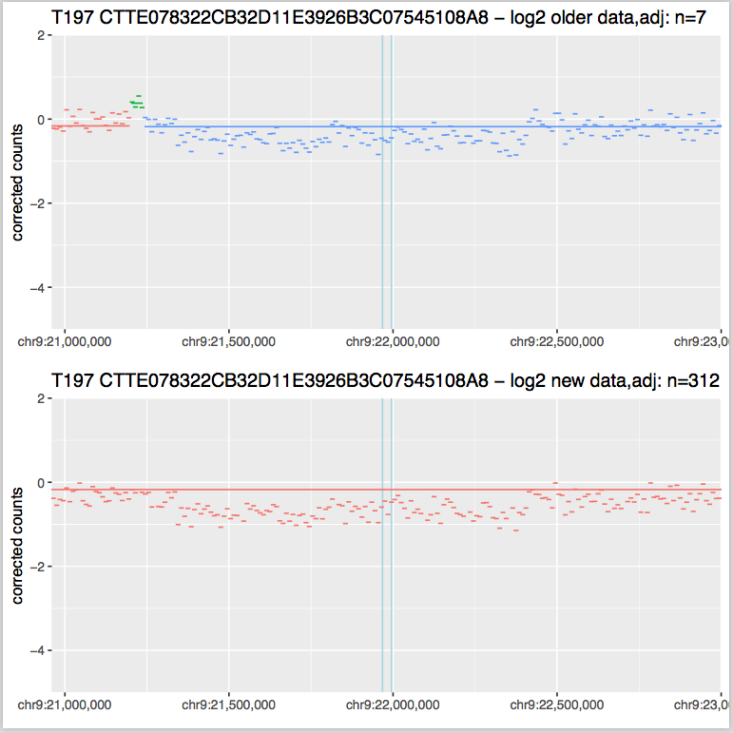


Figure C.5. The *CDKN2A* region for Sample 197

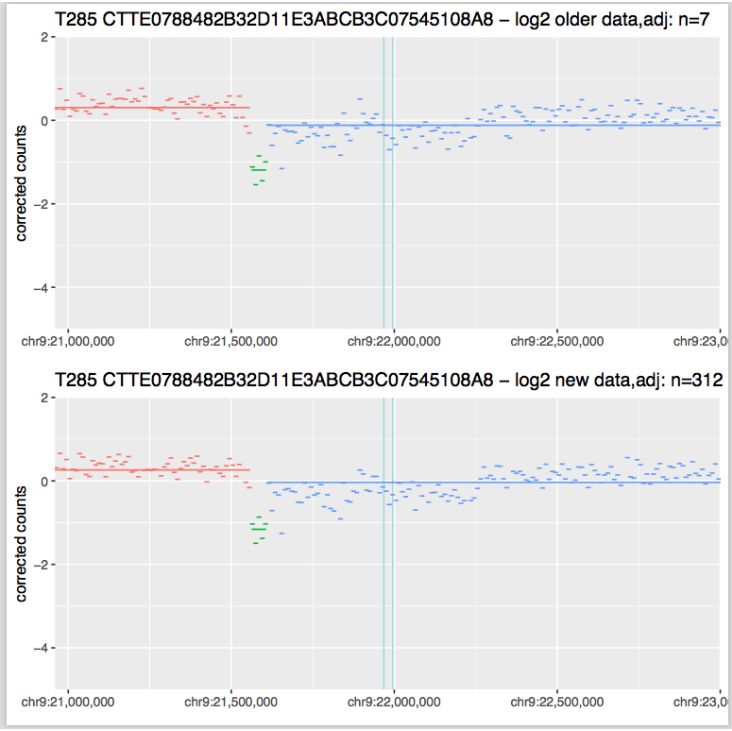


Figure C.6. The *CDKN2A* region for Sample 285

Appendix D

D.1 Whole Genome Comparison of LMC and TCGA Data

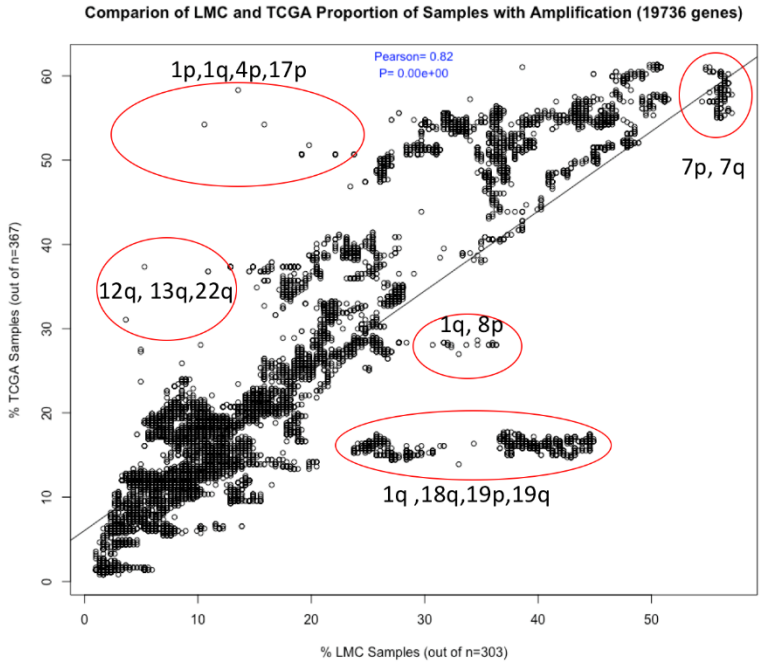


Figure D.1. Whole genome comparison of rates of amplification among the genes common to both TCGA and LMC lists

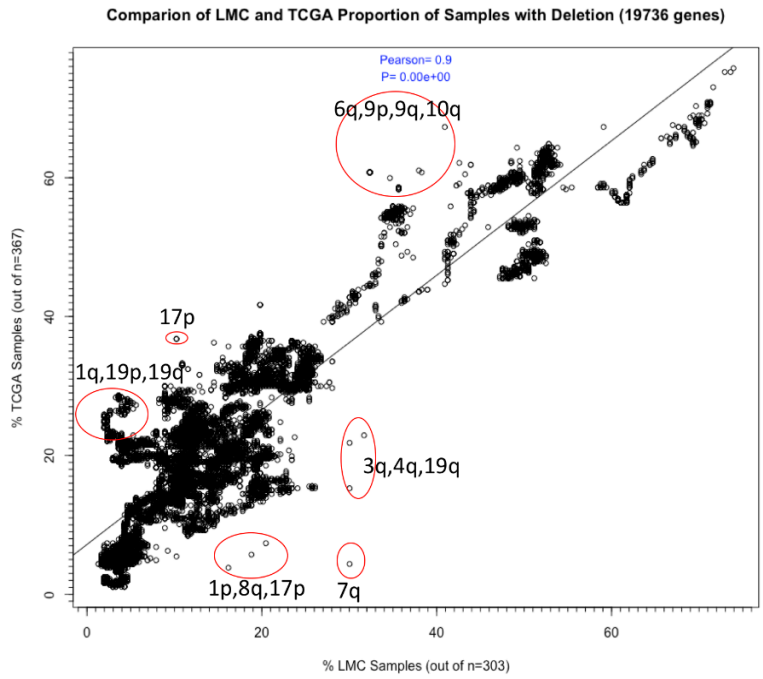


Figure D.2. Whole genome comparison of rates of deletion among the genes common to both TCGA and LMC lists

Appendix E

E.1 New Kundaje et.al. GRCh38 Blacklist

Table E.1. New GRCh38 Blacklist by Kundaje et al. (2020)

Location	Start	End	Size(kb)	Location	Start	End	Size(kb)
chr1	237945285	237946507	1.22	chr1	628903	635104	6.20
chr1	237948983	237949365	0.38	chr1	5850087	5850571	0.48
chr1	237951294	237951802	0.51	chr1	8909610	8910014	0.40
chr2	638427	638808	0.38	chr1	9574580	9574997	0.42
chr2	1087103	1087484	0.38	chr1	32043823	32044203	0.38
chr2	16271753	16272134	0.38	chr1	33818964	33819344	0.38
chr2	22316878	22317258	0.38	chr1	38674335	38674715	0.38
chr2	24644617	24644997	0.38	chr1	50017081	50017546	0.47
chr2	32916201	32916632	0.43	chr1	52996949	52997329	0.38
chr2	33767290	33767703	0.41	chr1	55372488	55372869	0.38
chr2	33964664	33965045	0.38	chr1	67971776	67972156	0.38
chr2	36276769	36277149	0.38	chr1	73258720	73259100	0.38
chr2	40784787	40785278	0.49	chr1	76971068	76971595	0.53
chr2	49229452	49230058	0.61	chr1	93936365	93936747	0.38
chr2	50588765	50589566	0.80	chr1	93937447	93937827	0.38
chr2	54451654	54452034	0.38	chr1	102160407	102160787	0.38
chr2	57648677	57649057	0.38	chr1	103620975	103621378	0.40
chr2	67953669	67954049	0.38	chr1	106803432	106803816	0.38
chr2	75063567	75063994	0.43	chr1	106804021	106804224	0.20
chr2	81666317	81666849	0.53	chr1	106804753	106805343	0.59
chr2	82814941	82815321	0.38	chr1	121609948	125063427	3453.48
chr2	82815451	82816236	0.79	chr1	125166231	125184683	18.45
chr2	82816261	82816647	0.39	chr1	143184599	143276861	92.26
chr2	82818378	82818748	0.37	chr1	146992422	146992802	0.38
chr2	82820800	82821005	0.21	chr1	158449073	158449453	0.38
chr2	85068666	85069046	0.38	chr1	158872114	158872494	0.38
chr2	87824709	87825530	0.82	chr1	159295111	159295493	0.38
chr2	89272789	89273133	0.34	chr1	169473895	169474338	0.44
chr2	89827607	89827706	0.10	chr1	170006204	170006584	0.38
chr2	89828636	89828710	0.07	chr1	172710350	172710732	0.38
chr2	89828842	89828942	0.10	chr1	181422611	181423158	0.55
chr2	89833685	89833793	0.11	chr1	191961694	191962163	0.47
chr2	89839592	89839709	0.12	chr1	195288048	195288429	0.38
chr2	89909317	89909789	0.47	chr1	199487949	199488149	0.20
chr2	90379778	90402456	22.68	chr1	214709795	214710175	0.38
chr2	92081223	92081398	0.18	chr1	215499615	215500014	0.40
chr2	92188125	94293463	2105.34	chr1	226652017	226652398	0.38
chr2	94499181	94570956	71.78	chr1	227699752	227700133	0.38
chr2	94898976	94899645	0.67	chr1	229019365	229019745	0.38
chr2	94900639	94900840	0.20	chr1	233139985	233140365	0.38
chr2	94901421	94901808	0.39	chr1	235520204	235520404	0.20
chr2	97189431	97189813	0.38	chr1	235537405	235537785	0.38
chr2	102482582	102482962	0.38	chr1	235538899	235540112	1.21
chr2	102505606	102505987	0.38	chr1	235540243	235540623	0.38
chr2	110072034	110072434	0.40	chr1	235540886	235541649	0.76
chr2	110299106	110299346	0.24	chr1	235870625	235871005	0.38
chr2	116751234	116751614	0.38	chr1	237940595	237940979	0.38
chr2	116752004	116752448	0.44	chr1	237941045	237941514	0.47
chr2	116752517	116752897	0.38	chr1	237941893	237942746	0.85
chr2	117020171	117020552	0.38	chr1	237943028	237943416	0.39
chr2	117021107	117022152	1.05	chr1	237943490	237945232	1.74

Appendix F

F.1 Plots of the normal samples

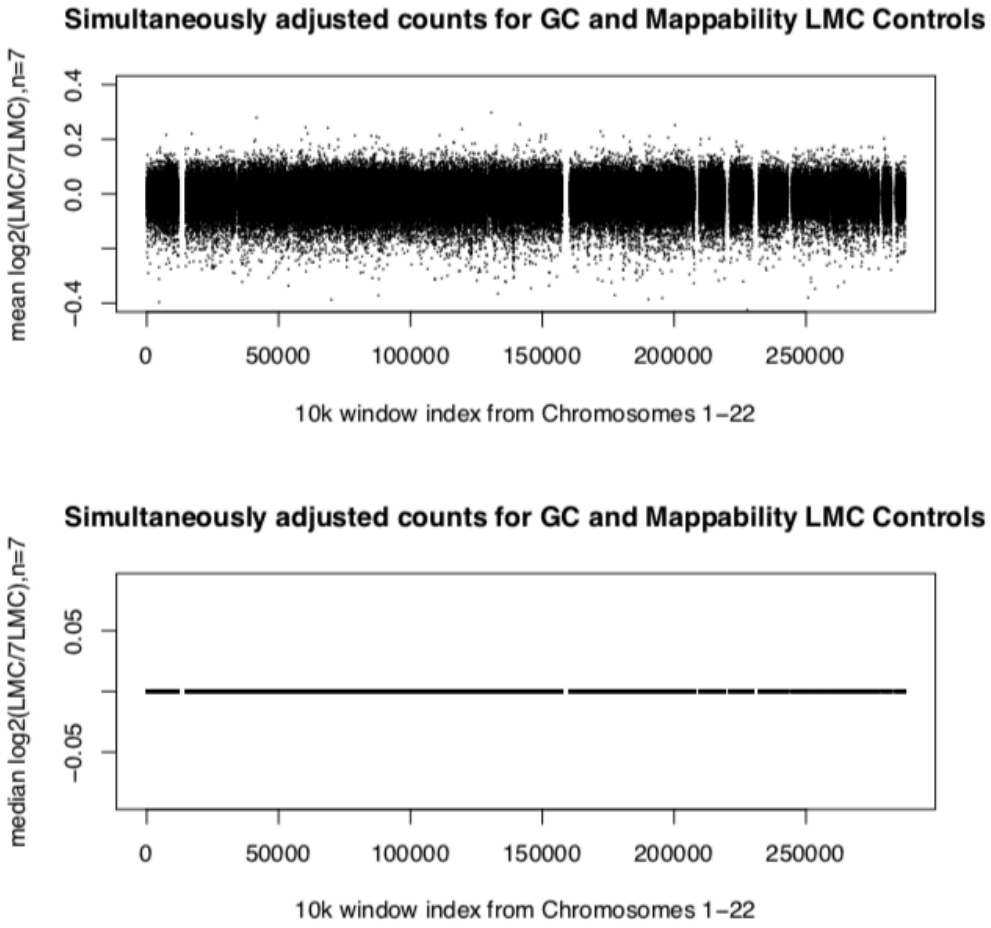


Figure F.1. Mean (top) and median (bottom) read counts per window of the 7 normal samples

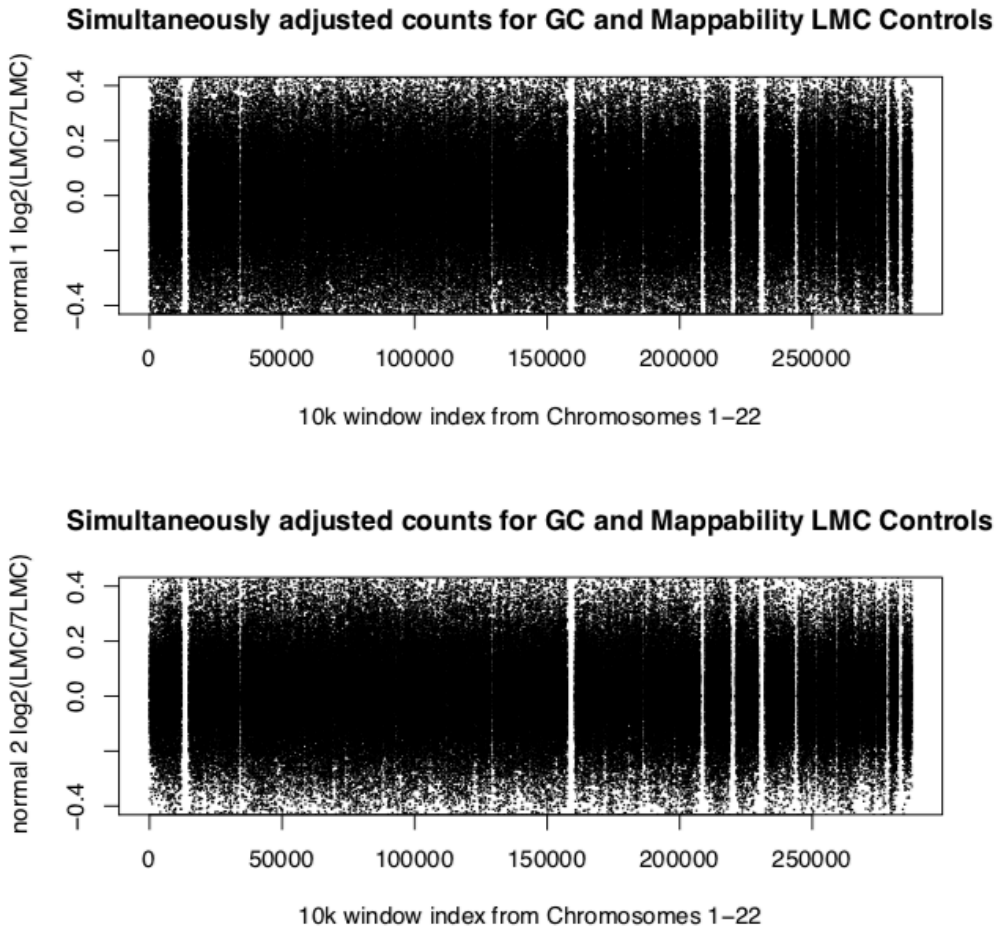


Figure F.2. Individual copy number profile for 2 normal samples.

References

1. Filia, A., et al., *High-Resolution Copy Number Patterns From Clinically Relevant FFPE Material.* Scientific Reports, 2019. **ISSN 2045-2322**(In Press).
2. Scheinin, I., et al., *DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly.* Genome Res, 2014. **24**(12): p. 2022-32.
3. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
4. Mermel, C.H., et al., *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol, 2011. **12**(4): p. R41.
5. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer Discov, 2012. **2**(5): p. 401-4.
6. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
7. *Picture of the Skin. Available at : https://www.webmd.com/skin-problems-and-treatments/picture-of-the-skin#1 (Accessed on March 21, 2019).*
8. *Layers of the Skin. Available at https://opentextbc.ca/anatomyandphysiology/chapter/5-1-layers-of-the-skin/ (Accessed on March 21, 2019).*
9. *How Much Does Your Skin Weigh? . Available at : https://www.livescience.com/32939-how-much-does-skin-weigh.html (Accessed: March 23, 2019).*
10. *What is the skin. Available at: https://www.clinimed.co.uk/wound-care/wound-essentials/structure-and-function-of-the-skin (Accessed : March 23, 2019).*
11. *Layers of the Skin : Anatomy & Physiology by Edited and Revised by Lindsay M. Biga, Sierra Dawson, Amy Harwell, Robin Hopkins, Joel Kaufmann, Mike LeMaster, Philip Matern, Katie Morrison-Graham, Devon Quick, Jon Runyeon Art edited and created by Leeah Whittier is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise. Avaialble at : http://library.open.oregonstate.edu/aandp/chapter/5-1-layers-of-the-skin/#fig-ch05_01_01 (Accessed: March 23, 2019).*
12. *Layers of the Skin. Available at https://visualsonline.cancer.gov/details.cfm?imageid=4366 (Accessed : March 22, 2019).*
13. *Cells in the dermis and epidermis. https://www.histology.leeds.ac.uk/skin/epidermis_cells.php (Accessed : March 28, 2019).*
14. Yung, A., *Structure of normal skin . Available at : https://www.dermnetnz.org/topics/the-structure-of-normal-skin (Accessed on March 28, 2019).* 2007.
15. Green, H. and P. Djian, *Consecutive actions of different gene-altering mechanisms in the evolution of involucrin.* Mol Biol Evol, 1992. **9**(6): p. 977-1017.
16. Yamaguchi, Y. and V.J. Hearing, *Melanocytes and their diseases.* Cold Spring Harb Perspect Med, 2014. **4**(5).
17. Simon, J.D., et al., *Current challenges in understanding melanogenesis: bridging chemistry, biological control, morphology, and function.* Pigment Cell Melanoma Res, 2009. **22**(5): p. 563-79.
18. Hearing, V.J., *Determination of melanin synthetic pathways.* J Invest Dermatol, 2011. **131**(E1): p. E8-E11.

19. Kondo, T. and V.J. Hearing, *Update on the regulation of mammalian melanocyte function and skin pigmentation.* Expert Rev Dermatol, 2011. **6**(1): p. 97-108.
20. Wasmeier, C., et al., *Melanosomes at a glance.* J Cell Sci, 2008. **121**(Pt 24): p. 3995-9.
21. Langerhans, P., *Uber die nerven der menschlichen haut.* Archives of Pathological Anatomy, 1868. **44**: p. 325–37.
22. Jolles, S., *Paul Langerhans.* J Clin Pathol, 2002. **55**(4): p. 243.
23. Silberberg, I., *Apposition of mononuclear cells to langerhans cells in contact allergic reactions. An ultrastructural study.* Acta Derm Venereol, 1973. **53**(1): p. 1-12.
24. Chomiczewska, D., et al., *[The role of Langerhans cells in the skin immune system].* Pol Merkur Lekarski, 2009. **26**(153): p. 173-7.
25. *Macmillan Cancer Support. (2017) Types of melanoma - understanding - Macmillan cancer support. Available at: http://www.macmillan.org.uk/information-and-support/melanoma/understanding-cancer/melanoma-types.html. (Accessed: 11 February 2017).*
26. Korchagin, K., MELANOMA BIOLOGY. [online] Cancerlink.ru. Available at: http://cancerlink.ru/enmelbiology.html. (Accessed 15 April 2020).
27. *Acquired melanocytic naevus. Available at : https://dermnetnz.org/topics/melanocytic-naevi-pathology/ (Accessed 13 May 2020). .*
28. *What is a nevus? . Available at: https://www.healthline.com/health/nevus (Accessed 13 May 2020).*
29. *What is a mole? Available at: https://dermnetnz.org/topics/mole/ (Accessed 13 May 2020).*
30. Urso, C., *Are growth phases exclusive to cutaneous melanoma?* J Clin Pathol, 2004. **57**(5): p. 560.
31. Clark, W.H., Jr., et al., *A study of tumor progression: the precursor lesions of superficial spreading and nodular melanoma.* Hum Pathol, 1984. **15**(12): p. 1147-65.
32. 2020), U.Y.P.R.M.A.a.h.w.o.o.c.s.m.t.u.-y.-p.-r.-m.A.o.M.
33. 2020), S.c.m.A.a.h.w.n.u.c.m.-s.-c.A.M.
34. Auslender, S., et al., *Lentigo maligna and superficial spreading melanoma are different in their in situ phase: an immunohistochemical study.* Hum Pathol, 2002. **33**(10): p. 1001-5.
35. Warren, M.P. and V.M. Harvey, *IMAGES IN CLINICAL MEDICINE. Acral Lentiginous Melanoma.* N Engl J Med, 2015. **373**(19): p. 1864.
36. *Skin Cancer. Available at : https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data (Accessed: March 29, 2019).*
37. Nikolaou, V. and A.J. Stratigos, *Emerging trends in the epidemiology of melanoma.* Br J Dermatol, 2014. **170**(1): p. 11-9.
38. *Cancer Research UK. (2019). Melanoma Skin Cancer Incidence Statistics. Available at : https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer#heading-Zero (Accessed: March 29, 2019).*
39. *WHO (2016) Health effects of UV radiation. Available at: http://www.who.int/uv/health/uv_health2/en/index1.html. (Accessed: 31, January, 2017).*
40. *LINC02206 long intergenic non-protein coding RNA 2206 [Homo sapiens (human)]. Available at : https://www.ncbi.nlm.nih.gov/gene/?term=LINC02206. (Accessed on July 4, 2020).*
41. *Estimated age-standardized incidence rates (World) in 2018, melanoma of skin, both sexes, all ages. Available at: http://gco.iarc.fr/today/online-analysis-map?v=2018&mode=population&mode_population=continents&population=9*

00&populations=900&key=asr&sex=0&cancer=16&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=5&group_cancer=1&include_nmsc=1&include_nmsc_other=1&projection=globe&color_palette=default&map_scale=quantile&map_nb_colors=5&continent=0&rotate=%255B10%252C0%255D . *(Accessed: April 22, 2019).*

42. Chang, C., et al., *More skin, more sun, more tan, more melanoma.* Am J Public Health, 2014. **104**(11): p. e92-9.
43. Gandini, S., P. Autier, and M. Boniol, *Reviews on sun exposure and artificial light and melanoma.* Prog Biophys Mol Biol, 2011. **107**(3): p. 362-6.
44. Marks, R. and D. Whiteman, *Sunburn and melanoma: how strong is the evidence?* BMJ, 1994. **308**(6921): p. 75-6.
45. Whiteman, D. and A. Green, *Melanoma and sunburn.* Cancer Causes Control, 1994. **5**(6): p. 564-72.
46. Fargnoli, M.C., et al., *Constitutional and environmental risk factors for cutaneous melanoma in an Italian population. A case-control study.* Melanoma Res, 2004. **14**(2): p. 151-7.
47. Whiteman, D.C., et al., *Melanocytic nevi, solar keratoses, and divergent pathways to cutaneous melanoma.* J Natl Cancer Inst, 2003. **95**(11): p. 806-12.
48. Cho, E., B.A. Rosner, and G.A. Colditz, *Risk factors for melanoma by body site.* Cancer Epidemiol Biomarkers Prev, 2005. **14**(5): p. 1241-4.
49. *Cancer.Net. 2021. Familial Malignant Melanoma. [online] Available at: <https://www.cancer.net/cancer-types/familial-malignant-melanoma> [Accessed 28 June 2021].*
50. Goldstein, A.M., et al., *Features associated with germline CDKN2A mutations: a GenoMEL study of melanoma-prone families from three continents.* J Med Genet, 2007. **44**(2): p. 99-106.
51. Goldstein, A.M. and M.A. Tucker, *Screening for CDKN2A mutations in hereditary melanoma.* J Natl Cancer Inst, 1997. **89**(10): p. 676-8.
52. Kefford, R.F., et al., *Counseling and DNA testing for individuals perceived to be genetically predisposed to melanoma: A consensus statement of the Melanoma Genetics Consortium.* J Clin Oncol, 1999. **17**(10): p. 3245-51.
53. Helgadottir, H., et al., *Germline CDKN2A Mutation Status and Survival in Familial Melanoma Cases.* J Natl Cancer Inst, 2016. **108**(11).
54. Balch, C.M., et al., *Final version of 2009 AJCC melanoma staging and classification.* J Clin Oncol, 2009. **27**(36): p. 6199-206.
55. *Clark and Breslow staging. Available at : https://www.cancerresearchuk.org/about-cancer/melanoma/stages-types/clark-breslow-staging. (Accessed: March 19, 2019).*
56. *TNM Staging. Available at : https://www.cancerresearchuk.org/about-cancer/melanoma/stages-types/tnm-staging . (Accessed : March 19, 2019).*
57. *Guide to Staging - Melanoma. Available at : https://www.skincancer.org/skin-cancer-information/melanoma/the-stages-of-melanoma/guide-to-staging-melanoma (Accessed : Marc 20, 2019)*
58. Jewell, R., et al., *The clinicopathological and gene expression patterns associated with ulceration of primary melanoma.* Pigment Cell Melanoma Res, 2015. **28**(1): p. 94-104.
59. *Sentinel Lymph Node Biopsy. Available at : https://www.cancer.gov/about-cancer/diagnosis-staging/staging/sentinel-node-biopsy-fact-sheet#what-are-lymph-nodes . (Accessed on 15 May 2020).*
60. *Lactate dehydrogenase (LDH) test. Available at: https://www.nhs.uk/conditions/ldh-test/. (Accessed : March 20, 2019).*
61. Finkelstein, J.A. and H. Kreder, *Spine stats. The Cox regression analysis.* Spine J, 2001. **1**(5): p. 382.
62. Salamzadeh, J., et al., *A Cox regression analysis of covariates for asthma hospital readmissions.* J Asthma, 2003. **40**(6): p. 645-52.

63. Zhang, W.J., D.H. Yang, and J.C. Shu, *[Cox regression analysis of predictive factors of hepatorenal syndrome].* Zhonghua Gan Zang Bing Za Zhi, 2003. **11**(10): p. 586-7.
64. Colombino, M., et al., *BRAF/NRAS mutation frequencies among primary tumors and metastases in patients with melanoma.* J Clin Oncol, 2012. **30**(20): p. 2522-9.
65. Ekedahl, H., et al., *The clinical significance of BRAF and NRAS mutations in a clinic-based metastatic melanoma cohort.* Br J Dermatol, 2013. **169**(5): p. 1049-55.
66. Bauer, J., et al., *BRAF mutations in cutaneous melanoma are independently associated with age, anatomic site of the primary tumor, and the degree of solar elastosis at the primary tumor site.* Pigment Cell Melanoma Res, 2011. **24**(2): p. 345-51.
67. Pollock, P.M., et al., *High frequency of BRAF mutations in nevi.* Nat Genet, 2003. **33**(1): p. 19-20.
68. Davies, H., et al., *Mutations of the BRAF gene in human cancer.* Nature, 2002. **417**(6892): p. 949-54.
69. Omholt, K., et al., *NRAS and BRAF mutations arise early during melanoma pathogenesis and are preserved throughout tumor progression.* Clin Cancer Res, 2003. **9**(17): p. 6483-8.
70. Hodis, E., et al., *A landscape of driver mutations in melanoma.* Cell, 2012. **150**(2): p. 251-63.
71. *The Cancer Genome Atlas. Skin Cutaneous Melanoma (TCGA, provisional). Available at: http://www.cbioportal.org/study?id=skcm_tcga#cna (Accessed: 1 February 2017)*.
72. Cancer Genome Atlas, N., *Genomic Classification of Cutaneous Melanoma.* Cell, 2015. **161**(7): p. 1681-96.
73. Stephens, P.J., et al., *Massive genomic rearrangement acquired in a single catastrophic event during cancer development.* Cell, 2011. **144**(1): p. 27-40.
74. Maher, C.A. and R.K. Wilson, *Chromothripsis and human disease: piecing together the shattering process.* Cell, 2012. **148**(1-2): p. 29-32.
75. Forment, J.V., A. Kaidi, and S.P. Jackson, *Chromothripsis and cancer: causes and consequences of chromosome shattering.* Nat Rev Cancer, 2012. **12**(10): p. 663-70.
76. Kloosterman, W.P., et al., *Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer.* Genome Biol, 2011. **12**(10): p. R103.
77. Parker, M., et al., *C11orf95-RELA fusions drive oncogenic NF-kappaB signalling in ependymoma.* Nature, 2014. **506**(7489): p. 451-5.
78. Anderson, N.D., et al., *Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors.* Science, 2018. **361**(6405).
79. McDermott, D.H., J.L. Gao, and P.M. Murphy, *Chromothriptic cure of WHIM syndrome: Implications for bone marrow transplantation.* Rare Dis, 2015. **3**(1): p. e1073430.
80. Nik-Zainal, S., et al., *Mutational processes molding the genomes of 21 breast cancers.* Cell, 2012. **149**(5): p. 979-93.
81. Lada, A.G., et al., *AID/APOBEC cytosine deaminase induces genome-wide kataegis.* Biol Direct, 2012. **7**: p. 47; discussion 47.
82. Burns, M.B., N.A. Temiz, and R.S. Harris, *Evidence for APOBEC3B mutagenesis in multiple human cancers.* Nat Genet, 2013. **45**(9): p. 977-83.
83. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-454.
84. Zarrei, M., et al., *A copy number variation map of the human genome.* Nature Reviews Genetics, 2015. **16**(3): p. 172-183.
85. Shah, S.P., et al., *Modeling recurrent DNA copy number alterations in array CGH data.* Bioinformatics, 2007. **23**(13): p. i450-8.

86. Wu, H.T., I. Hajirasouliha, and B.J. Raphael, *Detecting independent and recurrent copy number aberrations using interval graphs.* Bioinformatics, 2014. **30**(12): p. i195-203.
87. Shlien, A. and D. Malkin, *Copy number variations and cancer.* Genome Med, 2009. **1**(6): p. 62.
88. Li, W., A. Lee, and P.K. Gregersen, *Copy-number-variation and copy-number-alteration region detection by cumulative plots.* BMC Bioinformatics, 2009. **10 Suppl 1**: p. S67.
89. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.
90. Cachia, A.R., et al., *CDKN2A mutation and deletion status in thin and thick primary melanoma.* Clin Cancer Res, 2000. **6**(9): p. 3511-5.
91. Sauter, E.R., et al., *Cyclin D1 is a candidate oncogene in cutaneous melanoma.* Cancer Res, 2002. **62**(11): p. 3200-6.
92. Curtin, J.A., et al., *Somatic activation of KIT in distinct subtypes of melanoma.* J Clin Oncol, 2006. **24**(26): p. 4340-6.
93. Wilson, M.A., et al., *Copy Number Changes Are Associated with Response to Treatment with Carboplatin, Paclitaxel, and Sorafenib in Melanoma.* Clin Cancer Res, 2016. **22**(2): p. 374-82.
94. Li, W.L. and M. Olivier, *Current analysis platforms and methods for detecting copy number variation.* Physiological Genomics, 2013. **45**(1): p. 1-16.
95. Kalendar, R., et al., *Copy-number variation of housekeeping gene rpl13a in rat strains selected for nervous system excitability.* Mol Cell Probes, 2017. **33**: p. 11-15.
96. Bentley, D.R., et al., *The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X.* Nature, 2001. **409**(6822): p. 942-3.
97. Cheung, V.G., et al., *Integration of cytogenetic landmarks into the draft sequence of the human genome.* Nature, 2001. **409**(6822): p. 953-8.
98. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.
99. Teo, S.M., et al., *Statistical challenges associated with detecting copy number variations with next-generation sequencing.* Bioinformatics, 2012. **28**(21): p. 2711-8.
100. Tattini, L., R. D'Aurizio, and A. Magi, *Detection of Genomic Structural Variants from Next-Generation Sequencing Data.* Front Bioeng Biotechnol, 2015. **3**: p. 92.
101. Magi, A., et al., *Read count approach for DNA copy number variants detection.* Bioinformatics, 2012. **28**(4): p. 470-8.
102. Earl, D., et al., *Assemblathon 1: a competitive assessment of de novo short read assembly methods.* Genome Res, 2011. **21**(12): p. 2224-41.
103. Xie, C. and M.T. Tammi, *CNV-seq, a new method to detect copy number variation using high-throughput sequencing.* BMC Bioinformatics, 2009. **10**: p. 80.
104. Newton-Bishop, J.A., et al., *25-Hydroxyvitamin D2 /D3 levels and factors associated with systemic inflammation and melanoma survival in the Leeds Melanoma Cohort.* Int J Cancer, 2015. **136**(12): p. 2890-9.
105. Davies, J.R., et al., *Development and validation of a melanoma risk score based on pooled data from 16 case-control studies.* Cancer Epidemiol Biomarkers Prev, 2015. **24**(5): p. 817-24.
106. Conway, C., et al., *Gene expression profiling of paraffin-embedded primary melanoma using the DASL assay identifies increased osteopontin expression as predictive of reduced relapse-free survival.* Clin Cancer Res, 2009. **15**(22): p. 6939-46.
107. Wood, H.M., et al., *Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens.* Nucleic Acids Res, 2010. **38**(14): p. e151.

108. Craig, D.W., et al., *Identification of genetic variants using bar-coded multiplexed sequencing.* Nat Methods, 2008. **5**(10): p. 887-93.
109. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
110. *Broad Institute. (Accessed: 2018/02/21; version 2.17.8). "Picard Tools." Broad Institute, GitHub repository. http://broadinstitute.github.io/picard/.*
111. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11 10 1-11 10 33.
112. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
113. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-+.
114. Zheng-Bradley, X., et al., *Alignment of 1000 Genomes Project reads to reference assembly GRCh38.* Gigascience, 2017. **6**(7): p. 1-8.
115. Kendall, J. and A. Krasnitz, *Computational methods for DNA copy-number analysis of tumors.* Methods Mol Biol, 2014. **1176**: p. 243-59.
116. Gusnanto, A., et al., *Estimating optimal window size for analysis of low-coverage next-generation sequence data.* Bioinformatics, 2014. **30**(13): p. 1823-9.
117. Nagarajan, N. and M. Pop, *Sequencing and genome assembly using next-generation technologies.* Methods Mol Biol, 2010. **673**: p. 1-17.
118. Pop, M., *Genome assembly reborn: recent computational challenges.* Brief Bioinform, 2009. **10**(4): p. 354-66.
119. Alkan, C., et al., *Personalized copy number and segmental duplication maps using next-generation sequencing.* Nat Genet, 2009. **41**(10): p. 1061-7.
120. Chen, Y.C., et al., *Effects of GC bias in next-generation-sequencing data on de novo genome assembly.* PLoS One, 2013. **8**(4): p. e62856.
121. Scheibye-Alsing, K., et al., *Sequence assembly.* Comput Biol Chem, 2009. **33**(2): p. 121-36.
122. Rudd, M.K. and H.F. Willard, *Analysis of the centromeric regions of the human genome assembly.* Trends Genet, 2004. **20**(11): p. 529-33.
123. Benjamini, Y. and T.P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing.* Nucleic Acids Res, 2012. **40**(10): p. e72.
124. Boeva, V., et al., *De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis.* Nucleic Acids Res, 2010. **38**(11): p. e126.
125. Boeva, V., et al., *Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization.* Bioinformatics, 2011. **27**(2): p. 268-9.
126. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.* Bioinformatics, 2012. **28**(3): p. 423-5.
127. Kuan, P.F., et al., *A Statistical Framework for the Analysis of ChIP-Seq Data.* J Am Stat Assoc, 2011. **106**(495): p. 891-903.
128. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
129. Durrett, R., et al., *Evolutionary dynamics of tumor progression with random fitness values.* Theor Popul Biol, 2010. **78**(1): p. 54-66.
130. Derrien, T., et al., *Fast computation and applications of genome mappability.* PLoS One, 2012. **7**(1): p. e30377.
131. Pickrell, J.K., et al., *False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions.* Bioinformatics, 2011. **27**(15): p. 2144-6.
132. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 2004. **5**(4): p. 557-72.

133. Hsu, F.H., et al., *A model-based circular binary segmentation algorithm for the analysis of array CGH data.* BMC Res Notes, 2011. **4**: p. 394.
134. Venkatraman, E.S. and A.B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array CGH data.* Bioinformatics, 2007. **23**(6): p. 657-63.
135. *Alastair Droop (2017). mcnv: Melanoma CNV Project Analysis. R package version 1.5-22.*
136. *Structural variant: Esv3620012 - explore this SV - Homo sapiens - Ensembl genome browser 86(1000) Available at: http://www.ensembl.org/Homo_sapiens/StructuralVariation/Explore?r=9:23361912-23378571;sv=esv3620012;svf=50963477;vdb=variation (Accessed: 29 November 2016).*
137. *H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.*
138. Phillips, L.H., 2nd, et al., *The epidemiology of myasthenia gravis in central and western Virginia.* Neurology, 1992. **42**(10): p. 1888-93.
139. Durinck, S., et al., *BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.* Bioinformatics, 2005. **21**(16): p. 3439-40.
140. Durinck, S., et al., *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.* Nat Protoc, 2009. **4**(8): p. 1184-91.
141. Kato, S., et al., *Gain-of-function genetic alterations of G9a drive oncogenesis.* Cancer Discov, 2020.
142. Willmore-Payne, C., et al., *BRAF and c-kit gene copy number in mutation-positive malignant melanoma.* Hum Pathol, 2006. **37**(5): p. 520-7.
143. Lazar, V., et al., *Characterization of candidate gene copy number alterations in the 11q13 region along with BRAF and NRAS mutations in human melanoma.* Mod Pathol, 2009. **22**(10): p. 1367-78.
144. Gerami, P., et al., *Copy number gains in 11q13 and 8q24 [corrected] are highly linked to prognosis in cutaneous malignant melanoma.* J Mol Diagn, 2011. **13**(3): p. 352-8.
145. Tang, B., et al., *Palbociclib for treatment of metastatic melanoma with copy number variations of CDK4 pathway: case report.* Chin Clin Oncol, 2018.
146. Vizkeleti, L., et al., *The role of CCND1 alterations during the progression of cutaneous malignant melanoma.* Tumour Biol, 2012. **33**(6): p. 2189-99.
147. Bishop, D.T., et al., *Genome-wide association study identifies three loci associated with melanoma risk.* Nat Genet, 2009. **41**(8): p. 920-5.
148. Martin, J., et al., *The sequence and analysis of duplication-rich human chromosome 16.* Nature, 2004. **432**(7020): p. 988-94.
149. Udart, M., et al., *Chromosome 7 aneusomy. A marker for metastatic melanoma? Expression of the epidermal growth factor receptor gene and chromosome 7 aneusomy in nevi, primary malignant melanomas and metastases.* Neoplasia, 2001. **3**(3): p. 245-54.
150. Cross, N.A., et al., *Multiple locations on chromosome 3 are the targets of specific deletions in uveal melanoma.* Eye, 2006. **20**(4): p. 476-481.
151. Tsui, L.C., *Genetic markers on chromosome 7.* J Med Genet, 1988. **25**(5): p. 294-306.
152. McNamara, M., et al., *Assessment of chromosome 3 copy number in ocular melanoma using fluorescence in situ hybridization.* Cancer Genet Cytogenet, 1997. **98**(1): p. 4-8.
153. Lazar, V., et al., *Characterization of candidate gene copy number alterations in the 11q13 region along with BRAF and NRAS mutations in human melanoma.* Modern Pathology, 2009. **22**(10): p. 1367-1378.
154. Casper, J., et al., *The UCSC Genome Browser database: 2018 update.* Nucleic Acids Res, 2018. **46**(D1): p. D762-D769.

155. GRC, T.G.R.C., *Human Genome Overview: Modeled centromeres and heterochromatin regions.* 2017.
156. *RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.*
157. *R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.*
158. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.
159. Eichler, E.E., R.A. Clark, and X. She, *An assessment of the sequence gaps: unfinished business in a finished human genome.* Nat Rev Genet, 2004. **5**(5): p. 345-54.
160. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
161. Kundaje, A., *A comprehensive collection of signal artifact blacklist regions in the human genome.* 2016.
162. Gregory R. Warnes, B.B., Lodewijk Bonebakker, Robert, et al., *Package 'gplots'.* 2016.
163. Rieber, N., et al., *Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies.* PLoS One, 2013. **8**(6): p. e66621.
164. Arreaza, G., et al., *Pre-Analytical Considerations for Successful Next-Generation Sequencing (NGS): Challenges and Opportunities for Formalin-Fixed and Paraffin-Embedded Tumor Tissue (FFPE) Samples.* Int J Mol Sci, 2016. **17**(9).
165. Davis, C.A., et al., *The Encyclopedia of DNA elements (ENCODE): data portal update.* Nucleic Acids Res, 2018. **46**(D1): p. D794-D801.
166. Staaf, J., et al., *Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays.* Genome Biol, 2008. **9**(9): p. R136.
167. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* J Royal Stat Soc B (Methodological), 1995. **57**(1): p. 289-300.
168. Bishop, D.T., et al., *Copy number alteration in primary melanoma.* Cancer Research, 2017. **77**.
169. Mosen-Ansorena, D., A.M. Aransay, and N. Rodriguez-Ezpeleta, *Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data.* BMC Bioinformatics, 2012. **13**: p. 192.
170. Pozniak, J., et al., *Genetic and Environmental Determinants of Immune Response to Cutaneous Melanoma.* Cancer Res, 2019. **79**(10): p. 2684-2696.
171. Pouryazdanparast, P., et al., *The role of 8q24 copy number gains and c-MYC expression in amelanotic cutaneous melanoma.* Mod Pathol, 2012. **25**(9): p. 1221-6.
172. Scherer, S.W. and E.D. Green, *Human chromosome 7 circa 2004: a model for structural and functional studies of the human genome.* Hum Mol Genet, 2004. **13 Spec No 2**: p. R303-13.
173. Scherer, S.W., et al., *Human chromosome 7: DNA sequence and biology.* Science, 2003. **300**(5620): p. 767-72.
174. Pikor, L., et al., *The detection and implication of genome instability in cancer.* Cancer Metastasis Rev, 2013. **32**(3-4): p. 341-52.
175. Heitzer, E., et al., *Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing.* Genome Med, 2013. **5**(4): p. 30.
176. Taylor, A.M., et al., *Genomic and Functional Approaches to Understanding Cancer Aneuploidy.* Cancer Cell, 2018. **33**(4): p. 676-689 e3.
177. Domcke, S., et al., *Evaluating cell lines as tumour models by comparison of genomic profiles.* Nat Commun, 2013. **4**: p. 2126.

178. Luebker, S.A., W. Zhang, and S.A. Koepsell, *Comparing the genomes of cutaneous melanoma tumors to commercially available cell lines.* Oncotarget, 2017. **8**(70): p. 114877-114893.
179. Knijnenburg, T.A., et al., *Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas.* Cell Rep, 2018. **23**(1): p. 239-254 e6.
180. Wright, A.I., H.I. Grabsch, and D.E. Treanor, *RandomSpot: A web-based tool for systematic random sampling of virtual slides.* J Pathol Inform, 2015. **6**: p. 8.
181. Van Domelen, D.R., *dvmisc: Convenience Functions, Moving Window Statistics, and Graphics. R package version 1.1.4.*
. 2019.
182. Enninga, E.A.L., et al., *Survival of cutaneous melanoma based on sex, age, and stage in the United States, 1992-2011.* Cancer Med, 2017. **6**(10): p. 2203-2212.
183. Cox, D., *Regression Models and Life-Tables.* Journal of the Royal Statistical Society. Series B (Methodological), 1972. **34**(2): p. 187-220.
184. Rich, J.T., et al., *A practical guide to understanding Kaplan-Meier curves.* Otolaryngol Head Neck Surg, 2010. **143**(3): p. 331-6.
185. Kaplan EL and M. P., *Nonparametric estimation from incomplete observations.* Journal of the American Statistical Association, 1958. **53**(457): p. 81.
186. *Altman DG. London (UK): Chapman and Hall; 1992. Analysis of Survival times.In:Practical statistics for Medical research; pp. 365–93.*
187. Mantel, N., *Evaluation of survival data and two new rank order statistics arising in its consideration.* Cancer Chemother Rep, 1966. **50**(3): p. 163-70.
188. *Peto, Richard; Peto, Julian (1972). "Asymptotically Efficient Rank Invariant Test Procedures". Journal of the Royal Statistical Society, Series A. Blackwell Publishing. 135 (2): 185–207. doi:10.2307/2344317. hdl:10338.dmlcz/103602. JSTOR 2344317.*
189. *Harrington, David (2005). "Linear Rank Tests in Survival Analysis". Encyclopedia of Biostatistics. Wiley Interscience. doi:10.1002/0470011815.b2a11047. ISBN 047084907X.*
190. Bland, J.M. and D.G. Altman, *Survival probabilities (the Kaplan-Meier method).* BMJ, 1998. **317**(7172): p. 1572.
191. Yates, A.D., et al., *Ensembl 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D682-D688.
192. Korbel, J.O. and P.J. Campbell, *Criteria for inference of chromothripsis in cancer genomes.* Cell, 2013. **152**(6): p. 1226-36.
193. Rakosy, Z., et al., *Characterization of 9p21 copy number alterations in human melanoma by fluorescence in situ hybridization.* Cancer Genet Cytogenet, 2008. **182**(2): p. 116-21.
194. Gilbert, F. and N. Kauff, *Disease genes and chromosomes: disease maps of the human genome. Chromosome 9.* Genet Test, 2001. **5**(2): p. 157-74.
195. Stagni, C., et al., *BRAF Gene Copy Number and Mutant Allele Frequency Correlate with Time to Progression in Metastatic Melanoma Patients Treated with MAPK Inhibitors.* Mol Cancer Ther, 2018. **17**(6): p. 1332-1340.
196. Hellman, A., et al., *A role for common fragile site induction in amplification of human oncogenes.* Cancer Cell, 2002. **1**(1): p. 89-97.
197. Kwong, L.N. and L. Chin, *Chromosome 10, frequently lost in human melanoma, encodes multiple tumor-suppressive functions.* Cancer Res, 2014. **74**(6): p. 1814-21.
198. Miles, J.A., et al., *Relationship of Chromosome Arm 10q Variants to Occurrence of Multiple Primary Melanoma in the Population-Based Genes, Environment, and Melanoma Study.* J Invest Dermatol, 2018.

199. *LINC01917 long intergenic non-protein coding RNA 1917 [Homo sapiens (human)]. - Gene - NCBI. (n.d.). Available at https://www.ncbi.nlm.nih.gov/gene/101928167. (Accessed: June 23, 2020)."*

.

200. *LINC01919 long intergenic non-protein coding RNA 1919 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Available at https://www.ncbi.nlm.nih.gov/gene/102724651 (Accessed: June 23, 2020).*
201. *MDC1-AS1 MDC1 antisense RNA 1 [Homo sapiens (human)]- Gene - NCBI. (n.d.). Available at https://www.ncbi.nlm.nih.gov/gene/106478956 .(Accessed: June 23, 2020)."*

202. *LINC01090 long intergenic non-protein coding RNA 1090 [Homo sapiens (human)]- Gene - NCBI. (n.d.). Available at https://www.ncbi.nlm.nih.gov/gene/104355152#gene-expression (Accessed: June 24, 2020)."*

.

203. *MIR99AHG mir-99a-let-7c cluster host gene [Homo sapiens (human)]- Gene - NCBI. (n.d.). Available at https://www.ncbi.nlm.nih.gov/gene/388815. (Accessed: June 24, 2020).".*
204. Watahiki, A., et al., *MicroRNAs associated with metastatic prostate cancer.* PLoS One, 2011. **6**(9): p. e24950.
205. Glinsky, G.V., O. Berezovska, and A.B. Glinskii, *Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer.* J Clin Invest, 2005. **115**(6): p. 1503-21.
206. Nakanishi, M., et al., *NFBD1/MDC1 associates with p53 and regulates its function at the crossroad between cell survival and death in response to DNA damage.* J Biol Chem, 2007. **282**(31): p. 22993-3004.
207. Nakanishi, M., *[Regulation of p53 dependent apoptosis by functional binding of NFBD1/MDC1].* Hokkaido Igaku Zasshi, 2006. **81**(3): p. 221-6.
208. Stewart, G.S., et al., *MDC1 is a mediator of the mammalian DNA damage checkpoint.* Nature, 2003. **421**(6926): p. 961-6.
209. *LINC01680 long intergenic non-protein coding RNA 1680 [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/105371660/. (Accessed July 4, 2020).*
210. *LINC00626 long intergenic non-protein coding RNA 626 [Homo sapiens (human)]. Available at : https://www.ncbi.nlm.nih.gov/gene/?term=LINC00626. (Accessed on July 4, 2020).*
211. *LINC00626. Available at : https://www.ncbi.nlm.nih.gov/geoprofiles/?term=LINC00626. (Accessed on July 4, 2020).*
212. *OR2AJ1. Available at : https://www.proteinatlas.org/ENSG00000177275-OR2AJ1/tissue. (Accessed: July 4, 2020).*
213. *OR2T8. Available at : https://www.proteinatlas.org/ENSG00000177462-OR2T8. (Accessed on July 4, 2020).*
214. *RCSD1. Availabale at : https://www.ncbi.nlm.nih.gov/gene/?term=RCSD1. (Accessed on July 4, 2020).*
215. *PDE4B phosphodiesterase 4B [Homo sapiens (human)]. Available at : https://www.ncbi.nlm.nih.gov/gene/5142. (Accessed on JUly 4, 2020).*
216. *GRB14 growth factor receptor bound protein 14 [Homo sapiens (human)]. Available at https://www.ncbi.nlm.nih.gov/gene/2888. (Accessed on July 4, 2020).*
217. *MVB12B multivesicular body subunit 12B [Homo sapiens (human)]. Available at https://www.ncbi.nlm.nih.gov/gene/89853. (Accessed on July 4, 2020)*

.

218. *FMO6P flavin containing dimethylaniline monoxygenase 6, pseudogene [Homo sapiens (human)]. Available at https://www.ncbi.nlm.nih.gov/gene/388714. (Accessed on July 4, 2020).*
219. *CREG1. Available at : https://www.ncbi.nlm.nih.gov/gene/8804 .(Accessed on July 4, 2020).*
220. *MAPK10 Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=MAPK10. (Accessed on August 1, 2020).*
221. *FAM50B Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=FAM50B&keywords=FAM50B. (Accessed on August 1, 2020).*
222. *ZNF184 zinc finger protein 184 [Homo sapiens (human)]. Available at https://www.ncbi.nlm.nih.gov/gene/7738. (Accessed on August 1, 2020).*
223. *KRBA2 KRAB-A domain containing 2 [Homo sapiens (human)]. Available athttps://www.ncbi.nlm.nih.gov/gene/124751. (Accessed on August 1, 2020).*
224. *NOTCH4 notch receptor 4 [Homo sapiens (human)]. Available https://www.ncbi.nlm.nih.gov/gene/4855. (Accessed on August 1, 2020).*
225. *RPP40. Available https://www.proteinatlas.org/ENSG00000124787-RPP40/pathology. (Accessed on August 1, 2020)*
226. Diribarne, G. and O. Bensaude, *7SK RNA, a non-coding RNA regulating P-TEFb, a general transcription factor.* RNA Biol, 2009. **6**(2): p. 122-8.
227. *RIPK1 receptor interacting serine/threonine kinase 1 [Homo sapiens (human)]. Available https://www.ncbi.nlm.nih.gov/gene/8737. (Accessed on August 1, 2020)*
228. Sadovnick, A.D., et al., *Genetic modifiers of multiple sclerosis progression, severity and onset.* Clin Immunol, 2017. **180**: p. 100-105.
229. Furusato, B., et al., *CXCR4 and cancer.* Pathol Int, 2010. **60**(7): p. 497-505.
230. Miki, J., et al., *Identification of putative stem cell markers, CD133 and CXCR4, in hTERT-immortalized primary nonmalignant and malignant tumor-derived human prostate epithelial cell lines and in prostate cancer specimens.* Cancer Res, 2007. **67**(7): p. 3153-61.
231. Teicher, B.A. and S.P. Fricker, *CXCL12 (SDF-1)/CXCR4 pathway in cancer.* Clin Cancer Res, 2010. **16**(11): p. 2927-31.
232. Verone, A.R., et al., *Androgen-responsive serum response factor target genes regulate prostate cancer cell migration.* Carcinogenesis, 2013. **34**(8): p. 1737-46.
233. Klemke, M., et al., *High affinity interaction of integrin alpha4beta1 (VLA-4) and vascular cell adhesion molecule 1 (VCAM-1) enhances migration of human melanoma cells across activated endothelial cell layers.* J Cell Physiol, 2007. **212**(2): p. 368-74.
234. Zhong, W., et al., *DDX1 regulates alternative splicing and insulin secretion in pancreatic beta cells.* Biochem Biophys Res Commun, 2018. **500**(3): p. 751-757.
235. Cheng, S., et al., *Identification of a multidimensional transcriptome signature predicting tumor regrowth of clinically nonfunctioning pituitary adenoma.* Int J Oncol, 2020.
236. Xie, F., et al., *Overexpression of GPR39 contributes to malignant development of human esophageal squamous cell carcinoma.* BMC Cancer, 2011. **11**: p. 86.
237. Huang, W., et al., *SUN1 silencing inhibits cell growth through G0/G1 phase arrest in lung adenocarcinoma.* Onco Targets Ther, 2017. **10**: p. 2825-2833.
238. Jang, K., et al., *VEGFA activates an epigenetic pathway upregulating ovarian cancer-initiating cells.* EMBO Mol Med, 2017. **9**(3): p. 304-318.
239. Kim, M.S., et al., *Frameshift mutation of UVRAG, an autophagy-related gene, in gastric carcinomas with microsatellite instability.* Hum Pathol, 2008. **39**(7): p. 1059-63.

240. He, S. and C. Liang, *Frameshift mutation of UVRAG: Switching a tumor suppressor to an oncogene in colorectal cancer.* Autophagy, 2015. **11**(10): p. 1939-40.
241. Wilkinson, G.R., *Cytochrome P4503A (CYP3A) metabolism: prediction of in vivo activity in humans.* J Pharmacokinet Biopharm, 1996. **24**(5): p. 475-90.
242. Chen, C., et al., *ACER3 supports development of acute myeloid leukemia.* Biochem Biophys Res Commun, 2016. **478**(1): p. 33-38.
243. Lei, R., L. Feng, and D. Hong, *ELFN1-AS1 accelerates the proliferation and migration of colorectal cancer via regulation of miR-4644/TRIM44 axis.* Cancer Biomark, 2020. **27**(4): p. 433-443.
244. Yu, C.C., et al., *Genetic association analysis identifies a role for ANO5 in prostate cancer progression.* Cancer Med, 2020. **9**(7): p. 2372-2378.
245. Song, H.Y., et al., *Anoctamin 5 regulates cell proliferation and migration in pancreatic cancer.* Int J Clin Exp Pathol, 2019. **12**(12): p. 4263-4270.
246. Chang, Z., et al., *Anoctamin5 regulates cell migration and invasion in thyroid cancer.* Int J Oncol, 2017. **51**(4): p. 1311-1319.
247. Garrigos, C., et al., *Single nucleotide polymorphisms as prognostic and predictive biomarkers in renal cell carcinoma.* Oncotarget, 2017. **8**(63): p. 106551-106564.
248. Nakamura, H., et al., *Genomic spectra of biliary tract cancer.* Nat Genet, 2015. **47**(9): p. 1003-10.
249. *C7orf26. Available at: https://genevisible.com/cancers/HS/Gene%20Symbol/C7orf26. Accessed on July 9, 2020).*
250. Li, H., et al., *Genome-wide analysis of the hypoxia-related DNA methylation-driven genes in lung adenocarcinoma progression.* Biosci Rep, 2020. **40**(2).
251. Li, J., et al., *Genetic amplification of PPME1 in gastric and lung cancer and its potential as a novel therapeutic target.* Cancer Biol Ther, 2014. **15**(1): p. 128-34.
252. *NEBL. Avaialbale at https://www.proteinatlas.org/ENSG00000078114-NEBL/pathology. (Accessed July 23, 2020).*
253. *GGNBP1 gametogenetin binding protein 1 (pseudogene) [Homo sapiens (human)]. Available at https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=449520 .(Accessed on July 24, 2020).*
254. *RNF182 ring finger protein 182 [Homo sapiens (human)].Available at https://www.ncbi.nlm.nih.gov/gene/221687#gene-expression. (Accessed on July 24, 2020)*
255. Li, Z., et al., *Methylation profiling of 48 candidate genes in tumor and matched normal tissues from breast cancer patients.* Breast Cancer Res Treat, 2015. **149**(3): p. 767-79.
256. Li, W., et al., *Elevated MIR100HG promotes colorectal cancer metastasis and is associated with poor prognosis.* Oncol Lett, 2019. **18**(6): p. 6483-6490.
257. Zhang, D., et al., *Identification of hub genes related to prognosis in glioma.* Biosci Rep, 2020. **40**(5).
258. *ST3GAL4 Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=ST3GAL4&keywords=ST3GAL4. (Accessed on July 25, 2020).*
259. *GSEC Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=GSEC&keywords=GSEC . (Accessed on July 25, 2020).*
260. Eftedal, R., et al., *Alternative Interleukin 17A/F Locus Haplotypes are Associated With Increased Risk to Hip and Knee Osteoarthritis.* J Orthop Res, 2019. **37**(9): p. 1972-1978.
261. *MIR133B Gene (RNA Gene). Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIR133B . (Accessed on July 24, 2020).*

262. *MIR206 Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIR206&keywords=MIR206 . (Accessed on July 24, 2020).*
263. *IL17F Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=IL17F&keywords=IL17F. (Accessed on July 24, 2020).*
264. *ARHGEF12 Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=ARHGEF12&keywords=ARHGEF12 (Accessed on July 25, 2020).*
265. *DCPS Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=DCPS&keywords=DCPS. (Accessed on Jul5 25, 2020).*
266. *ALDH5A1 Gene, Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=ALDH5A1&keywords=ALDH5A1. Accessed on July 25, 2020).*
267. *KIAA0319 Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=KIAA0319&keywords=KIAA0319. (Accessed on July 25, 2020).*
268. Peltier, J., et al., *Activation peptide of the coagulation factor XIII (AP-F13A1) as a new biomarker for the screening of colorectal cancer.* Clin Proteomics, 2018. **15**: p. 15.
269. Watanabe, T., et al., *Prediction of response to preoperative chemoradiotherapy in rectal cancer by using reverse transcriptase polymerase chain reaction analysis of four genes.* Dis Colon Rectum, 2014. **57**(1): p. 23-31.
270. *LRRC4C leucine rich repeat containing 4C [Homo sapiens (human)] Gene ID: 57689, updated on 22-Jul-2020. Available at https://www.ncbi.nlm.nih.gov/gene/57689. (Accessed on July 25, 2020).*
271. *LINC02053 long intergenic non-protein coding RNA 2053 [Homo sapiens (human)]. Available at https://www.ncbi.nlm.nih.gov/gene/?term=LINC02053. (Accessed on July 25, 2020).*
272. Qin, X.G., et al., *Prognostic value of small nuclear RNAs (snRNAs) for digestive tract pan- adenocarcinomas identified by RNA sequencing data.* Pathol Res Pract, 2019. **215**(3): p. 414-426.
273. Roy, R., et al., *ADAM12 Is a Novel Regulator of Tumor Angiogenesis via STAT3 Signaling.* Mol Cancer Res, 2017. **15**(11): p. 1608-1622.
274. Craze, M.L., et al., *MYC regulation of glutamine-proline regulatory axis is key in luminal B breast cancer.* Br J Cancer, 2018. **118**(2): p. 258-265.
275. *VTI1A Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=VTI1A&keywords=VTI1A. (Accessed on July 27, 2020).*
276. Song, L., et al., *SORBS1 suppresses tumor metastasis and improves the sensitivity of cancer to chemotherapy drug.* Oncotarget, 2017. **8**(6): p. 9108-9122.
277. Fialka, F., et al., *CPA6, FMO2, LGI1, SIAT1 and TNC are differentially expressed in early- and late-stage oral squamous cell carcinoma--a pilot study.* Oral Oncol, 2008. **44**(10): p. 941-8.
278. Yuki, R., et al., *Overexpression of zinc-finger protein 777 (ZNF777) inhibits proliferation at low cell density through down-regulation of FAM129A.* J Cell Biochem, 2015. **116**(6): p. 954-68.
279. Yang, L., et al., *Common Nevus and Skin Cutaneous Melanoma: Prognostic Genes Identified by Gene Co-Expression Network Analysis.* Genes (Basel), 2019. **10**(10).
280. Novak, P., et al., *Agglomerative epigenetic aberrations are a common event in human breast cancer.* Cancer Res, 2008. **68**(20): p. 8616-25.
281. Liu, N., et al., *FAM163A, a positive regulator of ERK signaling pathway, interacts with 14-3-3beta and promotes cell proliferation in squamous cell lung carcinoma.* Onco Targets Ther, 2019. **12**: p. 6393-6406.

282. Li, Z., et al., *Identifying multiple collagen gene family members as potential gastric cancer biomarkers using integrated bioinformatics analysis.* PeerJ, 2020. **8**: p. e9123.
283. Kwon, O.K., et al., *Comparative Proteome Profiling and Mutant Protein Identification in Metastatic Prostate Cancer Cells by Quantitative Mass Spectrometry-based Proteogenomics.* Cancer Genomics Proteomics, 2019. **16**(4): p. 273-286.
284. *INTS7 Gene. Available at https://www.genecards.org/cgi-bin/carddisp.pl?gene=INTS7&keywords=INTS7. (Accessed on July 28, 2020).*
285. A, O.S., M.C. Parrini, and J. Camonis, *RalGPS2 Is Essential for Survival and Cell Cycle Progression of Lung Cancer Cells Independently of Its Established Substrates Ral GTPases.* PLoS One, 2016. **11**(5): p. e0154840.
286. Han, Y. and L. Zhou, *MiRNA-4665-3p Regulates Expression of PLD5 in Thyroid Cancer Patients and Predicts Death.* Journal of Biomaterials and Tissue Engineering, 2019. **9**(7): p. 871-880.
287. Trzeciak, M., et al., *Expression of Cornified Envelope Proteins in Skin and Its Relationship with Atopic Dermatitis Phenotype.* Acta Derm Venereol, 2017. **97**(1): p. 36-41.
288. Liu, Y., et al., *Systematic Identification and Assessment of Therapeutic Targets for Breast Cancer Based on Genome-Wide RNA Interference Transcriptomes.* Genes (Basel), 2017. **8**(3).
289. Lu, Y., et al., *Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer.* Gastroenterology, 2019. **156**(5): p. 1455-1466.
290. Lee, P., et al., *Phosphorylation of Pkp1 by RIPK4 regulates epidermal differentiation and skin tumorigenesis.* EMBO J, 2017. **36**(13): p. 1963-1980.
291. Piccirillo, S.G., et al., *Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells.* Nature, 2006. **444**(7120): p. 761-5.
292. New, M., et al., *MDH1 and MPP7 Regulate Autophagy in Pancreatic Ductal Adenocarcinoma.* Cancer Res, 2019. **79**(8): p. 1884-1898.
293. Tooze, S.A., et al., *MDH1 and MPP7 regulate autophagy in pancreatic ductal adenocarcinoma.* Cancer Res, 2019.
294. Chen, G., et al., *Upregulation of Circular RNA circATRNL1 to Sensitize Oral Squamous Cell Carcinoma to Irradiation.* Mol Ther Nucleic Acids, 2020. **19**: p. 961-973.
295. Cheng, G., et al., *Identification of PLXDC1 and PLXDC2 as the transmembrane receptors for the multifunctional factor PEDF.* Elife, 2014. **3**: p. e05401.
296. Zhang, F., et al., *A miR-567-PIK3AP1-PI3K/AKT-c-Myc feedback loop regulates tumour growth and chemoresistance in gastric cancer.* EBioMedicine, 2019. **44**: p. 311-321.
297. Huang, H., et al., *A qualitative transcriptional prognostic signature for patients with stage I-II pancreatic ductal adenocarcinoma.* Transl Res, 2020. **219**: p. 30-44.
298. Xu, Y., et al., *Identification of RNA Expression Profiles in Thyroid Cancer to Construct a Competing Endogenous RNA (ceRNA) Network of mRNAs, Long Noncoding RNAs (lncRNAs), and microRNAs (miRNAs).* Med Sci Monit, 2019. **25**: p. 1140-1154.
299. *LINC02645 long intergenic non-protein coding RNA 2645 [Homo sapiens (human)]. Available at https://www.ncbi.nlm.nih.gov/gene/105376351. (Accessed on July 29, 2020).*
300. Hijazi, M.M., et al., *NRG-3 in human breast cancers: activation of multiple erbB family proteins.* Int J Oncol, 1998. **13**(5): p. 1061-7.
301. Pinto, R., et al., *Pharmacogenomics in epithelial ovarian cancer first-line treatment outcome: validation of GWAS-associated NRG3 rs1649942 and*

BRE rs7572644 variants in an independent cohort. Pharmacogenomics J, 2019. **19**(1): p. 25-32.
302. Chapman, P.B., et al., *Improved survival with vemurafenib in melanoma with BRAF V600E mutation.* N Engl J Med, 2011. **364**(26): p. 2507-16.
303. Young, K., A. Minchom, and J. Larkin, *BRIM-1, -2 and -3 trials: improved survival with vemurafenib in metastatic melanoma patients with a BRAF(V600E) mutation.* Future Oncol, 2012. **8**(5): p. 499-507.
304. *LINC01339 long intergenic non-protein coding RNA 1339 [Homo sapiens (human)].Available at https://www.ncbi.nlm.nih.gov/gene/101929495 . (Accessed on July 30, 2020).*
305. Huang, J., et al., *Sox11 promotes head and neck cancer progression via the regulation of SDCCAG8.* J Exp Clin Cancer Res, 2019. **38**(1): p. 138.
306. Lennerz, V., et al., *The response of autologous T cells to a human melanoma is dominated by mutated neoantigens.* Proc Natl Acad Sci U S A, 2005. **102**(44): p. 16013-8.
307. Zhang, T., et al., *Low expression of RBMS3 and SFRP1 are associated with poor prognosis in patients with gastric cancer.* Am J Cancer Res, 2016. **6**(11): p. 2679-2689.
308. An, P., et al., *Genome-wide association studies identified novel loci for non-high-density lipoprotein cholesterol and its postprandial lipemic response.* Hum Genet, 2014. **133**(7): p. 919-30.
309. Patel, K., et al., *FAM190A deficiency creates a cell division defect.* Am J Pathol, 2013. **183**(1): p. 296-303.
310. Kang, S.U. and J.T. Park, *Functional evaluation of alternative splicing in the FAM190A gene.* Genes & Genomics, 2019. **41**(2): p. 193-199.
311. Lopes-Ramos, C.M., J. Quackenbush, and D.L. DeMeo, *Genome-Wide Sex and Gender Differences in Cancer.* Front Oncol, 2020. **10**: p. 597788.
312. Li, C.H., et al., *Sex Differences in Cancer Driver Genes and Biomarkers.* Cancer Res, 2018. **78**(19): p. 5527-5537.
313. Li, W.Q., et al., *Host Characteristics and Risk of Incident Melanoma by Breslow Thickness.* Cancer Epidemiol Biomarkers Prev, 2019. **28**(1): p. 217-224.