Phylogenetic Analysis of Cancer Evolution in Malignant Pleural Mesothelioma to Identify Driver Genes

by

Lee D. Brannan

Department of Health Sciences, University of Leicester

A thesis presented for the degree
Doctor of Philosophy

2021

Phylogenetic analysis of cancer evolution in Malignant Pleural Mesothelioma to identify driver genes

by

Lee D. Brannan

Malignant Pleural Mesothelioma (MPM) is a rare type of cancer which occurs in the mesothelium of the lungs, and is characterised by a long latency period followed by a highly aggressive phase once fully developed. The initial tumour is the result of exposure to asbestos, and the prognosis is very poor once diagnosis has taken place. Currently, several genes have been associated with MPM, but no drivers have been identified via phylogenetic analysis. In order to identify potential driver genes which cause the cancer to mutate into its aggressive state, three distinct phylogenetic pipelines were established to process whole-exome sequencing data taken from 25 MPM patients from the MEDUSA cohort. The first pipeline used copy number calls generated from the patient cohort and incorporated them into a phylogenetic inference software in order to generate trees displaying the evolution of the cancer across 4-5 samples per patient. The second pipeline used single-nucleotide variations generated from the patient cohort and incoporated them into a second phylogenetic inference software to generate a different set of trees. The third pipeline used the output from the first in order to generate a third set of trees and establish an order of mutation for events found in the truncal regions of the first tree set. Each pipeline provided strong evidence for the Neurofibromin 2/Merlin (*NF2*) gene as a potential driver in the progression of the cancer from its latent state. This phylogenetic inference is among the first times such a method has been used for MPM, with the findings possibly revealing candidates for potential drug-targeting in the cancer.

Acknowledgements

The first acknowledgements have to go to the British Lung Foundation and the University of Leicester Health Sciences Department for the funding for this project.

Thanks go to Prof Jacqui Shaw and Dr Ed Hollox for their additional guidance throughout the project.

A shoutout goes to my colleagues in Room 3.06 for the good times

Thanks to Novogene and the team in the Dean Fennell labs for data acquisition

My warmest gratitude to my best friend EC, for their support in keeping me sane for the last 4 years

And of course, a huge thanks to my supervisors Prof Frank Dudbridge and Prof Dean Fennell for not giving up me and for the continued support since the start

COVID Impact Statement

Even as a computer-based student, the impacts of COVID were broad, forcing all research to be done on a single laptop with limited internet access through home wifi. The loss of the computer station which had been used for the first half of the PhD was a major blow and resulted in the delaying of the final analysis, and subsequently the writing of the thesis.

Table of Contents

1) Introduction

1.1) Asbestos and Asbestos Usage

Asbestos is a term given to a group of fibrous rocky minerals which form long and thin fibre-like structures when they crystallize, and has been used in construction for over 100 years, mainly due to its flame retardant properties making it effective for insulating buildings whilst avoiding fire risks (Pira, Donato et al. 2018). It is now widely known that exposure to these fibrous structures can result in the development of diseases: asbestosis, lung cancer, and mesothelioma, which can develop upon inhalation of the asbestos particles (Mott 2012; Visonà, Silvia D et al. 2018).

In most Western countries and Japan, the use of asbestos is banned for construction purposes or anywhere within the public domain, though due to the long latency period the high prevalence of the diseases persists. In the United Kingdom for example, there is to be a predicted 91,000 deaths from asbestos-related diseases between 1968 and 2050 (with about two-thirds of these to occur between 2007 and 2022) assuming the year of peak exposure was 1968 (Tan, Warren et al. 2010). Similar scenarios can be seen across early industrial countries, such as the Lombardy region of Italy, which is heavily industrialised and had a high level of asbestos use throughout the 20th century (Mensi, De Matteis et al. 2016). The reason why the prevalence is expected to fall is due to the banning of asbestos resulting in it now being removed from any new builds, and thus massively decreasing the level of exposure which should occur.

However, it should be noted that as many older buildings still contain large amounts of asbestos, occasional exposure may still occur and result in mesothelioma cases later in the future.

Even though the use of asbestos has dropped massively in the Western world, which should lead to a large-scale decrease in the prevalence of asbestos-related disease, in some regions the use of asbestos is actually increasing (Choi, Lim et al. 2013; Leong, Zainudin et al. 2015). A large amount of the asbestos used in China is imported directly from Russia, which accounts for just under half of global supply at 49% (Stayner, Welch et al. 2013), and is used due to its widespread availability and low-cost, with its usage also widespread in Kazakhstan and India (Visonà, Silvia D et al. 2018). It is also the case that certain construction organisations and governing bodies in Russia and China do not support international findings that asbestos exposure results in cancer development (or at least certain types of asbestos)(Leong, Zainudin et al. 2015).

The content of this thesis focuses on the asbestos-related disease, Malignant Pleural Mesothelioma (MPM). Even though the prevalence of MPM is expected to be heavily reduced within the next couple of decades in a large proportion of the Western world, the continued and accelerated use of asbestos in Asia is likely to result in an epidemic within the next 30-40 years in these regions. The problem may be further amplified if the use of asbestos in the industrialisation of these places continues at a rapid rate, leading to thousands of exposure events to the workers and local communities. This results in a greater importance in understanding the pathophysiology of MPM and developing possible

treatments, as tumours will begin to develop in people exposed within the next few years, long after it has largely vanished in the Western world.

1.2) Malignant Pleural Mesothelioma

1.2.1) Introduction to Malignant Pleural Mesothelioma

Malignant Pleural Mesothelioma (MPM) is a rare type of cancer which occurs in the pleura of the lungs, a thin membrane which functions to separate the organ from the chest wall. As previously mentioned, the development of the cancer has been consistently linked to asbestos exposure via inhalation (Mott 2012). The cancer has a long latency period, meaning that it takes a long time for disease presentation to occur after the initial asbestos exposure event, which on average is around 40 years (Bibby, Anna C .2016). MPM is considered to be an aggressive type of cancer once it fully develops and is generally associated with a poor prognosis (Cao, Croce et al. 2012; Shavelle, Vavra-Musser et al. 2017; Taioli, van Gerwen et al. 2017).

MPM can be divided into three major types based on histological classification: epithelioid, which can be seen in clearly defined lumps of cells, and with pink cytoplasm; sarcomatoid, with overlapping irregularly shaped cells that have a large elongated structure; and biphasic, which is MPM comprising of both epithelioid and sarcomatoid cells. Histopathology pictures for each of the three types can be seen in Figures 1.1, 1.2 and 1.3. Epithelioid MPM accounts for around 60% of cases, with the other two types accounting for around 20% each (Inai, Kouki 2008).

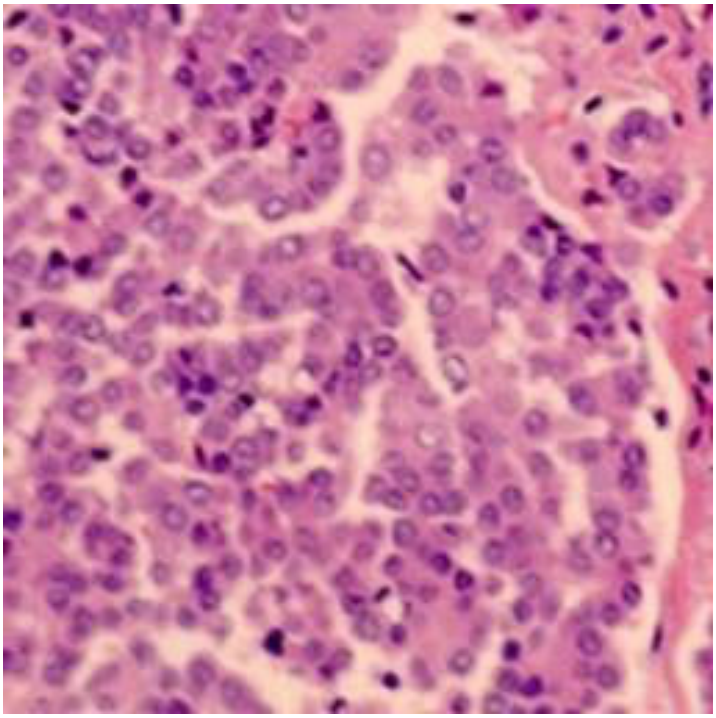


Figure 1.1: Histopathology picture of epithelioid cells in a MPM patient, taken from (Inai, Kouki 2008).

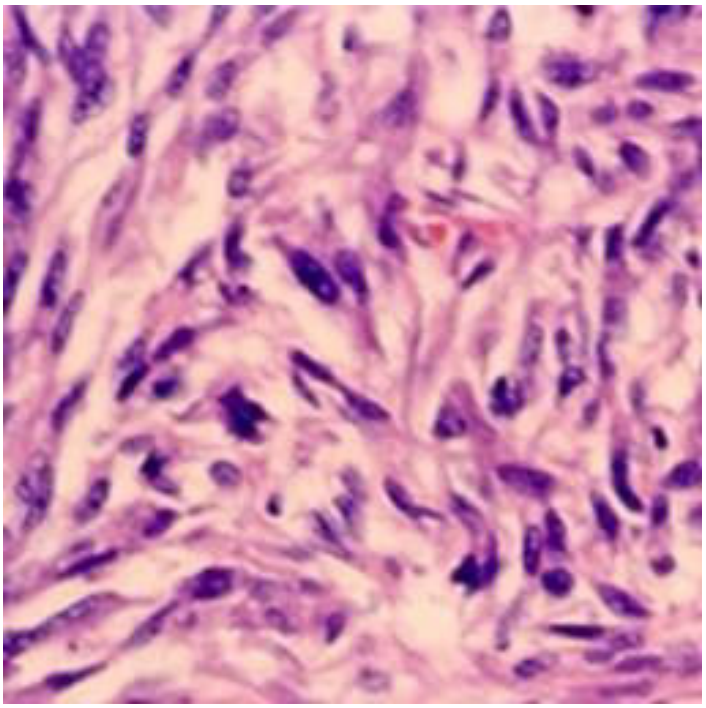


Figure 1.2: Histopathology picture of sarcomatoid cells in a MPM patient, taken from (Inai, Kouki 2008).

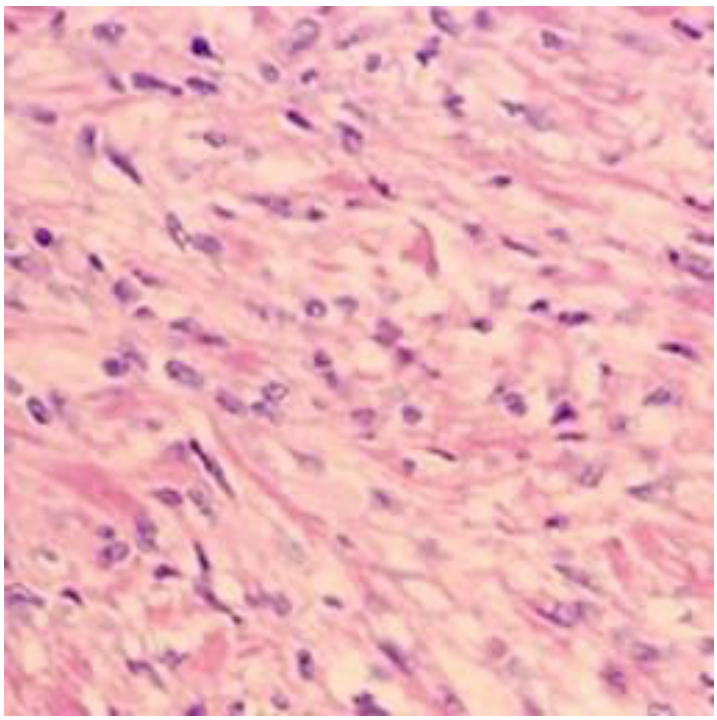


Figure 1.3: Histopathology picture of epithelioid and sarcomatoid cells in a MPM patient, resulting in biphasic MPM, taken from (Inai, Kouki 2008).

1.2.2) Risk Factors in Developing Malignant Pleural Mesothelioma

The single greatest risk factor in developing MPM is exposure to asbestos (Mott 2012; Robinson 2012), with studies showing evidence that greater intensities of exposure, for example, a long career in construction working with asbestos, lead to shorter latency periods in diagnosed patients (Bianchi, Bianchi 2007; Marinaccio, Binazzi et al. 2007). Further supporting evidence is the prevalence of MPM being higher in early industrial countries, with the highest rates being in the United Kingdom and Australia (with a rate of 30 cases per million), as well as Japan and the United States of America (Robinson 2012). The use of asbestos was widespread in the construction industries of these countries, especially in the late 1950s up to the late 1980s, resulting in high levels of exposure to the workers in these sectors (Marinaccio, Binazzi et al. 2007; Tan, Warren et al. 2010).

Age is also a risk factor in the development of MPM, with 69 years being the average age of diagnosis (Shavelle, Vavra-Musser et al. 2017). However, this is mostly due to the long latency period of the disease, as exposure to asbestos at between the ages of 25-30, would then only result in the initial tumour cells developing when that individual was in their 60s. It would then take additional time for diagnosis, especially if symptoms did not present for the first few months. Determining whether or not asbestos exposure has occurred, as well as the time period since the exposure and the intensity of the exposure, is very difficult as the only method for collecting this data is via retrospective interview. This relies on the individual to have a memory of a possible event where

exposure may have occurred, with level of exposure being almost impossible for them to determine. Individuals who worked in an environment where asbestos exposure occurred are likely to know the time period, but due to the continued exposure, intensity levels would be extremely difficult to estimate. This results in the possibility of a large amount of inconsistency when trying to quantify the overall effect of asbestos exposure in the development and outcome of MPM.

1.2.3) Survival, Staging and Treatment Options

Average survival time is a difficult statistic to measure in the case of MPM due to there being a myriad of factors which can affect the outcome of the disease. The type of MPM present will affect survival time, with epithelioid MPM generally having longer survival times than sarcomatoid MPM. This is due to the aggressive nature of sarcomatoid type MPM and its tendency to spread more quickly (Inai, Kouki 2008). As for Biphasic type MPM, the proportion of cell types can tilt survival time in either direction, with higher proportions of epithelioid cells resulting in bigger prognosis.

The staging of the cancer is also an important factor, with mesothelioma divided into 4 distinct stages known as the number system. Stage 1 is the earliest stage and is defined by mesothelioma cells being present in the pleura on a single side of the chest. Stage 1 can further be broken down into 1a and 1b stages, with stage 1a meaning the mesothelioma cells are only present in the outer layer of the pleura (parietal pleura), and stage 1b meaning the mesothelioma cells are present in the inner layer of the pleura (visceral pleura). Stage 2 refers to the mesothelioma having spread into both layers of the pleura, as well as either the diaphragm muscle or the lung tissue. Stage 3 is when the mesothelioma has begun to spread to the chest wall, pericardium (tissue surrounding the heart) or the lymph nodes. And stage 4 refers to the mesothelioma either growing through the diaphragm; spreading to the pleura on the opposite side; growing through the chest organs; or growing through the pericardium (Bonomi, Maria 2017).

Each subsequent stage reduces survival time, with 1 year survival time for each stage available from the Cancer Research UK (CRUK) website, and is shown in Figure 1.4. This value refers to the proportion of individuals who have survived after diagnosis for up to 1 year and is given as a percentage. Stage 1 and 2 both have a 1 year survival rate of 60%, stage 3 has a 1 year survival rate of 50%, and stage 4 has a 1 year survival rate of 30%. The general health of the patient also factors into the survival rate, though the variables defining this would be too numerous to list here. The overall 5 year survival rate for MPM (not considering stages) is around 12% (Shavelle, Vavra-Musser et al. 2017).

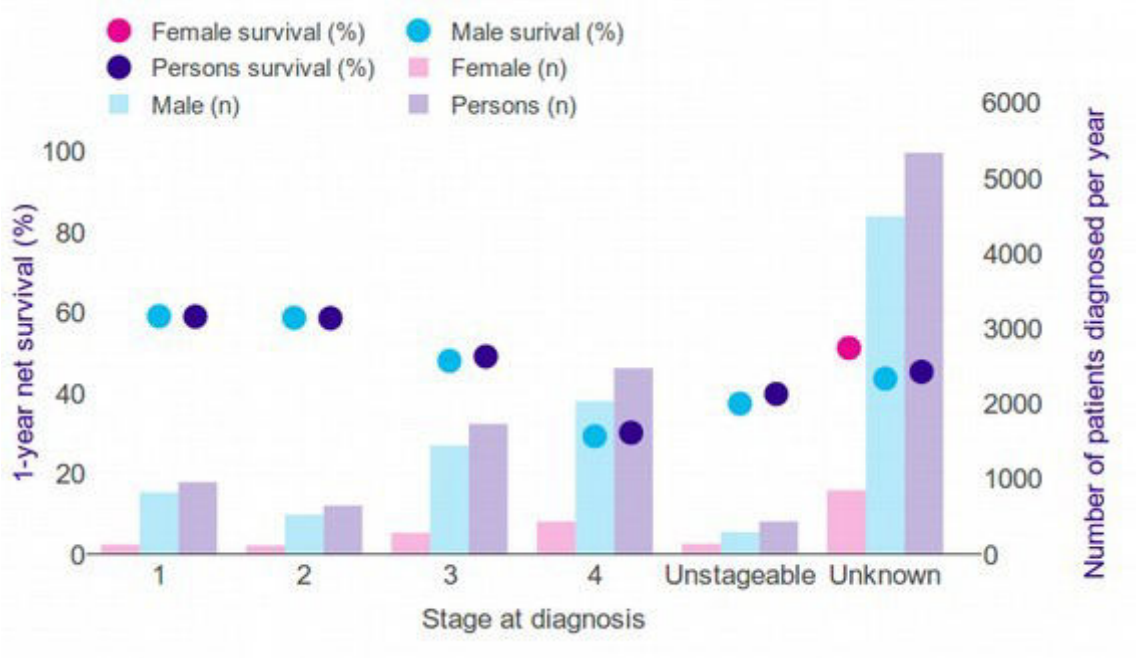


Figure 1.4: 1-year survival time for mesothelioma patients depending on what stage the cancer was when diagnosis occurred. The circles refer to the percentage of patients who survived after 1 year, and the bars indicate the total number of patients diagnosed at each stage. Unstageable cancer represents cases where staging tests did not take place as the invasiveness of the staging was thought to outweigh any benefit that could be obtained from the information. (taken from Cancer Research UK).

Treatment for MPM is also dependent various factors, including; stage of the cancer, location of the tumour, cell type of the mesothelioma (epithelioid, sarcomatoid, biphasic), and the general health of the patient. Generally, early stage mesothelioma is treated with surgery, though it is unlikely to permanently remove the cancer, and this is usually followed by treatment with chemotherapy or radiotherapy, or a combination of the two. For more advanced stage patients chemotherapy is offered to attempt to shrink the tumour, with radiotherapy as an alternative or used for combination treatment. In late stage mesothelioma surgery is usually ineffective due to the spread of the cancer over a large area or into multiple organs.

1.3) Known Genomic Features of Malignant Pleural Mesothelioma

1.3.1) Evaluation of Large Studies Using Whole Exome Sequencing to Analyse Malignant Pleural Mesothelioma

The data used in the analysis throughout the main body of the thesis was generated via whole exome sequencing (WES) of an MPM patient cohort, and as such, it is essential that the previous literature is reviewed to establish an understanding of what knowledge already exists in relation to this subject.

The most recent large-scale study performed on MPM data using WES was performed using The Cancer Genome Atlas dataset, which is a database containing datasets specifically to form cohorts for different types of cancer. The paper (Hmeljak, Julija et al. 2018) performed WES analysis on a cohort of 74 MPM samples in order to identify somatic mutations, or single-nucleotide mutations, as well as somatic copy number alteration events (SCNAs), which could be identified as “driver” events. Driver events can refer either to any mutation event which occurs in cancer, causing it to shift from a dormant state into a more aggressive one, or it can refer specifically to mutations in oncogenes, which are genes that cause positive proliferation of cancer cells when they are mutated (as opposed to suppressor genes which allow cancer to proliferate when they are inactivated). The study used various algorithms to assess the mutational burden of exome regions, both focused on finding single nucleotide mutations as well as changes in copy number.

The study identified large-scale copy number losses, which were recurrent between samples, that also sometimes spanned entire chromosome arms. This is reflective of previous studies, with large-scale copy number alterations known to occur throughout the genome in MPM and is considered a key feature of the genomic landscape of the disease (Furukawa, Toyooka et al. 2015; Hylebos, Van Camp et al. 2017). The prevalence of copy number loss in the absence of amplifications is what fuels the notion that MPM is likely to be caused by loss of function mutations in tumour suppressor genes as opposed to increased expression of oncogenes.

Key findings in the paper in relation to copy number variations were deletions or complete losses of both the *CDKN2A* and *NF2* genes, reporting homozygous loss of *CDKN2A* in 49% of samples and heterozygous deletions in 7%, and homozygous loss of *NF2* in 34% of samples with heterozygous deletions in another 30%. Both of these genes are reported at length throughout MPM literature with studies often reporting *CDKN2A* (Borczuk, Pei et al. 2016 Kettunen, Savukoski et al. 2014; Prins, Williamson et al. 1998) and *NF2* (Borczuk, Pei et al. 2016; Sato, Sekido 2018; Sekido, Pass et al. 1995) losses in copy number. Another frequently mutated gene was reported by the paper, *BAP1*, though it didn't have the same copy number alteration frequency as *CDKN2A* or *NF2*. However it did display a greater mutational burden overall, once somatic nucleotide mutations were taken into account. *BAP1* is also frequently reported in MPM literature (Bott, Brevet et al. 2011; Quetel, Meiller et al. 2020).

The heavier mutational burden displayed by *BAP1* in the paper was due to the large number of single nucleotide variations (SNVs) detected in the gene. These significant SNVs were detected using a software called MutSig2CV (Lawrence, Michael S. 2014) which detects when genes have a higher mutational burden than would be expected by chance, using the dataset as a whole to normalise the expected mutation rate. Increased numbers of SNVs were also detected in *CDKN2A* and *NF2*, as well as in three additional genes, *TP53*, *LATS2* and *SETD2*, all of which have also been frequently reported in MPM literature (de Assis, Isoldi 2014; Murakami, Mizuno et al. 2011; Sementino, Menges et al. 2018), although not to the same extent as *BAP1, CDKN2A* and *NF2*.

The study also reported an overall lower level of SNVs when compared to other cancers, a finding consistent with MPM literature, where it is reported that the mutational burden represented by single nucleotide mutations is low (Martincorena, Iñigo. 2015). This may imply that SNVs are not as influential in the development of MPM when compared to CNVs, which is supported by the fact that CNVs a more likely to remove the functions of entire genes and their surrounding regions. However, frameshift and nonsense mutations are still often reported in the literature in relation to SNVs in identified driver genes in MPM, including in the study being reviewed, suggesting that SNVs do likely play a role in MPM. Furthermore, SNVs can result in biallelic deletions where heterozygous copy number loss has occurred, i.e. they can cause a homozygous deletion at a single base loci possibly resulting in a complete loss of function of any gene they are present in. These will be important points to consider when assessing the role of SNVs in MPM throughout the main body of the thesis.

Another article published in 2016 (Bueno, Raphael. 2016) had a cohort size of 206 MPM samples, though only 99 were WES based. The findings of this study are largely parallel to the TCGA paper (Hmeljak, Julija et al. 2018) though interestingly *CDKN2A* was not reported as a significant finding.

At this point, the 2021 Zhang paper should be mentioned as the thesis author is also a supporting author in that study (Zhang, Jin-Li et al. 2021). It should be mentioned that no data produced for this thesis was used in the Zhang paper, though calls made for the cohort used in that paper were used in the thesis in Chapters 3 and 4 (see Methods). The Zhang paper used WES to extract data from a 22 patient cohort from the MEDUSA main cohort, and used phylogenetic inference to identify driver events which occurred early in the development of the cancer (more on this discussed in Section 1.5). The main findings of this paper were large scale heterozygous losses in *BAP1, CDKN2A* and *NF2* with *NF2* in particular reporting losses in 82% of patients.

It is clear that *NF2, BAP1* and *CDKN2A* are key genes in the development of MPM. The next section will focus on these three genes.

1.3.2) Key Driver Genes in Malignant Pleural Mesothelioma

BAP1, or BRCA1 associated protein 1 is a gene that has been consistently associated with MPM, both through CNVs and SNVs. The gene encodes a protein called ubiquitin carboxyl-terminal hydrolase BAP1, which is shortened to BAP1, that functions to remove ubiquitin from other proteins as a form of regulation. The BAP1 protein is known to operate in the processes of cell proliferation and cell death, both key features in the development of cancer (Carbone, Michele et al. 2020).

The cBioPortal (Cerami, Ethan et al. 2012; Gao, Jianjiong et al. 2013) was used to examine known point mutations in the *BAP1* gene, with results displayed in the lollipop plot in Figure 1.5. cBioPortal is a database of cancer genomics datasets and annotations which can be used to easily access mutational information about given genes. The same plots were also generated using cBioPortal for *NF2* and *CDKN2A* in Figure 1.6 and Figure 1.7 respectively. There were 18 listed known mutations in *BAP1* associated with MPM, all of which are non-synonymous and so should result in protein changes.

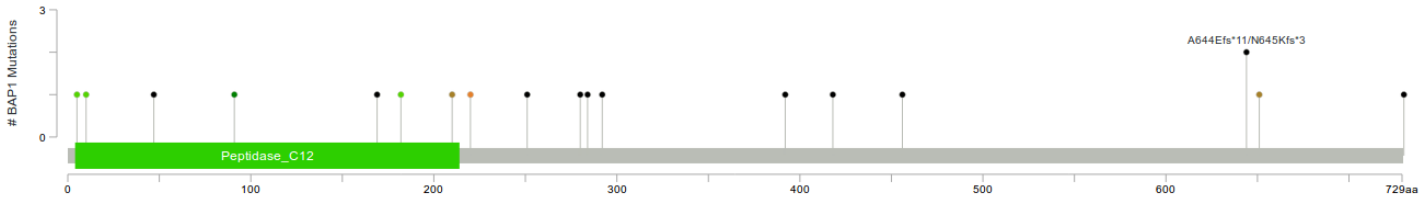


Figure 1.5: Lollipop plot generated by cBioPortal by filtering for *BAP1* and Pleural Mesothelioma. The plot displays 18 known point mutations in the gene, all of which are non-synonymous. The green section to the left of the plot is the splice variant of BAP1, Peptidase_C12.

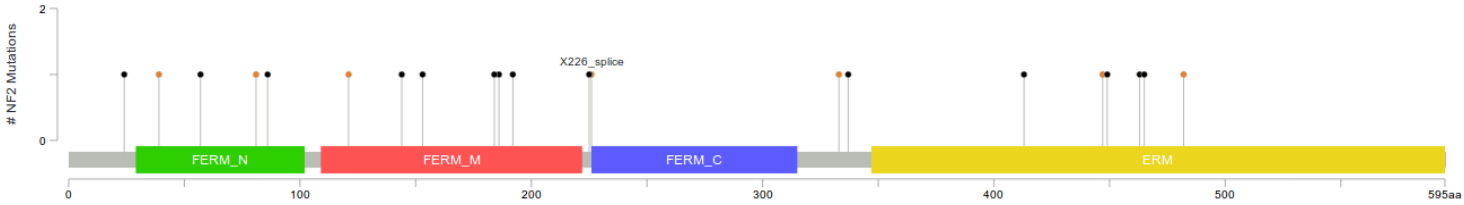


Figure 1.6: Lollipop plot generated by cBioPortal by filtering for *NF2* and Pleural Mesothelioma. The plot displays 21 known point mutations in the gene, all of which are non-synonymous. The splice variants are displayed in the coloured boxes.

Figure 1.7: Lollipop plot generated by cBioPortal by filtering for *CDKN2A* and Pleural Mesothelioma. The plot displays no known point mutations in the gene. The splice variants are displayed in the coloured boxes.

NF2, or moesin-ezrin-radixin like (MERLIN) tumor suppressor, encodes for a protein called merlin, which functions as a tumour suppressor protein in order to prevent cells from proliferating too quickly. It is clear that mutations in this gene that result in loss of function would be likely candidates when searching for cancer causing genes. Due to its role as a tumour suppressor, it isn't surprising that *NF2* has been associated in the development of other types of tumour, including breast, colorectal and skin cancers (Petrilli, A M. 2016). 21 known mutations are reported on cBioPortal for *NF2* in association with MPM, though none are reported more than once.

CDKN2A, or cyclin dependent kinase inhibitor 2A, which is a gene that encodes several proteins, two of which, the p16(INK4A) and the p14(ARF) proteins are tumour suppressors (He, Shenghui 2017). As with *BAP1* and *NF2*, loss of function in *CDKN2A* can clearly lead to unwanted cell proliferation, and thus result in tumours. Also much like the other two genes, *CDKN2A* has been associated with other cancers including breast cancer and pancreatic cancer (Goldstein, Alisa M et al. 2006). Interestingly, the cBioPortal reports no SNV mutations for *CDKN2A* in association with MPM, which may seem surprising due to the frequency of *CDKN2A* throughout the literature. However, this could be indicative that CNV based mutations are much more prevalent in the disease than SNVs, especially in the case of *CDKN2A* as was indicated in the results of the TGCA paper discussed in the previous section.

The relatively low number of reported mutations on cBioPortal for all three key driver genes could simply be due to the low number of cases present in the

database. Despite being one of the more common mesotheliomas, MPM is still classified as a rare cancer and has a low incidence rate, as discussed in Section 1.2. This means that there is a relatively low number of patients samples available to form cohorts and perform analysis on.

Though this section only briefly described the function of the most well established MPM associated genes, it is important for biological function to be acknowledged. Associating genes with disease is an important step in developing possible treatments, but an understanding of the biological function and pathways which the proteins encoded by these genes operate in is also paramount in successfully treating disease.

1.4) Biological Mechanisms in the Development of Malignant Pleural Mesothelioma

1.4.1) Frustrated Phagocytosis

The mechanism by which the fibrous strands of asbestos inhaled during an exposure event results in the development of MPM is regarded to be caused by a process called “Frustrated Phagocytosis”, referring to an incomplete attempt at phagocytosis by a macrophage (Gualtieri, Alessandro F. 2021). The normal process of phagocytosis involves a phagocyte cell expanding its cell membrane to engulf a foreign particle, followed by internally breaking down the target particle they have engulfed. It is an essential process in maintaining cell balance and protecting the host via the immune system.

Frustrated Phagocytosis occurs when a phagocyte cell in unable to engulf its target, and in the case of MPM development, that phagocyte cell is likely to be a macrophage. This is because part of the additional function of a macrophage cell in to engage the adaptive immune system. In Frustrated Phagocytosis, as the particle is too big to be engulfed by the macrophage, this leads to inflammation due to a prolonged immune response, which in turn leads to the production of ROS free radicals (Liu, Cheresh et al. 2013; Pietrofesa, Velalopoulou et al. 2016). These can then cause DNA damage in the surrounding cells which is what can then lead to mutation during DNA repair as the DNA repair system is eventually overwhelmed by mutation. This process can be observed in Figure 1.8.

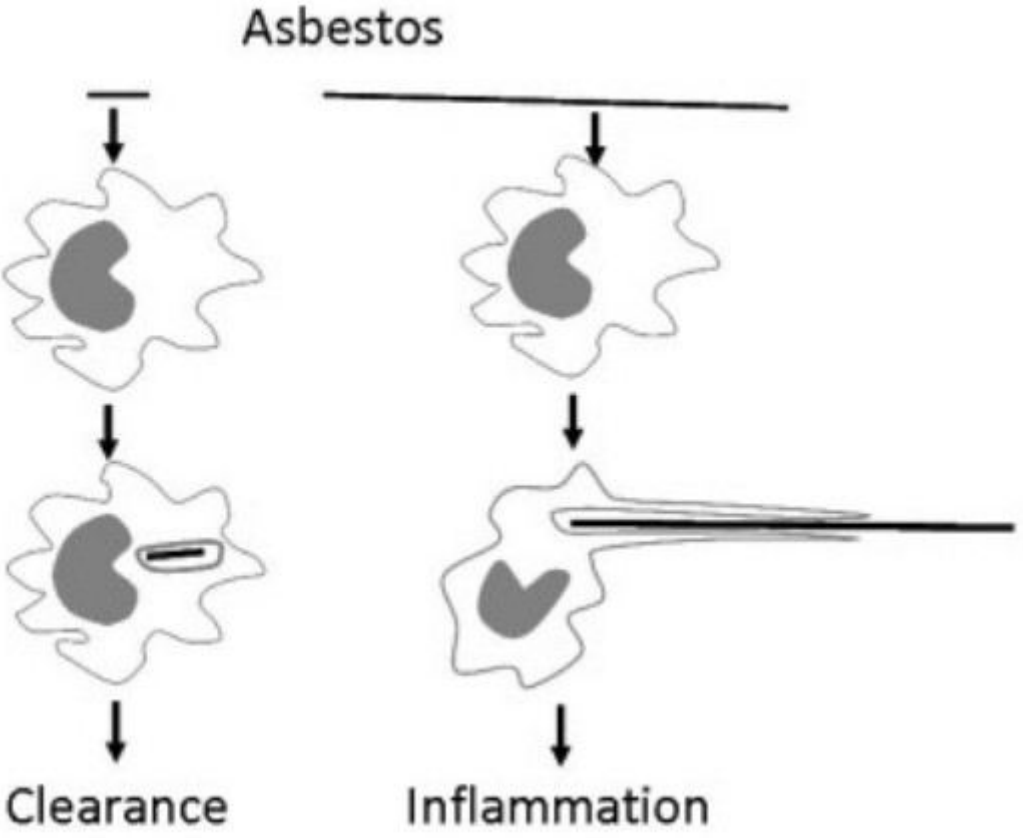


Figure 1.8: The process of normal and Frustrated Phagocytosis. The left of the figure represents a macrophage cell operating normally, engulfing a target, breaking it down internally and then releasing it. The right of the figure shows the process of Frustrated Phagocytosis. The asbestos fibre is too large for the macrophage to engulf, leading to prolonged inflmmation response. Figure taken from (Donaldson, Ken et al. 2010)

There is a lot of evidence supporting the inflammation response in the carcinogenesis of MPM due to asbestos exposure (Plato, Martinsen et al. 2016; Shukla, Gulumian et al. 2003) . It should be noted that whilst a basic understanding of what could be causing asbestos exposure to result in the development of MPM is useful, this is not the question which this project aims to answer. It could be that the findings of this project help to answer this question by identifying early genetic modifications which occur during the process, but the methods used do not specifically aim to decipher the specific cause of initial carcinogenesis. However, it was still important to include these theories as

understanding what may be causing the initiation of genetic changes could be useful in then analysing those genetic changes which do occur.

1.4.2) Other Possible Causes

Around 80% of MPM patients have had confirmed exposure to asbestos (based on retrospective interview), leaving 20% who developed the disease for unexplained reasons (Kroczynska, Cutrone et al. 2006; Murthy, Testa 1999; Pershouse, Heivly et al. 2006).

There has been evidence presented in the literature, that the SV40 virus is linked to the development of MPM without the requirement of asbestos exposure. How the virus causes carcinogenesis in unknown, though there is evidence that it results in the inactivation of certain tumour-suppressor genes such as p53 (activator of apoptosis) (Ahuja, Saenz-Robles et al. 2005). It could also be the case that the SV40 virus causes carcinogenesis in a similar way to how asbestos exposure is thought to, with the presence of the virus in cells resulting in a prolonged immunity response which then produces ROS resulting in DNA damage and inducing the eventual formation of tumour cells. It should be noted however, that the prevalence of SV40 traces in MPM patients is far below 20% and so cannot totally explain (if at all) how the remaining non-asbestos cases developed. Though it is still important to consider, as asbestos may not be the only cause of MPM development and so cannot be attributed to every case conclusively, it is still highly likely it is the most significant agent in the proliferation of the process.

There is also evidence that other types of foreign particle could result in Frustrated Phagocytosis in much the same way as asbestos, in particular, carbon nanotubes have multiple studies linking them to this very phenomenon (Benedetti, Serena et al. 2015; Donaldson, Ken et al. 2010). The process would

work in much the same way as with asbestos, with the macrophage cells unable to engulf the carbon nanotubes resulting in prolonged inflammation and eventually DNA damage and mutation.

It is important to consider the other possible causes of MPM besides asbestos, as this may reveal additional findings when analysing the results of the research, however, as asbestos is the firmest accepted agent in the development of MPM, and also appears to account for the majority of cases, it is the most significant focus when considering possible carcinogenic agents. It is also important to remember that determining asbestos exposure is done via retrospective interview and this method lends itself more to under-calling the true number of asbestos exposed cases. This is because it is logically more likely that an individual would have forgotten possible exposure events after 30-40 years (the time when diagnosis and the interview will have taken place), than someone falsely remembering an event where they think they were exposed. It can also be assumed that someone who doesn't remember whether or not they were specifically exposed, but has had a history in the construction industry during the time period when asbestos was heavily used, is considered to have been asbestos exposed.

1.5) Phylogenetics and Evolution of Cancer

Phylogenetics is the study of evolutionary relationships between defined groups of biological units, be that individual organisms, entire species, or sets of genes. Phylogenetic methodology usually results in the output of a phylogeny (tree), which explains the relationship between the units being assessed. Different methods of phylogenetic inference include maximum parsimony, which connects units based on the lowest number of changes between them; neighbour-joining which is an iterative method that joins together the most likely pairs step-by-step; and Bayesian methods which work out the best overall tree by comparing hundreds or thousands of trees with different assigned likelihood values.

The use of phylogenetics for the analysis of cancer has been a concept for over two decades (Desper, Jiang et al. 1999), with the ability to identify the order of mutations an attractive feature when considering the development of treatments. A variety of phylogenetic methods have been used in various analysis of cancers (Alexandrov, Nik-Zainal et al. 2013; Gerlinger, Horswell et al. 2014; Jamal-Hanjani, Wilson et al. 2017; Schwartz, Schäffer 2017). In particular, the TRACERx Consortium (TRAcking non-small cell lung Cancer Evolution through therapy (Rx)) study of phylogenetic analysis in lung cancer (Abbosh, et al. 2017) employed solid methodology to identify clonal single-nucleotide variants using circulating tumour DNA, based on the following logic.

The basis for the ability to generate phylogenetic analyses for these tumours is the heterogeneous nature of cancer, meaning different regions within the same

patient can have vastly different genetic architecture, even those present within close proximity in the same tissue (Curtis, Shah et al. 2012). Whilst this has a confounding effect for treatment, as a sample taken is not necessarily representative of the entire tumour (and is in fact, unlikely to be) and means it can be hard to direct treatment options optimally (Merlo, Pepper et al. 2006), replicating and migrating cells, the process of tumour development has strong evolutionary signatures in the form of large-scale genetic changes which occur and can then be traced back. The idea is, that samples from different tumours within the same patient should have a significant amount of genetic difference, and the amount of difference that exists between compared samples is the key to phylogenetic analysis (Somarelli, Ware et al. 2017).

For example, two samples with a large amount of genetic difference would be assumed to have diverged early in the development of the cancer, and possibly even at the first stages near to where the first cancer cells developed. Those samples that are more similar will be assumed to have diverged later on in the process. The method also works for looking at specific genetic changes and provides valuable information into when those changes may have occurred. If there is a common genetic alteration present in all tumours, the assumption would be that this particular genetic alteration occurred in the original subclone of the cancer (the region of cancer cells which then proliferated and lead to the development of other tumour sites via the spread of the cancer through the tissue and bloodstream).

The understanding of this is paramount to the research intentions of the project, as it is these early events (genetic alteration events seen to be recurring across

all samples in a patient) which are thought to harbour the driver mutations for MPM. The reasoning behind thinking this is that shared genetic events across all samples will appear in the “trunk” of the tree, or the first branch between the normal tissue and the first common precursor. Any changes which occurred during the latency period, and before the cancer began to develop and proliferate will appear in this region of the tree. It is therefore logical to assume that changes here are responsible for the transition of the cancer into its non-latent, aggressive state. With this in mind, identifying candidate regions in this area, and possibly even specific candidate genes, can then allow for further analysis in the form of functional and comparative studies, which may then lead into drug design and trials on the appropriate candidates. By arresting the cancer before it begins to spread, the disease will be less serious and surgical treatment may then be more effective. Databases such as The Catalogue of Somatic Changes in Cancer (COSMIC) (Forbes, Bhamra et al. 2008) to catalogue observations like this. Comparing the results of the phylogenetic analysis to databases such as this is extremely useful, both in terms of investigating whether recurring results have been seen to be oncogenic in other experiments, and to confirm that findings which do not appear in the database may be novel and specific either to MPM or just to the individual patient based on their level of exposure, method of exposure or germline genetic variation.

This project focuses on the generation of phylogenies from tumour samples taken from MPM patients in order to determine; a recurring pattern in which the tumour originates and then spreads, driver mutations which occur during the latency period of the tumour and cause it to develop into a full cancer, and to determine whether specific mutations result in significant differences in patient

outcome (as well as whether these mutations are more likely to produce recurring mutations further downstream).

Limitations of this methodology for MPM samples include; the assumption that all samples taken from separate anatomical positions will be derived from the same original tumour, and that germline DNA is used as the precursor or rather, root of the tree. In answer to the first limitation, it could be argued that there is a high likelihood of all tumour tissue in a single patient being derived from a single original clone. There is a long latency time before the cancer enters its aggressive stage, triggered by a mystery event that the analyses is trying to uncover, and during this latency period mutations are accumulating throughout the DNA across the cells in the tissue. However, once the event occurs the expansion of an individual clone is rapid, and the likelihood that two clones would hit this mystery driver event in a close enough time period to compete seems like an unlikely event.

In regards to the second limitation, it is true that the germline DNA may not be the best precursor, as the driver event will be occurring on an already heavily mutated genomic landscape. However, the driver event will undergo rapid positive selection in the cancer cell population, resulting in the highly variable mutated landscape being statistically insignificant when simply compared to the germline. There's also a problem with practicality, MPM is almost never identified until after the driver event has occurred and the tumour has begun to proliferate, by which point it would be too late to collect the pre-driver tissue without contamination.

The methods employed by other phylogenetic studies in cancer, and in particular the methods performed by TRACERx in lung cancer are a suitable template in MPM for these reasons.

1.6) Aims of the Thesis

This thesis was founded on the hypothesis that the use of phylogenetic inference methods would allow “driver” events to be identified in Malignant Pleural Mesothelioma, and that the identification of these events would be beneficial in the selection of future drug targets to treat the disease.

The aims of the thesis are: to establish a pipeline which can handle both copy number and single nucleotide variant data in order to generate phylogenetic trees in order for truncal (or clonal) events to be identified; test the robustness of this pipeline via comparisons of different software; identify possible trajectories between any identified truncal events to try and establish true “driver” variants.

2) Methods and Materials

2.1) Introduction to Project Pipelines, Methods and Materials

This chapter covers the development and deployment of the three bespoke pipelines used in this project to detect driver events in MPM. It covers the justification and description of the methods used throughout the project as well as information about how the data was initially acquired and processed prior to the start date of the project. An outline of each of the pipelines is given here as well as reference to the relevant section of the chapter which addresses the methods used for each.

The first pipeline focused on copy number variation in MPM, which as discussed in Chapter 1, is well established in the genomic landscape of the cancer. As such, the methods employed for this pipeline focused solely on copy number variants found in the data and used only these in the subsequent phylogenetic analysis. The pipeline consisted of two copy number estimations softwares, Sequenza (Favero, Joshi et al. 2015) and ABSOLUTE (Carter, Cibulskis et al. 2012), to be compared using an orthogonal approach, with the results then being used in a phylogenetic generation software, TuMult (Letouze, Allory et al. 2010), which uses only copy number data to generate trees. A further copy number dataset generated by Novogene, using ASCAT (Van Loo, Nordgard et al. 2010), was also used in this pipeline to generate a second set of phylogenetic results, which were then compared to the original as an orthogonal approach. The last objective of the first pipeline was to attempt to infer anatomical spread of MPM on the pleura based on event sequence in the

phylogeny. Section 2.3 of this chapter covers this pipeline with results shown in chapter 3 (from section 3.3 onwards).

The second pipeline is focused entirely on single-nucleotide variations in the dataset (though does use copy number calls for correction), and was established to explore the impact of these SNVs in the development of MPM as discussed in Chapter 1. Similar to the structure of the first pipeline, this one used two single-nucleotide variant callers, VarScan2 (Koboldt, Zhang et al. 2012) and MuTect2 (Cibulskis, Lawrence et al. 2013), to establish an orthogonal approach and provide a measure of quality control in the selection of individual SNVs. Similar to ASCAT, the MuTect2 data was generated by Novogene and then used in this project. The data generated by this method was then analysed using another phylogenetics software which is designed to operate using SNV input. Section 2.4 covers the methods for this pipeline with chapter 4 containing the results.

The final pipeline used a single piece of software, Revolver (Caravagna, Giarratano et al. 2018), alongside results which were generated by the previous pipelines to establish any recurrent trajectories observed, i.e. it compared the phylogenies generated to attempt to find patterns in key events called by the other pipelines. Section 2.5 of this chapter covers this final pipeline with result shown in chapter 5.

Chapter 2 aims to establish an understanding of how the initial data was extracted and processed from the patients, as well as a general understanding of the structure of each pipeline and the software/computational steps taken to

generate the project results.

2.2) Sample Extraction, Sequencing and Assembly

2.2.1) Sample Selection and Extraction

The raw data for this project was obtained from 25 patients who were part of the MEDUSA cohort established by the Mesothelioma Research Programme at the University of Leicester. The patients were recruited into this cohort prior to scheduled Extended Pleurectomy Decortication surgery, which involves the removal of the pleura and subsequent removal of tumour tissue, and is considered a lung-sparing surgery due to it preserving the lung. Tumour tissue samples were taken from 5 anatomical locations, as described below, with additional steps then taken to ensure sufficient tumour content and prevention of cross contamination between each sample. 23 blade scalpels were used to cut 10 pieces of tumour tissue (or sometimes less if the tissue samples were too small) measuring approximately 1.5cm x 0.5cm in order to avoid cross contamination of samples. The samples were reviewed via a histopathologist via assessment of hematoxylin and eosin slides to ensure there was sufficient tumour content present in each sample. Prior to each surgery, multiple blood samples were taken from each patient and frozen, with DNA extracted from these blood samples acting as matched normal samples to provide a germline reference for each patient. These steps were performed by the MEDUSA surgical team.

The samples for each patient were numbered 1 to 5, with sample 2 only being present in those 6 patients who had 5 samples extracted. Sample 1 was taken from the apex of the lung which is located at the top of the structure where it appears to form a point. Sample 2 was taken from near the pericardium, which

is the tissue that surrounds the heart. Sample 3 was taken from the anterior costophrenic angle, which is located at the bottom tip of the lung structure and on the ventral side. Sample 4 was taken from the posterior costophrenic angle, located at the bottom tip of the lung but on the dorsal side. Sample 5 was taken from the oblique fissure, a fissure on either lung present towards the bottom of the structure. The oblique fissure was selected over the horizontal fissure as it is present on both lungs and so gives better options when selecting tissue to extract, whereas the horizontal fissure is only present on the right lung and so may not be suitable for extraction in all cases. Figure 2.1 demonstrates the approximate regions from which the samples would have been taken.

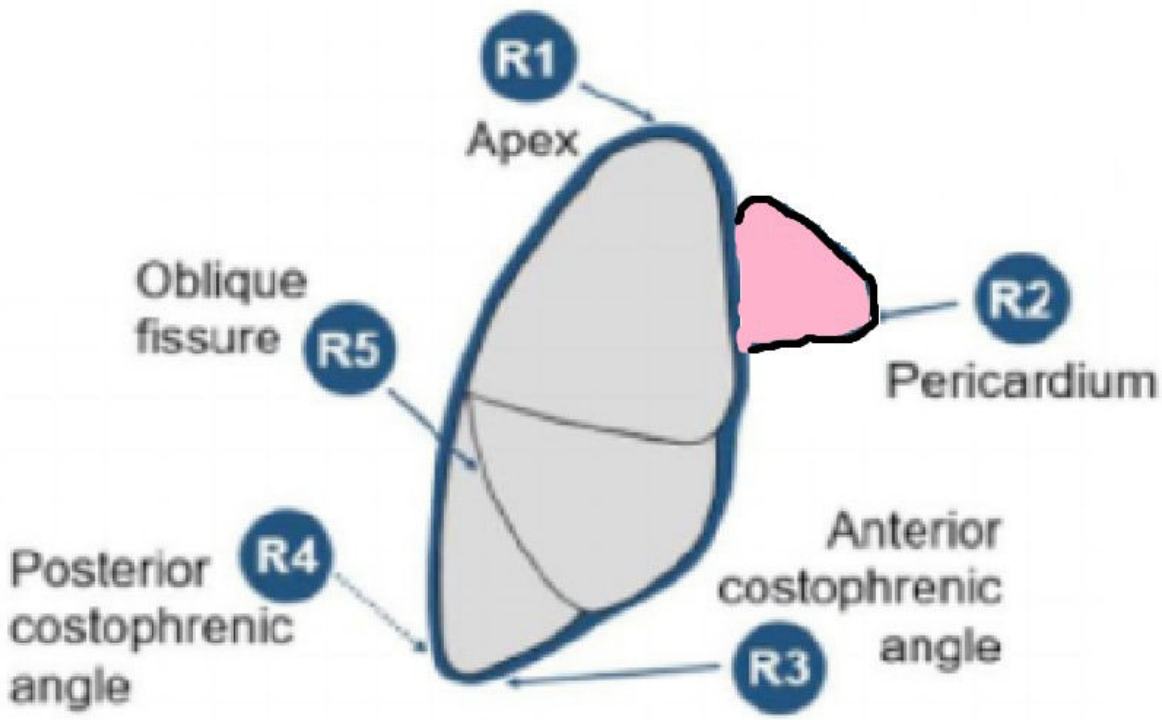


Figure 2.1: The approximate locations where the biopsy samples were taken from MPM patients. 1) the apex of the lung. 2) the pericardium or surrounding tissue. 3) the anterior costophrenic angle of the lung (front-facing). 4) the posterior costophrenic angle (rear-facing). 5) the oblique fissure (present on both the left and right lungs, unlike the horizontal fissure only present on the right lung). In patient 27, sample 2 was extracted from the anterior chest wall instead of the pericardium, but the relative position would be similar to that shown in the diagram. Diagram adapted from (Zhang, Jin-Li et al. 2021).

The histopathologist assessment led to the Sample 2 (pericardium) samples being excluded from 19 patients, due to lack of tumour content, with only 6 patients retaining these samples for the purposes of further analysis (1, 12, 23, 24, 27 and 34). (It should be noted that the numbering of the patients is arbitrary and not reflective of any patient data). All samples were flash frozen using liquid nitrogen in order for DNA extraction and sequencing at a later date.

2.2.2) DNA Extraction and Exome Sequencing

DNA was extracted from the tumour tissue samples and the blood samples using the QiAamp mini kit (Qiagen) and QiAamp DNA blood mini kit (Qiagen) respectively, following the manufacturer's instructions each time. After the DNA was extracted from the tissue, whole-exome sequencing was performed using Illumina sequencing and an exome sequencing library, giving a final mean coverage of 276X in both the normal and tumour samples. In order to perform sequence alignment, the raw reads underwent quality control to remove low quality reads and reads that were contaminated by an adaptor. These quality controlled reads were then used in the Burrows-Wheeler Aligner (BWA) (version 0.7.17), which is the most well established sequence alignment software, to map the reads to a human reference genome (Li, Durbin 2009). The hg19 human genome taken from the UCSC website (Church, Schneider et al. 2011) and was used as a reference for the alignment. The reads and human reference genome were then run through the mem algorithm in the BWA with all the parameters set to default values. Once the alignment was completed, the newly generated BAM files were sorted using the Sambamba tool (version 0.6.7) (Tarasov, Vilella et al. 2015) which is a software used to manipulate SAM and BAM files, and indexed. Duplicate reads in the BAM files were then marked using Picard tools (version 2.18.9). It should be noted here, that all prior steps discussed so far in this chapter were not performed by the author of this thesis, and that the final BAM files were the starting point in terms of data for this project.

2.3) Copy Number Based Pipeline Methods and Software

2.3.1) Selection of Sequenza as Copy Number Calling Software

Copy number variations (CNVs) are structural mutations which occur in the genome and are defined as the multiplication or deletion of DNA segments which have a length of greater than 1kb bases (Zare, Dow et al. 2017). The presence of these CNVs have been associated with various complex diseases, including schizophrenia, autism and multiple cancers (Zhang, Gu et al. 2009) though the specific biological mechanisms involved can be difficult to decipher due to the large allelic variance in CNVs, i.e. whereas a single-nucleotide polymorphism can exist in one of three states, the amount of differences that can occur within a CNV segment over a certain region are vast (Lupski 2007). Significant association has been specifically found between CNVs and cancers (Greenman, Stephens et al. 2007; Meyerson, Gabriel et al. 2010; Speleman, Kumps et al. 2008). It has been observed that cancer is a disease mainly consisting of somatic mutations which occur in the genomes of normal cells and thus cause them to develop into tumours (Shlien, Malkin 2009) though susceptibility to developing cancer can be found in the germline. Of these somatic mutations, CNVs are widely considered to be one of the most significant in the development of a tumour, due to the inactivation of tumour suppressor genes from deletions (both heterozygous and homozygous) or the activation of oncogenes due to the amplification caused by increases in copy number (Albertson, Collins et al. 2003; Beroukhim, Mermel et al. 2010). As such, it follows that identifying somatic CNVs and the genomic regions they incorporate is essential in elucidating the biological mechanisms involved in tumour development and in improving cancer treatment.

Whole-exome sequencing (WES) is a popular method for identifying mutations in cancer tumours, as it can provide a high coverage compared to whole-genome sequencing (WGS) and can be performed at a lower cost (Rabbani, Tekin et al. 2014). Unique methods are required for estimating copy-number in WES data, as software already developed for the analysis of WGS is unsuitable due to the hybridisation problem. There are several factors that make detecting CNVs in WES cancer data difficult, firstly somatic CNVs tend to have a much greater coverage of the genome than germline variations would (Kidd, Cooper et al. 2008), meaning the same techniques cannot be used for both and novel software must be made. Furthermore, the unknown ploidy within tumours means that this value must also be calculated when searching for CNVs (ploidy refers to the number of total copies of the genome, with two being the normal amount in diploid organisms). The most major difficulty factor is the cellularity of the cancer tissue, which is further complicated by possible contamination from normal tissue in the sample, and that there is a possibility that several distinct cancer clones could also be present in a single sample. Despite these challenges, a plethora of software has been developed to accomplish the detection of CNVs in WES cancer data (Zare, Dow et al. 2017; Zhang, Bai et al. 2019).

Detecting CNV mutations from sequencing data is performed in one of three ways; using a pair-end approach, using an assembly based approach, or using a read-depth approach. As the reads required by the pair-end and assembly based approaches need to be longer than the very short reads produced by WES, these methods cannot be used in the identification of CNVs unless WGS is used. Therefore, all methods developed to handle WES must use the read-

depth approach when estimating copy-number (Zare, Dow et al. 2017). The read-depth approach basically functions by comparing the number of reads present across every region of the genome, and then assigning a copy-number to that region based on how the ratio of total reads compares to neighbouring regions as well as to a matched normal sample in most cases.

There are several published software packages which have been specifically designed for use with cancer data, with a portion of them also specifically catered towards dealing with whole exome sequencing data. ASCAT (Van Loo, Nordgard et al. 2010) is a CNV-caller which determines copy number in cancer by comparing matched normal and tumour samples to a list of already known single nucleotide polymorphisms (SNPs) and compares the read depth of the SNP loci between the three. It also determines ploidy and cellularity estimates for a given patient sample, and then uses the values calculated to call a copy number profile across the regions provided. CNVkit (Talevich, Shain et al. 2016) targets both targeted and random non-targeted reads to infer copy number by applying corrections based on the notion that the random off target reads have less bias. For whole exome sequencing data, there's ExomeCNV (Sathirapongsasuti, Lee et al. 2011) designed to call copy number and loss of heterozygosity, based on short mapped reads, providing read depth and B-allele frequencies. Considering multiple methodologies for the calling of copy number variants is essential is ensuring that the one used is the most suitable for the pipeline and downstream software.

The first major analysis pipeline used in the project was assembled using software which could call copy-number changes from exome data, and

phylogenetic software which could then use the resulting data to build a tree for each patient. The first step was to select a copy-number calling software with the following criteria; the ability to accept exome data as an input, the ability to accept data generated from cancer tissue which would be expected to have lower cellularity, and finally the ability to distinguish between changes in ploidy and large-scale copy-number alterations. The last criterion was necessary due to the possibility of MPM causing very large copy-number alterations that may effect the entire length of a chromosome arm but were not actually caused by a full chromosome replication (or deletion). Multiple options were considered for a software to fulfil this role, and eventually Sequenza (Favero, Joshi et al. 2015) was chosen. A further software was selected called ABSOLUTE (Carter, Cibulskis et al. 2012) which would be used to support the Sequenza software (details discussed later in the chapter). Sequenza would be used in the main analysis pipeline to generate the copy-number input, whilst ABSOLUTE would be used as a comparison to the ploidy and cellularity estimates made by Sequenza.

Sequenza is a software package specifically made to incorporate matched normal and tumour exome data (or whole genome data) to infer copy-number profiles based on estimates of cellularity and ploidy in the tumour samples (Favero, Joshi et al. 2015). It is divided into two major software components; a Python script used to process the sequencing data and create the input for the second component, and an R package which uses a probabilistic model to generate the copy number profiles for each sample. Sequenza was selected over other options due to it being developed specifically for use with tumour data, and the ability to work with samples with very low cellularity with copy

number calls being accurately reported at cellularity values as low as 25%. Furthermore, it isn't limited to genome data and can readily accept exome data as well, although it cannot take BAM files as raw input. It was selected as the primary copy-number calling software over others due to having reported higher accuracy, especially at lower cellularity values (Favero, Joshi et al. 2015). Sequenza uses B-allele frequency (BAF) as well as depth ratio in order to establish the parameters for its probabilistic model. The BAF refers to the ratio of allele frequency between the A and B alleles, such that a BAF of 0.5 would indicate an equal proportion of both alleles being present. The depth ratio indicates the sequencing depth of any given segment of the exome compared to normalised values. This can be used to identify regions of deletion or gain where an unexpected read depth is present. Sequenza also produces output in a tabulated text format, making it useful when converting the data to generate input for the subsequent step of the pipeline.

2.3.2) Generation of Sequenza Input

As stated in section 2.2.2, the sequencing data was assembled and then stored in BAM format, with each patient consisting of 4 or 5 tumour samples and 1 normal tissue sample, meaning there were 5 or 6 large BAM files to process per patient. The first step in using Sequenza required the sequencing data to be in pileup format, so the pipeline required some preliminary processing before Sequenza could be used. Pileup format is similar to BAM format in that they both contain sequence alignment data in a text format, though pileup files are not limited to segment sizes of 128mb and can also facilitate the visual display of this data. The generation of the pileup files is the most time-consuming and computationally heavy step in the Sequenza pipeline, and acts as a practical limitation in the software. This can be overcome by splitting the BAM files before conversion into pileup format, and then re-merging the files at the end of the pre-processing steps before the R-based analysis component of Sequenza is performed. Sequenza includes a function in its Python script to allow for this.

Pileup files were created from the already generated BAM files using the *mpileup* function in SAMtools (Li, Handsaker et al. 2009), which is a software suite widely used for the manipulation of alignment data. Version 1.9 of SAMtools was used for all files. Creating a pileup file required a reference genome in FASTA format , so the same UCSC reference file that was used previously to generate the BAM files was also used here (as mentioned in chapter 2.2). The -Q value for the *mpileup* function was set as 20, meaning that any reads with a quality below that value would not be considered in the assembly of the final file. Sequenza also required a GC content file alongside the pileup files in order to normalise the depth ratio calculations used in the first

analysis step. This file is required to be in *wiggle track* format (Kent, Zweig et al. 2010) and was produced using the Python script included in the Sequenza package, with the same genome reference file mentioned previously used as the source. A -W value of 50 was selected as recommended by the Sequenza documentation, which enforced a window size of 50 bases in the creation of the file.

The pileup files then needed to be converted into the unique Sequenza format, *seqz*, which were made using the Python script, and for which the GC content file is required, as well as a normal pileup file and a tumour sample pileup file. This script extracted sequencing depth from the pileup files, with a minimum requirement of at least 20 reads across both the tumour and normal pileup files for each genomic region. It determined homozygous and heterozygous loci in the normal sample and then called variant alleles in the tumour sample followed by the allelic frequency. The output of this script is a tab-delimited text file, which can be easily imported into R for the analysis step. However, an optional step was taken to bin the data using a further function provided in the Python script. This was done in order to improve performance in R, and is also recommended in the Sequenza user guide to mitigate the long processing times and is stated to have a negligible effect on the results. A window size of 50 was selected for the binning function. This binned output was then ready for analysis in the R component of Sequenza.

It should be noted that initially, version 2.1.0 of Sequenza was used to process and analysis the data in this stage of the pipeline. However, in May 2019 a new version of Sequenza (3.0.0) was released and so the samples were re-

analysed. The updated release made few changes to the probabilistic model and were mostly focused on improving the performance of the software. Although runtime was greatly reduced, there were no noticeable changes in the copy-number profiles produced by Sequenza.

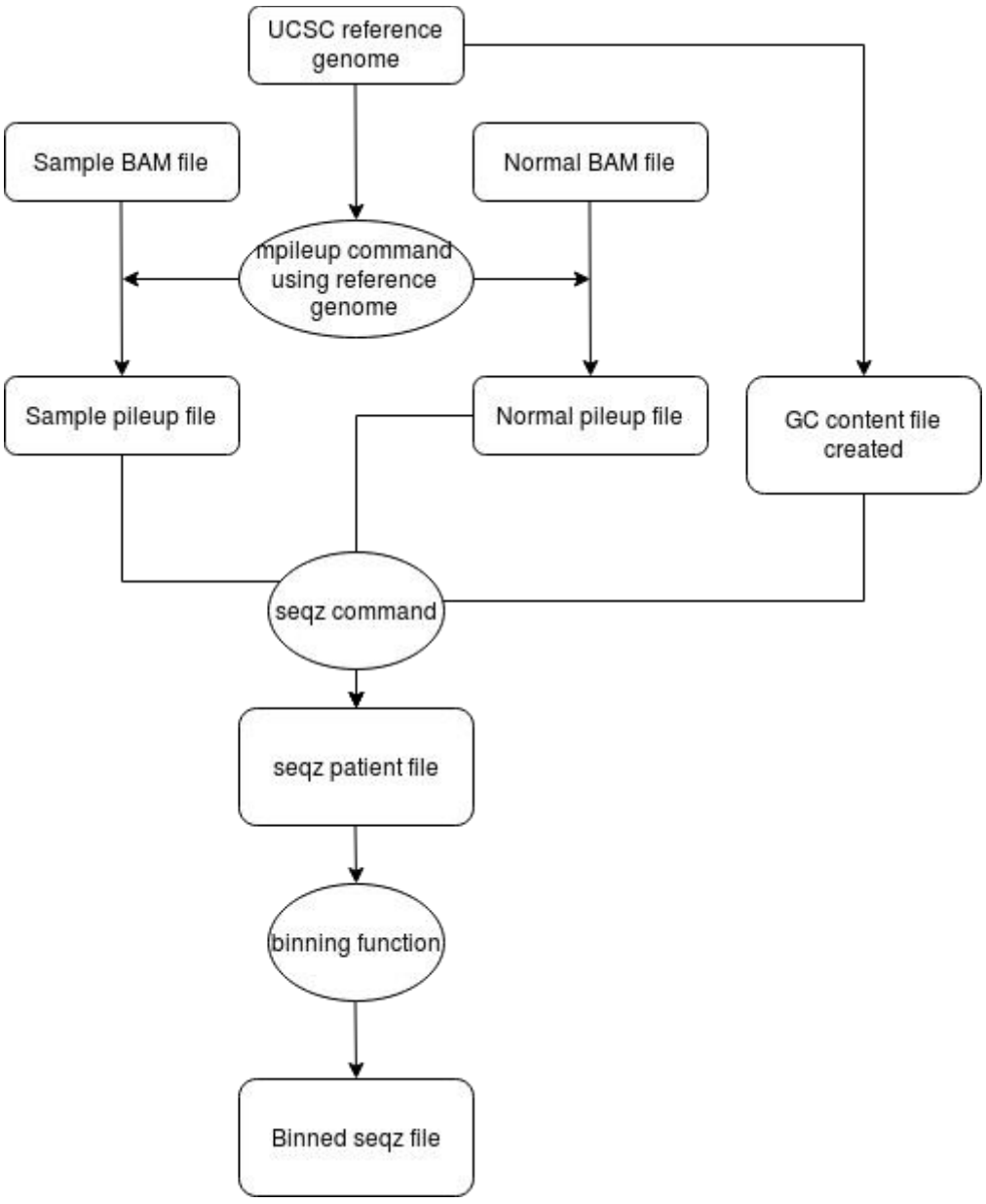


Figure 2.2: Flowchart displaying the general pipeline from the BAM files generated by Novogene up to the point where the binned seqz files were ready to be imported to R for use as input.

2.3.3) Sequenza Copy Number Analysis in R

All 106 *seqz* files were then analysed using the R component of the Sequenza package, which is split into 3 major functions. The *sequenza.extract* function is run initially in order to read the input files into R and normalise the normal versus tumour depth ratio using the GC content file. This function can then use the normalised depth ratio to establish distinct allele-specific copy number segments in the exome by establishing breakpoints where the depth ratio fluctuates. The second function is *sequenza.fit* which applies the probabilistic model for estimating copy-number profiles. The model incorporates the copy number segments calculated in the previous step, as well as the copy number of the B allele (which is defined as the allele with minor frequency). The function then infers both ploidy and cellularity estimates and uses these 4 parameters to generate copy-number profiles for each sample. The model defines prior probabilities for copy number so that 2 copies will be preferred by default. This function also provides alternate solutions along a range of ploidy and cellularity values. There is a secondary implementation of this function which is available, that allows for cellularity and/or ploidy values to be given to the model from the onset, enforcing those values and removing their estimation from the work flow. This can be done in cases where two alternate solutions are extremely close in probability but other external data can provide evidence towards one. A use of this secondary implementation will be discussed later in this chapter. The final function of the R component is *sequenza.results* which simply formats and returns the results of the previous functions and allows for visualisation of these results. Each function feeds directly into the next, with all output being produced by the final function.

2.3.4) ABSOLUTE Copy Number Estimation

Although Sequenza was selected as the software to be used in order to estimate copy-number segments in the exome data, it was decided that additional software should be incorporated into the pipeline in order to validate the ploidy and cellularity estimates made by Sequenza, and thus, provide an orthogonal approach in determining copy number estimates for the pipeline. A suitable software would need to be able to handle Sequenza output or WES input, and would need to have been developed with cancer data specifically in mind. An immediate obvious choice was ABSOLUTE (Carter, Cibulskis et al. 2012), which was a well established software that has been used in numerous cancer-based studies to provide estimates of cellularity and ploidy (Hu, Estecio et al. 2021; Krause, Roma et al. 2021; Yu, Chen et al. 2019). ABSOLUTE is also mentioned directly in the Sequenza paper (Favero, Joshi et al. 2015), where it is used to test against the cellularity and ploidy values estimated by Sequenza. Although in the paper, Sequenza is reported to produce more accurate values when compared to a range of other estimates from the same samples, it should be noted that ABSOLUTE was not using Sequenza generated output for these calculations, with both softwares using pre-computed segment files. Further to this, in this project, ABSOLUTE is used only as a comparison to the predictions made by Sequenza, via comparison of results from both pieces of software.

Other software was looked at including PyLOH (Li, Xie 2014) and qpure (Song, Nones et al. 2012), though both were unable to determine accurate estimates from exome-based data, and only worked with whole-genome results. There

was in fact, only a single other software available at the time which could handle exome data, AbsCN-seq (Bao, Pu et al. 2014). The AbsCN-seq paper refers to ABSOLUTE at the time as the “gold standard” for estimating ploidy and cellularity in cancer data, and whilst this is not validation enough by itself, it is certainly a strong recommendation. The main difference between the two softwares is that ABSOLUTE only uses the segment file input in its calculations, whereas AbsCN-seq uses independent data alongside the segment file input. As the selected software was to be used for comparison purposes, it was decided that ABSOLUTE would be the better choice as the lack of any independent data input for AbsCN-seq may influence the results and create bias in favour of the estimates matching those produced by Sequenza.

ABSOLUTE is a computational method which functions to infer both cell ploidy and tumour purity (cellularity) using pre-computed copy-number segments (in this case, produced by Sequenza) and pre-computed recurrent cancer-karyotype models which are provided by ABSOLUTE itself. It can also take an optional additional input file, called a point mutation file, which is essentially a file containing single-nucleotide polymorphisms called on the cancer data, though, at the time of analysis this additional data was not available. ABSOLUTE first estimates ploidy and cellularity from the copy-number profiles provided in the segments file using a probability model. It then uses the karyotype models from a large internal collection to resolve samples where an estimation could not be reached using the probability model alone. It selects the most simple karyotype solution (or rather the one that appears most commonly) that fits the data, drawing from a large range of karyotype models, from haploid genomes to hyper-aneuploid genomes (>6 whole copies).

2.3.5) ABSOLUTE Input and Copy Number Analysis

The input for ABSOLUTE was generated using a perl script to extract the relevant information from the segments file produced by Sequenza. The copy-number segments file could be provided either in HAPSEG format (a specific format created for use with the HapSeg software), or as a simple tab-delimited segmentation file. The basic file was chosen as it would be easier to convert the Sequenza format into that format than into the HAPSEG format, and either formats should produce identical output. The input format required the chromosome number, start position and end position, and the copy number value for each segment. All of this information was extracted using a simple Perl (v 5.18.2) script, though segments with a size of less than 10kb or a total copy number above 10 were not included and were discarded at this stage. This was decided as very small segments are likely to be artifacts from the copy-number calling software, with the same logic applied to segments with a total copy-number size of over 10. In fact, the only segments removed for having a copy-number size of above 10 were also segments which were smaller than 10kb. Once the input was generated, it could be used with the R-based ABSOLUTE package. R version used was 3.4.4. ABSOLUTE version used was 1.0.6.

ABSOLUTE runs with a single command, and the majority of the arguments were kept as the default values. There were exceptions to this though, as the “min.ploidy value”, which discards any solutions that estimate a lower ploidy, was set to 1. Genome was set to “hg19”, and the platform was set as “Illumina_WES”. The only other argument altered was “copy_num_type” was set to “total”, which was required as the generated input used the tab-delimited

format. All 106 samples were successfully run through ABSOLUTE, providing a plot displaying the estimated solutions, a plot displaying copy number, genomic fraction and copy ratio, and an Rdata file containing all the values used to plot the graphs.

2.3.6) Selection of TuMult as a Phylogenetic Analysis Software

As discussed, a major aim of this project is to identify driver mutations in the evolution of MPM which occur early in the process through the use of phylogenetics, with the ultimate aim of providing possible druggable targets which can arrest the process before the cancer reaches its aggressive form. As stated previously, MPM has been shown to be mainly driven by copy number alteration events as opposed to single nucleotide variations (though these events may still be involved). As such it follows that a phylogenetic software which focuses on copy number events should be selected for use in the initial pipeline. Several algorithms incorporating different methods were considered, though TuMult (Letouze, Allory et al. 2010) was chosen as the software to be used in conjunction with Sequenza to generate the first phylogenetic results. TuMult is an R-based software which consists of a single R script containing multiple functions, and calculates tree topologies and maps copy-number variation events based on copy-number estimation profiles (Letouze, Allory et al. 2010).

TuMult functions by mapping common breakpoints between samples in each patient and follows the assumption that if a breakpoint has occurred in an identical location (or a location which is extremely close) on the exome, then it is unlikely that this event occurred on two separate occasions in the evolution of the cancer and so must be representative of a single event which happened prior to the samples diverging. This is in contrast to other methods which use regions of corresponding copy number between patients and assume that they occurred from a single event regardless of the segmental breakpoints. This method can mistakenly call convergent events as single events and so has a

tendency to place events earlier on phylogenetic trees than where they actually occurred. The use of breakpoints by TuMult is more reliable and allows for a more accurate prediction of early events. TuMult uses a neighbour-joining inference method to calculate tree topologies, assigning events to branches based on the lowest available evolution distance. This means it tries to establish the lowest total overall distance in the output tree and will always produce the same tree from the same input data. This basic assumption behind this is that samples which have more similar events in common diverged more recently than samples which are more different, and that the more frequently an event occurs (assuming the same breakpoints) the earlier in the tree that event will be placed. Unlike certain neighbour-joining methods TuMult does make the assumption that single events are all equal in terms of evolutionary time and so branch length on the outputted trees does not correspond to time and instead simply reformats to allow for the visualisation of all events which have occurred on any given branch. Due to this assumption, by default, TuMult treats every single event as significantly as any other (Kannan, Wheeler 2012).

TuMult was also selected as its required input was easily generated from a text tabulated format provided by a software like Sequenza. In terms of methodology, TuMult uses a neighbour-joining method, which is compatible with each patient containing multiple samples. TuMult would not be suitable for analysis on patients with only 1 or 2 samples, although it would still be able to provide some information in the case of 2 samples as it could call events which occurred prior to the samples diverging. Another major advantage of using a neighbour-joining method is that it allows for extremely quick runtimes when compared to alternative methods such as maximum parsimony or maximum

likelihood (Kuhner, Felsenstein 1994). For these reasons TuMult was selected as the most suitable software for generating the phylogenies for this component of the project. The limitations of TuMult were outweighed by its suitability to the dataset and for the practicalities of the software which allowed for quick turnover despite limited available computational resources. TuMult version 1.0 was used for all analyses.

2.3.7) Generation of TuMult input

The input for TuMult comprises of 4 distinct components: a text file containing positional information of all segments per patient to be analysed including cytogenetic bands (probe file), a text file containing copy number estimates per sample of all segments to be analysed (profile file), an Rdata reference file containing a reference dataset derived from unrelated patients, and an integer value which defines the number of probes below which two breakpoints from two samples are considered identical (breakpoint value). The two text files could mostly be generated directly from the segments file produced by Sequenza for each sample, though several steps were required in order to build the correct input format. This is the step in the analysis where all samples for a patient are combined into a single input, as the profile file contains the copy number data for all samples per patient. The first major problem to overcome was that both the probe file and the profile file needed to have an identical number of lines as each line in each file need to be paired with the corresponding line in the other file. This was problematic, as the segments called by Sequenza were not of equal length and varied greatly between patients, meaning all profile files would not be able to directly correspond to a single probe file containing defined positions. The solution to this problem is addressed in the TuMult user guide, the data could be split into equally sized bins and the copy number of each new binned segment assigned a value based on the segments file produced by Sequenza.

Choosing the size of the bins which the segments would be split into was a crucial step in generating the TuMult input. Initially a value of 10kb was chosen, though it was decided that the subsequent trees produced contained a lot of

noise from smaller segments which tended to have much higher copy number values than the average for the same sample. This is demonstrated in Figure 3.20 which is the TuMult tree produced using a 10kb bin size. A bin size of 100kb was then selected, as the majority of high confidence events had a total size greater than this value, whilst it also allowed for many to be ignored. Setting the bin size to 100kb meant that any segments with a total length below that value would be removed from the dataset at this stage and so would not be represented in the TuMult input. A comparison between the two bin sizes can be seen from Figures 3.19 and 3.20, displaying trees for Patient 1 Sample 1 with a 100kb bin size and a 10kb bin size respectively.

The binning method was performed using a small collection of scripts generated using the programming language Perl (Wall, Christiansen et al. 2000) version 5.18.2. In order to generate equal bins for each patient, the lowest positional value for each chromosome from each sample was set as the starting point of the bins, and the highest positional value for each chromosome from each sample was set as the ending point. This was necessary as the chromosome lengths between samples was not always the same for different samples in the same patient. Once the start and end points for each chromosome had been selected, the probes files could the be created by generating a 100kb segment (or probe) from the start point, and adding a new segment in 100kb increments until the highest positional value on the chromosome was reached or exceeded. Using these 100kb probes, the copy number information could then be extracted from the segments file produced by Sequenza. The copy number value in each region defined by the probes was then written to a new file, which would act the profiles file for TuMult. In segments where there were two copy

number values (essentially, where a probe would overlap the breakpoint of two segments), the copy number value was set as the non-normal (not 2) value for the entire probe, or if both were non-normal it would keep the same copy number value as the previous probe.

This allowed for the profiles and probes files to generated with matching data on corresponding lines as required by TuMult, though there is an obvious drawback to this method in that the absolute breakpoint value is lost and rounded to the next 100kb artificial breakpoint. This seems like a major limitation initially as it opposes TuMults unique selling point of using unique breakpoints to better define single mutation events. However, the breakpoint value (4^{th} input for TuMult) is implemented in order to account for this issue. The default breakpoint value integer given by TuMult is 2, which is the value that was used for all patients in this pipeline. A value of 2 means that as long as a breakpoint was within 2 probe lengths of a breakpoint in another sample, then they would be counted as identical breakpoints. This is an acceptable compromise, as in order to be counted as a single event by TuMult, the breakpoints need to be identical at both ends of the event, and breakpoints occurring within 2 probe lengths at both ends of a single event should be an extremely unlikely event. The breakpoint value was adjusted from a range between 2-10 to see what effect it would have on the produced trees, though no difference was seen with any value. This implies that there were no events in any patient with occurred close enough that increasing the acceptable number of probe lengths from 2 to 10 allowed for any other events to be included.

It should be noted that TuMult does not directly record the specific copy number

for each of the binned segments. TuMult uses its own numbering system which reflect copy number status; with 0 defining a normal segment, 1 defining a heterozygous gain (or a copy number increase of 1, meaning a total copy number of 3), 2 defining a gain with a total copy number value of 4 or any value higher, -1 defining a heterozygous loss (copy number of 1), and -2 defining a homozygous loss (copy number of 0). This means that TuMult does suffer from a loss of specificity in regard to total copy number, but this can also lead to improved tree calculations in the distance matrix as it will allow for more matches to be made. It could be argued because of this that TuMult has a slight bias in reporting more increasing copy number events.

The final input file required for TuMult is the Rdata reference set, which requires a dataset generated from unrelated patients with the same type of tumour. Unfortunately, no reference datasets were available for this analysis and so an artificial reference set was produced where every probe segment was reported as having a normal copy number value. It had to be ensured that the reference dataset contained an identical number of segments as the probes file in order for the TuMult script to accept it. This approach was decided upon as it seemed a completely normal artificial dataset would have the least impact on the results in the situation where no actual reference dataset was available.

The phylogenetic trees produced by TuMult are outputted as .dot files, which are non-human-readable files which need to be processed by additional software in order to produce actual images. A visualisation software called GraphViz (Gansner, North 2000) was used in order to convert the .dot files into png images.

2.3.8) ASCAT/TuMult Methods

It was suggested that a good way of validating the Sequenza/TuMult pipeline would be to run it using a different copy number calling software in the place of Sequenza, and see whether the results produced showed any concordance. Initially ABSOLUTE was considered as a candidate for this, but as ABSOLUTE does not produce its own set of copy number segments and instead adjusts a copy number profile given to it, a further copy number software would be needed anyway (as obviously Sequenza data could not be used in order to avoid bias for this analysis). Due to this ABSOLUTE was disregarded, and fortunately, the Novogene team was working in parallel on the same patient cohort using alternative pipelines and methods than the ones discussed in this project. One such software in the Novogene copy number pipeline was ASCAT (Allele-Specific Copy number Analysis of Tumors) (Van Loo, Nordgard et al. 2010).

ASCAT works by calculating copy number profiles, as well as estimates of cellularity and ploidy in a sample, based on sequencing read depth at specific single nucleotide polymorphisms (SNPs). These SNPs need to known prior to the use of the software and stored in a list that is given to ASCAT as one of the inputs, along with matched tumour and normal BAM files. It should be noted here, that the generation of ASCAT output was performed entirely by the Novogene team, and although the author of this thesis did then use this output to create input for TuMult, no collaboration on the running of ASCAT was done. Novogene generated the list of SNPs using AlleleCount (Raine, Van Loo et al. 2016) which is part of the ascatNgs software package, that was created in order to aid researchers in the use of the ASCAT software. They did this using a

reference data set taken from the 1000 genomes project (1000 Genomes Project Consortium, Auton et al. 2015) an initiative established to generate the largest catalogue of human genetic variation for use in genetic research, and used the AlleleCount tool alongside the matched normal and tumour BAM files generated for each patient and sample. This produced a normalised log transfer of read depth (LogR) from the tumour sequence data compared to the normal, as well as B-allele frequencies (BAF). After a GC-correction step, these values are then used in the ASCAT algorithm to generate copy number segment estimations, in a similar way as the Sequenza software does, as described in Section 2.3.1, though obviously the two algorithms are different and so perform differently. ASCAT was performed for all patients, with copy number profiles generated for the 17 patients in the established cohort of this project, producing profile graphs for each patient and the equivalent of a segments file, which contained the start and end positions for each segment, the major and minor copy number, as well as the B allele frequency. Results for this methodology can be found in Section 3.3.6.

2.3.9) Use of Branch Length to Infer Anatomical Spread of MPM

One of the objectives of this pipeline was to attempt to establish a possible recurrent anatomical trajectory of MPM as it spreads across the pleura. The average branch lengths were calculated for each sample in an attempt to try and determine if there was a clear order of divergence which could then be mapped to the physical space of the pleura. This was done by simply counting the number of events from the patient node up to the root of the tree. As such, the value of all branch lengths will exceed the total number of events called, as events on shared branches will be counted for all the samples which stem from that branch. The branch lengths can be used as a loose method to determine evolutionary distance for each sample, and due to the assumptions made by TuMult discussed earlier, lower values for branch length assume that the particular cancer population associated diverged more recently.

Unfortunately no results are present directly answering the question of this objective, though a table (Table 3.10) does display the total number of events and different branch lengths for each patient in section 3.4.3. Detailed discussion about the reasoning for this can be seen in section 3.5.2.

2.4) Single Nucleotide Variant Calling Software

2.4.1) Selection of VarScan2 as the SNV Calling Software

Single nucleotide variations (SNVs) are mutations that occur in the genome when a nucleotide is altered, resulting in a change in base. As opposed to CNVs, SNVs are not well associated with driving the progression of cancer, and instead tend to confer susceptibility (Deng, Zhou et al. 2017). Nevertheless, there are still a large amount of studies which find association in specific cases (Chen, Zhang et al. 2021; Gan, Carrasco Pro et al. 2018) and many variant callers designed to be used with cancer derived sequence have been developed within the last decade (Xu 2018). There are unique challenges in calling SNVs from cancer-derived sequencing data, with a major difficulty being the ability to distinguish between genuine low-frequency variants and sequencing artifacts. Somatic SNV calling generally follows a three-step process: a pre-processing step in which low-quality reads are filtered out of the sequence; a variant calling step which is the main part that different calling software will vary; and a quality control step which filters out calls that don't meet the threshold criteria of the calling software.

Published software which acts to call SNVs in cancer sequence include Shimmer (Hansen, Gartner et al. 2013) which uses statistical hypothesis testing to call somatic SNVs in either exome or whole genome sequencing data extracted from tumours. It is also specifically designed to work with samples that are highly contaminated, or have a high level of heterogeneity. MuTect2 (Cibulskis, Lawrence et al. 2013) uses joint allele frequencies alongside a probability model to calculate differences in Variant Allele Frequency (VAF), and

is also designed to work with samples that have very low cellularity values (<0.1). VarDict (Lai, Markovets et al. 2016) is able to detect variants in indels as well as in SNVs, with indels being small insertions or deletions in the sequence (of up to 50bp).

Although not considered to be as influential or frequent in cancers, there is still a large demand for SNV calling software that can be used with cancer-derived sequence. It is critical that the unique aspects of SNV calling software are examined before selecting one for a pipeline, in order to ensure that it complements the other elements, allowing for greater accuracy in the final output. Most studies into the genes associated with MPM development are mainly focused on copy number variation, with single-nucleotide alterations often only being used to call copy number segments by software such as ASCAT, and do not consider the possibility that single base changes at specific sites could also be a hidden factor in the transformation of the cancer from its latent state to its aggressive state. It should be noted that there has been no studies published at the time of writing, which explore the idea of single-nucleotide drivers in MPM that have been assessed using a phylogenetic strategy. A key feature of using phylogenetics to identify drivers is that it can be established whether any particular single-nucleotide variation (SNV) occurred early in the evolution of the cancer, or at least whether it took place before the initial divergence of cancer cell populations. It is these mutation events which have the potential to effect cellular function in such a way to cause the rapid replication of cancer cells. Though it is obviously unlikely that the occurrence of a single SNV could cause the same level of impact as a large copy number loss, it is possible that certain positions or certain combinations of SNVs are

significant in driving MPM.

(It should be noted here that the patient cohort was reduced from 25 to 17 due to the comparative analysis of Sequenza and ABSOLUTE as to be discussed in Chapter 3, and that the same 17 patients would be used in this second pipeline due to those findings. That is to say, the extremely low cellularity values reported for the excluded samples by Sequenza and ABSOLUTE would still have an adverse effect in this pipeline as it would have done in the Sequenza/TuMult pipeline.)

In order to explore this concept, an SNV caller would need to be selected to generate a list of variants present in the tumour samples from the patient cohort. The selected variant caller would need to fulfil the following criteria: be able to use whole exome sequencing data as input; be able to use data that was generated from cancer tissue samples, or rather data that was known to have lower expected values of cellularity; and be able to use matched normal tissue data in order to distinguish between germline variants that may have been present prior to the initial formation of cancer cells. There are several bespoke SNV calling software designed specifically to deal with cancer tissue sequence available. Of these, VarScan2 (Koboldt, Zhang et al. 2012) was selected to generate SNV calls for the purposes of this pipeline. VarScan2 could meet all of the listed criteria mentioned previously that are required to operate in this position, and uses a different calling framework than most other published softwares which tend to use Bayesian statistical probabilities to detect and evaluate variants. VarScan2 uses a heuristic statistical approach which is designed to operate better at higher read depths and samples with lower purity

(either through decreased cellularity or admixture of multiple cell populations), which are factors that often cause variant callers to struggle. Furthermore, a study published in 2013 evaluated different somatic variant callers, and found that VarScan2 performed the best overall, particularly at read depths of between 200-500 (Stead, Sutton et al. 2013).

VarScan2 accesses the tumour and normal sample data simultaneously in order to make pairwise comparisons of the read depth and and nucleotide calls at each position in the genome (or exome). To detect variants, the software uses a heuristic algorithm to independently determine the genotype at individual positions based on the variant allele frequency (VAF) (the proportion of non-germline alleles reported in the data). A heuristic algorithm is one which is designed to solve a problem efficiently at the cost of precision, and so does not guarantee a perfect solution but completes the problem much more quickly. VarScan2 also makes copy number alteration predictions based on differences in normalised read depth simultaneously. Based on which version of the algorithm is performed, VarScan2 can report either germline, somatic or loss of heterozygosity (LOH) variants in the data. LOH variants refer to positions which are heterozygous in the normal sample (i.e. have two alleles present at a given genomic position) which are then found to be homozygous in the tumour sample.

2.4.2) VarScan2 Input Generation and Throughput

For input, VarScan2 requires a pileup file for each patient sample and a pileup file for the matched normal sample in each case. These files had already been generated for use with Sequenza and so could be used again here without the need to recreate the pileups in SAMtools. The methodology for creation of the pileup files can be read in section 2.3.2. However, the pileup files were slightly adjusted as they were sorted using SAMtools so as to ensure the that chromosome positions were listed in ascending order, which is a requirement for the VarScan2 algorithms.

VarScan2 somatic algorithm (version 2.3) was used to call variants for each patient and sample in the cohort, initially using all default parameters except for the tumour purity, which was set to 0.5. This was selected at first as it was recommended in the VarScan2 user guide to use this value when actual values were unknown, however, cellularity estimates were available both from ABSOLUTE and Sequenza. To take advantage of the availability of this data, the ABSOLUTE and Sequenza cellularity estimates were combined for each sample, to give an average estimate of the tumour cellularity. These combined values were then used for each corresponding VarScan2 run, with the tumour purity parameter changed each time. A comparison of the Sequenza values, ABSOLUTE values, and leaving the parameter at its default setting was considered here, but the resources required, both in terms of time and computational power, were too large to justify it, though it would be an interesting route for further analysis in this pipeline and to assess the effect of different values in this parameter on the overall results.

Of the 106 total samples, 74 were successfully run through VarScan2 somatic, with the remaining 32 samples removed based on results of the comparative analysis between Sequenza and ABSOLUTE, leading to the removal of 8 patients as mentioned in the previous section. Each sample gave two output files in the form of tab-delimited lists, one for the variants called and one for indels reported. Indels refer to small scale insertion or deletion in the genomic sequence, of between 1 and 50 nucleotide bases. These indel calls were not used for any further analysis in this project though examining the potential of indels in the pathophysiology of MPM could be an interesting expansion in future research. The variants file contained SNVs detected by the VarScan2 somatic algorithm, though these were then processed using the processSomatic command which is part of the VarScan2 software package. This resulted in 4 output files: one containing germline variants, one containing LOH variants, and two containing somatic variants, with high confidence calls and low confidence calls stored in each file respectively. A somatic call was considered to be high confidence if it had a p-value of less than 0.05 with adjustments for multiple comparisons and false discovery rate performed by the somatic algorithm. The high confidence somatic files were the ones used throughout the subsequent steps of this pipeline, as germline variants and LOH variants were not the target of this investigation. That is to say, somatic variants are the ones where the most interest lies as they are the variants that have occurred within the cancer cells, whereas germline and LOH variants would be more suited to when searching for cancer susceptibility.

2.4.3) Quality Control Using MuTect2

Further quality control steps were then taken to ensure the variant calls were of the absolute highest standard for the rest of the pipeline, the first of which was a comparison to variant calls made by a separate software called MuTect2 (Cibulskis, Lawrence et al. 2013). MuTect2 was the software of choice of Novogene, who were using it for variant calling in their own pipeline which was using data from the same patient cohort, as with ASCAT mentioned previously. MuTect2 was also briefly discussed in Section 2.4.1, as a variant caller which uses a Bayesian statistical model to detect somatic alterations in cancer tissue derived sequence. MuTect2 was specifically designed to work with data where there was very low cellularity (as low as 0.1) and still produce accurate calls, making it useful in the analysis of cancer data. The reason MuTect2 was not originally chosen for use as the primary SNV caller in this pipeline is that the cellularity, although low, is not as extremely low as the cases for which MuTect2 was designed in mind of. Further to that, as mentioned previously, VarScan2 does not use Bayesian statistics in its approach to variant calling which is stated to help avoid certain pitfalls in the process. The MuTect2 results used for this quality control were produced full independently by Novogene, with the author of this project having no input into producing them. MuTect2 version 4.0.5.1 was used by Novogene and was run using default parameters.

The initial step of the additional quality control of the high confidence somatic VarScan2 calls was a direct comparison with the MuTect2 results, produced from the same BAM files but independently produced pileup files (though in theory the two sets of pileup files should be identical) and each patient and sample call file was compared to its complementary file in MuTect2. This was

done to create a list of consensus calls, i.e. the only calls carried forward through the rest of the pipeline had to be present in both the VarScan2 and MuTect2 outputs. The consensus calls were then processed to remove entries with a reported read depth of less than 30 in total or less than 5 in the tumour sample. Entries were also removed if the variant was detected in the germline more than 4 times, if the variant allele frequency in the tumour sample was less than 5, or if the variant allele frequency in the normal sample was greater than 1. This was to ensure that the variant calls labelled by VarScan2 as somatic were valid, as recurrent variants detected in the germline or in the normal sample would imply that the mutation is not fully somatic and was present in the genome prior to the first cancer mutations in the pleura.

2.4.4) Selection of PhyloWGS as an SNV-based Phylogenetic Analysis Software

In this section the selection of a phylogenetics software that specialises in inferring trees based on SNV calls will be discussed. The chosen phylogenetics software would need to able to infer clonality using SNV calls as input and also be able to work with data derived from whole exome sequencing.

A clear frontrunner was found with the PhyloWGS software (Deshwar, Vembu et al. 2015) which had the additional attractive feature of being able to incorporate CNV calls alongside SNV calls to account for changes in copy number and correct for this when considering the read depth in the SNVs. Due to MPM having a large amount of copy number variation, which had also been confirmed in this patient cohort, this feature was basically a necessity when trying to establish clonal SNV events, which could otherwise be obscured due to changes in read depth caused by gains or losses in copy number in the region incorporating the SNV. PhyloWGS had been created with exactly this issue in mind (though not specifically for use with MPM), and so was the perfect candidate for this pipeline. It should be noted that at the time of analysis, no other published software was available which contained this feature, making PhyloWGS essentially the only choice. Other SNV based phylogenetic software such as PhyloSub (Jiao, Vembu et al. 2014) or PurBayes (Larson, Fridley 2013) were also considered, but without addressing the issue of how CNVs will effect the read depth of SNVs, the results from software such as these would have likely been inaccurate. The importance of this issue can be expanded on even more when it is considered that SNVs that are present within important copy number change regions, are themselves more likely to be significant in the

progression of MPM than the majority of SNVs, and that they would likely be completely misplaced in any phylogenetic tree if the copy number effect of said region was not taken into account. It is important to note that despite the name and the user guide referring exclusively to whole genome sequencing (WGS), PhyloWGS is fully compatible with WES data, and this is confirmed in the frequently asked questions section of the online hub that supports the software.

PhyloWGS operates using a probabilistic Bayesian statistical model, and uses the read depth of the SNV input calls, corrected for using the CNV calls, to generate a distribution of trees. PhyloWGS then samples from this distribution using a MCMC (Monte Carlo Markov Chain) and reports the tree or trees which maximise the likelihood of the sampling run data. A MCMC is a method of sampling, used on a probability distribution, in which multiple chains are given random start points and through the directions of an algorithm which gives states that would increase the overall likelihood of said chain. The random simulation of data points, when run hundreds or thousands of times will begin to reveal the set of states which has the likelihood, i.e. the set of states (in this case, the tree) which appears most commonly. Although PhyloWGS reports quicker runtimes than other phylogenetics software in the SNV arena, it still requires a large amount of time and computational power to be used, which is a limitation with the method, though one that cannot be easily overcome.

2.4.5) Generation of PhyloWGS Input and Work-flow

PhyloWGS only requires two files as input, an ssms file (“simple somatic mutations file”) containing the SNV calls and supporting data, and a cnvs file containing the CNV calls used to adjust during the running of the software, however, multiple stages of pre-processing had to done in order to correctly format this input for PhyloWGS. The ssms file required a unique id identifier for each unique ssms reported in a patient, the total number of reference-allele reads at the loci reported for each sample in the patient, and the total number of reads (both reference and variant) at the loci, which requires recurrent SNVs between samples in a patient to be identified beforehand. Both the reference-allele reads and the total reads sections of the input had to be divided via a comma to indicate which value was from which sample, and they required to be kept in the same order throughout all entries.

This step was done using a Perl script, which took all samples, in the form of the consensus calls generated from VarScan2 and MuTect2. The script had to be able to identify when an identical SNV was present in multiple samples, and then record the information as a new single event for the ssms file. Here is where the first problem arose, as most of the SNV events reported by VarScan2 were not present in all of the samples in any given patient, meaning the information could not be extracted from the consensus calls file (as this only had variant calls). This meant that a second step would have to take place, where the read depth for samples that had did not have a variant that another sample did, would need to be extracted from the BAM file using the bam_readcount tool. This resolved this issue and meant that accurate read counts could be provided for all samples for each unique variant, even when

that variant was not present in a sample. To indicate that this was the case, both the reference-allele reads and total reads for that sample were set to the same value. An additional pair of values was also required in the ssms files, which were an mu_r value and a mu_v value. The mu_r value is the expected fraction of reference alleles in the reference population, and so ideally would have a value of 1. The value was set to 0.999 as recommended by PhyloWGS to account for the error-sequencing rate of Illumina. The mu_v value is the expected fraction of reference alleles in the variant population, and so would be expected to have a value of 0.5. The value was set to 0.499 as recommended by PhyloWGS to account for the error-sequencing rate of Illumina. These steps were all repeated for all 17 patients in order to produce ssms files for each.

The cnvs files were generated using a Perl script that used the generated ssms file and a Sequenza segments file to determine which SNVs were within the region effected by a copy number change event. The cnvs input file itself required a unique id identifier as with the ssms, as well as the region the CNV covered, the estimated copy number of the CNV, and the SNV entries present in the ssms file that are located within the CNV. It should be noted that PhyloWGS only considers CNVs that have SNVs within them for the purpose of analysis, and that the ssms field in the cnv file is a requirement, or the software will throw an error. All of these values could be acquired from the Sequenza output, though there were issues in the creation of the cnvs file. The main issue, was that if multiple CNVs in a patient effected the same SNVs, which was fairly common, then there was no way to indicate this in the cnvs file. For example, given an SNV that is present in 4 samples in the ssms file, if in 2 of the samples the SNVs are effected by a copy number loss from a particular CNV, and the

other two samples were effected by a copy number loss from a different CNV (that covered the same region but was distinct based on its start and end points) then there was no way to inform PhyloWGS that this was the case based on the required format of the input. This actual phenomenon occurred frequently during the creation of the cnvs input, but there was no solution to resolving the issue and so all CNVs simply had to be included with PhyloWGS presumably assuming that all CNVs were effecting all SNVs in the ssms file for that particular entry. Patient 12 in particular was very difficult to deal with because of this limitation, as the genome doubling event resulted in CNVs being associated with almost every SNV entry. This process was repeated for all 17 patients, giving the required input for PhyloWGS to run for the entire cohort.

A general comment on the input format for PhyloWGS is that it is quite challenging to generate and has functional limitations as listed above.

All 17 patients were then run through the PhyloWGS software, which consists of a series of Python scripts, with all default parameters used and the number of chains set to 14. This chain number was selected due to the structure of the High Performance Computer Cluster at the University of Leicester, which was used for all computational heavy tasks in this project. Computational nodes in the cluster each have 28 cores, and so a PhyloWGS run for a single patient could be run using half an entire node. The default number of samples generated by the probability distribution of the PhyloWGS statistical model is 2500, meaning that each patient had a total of 35000 trees generated by PhyloWGS (14 chains multiplied by 2500 samples).

Once the runs were complete the write_results.py script was run with default parameters apart from the –include-ssm-names flag, which was activated so that ssm names would be included in the different nodes of the trees produced. This generated a json file containing all the sampled trees, a tree summary file, a mutation summary file and a list of each ssm assigned to each subclone in the trees.

2.4.6) PhyloWGS Output Workaround

PhyloWGS produces output in the form of a json file that is to be used with the index_data.py script provided by PhyloWGS to host the data on a local server and allows for visualisation of all sampled trees for each patient. Unfortunately, this is another limitation of the PhyloWGS software, as this is the only way in which the results can be viewed using the methods provided by the package. The visualisation page struggled to perform with the memory limitations of a localised machine, due to the huge amount of data it had to display at any one time. This resulted in the page essentially loading indefinitely with all menu or figure actions to irresponsible to function, whilst also crippling the local machine due to a surplus of RAM usage. In order to escape this limitation, it was determined that the json file itself could be targeted and the data be acquired by simply pulling it directly from the file using the Linux command line.

The tree summary file was the first to be targeted in an attempt to extract the most likely tree for each patient, however, upon sampling some of the patient json files, it was clear that all patients had multiple trees of highest likelihood. This was interesting, as in these trees, the separation of ssms in subgroups was always the same, but the tree topology could be vastly different, with the software unable to determine how certain ssms were related to other in the evolutionary process of the cancer. Every patient had a minimum of 5 different tree topologies awarded highest likelihood, with Patient 64 having the highest number of tree topologies at 18. The mean average number of highest likelihood tree topologies was 8.2. Whilst the lack of useful visualisation of highest likelihood trees for each patient was unfortunate, that each patient had the same subgroups in these trees was a promising finding, indicating a good

degree of robustness in the PhyloWGS method. The content of these subgroups were inspected by extracting the ssm list from the json file of each patient.

As the truncal region of the tree is where clonal ssms should be located, this the region that was focused on for the purposes of analysis. This was done by using the following command to extract the ssms that were present in the truncal region of every sampled tree for each patient:

```
jq '.["mut_assignments"]["1"]["ssms"]' > "${file/%json/txt}";
```

This generated a text file of 35000 lines, with each line being the truncal region of one of the sampled trees produced by PhyloWGS. The logic behind this was that each ssm (which was then associated back to its original SNV) could be counted to see how many times it was classified as truncal by PhyloWGS, giving a primitive source of likelihood for that SNV being truncal in the patient. For example, if an ssm appeared in all 35000 lines of the file, then it would have a likelihood of 1 for being truncal in the patient. The SNV the ssm was matched with could then be viewed to see its position in the exome as well as the genotype viewed. This method was performed for every patient and it was decided that any SNV which appeared in more than 50% of sample trees, i.e. had a likelihood of at least 0.5, would be classified as clonal for that patient. The basic logic for this was that a likelihood of greater than 0.5 meant that the variant was appearing in the trunk of the tree more often than not, though this is not an ideal measure of clonality. This was certainly a less than perfect measure in determining which SNVs were classified as clonal by PhyloWGS, as it is possible that certain biases in the MCMC could result in a variant appearing

more often than would be expected in the truncal region of a given patients sample trees. One method of correcting for this would be to run the PhyloWGS software several times, as the output would theoretically be slightly different each time due to the random aspect of the probability distribution and MCMC, with the results then being compared to establish consensus in the patients. However, due to computational and time limitations, this unfortunately could not be done. It is for this reason that there were no trees generated for this pipeline as the computational restraints were just too great, with the work described above being all that could be done.

2.5) Revolver Pipeline

2.5.1) Revolver as a Tool to Measure Driver Event Trajectories

The third and final pipeline used differs from the previous two in that: it uses data from an entire cohort in a single run; and, whereas the concept of the previous pipelines was established first before suitable software was then chosen, this third pipeline was decided upon due to the existence of the software package it uses. In this chapter, the Revolver (Repeated EVOLution in cancER) (Caravagna, Giarratano et al. 2018) software suite will be used to determine driver trajectories within the reported driver mutations of the 17 MPM patient cohort.

Revolver was created to address the issue of trunk order in cancer phylogenetics, which refers to the ability of various phylogenetics to determine truncal events accurately, but not be able to then arrange those mutation events in the order in which they occurred. Naturally, this is a highly complementary concept in terms of the project aims, as the pipelines have been able to identify truncal events in each patient, with the only methods of calling drivers being looking for recurrence between patients and previous associations with the gene contained in the truncal region, as well as assessing the regions for their biological plausibility to explain cancer progression (i.e. whether it makes biological sense that a particular gene could be a driver). In combination with a method like Revolver, strong evidence could be provided in calling particular genes as drivers if they were seen to be occurring later in the trunk, just prior to the progression of MPM into the highly aggressive metastatic state.

Revolver works by taking a full cohort of pre-computed trees, generated using another phylogenetic software and then assessing them based on maximum likelihood estimation to determine recurrent instances of driver trajectory. Briefly, the Revolver algorithm observes patterns of evolutionary trajectory in driver mutations that are recurrent in multiple patients. i.e. it records how often a given driver event is placed subclonally when it isn't part of the truncal branch. It can then use these observations across the entire cohort to determine the most likely trajectories for the drivers when they do appear truncally in a different patient. It gives higher likelihoods for trajectories which appear at a greater rate than others, and then use this information to form an order of trajectories in the truncal branch for any given drivers. It does this by computing artificial trees using the pre-computed phylogenetic input as a basis, and then calculates the likelihood of different groups forming between drivers. It does this a number of times until a likelihood threshold has been met, determined by the total number of patients and drivers.

2.5.2) Generation of Revolver Input

As an R package, Revolver takes input in the form of a data frame consisting of 7 columns, with each row representing a single genetic alteration event in the patient cohort. Revolver expects the input data to have been produced using a phylogenetic pipeline (though the option also exists for Revolver to do this step itself, however, as output from Sequenza/TuMult was readily available, this option was not explored), and then formatted into a data frame with the following columns; 'misc' a custom field that wasn't used in revolver analysis, this value was left blank for all entries; 'patientID' a string containing the patient number; 'variantID', a string containing the cytoband and copy number alteration state of the given event, -22p for example. The variantID field actually required some additional work to generate, as many alterations in the subclonal branches of the TuMult trees had not been compared from patient to patient. This was because the truncal branch of the tree was the focus of the Sequenza/TuMult pipeline, and whilst the events on each tree were catalogued to compare to other events within the same tree, comparison of non-truncal branch events was never carried out between patients. This involved checking the segments file for each event and manually comparing them to events present on the other patient trees based on whether the cytoband labels were similar. The specific region the events covered could then be compared to determine whether events from separate trees could be counted as recurrent. For the majority of cases, the choice was clear but there were some confounding factors in some comparisons, for instance, where two events overlapped but had substantial non-overlapping regions. In cases where a clear distinction could not be made, the events were listed separately.

The 4th column required in the Revolver input data frame was a clusterID, which was a integer value to be set to determine how each event was grouped on each patient tree. The clusterID was linked to the patientID, meaning each patient could use the same values without convoluting the input. For each patient, the truncal branch was listed as with a cluster value of '1', with subsequent shared and node branches being given cluster values between 2 and 7. The clusterID is how the Revolver algorithm determines how events were placed on the phylogenetic trees. The 5th column in the data frame was 'is.driver' and required a TRUE or FALSE value of whether the listed event was a driver or not. In order to be considered a driver, an event had to be have appeared in the truncal region of at least 2 patients, the full list of drivers can be seen in Chapter 3.7.3 in Figure 3.26. The 6th column for the data frame was 'is.clonal' and required TRUE or FALSE value on whether the event was in the truncal region of the tree. This was set to TRUE for every event where the clusterID value was set to 1, as they were already classified as the truncal branch in each patient.

The final column is the 'CCF' column and expects either a parsable list of cancer cell fraction (CCF) values for each sample in the patient that the event line is in, or a binary value of either 1 or 0, representing whether the event was present in each sample. CCF values represent what proportion of the cancer cells in a given sample harbour a specific variant or copy number event. As CCF values had not been generated, the binary values were used, which unfortunately would result in an overall lower resolution from Revolver and limit the plots which could be produced. However, the main Revolver drivers plot would still be available and be able to portray the main findings of the analysis.

2.5.3) R Analysis of Revolver Cohort

After construction of the data frame, it was then passed into Revolver via the revolver_cohort command in R. The parameters were set so that the CCF_Parser was used, in order to parse the binary values in the 'CCF' column of the data frame, the ONLY.DRIVER flag was set to FALSE and the minimum number of clusters was set to 0. This command produces a revolver cohort object, which could then be used as input for the following computation step.

The revolver cohort object was then used in the compute_mutation_trees command, as the 'CCF' values being used were binary. Default parameters were used for this command. This modified the revolver cohort object so that it now also contained the computed mutation trees.

The revolver cohort object was used as input for the revolver_fit command, which is the implementation of the main fitting function, where Revolver performs its 2-step algorithm and calculates the evolutionary distance between drivers, and determine trajectories based on the computed mutation trees. The default parameters for this command were used, except that the max iterations were set to 10 as the cohort was only made up of 17 patients and so would not require more than that to establish accurate trajectories between drivers.

Once the revolver_fit command was completed, it produced a revolver fit object which could be use to generate a drivers graph, which allows for the visualisation of driver trajectories in a patient cohort. Results can be seen in Section 5.2.

3) Copy Number Based Pipeline, Results and Discussion

3.1) Introduction to Copy Number Pipeline, Methods and Materials

This chapter contains the results and discussion for the copy number based pipeline described in Section 2.3. It also contains a patient cohort results section which displays the characteristics of the 25 MEDUSA patients from whom the DNA was extracted. It opens with that section, followed by: the Sequenza results, including examples of the plots generated for each patient and sample; the ABSOLUTE results, also including examples of plots generated for each patient and sample; a section discussing comparison between the Sequenza and ABSOLUTE results and justification in how this caused patients to be removed from the cohort; a section discussing comparison between the Sequenza and ASCAT results; a section displaying example TuMult results for an individual patient; TuMult summary results, including tables displaying the results across the cohort. This is then followed by a discussion of these results which includes certain interesting cases which occurred, the meaning of key findings in the chapter, and the limitations of the pipeline.

It has already been established several times in the previous chapters that copy number variation is the major feature in the development of MPM, with the methods employed by this pipeline focused solely on deciphering key CNV events that occur early in the process. The aims of this chapter are: to identify important truncal regions where recurrent copy number events occur across the

patient cohort, and suggest possible MPM associated genes which they contain; critically evaluate the methods employed, noting why they are suitable for this type of analysis as well as limitations in the software/methods; to discuss why the anatomical spread of MPM may not be possible to infer using phylogenetic methods.

3.2) Patient Cohort Characteristics

This section describes the characteristics of the 25 MEDUSA patient cohort which was analysed during the course of the project. The findings as reported were obtained from a summarised clinical report and are displayed in Table 3.1.

MEDUSA ID	Progression free survival (days)	overall survival (days)	status	asbestos exposure (exposed=1)	histology (0=epithelioid,1 biphasic, 2 sarcomatoid)	gender (female=1)	age >median *	laterality
MED01	639.0	875.0	1	1	0	0	0	Right
MED03	112.0	112.0	1	1	0	0	1	Left
MED06	143.0	174.0	1	1	0	0	1	Left
MED07	98.0	242.0	1	1	0	0	0	Right
MED08	117.0	404.0	1	1	0	0	0	Right
MED09	82.0	124.0	1	1	1	1	1	Right
MED12	433.0	622.0	1	1	0	0	0	Left
MED16	145.0	214.0	1	1	0	0	1	Right
MED18	331.0	455.0	1	1	1	0	0	Right
MED20	87.0	186.0	1	1	0	0	0	Left
MED23	224.0	578.0	1	0	0	0	0	Right
MED24	82.0	389.0	1	1	1	0	1	Right
MED27	108.0	186.0	1	1	0	0	0	Right
MED32	219.0	461.0	1	1	0	0	0	Right
MED33	218.0	218.0	1	1	0	0	1	Left
MED34	616.0	616.0	1	1	1	0	0	Right
MED35	95.0	95.0	1	1	0	0	1	Right
MED37	196.0	196.0	1	1	0	0	1	Right
MED62	102.0	203.0	1	1	0	0	1	Right
MED64	127.0	772.0	1	1	0	0	0	Right
MED75	128.0	385.0	1	1	0	0	1	Right
MED78	104.0	191.0	1	1	0	0	0	Right
MED84	32.0	206.0	1	0	0	0	0	Right
MED85	90.0	341.0	1	0	0	0	1	Right
MED91	96.0	109.0	1	1	0	0	1	Right

Table 3.1: Table displaying characteristics for the 25 MEDUSA cohort, from left to right the columns are as follows: MEDUSA ID displays the patient number for each individual; Progression free survival (days) displays the number of days the patient had no further cancer progression post surgery; overal survival (days) displays the number of days the patient survived post surgery; status refers to the survival of the patient with 0 = survived (up to the date), 1 = deceased; asbestos exposure displays whether the patients were able to report exposure to asbestos previously in life 0 = no exposure, 1 =exposure; histology refers to the type of MPM each patient was diagnosed with, 0 = epithelioid, 1 = biphasic, 2 =sarcomatoid; gender refers to whether the patient was male = 0 or female = 1; age > median* displays whether the patient was above or below the median age of the cohort, 0 = younger, 1 = older; laterality refers to which lung the cancer was present on. *Individual ages were not displayed for each patient, however the median age was 70 years old with an IQR of 66-74 years old (min and max age range was 53-78 years).

Each patient was longitudinally tracked following their surgery up until disease progression was detected, with a large range of values in progression free survival and overall survival, though unfortunately all patients were deceased prior to the conclusion of the project. All but 3 of the patients had reported some form of asbestos exposure previously in life, though it is a difficult characteristic

to identify in the case of non-exposed individuals, as it would be impossible to evaluate every aspect of their past life for possible exposure events. Most of the patients have epithelioid type mesothelioma, with 4 patients identified as biphasic and none with sarcomatoid type. All but 1 of the patients were male, an expected result giving the higher incidence of MPM in men. Individual ages for each patient were not provided in the documentation, but median age (70 years) was included and values given as to whether patients were above or below this age. Only 5 patients had the cancer present on their left lung, with the other 20 having it on their right.

3.3) Copy Number Estimation Results

3.3.1) Sequenza Individual Sample Results

All of the 106 samples were successfully run through Sequenza, producing a range of results as discussed below. Data from Patient 1 will be used as examples throughout this chapter and the rest of the report to demonstrate the output of the various software, as fully displaying all results would not be possible in the main body of text. Patient 1 was selected as it was one of the 5 sample patients and reported high cellularity values, whilst also showcasing some interesting results which can be better discussed alongside figures.

Sequenza produces the following figures for each sample; depth ratio (before and after the normalisation step) per chromosome (Figure 3.1), mutations, B-allele frequency and the depth ratio per chromosome (Figure 3.2), proportion of the exome with different copy number values (Figure 3.3), genome wide view of total copy number, allele-specific copy number and a depth ratio and allele frequency overview (Figure 3.4), the likelihood of each cellularity/ploidy solution (Figure 3.5), a visualisation of the highest likelihood model fit (Figure 3.6), and visualisations of the alternate model fits (Figure 3.7).

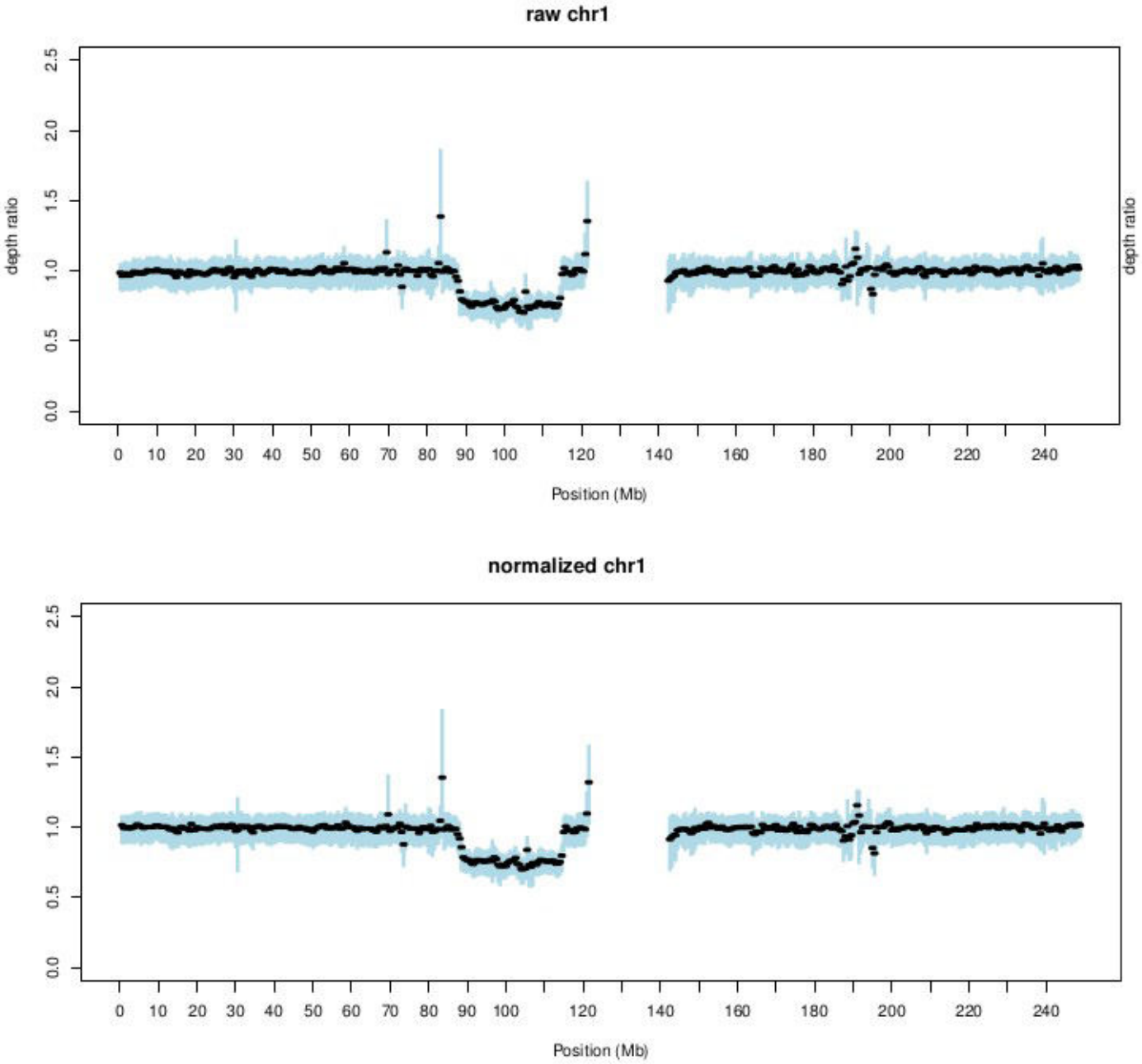


Figure 3.1 : Figure produced by Sequenza from Patient 1 Sample 1, and displays only the results from chromosome 1 of the sample. The two graphs display the depth ratio between the normal and tumour samples, raw on the top and normalised on the bottom. The normalised results on the bottom are the results of interest in terms of analysis, though comparing them to the raw results displays the effects of normalisation. The empty space visible in the middle of each graph represents the centromere of the chromosome.

In Figure 3.1, the value of 1.0 in the normal and tumour depth graphs is calculated from the average read depth across the entire chromosome. Only segments with a read depth of at least 20 are included. The depth ratio displays how the average read depth differs between the normal and tumour samples, a drop in read depth from normal to tumour can be seen between around positions 86mb to 115mb indicating there may be deletion events in the region. The empty “gap” in the 120mb to 140mb region is where the centromere of the chromosome is located, and so is not included in sequencing data.

The mutant allele frequency shown in Figure 3.2 gives a good indication of point mutations detected by Sequenza, although this output is not actually used in subsequent steps of the analysis. The B-allele frequency graph displays the frequency of the minor allele at every position on the chromosome. Segments of lower B-allele frequency indicate a heterozygous deletion in the same region. The depth ratio graph is essentially the same as in Figure 3.1, although also includes the estimated copy number value.

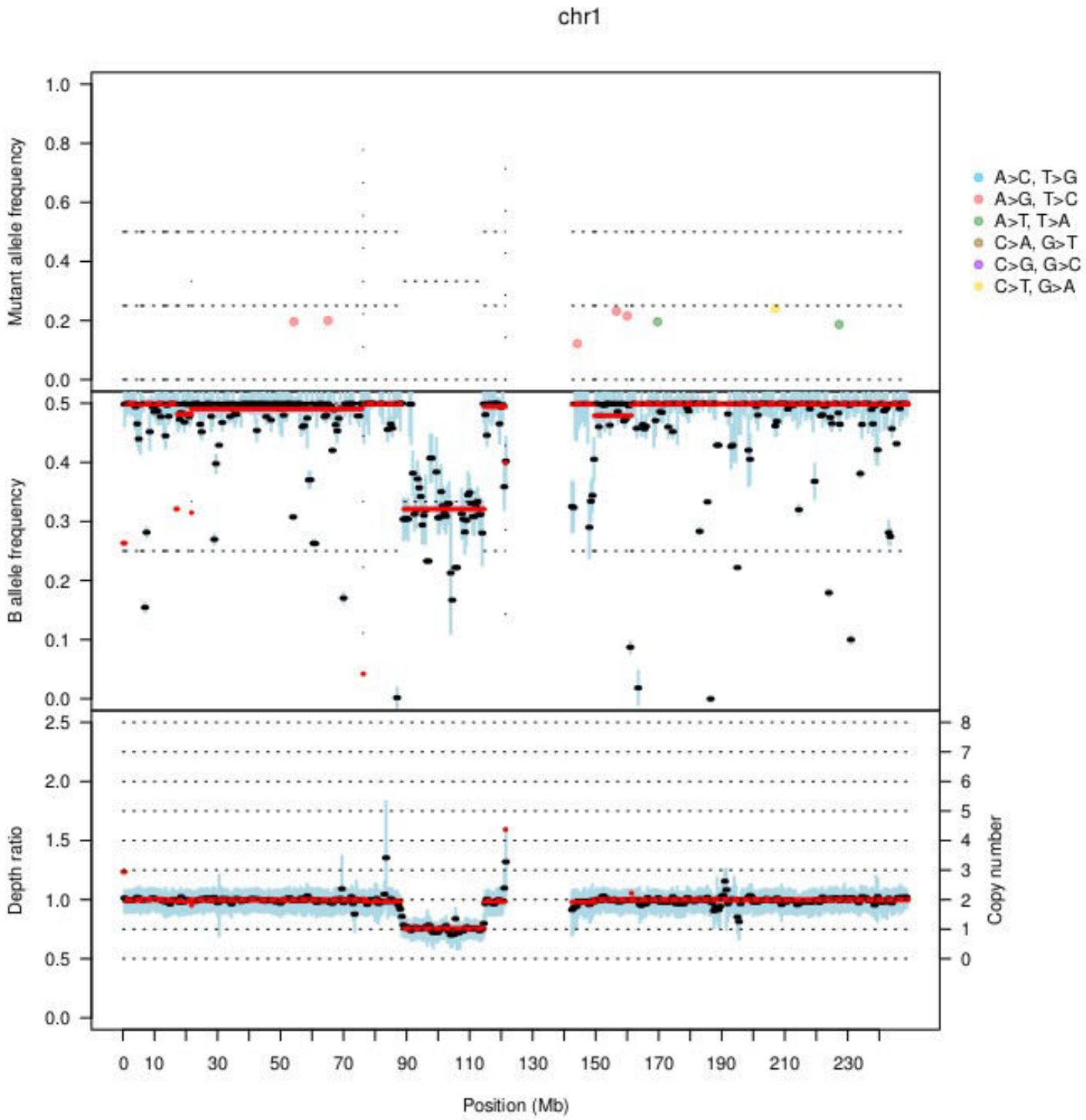


Figure 3.2: Figure produced by Sequenza from Patient 1 Sample 1, and displays only the results from chromosome 1 of the sample. The top graph displays the location of mutant alleles, with colour indicating the change in genotype. The middle graph shows the B-allele frequency across the chromosome, the drop in B-allele frequency between 86mb and 115mb indicates a possible heterozygous deletion in that region. The bottom graph displays depth ratio between the normal and tumour samples, with the addition of estimated copy number value along the right-hand side of the graph. The gap in the 120mb to 140mb region indicates the centromere in each graph. The B allele frequency in the middle graph only has a range of up to 0.5 due to the defining feature of B alleles being that they can be no greater than 0.5 in frequency as they represent the minor allele.

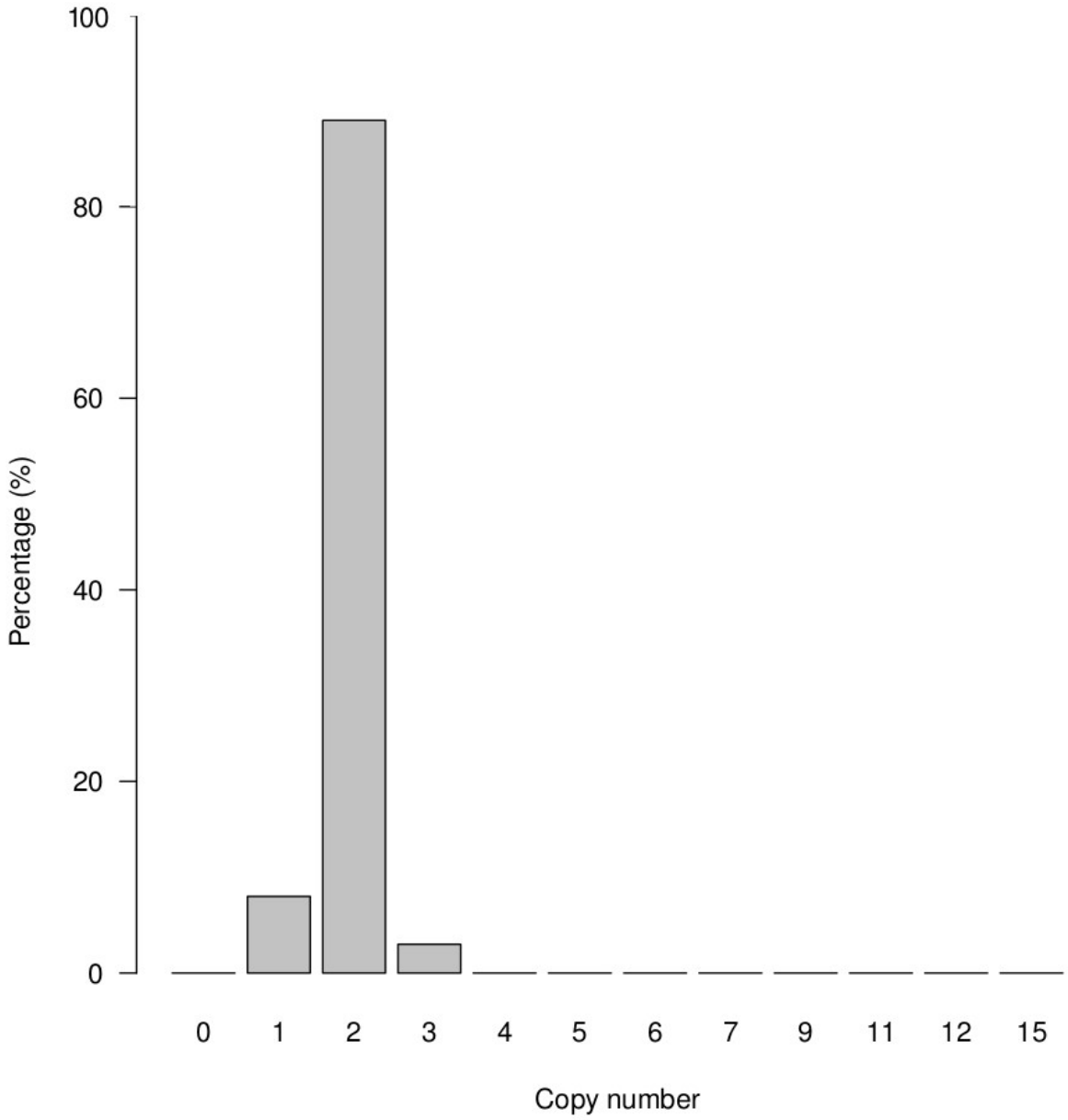


Figure 3.3: Figure produced by Sequenza from Patient 1 Sample 1. A bar plot of the proportion of the exome with different copy number values.

Figure 3.3 is a bar plot indicating that approximately 90% of the exome has a copy number of 2, 8% has a copy number of 1 and 2% has a copy number of 3. With only approximately 10% of the exome effected by copy number alteration events, Patient 1 has a high level of stability compared with other patients. It should also be noted that subsequent alteration events across the same region may cause multiple changes in the copy number value, and so 10% isn't a conclusive value for the amount of change which has occurred.

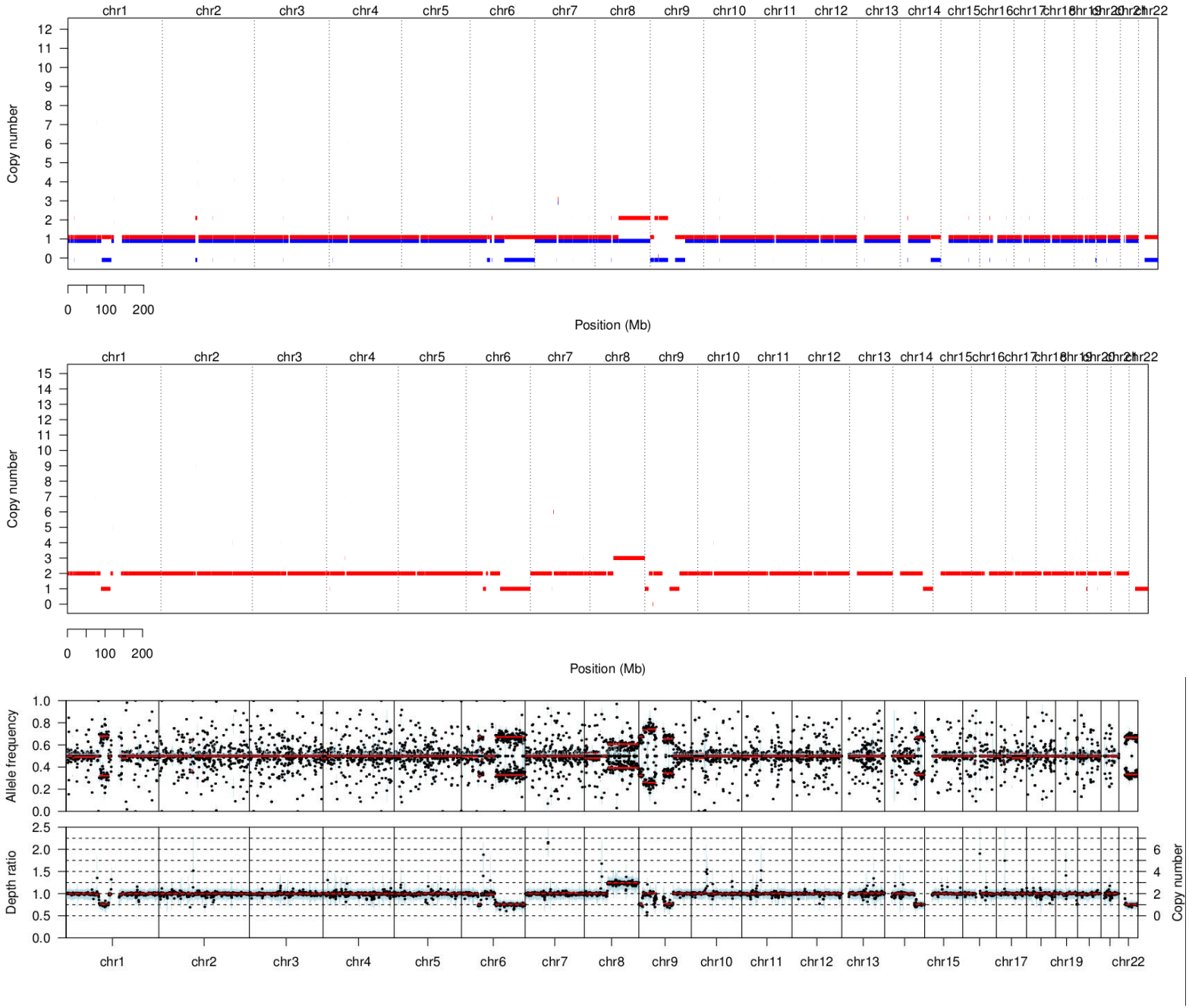


Figure 3.4: Figure produced by Sequenza from Patient 1 Sample 1. The top graph displays allele-specific copy number values across the entire exome, with red signifying the major allele and blue signifying the minor (or B) allele. The middle graph is the absolute copy number values across the exome. The bottom graph displays both the specific allelic frequency and the depth ratio across the entire exome.

Figure 3.4 (seen rotated) allows for the viewing of the copy number values across all chromosomes, as well as the allele frequency values for these same regions. Large scale deletions can clearly be seen in chromosomes; 1, 6, 9, 14 and 22 whilst a large scale gain can be seen in chromosome 8. Closer inspection of chromosome 9 will also display the only homozygous deletion event in the exome. No chromosomes reported a copy number above 3.

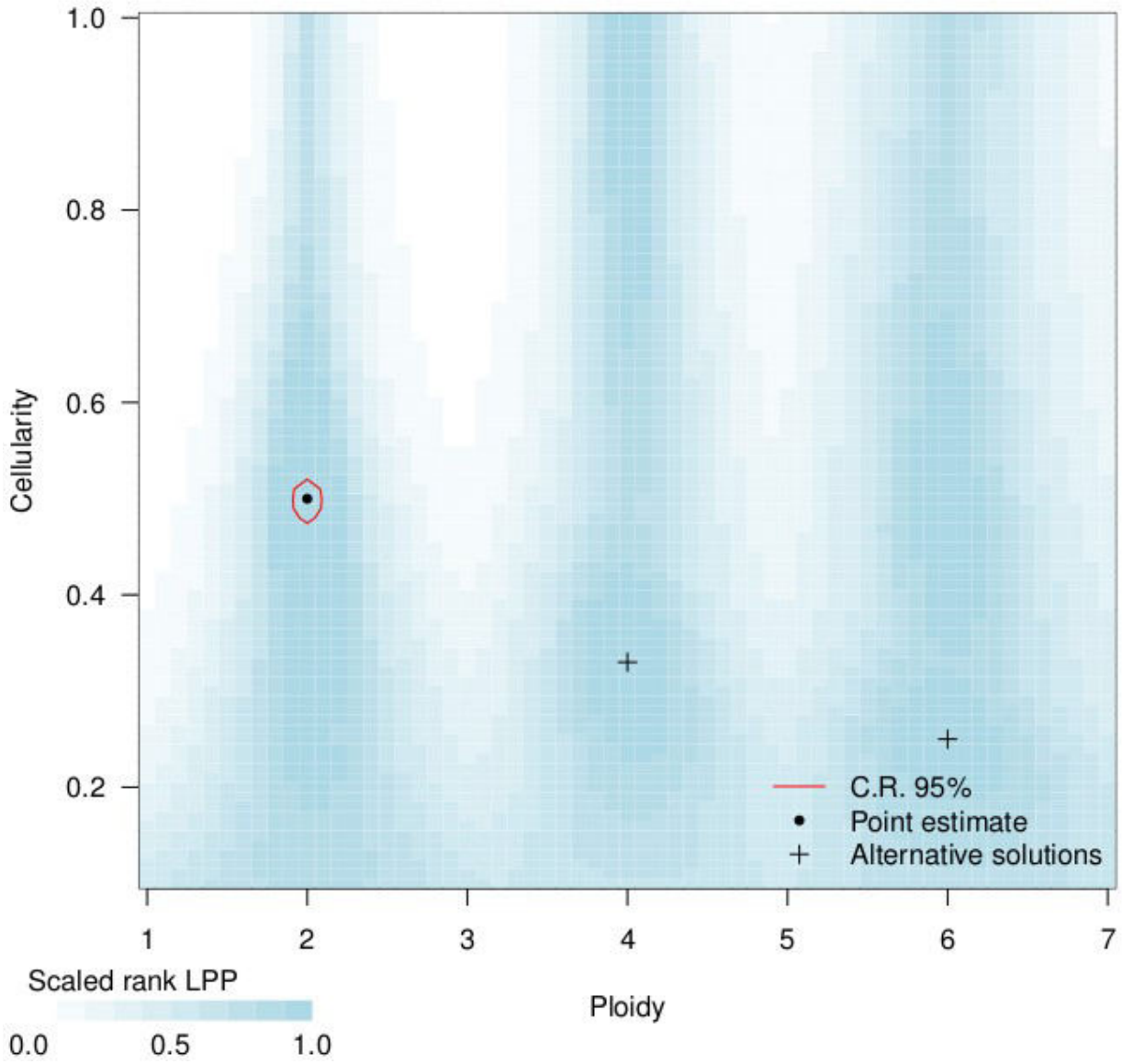


Figure 3.5: Figure produced by Sequenza from Patient 1 Sample 1. Model parameter values of cellularity and ploidy with likelihood values for each combination. Darker blue indicates higher log posterior probability (LPP), with the red line incircling the smallest number of points with a LPP >0.95. The crosses are combinations of ploidy and cellularity values of alternative solutions proposed by Sequenza.

Patient 1 Sample 1 provides a good example of the Sequenza model parameters, with an estimated cellularity of 0.5 and a ploidy of 2 used to call copy number values due to the highest confidence value displayed in Figure 3.5. It can also be seen the higher values of cellularity still maintain high LPP values when matched with ploidy samples which are multiples of 2. This is due to the relatively low level of copy number alterations present in Patient 1. Higher cellularity values matched with odd number ploidies give much lower LPP

values eventually reaching 0, although lower cellularity values give higher probabilities in these regions due to difficulty of generating good estimates at extremely low cellularity. The alternate models also require lower cellularity levels in order to match higher ploidy values in the model parameters.

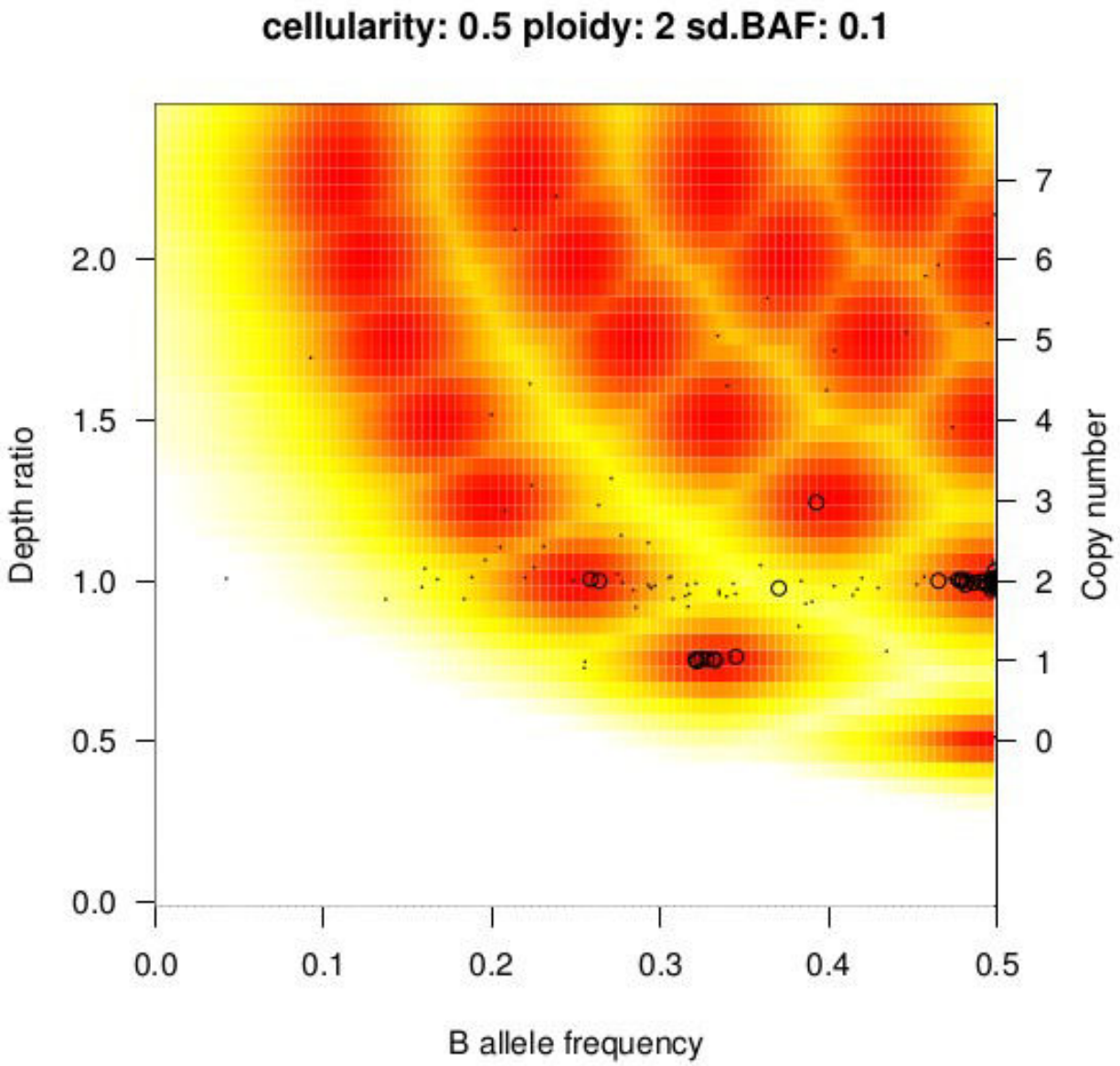


Figure 3.6: Figure produced by Sequenza from Patient 1 Sample 1. B allele frequency values and observed depth ratio for each segment, defined by black circles and dots. The density of the colour is representative of the joint LPP values.

The density of the colour in Figure 3.6 is stronger for LPP values that are higher. These values are calculated for the ploidy and cellularity estimates which are displayed in Figure 3.5. This value is calculated for hypothetical 10Mb segments due the actual segment length causing a change in the LPP value. Segments which can be seen in regions of weaker colour density can be indicative of subclonal events, though can also be attributed to errors in the model parameters. A small event can be seen in the stronger colour region at a copy number value of 0, which corresponds to the homozygous deletion event

in chromosome 9 displayed in Figure 3.4.

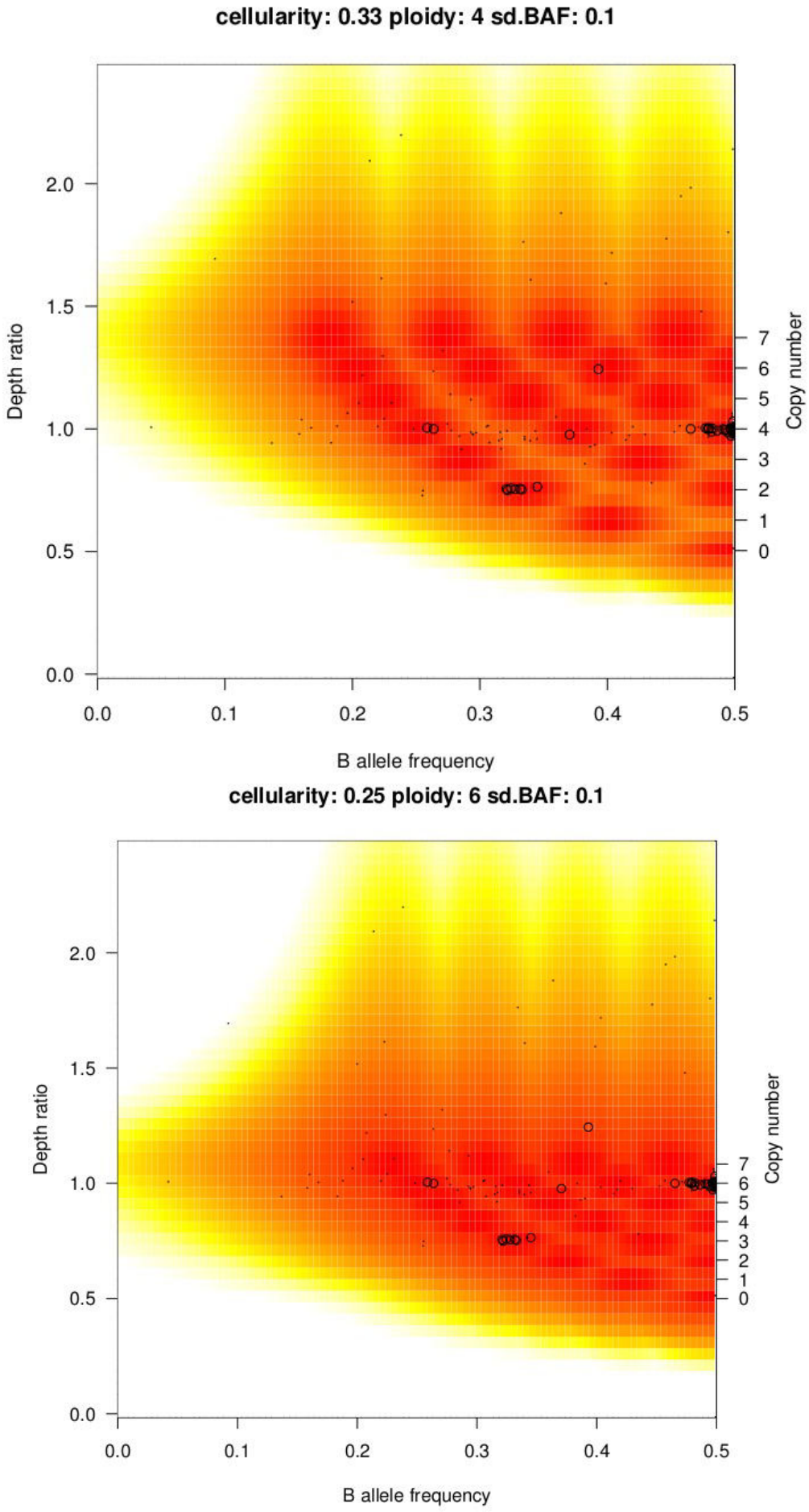


Figure 3.7: Figure produced by Sequenza from Patient 1 Sample1. B allele frequency values and observed depth ratio for each segment in the alternatie models, defined by black circles and dots. The density of the colour is representative of the joint LPP values.

The alternate models provided by Sequenza as seen in Figure 3.7, can be compared to copy number estimates generated by other software or methods, where differences in ploidy values can explain unexpected copy number values. All but two samples gave a primary model with a ploidy of between 1.8 and 2, both present in Patient 12. This implies that there may be a lack of genome doubling in MPM, which is a phenomenon known to occur in several cancers, as discussed in chapter 1. Even in Patient 12, only 2 of the 5 samples displayed signs of genome doubling, indicating that the possible doubling event occurred after the cancer had already developed and spread. With only 2 of 106 total samples displaying any evidence of genome doubling, it is possible that this type of event is not important for the development of MPM, or at least is not significant in the majority of patients.

Patient 12 Samples 1 and 4 both reported higher ploidy numbers, with the most likely model having a ploidy parameter of 3.4 in both cases. This is an unusual ploidy, though as it's representative of the entire exome, the number can be explained due to different total ploidy on different chromosomes. Figure 3.8 displays both the whole exome copy number profiles of both these samples.

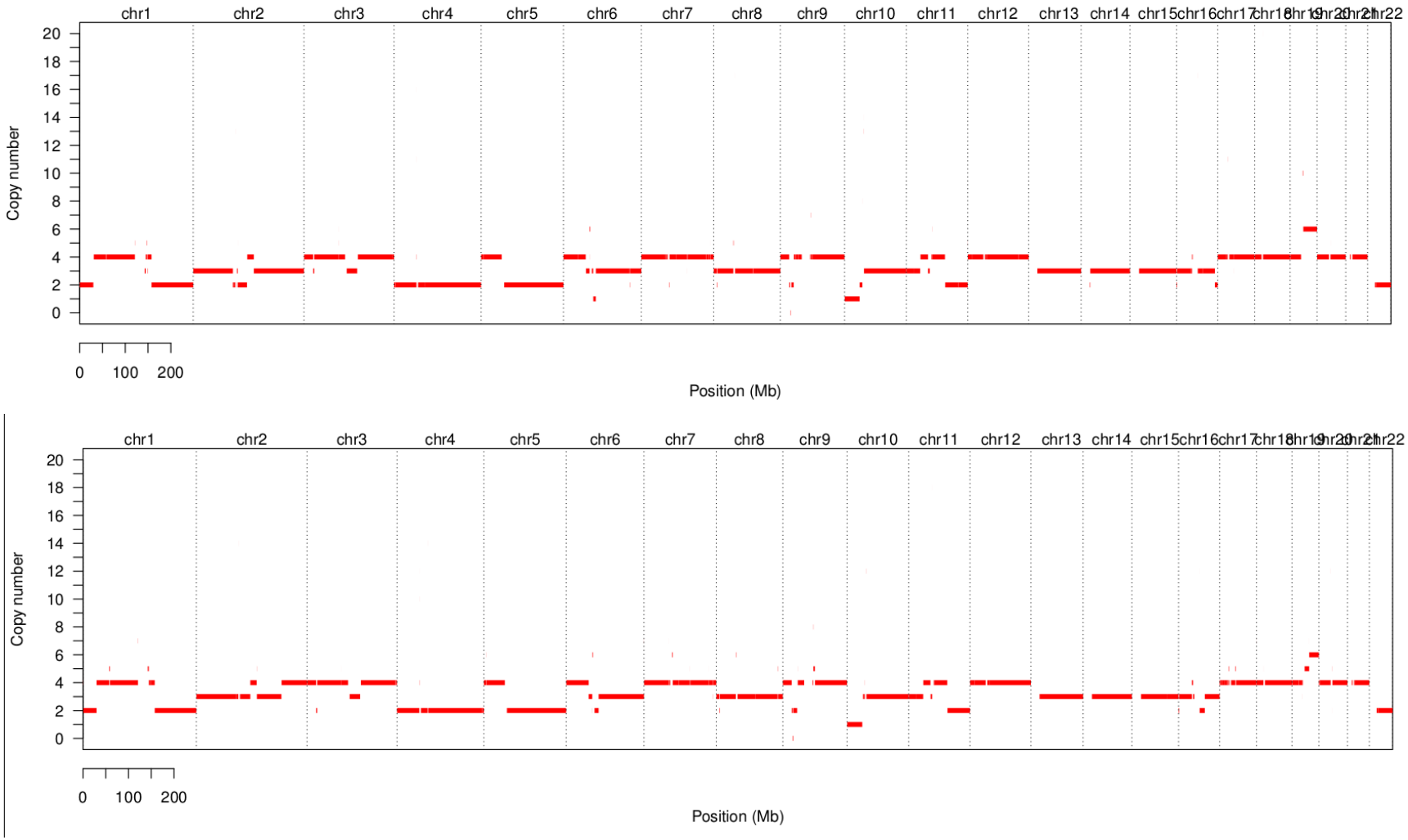


Figure 3.8: Produced by Sequenza from Patient 12 Sample 1 (Above) and Sample 4 (Below). The absolute copy number across the entire exomes of samples 1 and 4, both of which display ploidy values of 3.4 in the most likely model.

Large scale copy number alterations can be seen in both samples, with a range of values from between 1 and 6. Entire chromosomes can be seen to be reported with different copy numbers, though what is noticeable is that both samples share a similar pattern of change, with only minor differences between the two. An initial look at Figure 3.8 might imply that the data quality of these samples was lower and thus made it difficult for depth ratio to be accurately calculated. However, the similar copy number states between the two samples imply that these events may have happened prior to samples 1 and 4 establishing themselves, but after divergence had already occurred for the other three samples. It was decided the samples should remain in the dataset despite the ploidy value for both being an outlier. This is because it could be interesting to test the impact of this situation on the phylogenetic software to be used later in the pipeline. The implication if the data for these two samples is accurate, is that there may be rare cases where MPM results in mass genome instability, resulting in fluctuations in copy number.

As well as generating figures, Sequenza also produces results in tabulated text format. The mutations file contains the positional information for mutated alleles and the specific genotypic change which has occurred. The mutations file was never parsed for use in further pipelines but was necessary to generate summary results from Sequenza. Another tabulated text file Sequenza generates is the segments file. This file contains all reported copy number segments, giving both start and end positions, and copy number, both total and for each allele. This is the file which is used for generating the input of the next step of the pipeline, and is the only output directly necessary for this. The file must at least contain as many segments as there were chromosomes, as a

chromosome with no reported copy number alterations would be recorded as a single uninterrupted segment incorporating the entire length of the chromosome in question.

3.3.2) ABSOLUTE Results

For the purposes of displaying an example of the ABSOLUTE results per sample, both patient 1 sample 1 and patient 12 sample 1 will be used. Patient 1 as a standard and patient 12 to show how ABSOLUTE also predicted that the ploidy was increased for this sample, as it was for Sequenza, and this gives a good basis for comparison.

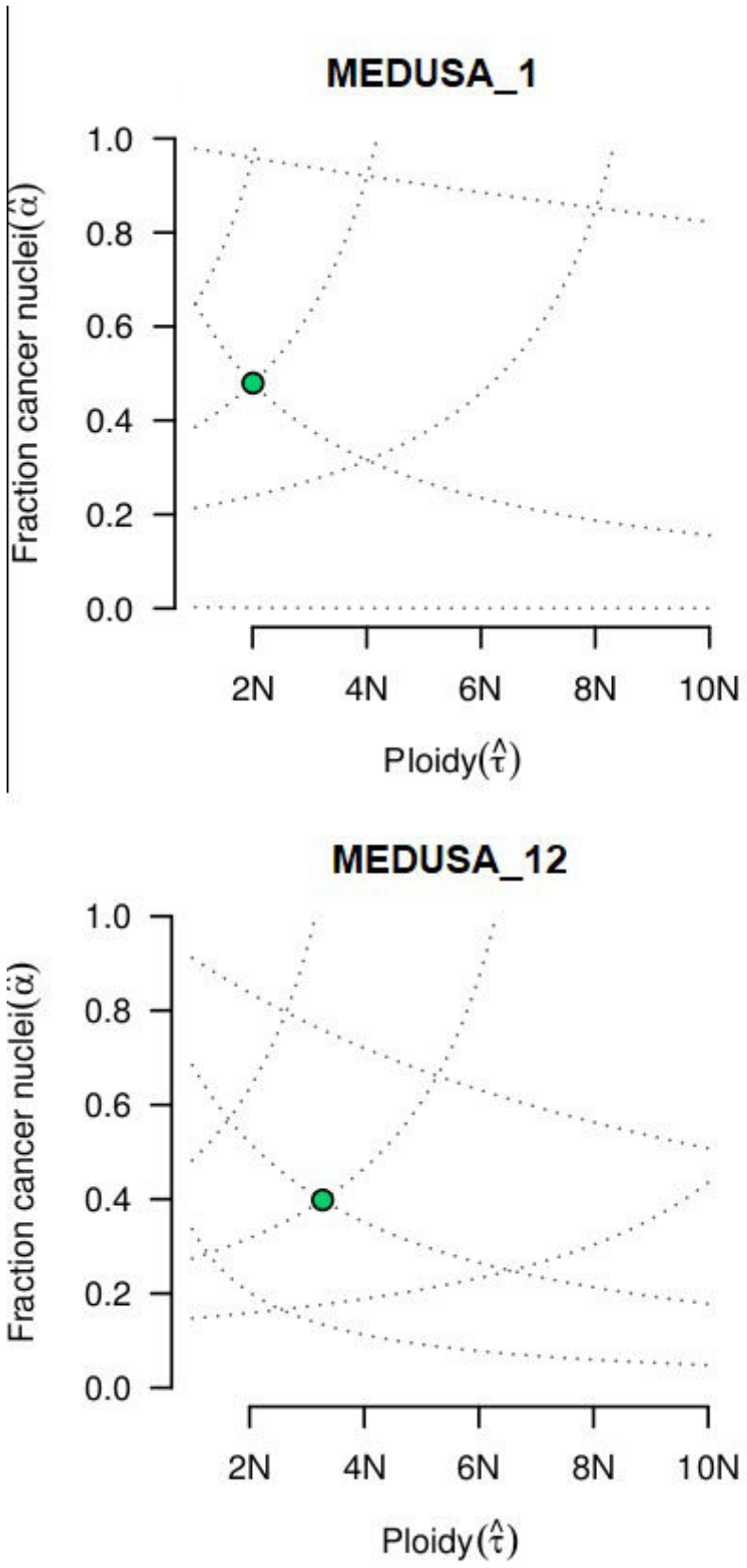


Figure 3.9: Figure produced by ABSOLUTE for patient 1 sample 1 (top) and patient 12 sample 1 (bottom). The green dot represents the position of the most likely solution in terms of ploidy and cellularity (Fraction cancer nuclei) values.

Figure 3.9 is the solutions plot generated by ABSOLUTE for patient 1 sample 1 and for patient 12 sample 1, with a range of ploidy estimates along the x-axis and the full range of cellularity plotted on the y-axis (cellularity is labelled as fraction cancer nuclei). The green dot represents the solution which had the greatest likelihood and in the case of patient 1 sample 1 can be seen at the intersection of ploidy = 2.01 and cellularity = 0.48 (these precise values were taken from the Rdata output also produced by ABSOLUTE). Alternative solutions can be seen where other intersections occur, at higher ploidy and cellularity values such as ploidy = 4.00 and cellularity = 0.32. Patient 1 sample 1 displays a very stable result and is almost identical to the estimates provided by Sequenza (see Figure 3.5).

The solutions plot for Patient 12 sample 1 produced by ABSOLUTE, was one of the two samples reported by Sequenza as having a higher than average ploidy of around 3.4 (the other being sample 4). It can be seen in the figure that ABSOLUTE predicted a similar best solution as Sequenza, with the ploidy being reported as 3.27 (pulled from Rdata output), compared to the 3.4 estimated by Sequenza. A similar pattern was observed for Patient 12 sample 4, with Sequenza and ABSOLUTE estimating ploidy values of 3.4 and 3.37 respectively. The alternative solutions for these samples all fell significantly away from a ploidy of 2, with solutions the closest to this having either very high or very low cellularity estimates. This provides evidence that the amplified ploidy seen in these samples is not an artifact caused by low resolution in sequencing, or by mistakes made during Sequenza copy-number calls.

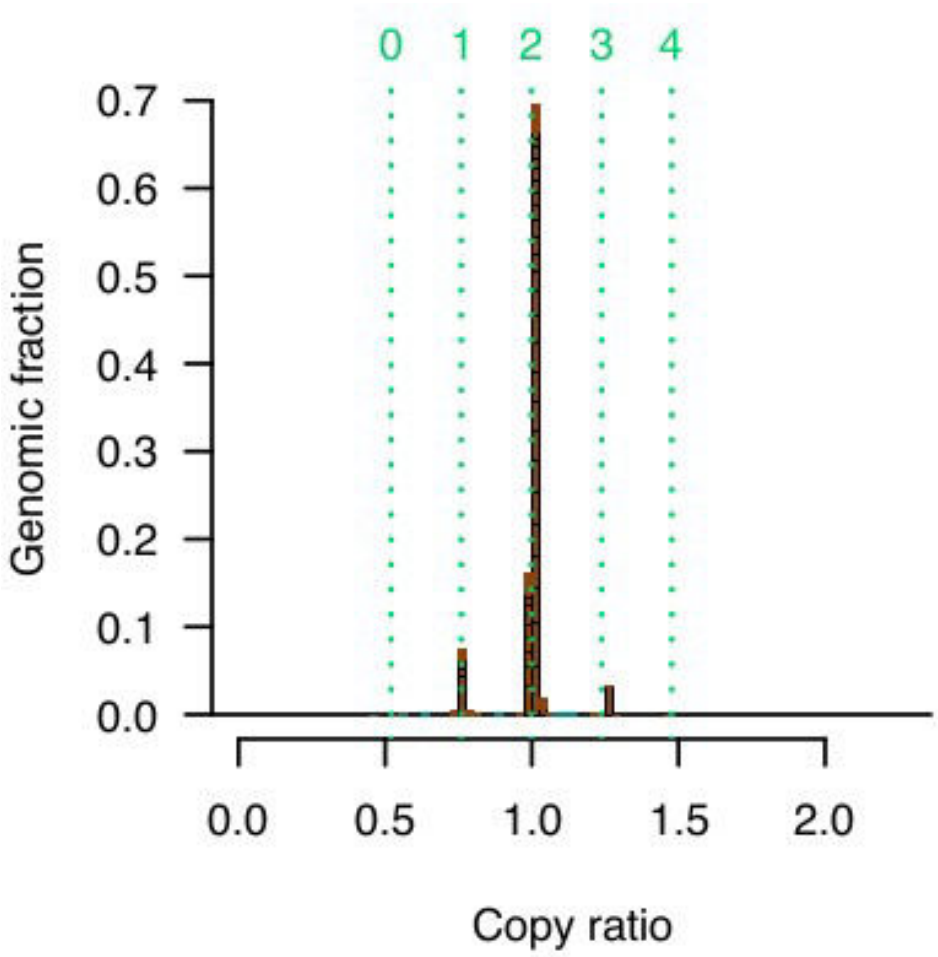


Figure 3.10: Figure showing the genomic fraction for each copy-number value for Patient 1 sample 1, as estimated by ABSOLUTE. Most of the "genomic fraction" (though the data only represents the exome), can be seen to have a copy ratio of 2, with small amounts indicating both the loss and gain of sequence, possibly due to mutation events.

Figure 3.10 is the genomic fraction for each copy-number value (labelled copy ratio) estimated by ABSOLUTE across the segments provided for Patient 1 sample 1. It agrees strongly with the Sequenza estimates, shown in figures 3.3 and 3.4, where most of the "genomic fraction" (though it actually represents exome sequence) is estimated to have a copy-number of 2, with small sections having slightly lower or higher copy-number. These deviations from a copy-number of 2 are representative of loss or gain events in the exome where a mutation has taken place and altered the copy-number of the region. This figure indicates that Patient 1 sample 1 has a larger amount of exome affected by

deletion events than duplication events, though doesn't display whether deletion or duplication events are more common or not, as it could be the case that deletion events simply affect a larger region per event on average.

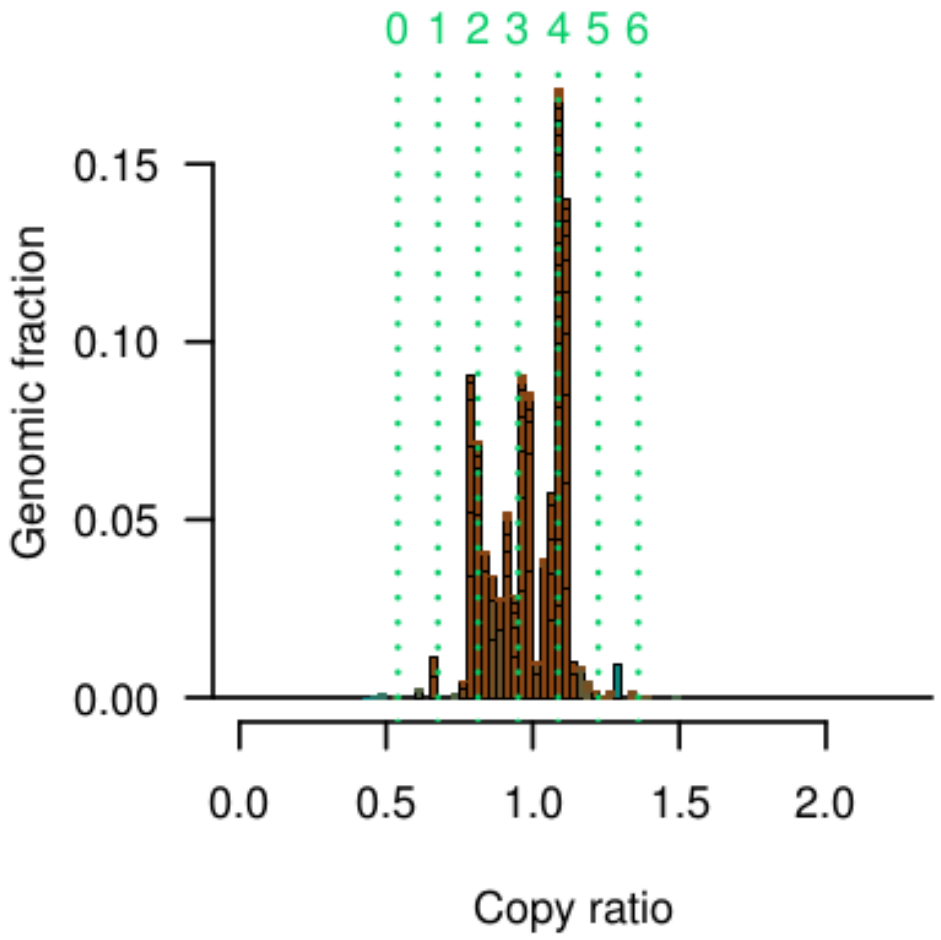


Figure 3.11: Figure showing the genomic fraction for each copy-number value for Patient 12 sample 1, as estimated by ABSOLUTE. Unlike for the Patient 1 sample, Patient 12 sample 1 displays a significant amount of the genomic fraction with a copy-number of 4, with much of the rest accounted for by values of between 2 and 4.

Figure 3.11 shows the coverage of different copy-number values across the exome sequence for Patient 12 sample 1, as estimated by ABSOLUTE. Unlike with Figure 3.10, this figure displays a larger range of copy-number values, with the majority of the exome being estimated as having a value of between 2 and 4, with small amounts both above and below this range. This gives greater insight into why a ploidy value of 3.27 was predicted for this sample by ABSOLUTE, and agrees with the Sequenza estimate for the same sample (3.4), with the actual break down of the fractions of the exomes estimated at different copy-number values providing interesting possible causes.

The ploidy for any given sample is essentially calculated using the average value of all of the copy-number values estimated by either Sequenza or ABSOLUTE, though in the vast majority of cases, a high proportion of the exome is reported to have the same copy-number, with outliers being caused by mutation events causing the value to rise or decrease based on whether a gain or deletion has happened respectively (see later summary figures for a more detailed explanation on this). However, for Patient 12 samples 1 and 4, this is clearly not the case, and although one possible explanation of this amplified ploidy is genome doubling (as mentioned earlier in the chapter) the ABSOLUTE results indicate that this may not be the case.

If genome doubling was responsible for the ploidy being reported by both Sequenza and ABSOLUTE, it would be expected for a large proportion of the genomic fraction to be estimated at a copy-number of 4 (or close to 4) with minor outliers at distinct copy-number values such as 3 or 5 which would indicate mutation events in the samples. However, this is not what Figure 3.12 displays, instead showing a wide spread of estimates across the copy ratio. One possible explanation is simply that this is the result of a massive amount of mutation which has taken place in the genome (chromothripsis) of the cancer cells that comprise samples 1 and 4, which occurred prior to the two samples separating in physical space and thus is now detectable in both samples. Another explanation is that the two samples appear this way due to the result of admixture of multiple cancer cell populations within the tumour, a phenomenon which was discussed in both chapters 1 and 2, and that the differing amounts of reads for any given section of the exome has caused the depth ratio to fluctuate and resulted in amplified copy-number estimates.

It is difficult to differentiate between these possibilities, as either could cause the effect of increased ploidy estimation using depth ratio based methods. However, further discussion of this is present later on in this chapter after other results have also been discussed, providing extra evidence to the arguments.

	ABSOLUTE Values					
P Number	S1	S2	S3	S4	S5	Avg Cellularity
1	0.48	0.66	0.43	0.61	0.59	0.58
3	0.17	####	0.19	0.55	0.2	0.277
6	0.48	####	0.54	0.46	0.3	0.445
7	0.24	####	0.25	0.24	0.2	0.23
8	0.2	####	0.17	0.16	0.23	0.19
9	0.22	####	0.21	0.23	0.25	0.23
12	0.4	0.46	0.29	0.34	0.52	0.402
16	0.19	####	0.2	0.24	0.21	0.21
18	0.6	####	0.29	0.43	0.64	0.49
20	0.19	####	0.25	0.2	0.17	0.2
23	0.32	0.27	0.22	0.39	0.37	0.314
24	0.77	0.36	0.9	0.6	0.81	0.688
27	0.57	0.73	0.51	0.51	0.52	0.568
32	0.19	####	0.28	0.21	0.2	0.22
33	0.36	####	0.39	0.49	0.29	0.383
34	0.63	0.69	0.71	0.47	0.66	0.632
35	0.18	####	0.2	0.22	0.19	0.198
37	0.53	####	0.48	0.56	0.48	0.513
62	0.26	####	0.19	0.17	0.18	0.2
64	0.24	####	0.28	0.24	0.25	0.253
75	0.74	####	0.56	0.78	0.31	0.598
78	0.45	####	0.49	0.37	0.33	0.41
84	0.46	####	0.33	0.69	0.36	0.46
85	0.28	####	0.44	0.5	0.5	0.43
91	0.55	####	0.48	0.27	0.35	0.413

Table 3.2: Table displaying the cellularity for each patient and sample estmated by ABSOLUTE. The average cellularity refers to the mean value. The #### symbols in entries where sample 2 data would be expected represents that their was no data for sample 2 in these patients, as discussed in Methods.

Table 3.2 is a table displaying the estimated cellularity values for all 106 samples, as well as the mean average cellularity for each of the patients. The values cover a large range, with Patient 8 having an average cellularity value of 0.19 and Patient 24 having an average cellularity of 0.688. In terms of individual sample values, the lowest value is 0.16 and is for Patient 8 sample 4, whereas the highest value is 0.9 and is for Patient 24 sample 3. These values were compared to the equivalent values produced by Sequenza (discussed in the next section) and the combined results were used to determine which patients would be pushed forward along the pipeline and used for further analysis.

	ABSOLUTE Values					
P Number	S1	S2	S3	S4	S5	Avg Ploidy
1	2.01	2	1.96	2.01	1.98	1.99
3	2.15	####	1.91	1.8	2.09	1.99
6	1.91	####	1.83	1.88	5.34	2.74
7	2.04	####	2.06	2.1	2.12	2.08
8	2.14	####	2.21	2.11	2.1	2.14
9	2.05	####	2.05	2.06	2.06	2.06
12	3.27	1.7	1.77	3.37	1.71	2.36
16	2.12	####	1.99	2.06	2.07	2.06
18	1.83	####	1.81	1.77	1.81	1.81
20	2.03	####	2	2.04	2.15	2.06
23	2.06	1.98	2.08	1.94	2.03	2.02
24	1.73	1.77	1.79	1.76	1.82	1.77
27	1.95	1.9	1.95	2.07	1.91	1.96
32	2.08	####	1.94	2.01	1.84	1.97
33	1.88	####	1.9	1.87	1.91	1.89
34	1.85	1.84	1.86	1.84	1.87	1.85
35	1.76	####	2.16	2.14	2.12	2.05
37	1.85	####	1.83	1.87	1.83	1.85
62	2.08	####	2.1	2.13	2.18	2.12
64	2.04	####	2.03	2	2	2.02
75	1.89	####	1.87	1.93	1.91	1.9
78	1.85	####	1.87	1.89	2.08	1.92
84	1.96	####	1.94	1.94	1.99	1.96
85	2.03	####	2.03	1.97	1.95	2
91	1.87	####	1.91	1.94	2.01	1.93

Table 3.3: A table displaying the ploidy for each patient and sample as estimated by ABSOLUTE. The average ploidy refers to mean value. The #### symbols in entries where sample 2 data would be expected represents that their was no data for sample 2 in these patients, as discussed in Methods.

Table 3.3 displays the estimated ploidy values for all 106 samples and the mean average values for each patient. The range of the average ploidy goes from a low of 1.77 for Patient 24 and a high of 2.74 for Patient 6. In terms of individual samples, the lowest value is 1.7 for Patient 12 sample 2 and the highest value is 5.34 for Patient 6 sample 5. Besides Patient 6 sample 5 the other obvious outliers are Patient 12 samples 1 and 4 which have been discussed previously. The average ploidy values are much more cohesive than the average cellularity values, or rather, they show less deviation between patients. Whereas the cellularity values are distributed more widely, the average ploidy is mostly within 0.2 of a ploidy of 2. This is expected as the model parameters for Sequenza

and ABSOLUTE both push towards a ploidy of 2 by default and only deviate from this if a different ploidy and cellularity combination have a higher likelihood. Furthermore, a ploidy of 2 is the default ploidy for sequence in the genome, so it should only differ when a mutation event has occurred. Of the 106 individual samples, 43 had an estimated ploidy of above 2, with a further 4 having an estimated ploidy of exactly 2. The rest of the samples (59) were all estimated to have a ploidy of less than 2, implying that overall, there is a greater amount of loss in the sample exomes than there is gain. There is no apparent correlation between the cellularity and ploidy estimates for each sample, possibly due to higher cellularity values only improving the ability of the models to estimate an accurate ploidy value.

One concern when using ABSOLUTE to validate results from a copy-number calling software, is that the results may be biased towards agreeing with the input, especially in the case of Sequenza where both algorithms use a depth-ratio based estimation system. However, the differences in the models will result in slightly different results in most cases, and this can be seen when comparing the summary figures for Sequenza and ABSOLUTE (Tables 3.13 to 3.16). Another major feature of ABSOLUTE is made to deal with just this possible problem, as when a estimation cannot meet the threshold likelihood ABSOLUTE uses its large database of karyotype examples to match the input with a solution already established.

3.3.3) Sequenza and ABSOLUTE Comparison

The following section contains 3 Tables and 1 Figure, and aims to compare the copy number estimation values generated by both Sequenza and ABSOLUTE, and then evaluate them to determine which values would be suitable for further analysis through the later steps of the pipeline. The 3 Tables display the cellularity values predicted by ABSOLUTE per patient per sample, the ploidy values predicted by ABSOLUTE per patient per sample, and a comparison table representing the mean average cellularity per patient calculated by Sequenza and ABSOLUTE. The Figure is a graphical interpretations of comparisons between the cellularity values produced by the two methodologies and is in the form of a box plot.

	Sequenza Values					
Patient Number	S1	S2	S3	S4	S5	Average Cellularity
1	0.5	0.66	0.5	0.61	0.59	0.572
3	0.22	####	0.5	0.56	0.36	0.41
6	0.49	####	0.54	0.46	0.61	0.525
7	0.19	####	0.17	0.18	0.19	0.19
8	0.15	####	0.18	0.18	0.15	0.165
9	0.16	####	0.21	0.1	0.23	0.175
12	0.4	0.47	0.27	0.38	0.25	0.354
16	0.33	####	0.33	0.27	0.24	0.293
18	0.63	####	0.29	0.45	0.64	0.503
20	0.17	####	0.22	0.24	0.22	0.213
23	0.32	0.26	0.17	0.41	0.36	0.304
24	0.8	0.49	0.9	0.6	0.86	0.73
27	0.57	0.73	0.51	0.56	0.53	0.58
32	0.14	####	0.31	0.23	0.23	0.233
33	0.42	####	0.41	0.53	0.29	0.413
34	0.63	0.7	0.72	0.48	0.67	0.64
35	0.1	####	0.3	0.32	0.16	0.22
37	0.52	####	0.48	0.55	0.47	0.505
62	0.12	####	0.1	0.37	0.1	0.173
64	0.13	####	0.25	0.22	0.19	0.198
75	0.74	####	0.57	0.79	0.32	0.605
78	0.48	####	0.49	0.43	0.31	0.428
84	0.47	####	0.37	0.7	0.36	0.475
85	0.28	####	0.48	0.51	0.49	0.44
91	0.55	####	0.49	0.25	0.33	0.405

Table 3.4: Table displaying the cellularity of each patient and sample as estimated by Sequenza, as well as the mean average cellularity per patient. The #### symbols in entries where sample 2 data would be expected represents that their was no data for sample 2 in these patients, as discussed in Methods.

Table 3.4 is a table displaying the Sequenza estimates of cellularity for all samples across all patients, as well as a mean average of the cellularity per patient. The range of the average values is greater than with ABSOLUTE, with the lowest value being 0.165 for Patient 8 and the highest value being 0.73 for Patient 24. Even though the range is slightly larger, it is the same patients reported at each end of the range scale. The range for individual samples is 0.15 to 0.9, very similar to ABSOLUTE, and is represented by Patient 8 sample 1 and Patient 24 sample 3 respectively. There is a slight difference in that it is Patient sample 1 rather that has the lowest value estimated by Sequenza, but Patient 8 sample 4 that has the lowest value estimated by ABSOLUTE.

Calculating the average cellularity per patient is important for use in quality control before the next step of the pipeline, with cellularity being the determining factor for which patients can be pushed forward in the pipeline.

	Sequenza Values					
Patient Number	S1	S2	S3	S4	S5	Average Ploidy
1	2	1.9	2	2	2	2
3	1.6	####	1.5	1.5	1.5	1.5
6	2	####	2	2	1.9	2
7	1.9	####	2	1.9	2	2
8	2.3	####	2.1	1.9	2	2.1
9	1.7	####	1.8	1.9	1.9	1.8
12	3.4	1.8	1.8	3.4	1.6	2.4
16	1.9	####	2.1	1.9	1.8	1.9
18	1.8	####	1.8	1.8	1.8	1.8
20	1.6	####	2.1	1.8	2	1.9
23	1.9	1.9	1.8	1.9	1.9	1.9
24	1.8	1.9	1.9	1.9	1.9	1.9
27	1.9	1.9	1.9	2	2	1.9
32	1.9	####	1.9	2	2	2
33	1.9	####	2	1.9	1.9	1.9
34	1.8	1.8	1.8	1.8	1.8	1.8
35	1.9	####	1.9	2	1.5	1.8
37	1.8	####	1.7	1.8	1.8	1.8
62	2	####	1.9	2	1.9	2
64	2.4	####	2	2.1	2	2.1
75	1.9	####	1.9	1.9	1.8	1.9
78	1.8	####	1.8	1.7	1.8	1.8
84	2	####	2	2	2	2
85	1.7	####	1.8	1.8	1.8	1.8
91	1.7	####	1.7	1.6	1.5	1.6

Table 3.5: Table displaying the ploidy of each patient and sample as estimated by Sequenza, as well as the mean average ploidy per patient. The #### symbols in entries where sample 2 data would be expected represents that their was no data for sample 2 in these patients, as discussed in Methods.

As discussed previously in the chapter, the Sequenza probability model opts for a default solution of 2 copies when estimating the copy-number of any predicted segment. The logic of this assumption is that the expected copy-number in normal tissue exome sequence should have a value of 2, representing the 2 copies of the genome present in every cell. As such, the mean average values seen in Table 3.5 all have values of near 2, with a range from 1.5 for Patient 3 and 2.4 for Patient 12. The lowest ploidy value for a single sample is 1.5 and is shared by Patient 3 samples 3, 4 and 5, and Patient 91 sample 5, with the highest ploidy value for a single sample being 3.4 for Patient 12 sample 1 and 4. Overall, the values estimated by Sequenza tend to be closer to a ploidy of 2

than those that were estimated by ABSOLUTE, which should be due to the inherent bias of the Sequenza model to default the copy-number to 2. However, of the 106 samples, only 8 were predicted to have a ploidy of over 2 by Sequenza, compared to the 43 predicted by ABSOLUTE. There are 22 samples with a Sequenza predicted ploidy of exactly 2 compared to the 4 predicted by ABSOLUTE. Meaning, that there are 76 samples that Sequenza estimated had a ploidy of less than 2. As with the ABSOLUTE estimates, this implies that there is a greater amount of exome sequence loss in the cohort than gain for the majority of patients. This contributes to the findings of this chapter by validating the copy number losses reported across the patient cohort.

Overall, despite the higher number of lower ploidy values predicted by Sequenza, both sets of results generally show a high level of concordance, with the ploidy estimates having higher similarity than the cellularity estimates in most cases. Even though the ABSOLUTE calls often reported the ploidy as above 2, the similarity in the values was still strong. Two clear exceptions to this are Patient 3 in terms of cellularity and Patient 6 sample 5 in terms of both ploidy and cellularity. The issues with the Patient 3 data are discussed in the next section. For Patient 6 sample 5, it is clear that ABSOLUTE has chosen a model that has a lower cellularity (of 0.3) and a larger ploidy 5.34, whereas Sequenza opted for the model where the ploidy was closer to 2 and there was a much higher cellularity for the sample (of 0.61). This was probably caused by the bias in the Sequenza model to opt for solutions with a ploidy close to 2 by default, with an alternative solution having values more similar to the ABSOLUTE estimates. When viewing the alternative solutions for Patient 6

sample 5 it can be seen that this was the case, where a solution was proposed with a cellularity of 0.37 and a ploidy of 0.57. Both solutions had a similar LPP of near 1, which explains the discrepancy seen here between the two softwares.

Patient Number	Average Cellularity/Sequenza	Average Cellularity/ABSOLUTE
1	0.572	0.58
3	0.41	0.277
6	0.525	0.445
7	0.19	0.23
8	0.165	0.19
9	0.175	0.23
12	0.354	0.402
16	0.293	0.21
18	0.503	0.49
20	0.213	0.2
23	0.304	0.314
24	0.73	0.688
27	0.58	0.568
32	0.233	0.22
33	0.413	0.383
34	0.64	0.632
35	0.22	0.198
37	0.505	0.513
62	0.173	0.2
64	0.198	0.253
75	0.605	0.598
78	0.428	0.41
84	0.475	0.46
85	0.44	0.43
91	0.405	0.413

Table 3.6: A table of the average values of cellularity for each patient as estimated by both Sequenza and ABSOLUTE. All patients comprise of 4 samples apart from 1, 12, 23, 24, 27 and 34, all of which have 5.

The cellularity of a tumour sample is reflective of its quality in respect to how valuable it is for analysis. As discussed in Chapter 2, low cellularity values are a major problem that needs to be overcome when working with cancer sequencing data. As such the values displayed in Table 3.6 were used to determine if samples were of sufficient quality to produce accurate results further down the pipeline. As Sequenza can reportedly estimate accurate results with cellularity values as low as 0.25, this was the value decided upon for whether a patient was sufficient for advancement to the next analytical step. However, in order to utilise the two different copy number estimation methods used on the data, it was decided that a patient would only be removed if both estimates fell below 0.25, or if the average values for cellularity between the two

methods was greater than 0.1. This means that 7 patients did not meet the quality threshold and so were not used for further analysis, these patients are; 7, 8, 9, 20, 32, 35 and 62. Figure 3.12 displays the comparative data for cellularity between the two methods and helps to visualise which patients should be removed from the cohort.

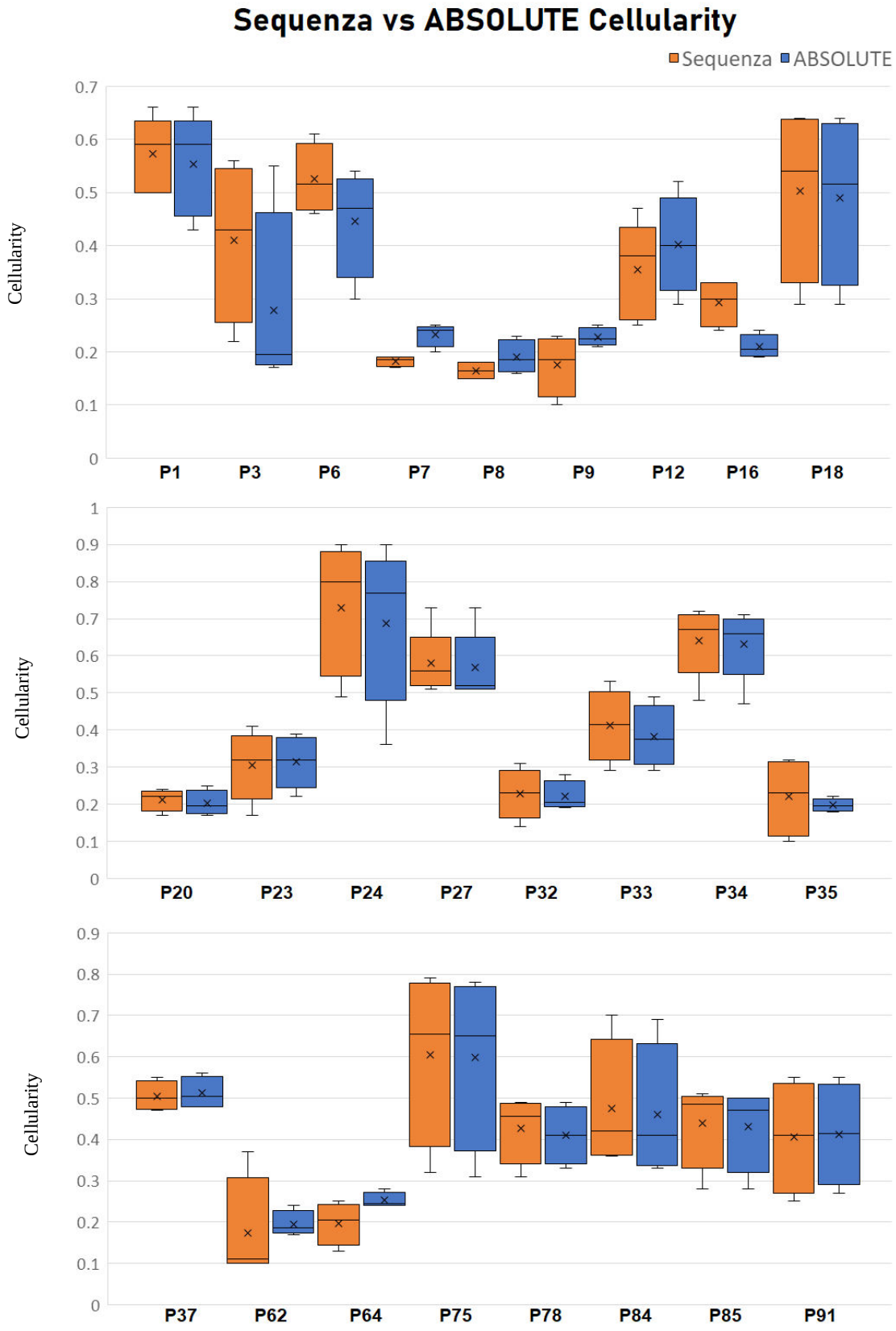


Figure 3.12: Box and Whisker plot of cellularity covering all 25 patients in the initial cohort, with values provided by Sequenza (orange) and ABSOLUTE (blue). The plots are separated into 3 graphs due to convenience in displaying the figure and not for any analytical reasons. The plots are grouped into pairs for each patient so they can be directly compared. The upper and lower values of each plot, the "whiskers", represent the max and min range of each set of values. The top and bottom of the "box" display the upper and lower interquartile range respectively, with the inner line showing the median value. The x in the box shows the mean value of the data.

Figure 3.12 helps to more clearly display the range of values between patients and why 7 of the patient cohort were removed from any additional analysis, all removed due to having cellularity values below the required threshold of 0.25.

As well as the 7 patients removed for low cellularity, Patient 3 was also removed from the cohort. Even though Patient 3 did meet the necessary quality threshold for cellularity from both Sequenza and ABSOLUTE, the difference between the mean values exceeded the limit of 0.1, as can be seen in Table 3.6 and Figure 3.12. It displayed a huge range of copy number values, ranging from 0 to 20, across large segments of the exome and also reported much greater numbers in the amount of segments predicted, especially in samples 3 and 4. Further inspection of the Patient 3 sequencing data revealed large amounts of low quality reads, implying an error during sequence assembly or low quality samples. It is also worth noting that Patient 3 had the biggest difference in average cellularity between Sequenza and ABSOLUTE, which by itself doesn't count for much, but when observed it conjunction with other evidence may also hint towards the Patient 3 sequencing data being unstable. For these reasons it was decided that there was enough evidence to remove Patient 3 from the cohort alongside the 7 patients who didn't meet the cellularity threshold. This left a cohort of 17 patients which could be analysed in subsequent steps of the Sequenza/TuMult pipeline, as well as in other pipelines established in later chapters.

It may be a concern that only the average cellularity was compared between Sequenza and ABSOLUTE, and that the cellularity of individual samples was

overlooked when deciding which patients and samples would be carried forward in the pipeline. For example, Patient 23 sample 3 did have an estimated cellularity of below 0.25 in both software outputs, with a value of 0.17 from Sequenza and a value of 0.22 from ABSOLUTE, so by the criteria used above it would be excluded from any further analysis. However, it was suggested that as long as the average cellularity for a patient met the 0.25 threshold, then the full patient should be included. This was decided in order to maintain the size of the cohort and avoid shrinking it more than necessary, and that the effect of a single sample may be minor in the overall goal of finding driver mutations in the tumour data. A counterpoint to this is that in order for any single event to be called as clonal, it would have to be present in all samples of a patient, and that one sample with low cellularity may cause certain events to be missed. However, not only would this effect be easy to observe in the output of each pipeline, there were only 4 samples total where this was a consideration: Patient 23 sample 3, Patient 16 sample 5. and Patient 64 samples 1 and 4. An event which is referred to as “clonal” means that it originally occurred in the first clone, i.e. the cancer cell population from which the current populations have all descended from.

A final note on this quality control step is to acknowledge that Patient 64 was very close to being excluded from the cohort. With the ABSOLUTE average cellularity estimate putting it only just above the threshold, and it falling significantly below the threshold when estimated by Sequenza, even though it technically passed it could be the case the cellularity is too low for the patient to yield valid results. This is especially true as Sequenza is the software with which it fell below the threshold. However, it was decided that it was better to

keep Patient 64 in the patient cohort as it had technically passed, but to keep in mind how close it was to being excluded when considering results.

3.3.4) Sequenza and ASCAT Comparison

In a similar fashion to the previous section, the Sequenza results were also compared to results produced by the software ASCAT, which was run by Novogene as described in chapter 2. This ASCAT data is what was used to provide copy number estimates in the Zhang, M paper (Zhang, Jin-Li et al. 2021) discussed in chapter 1 and present in the publication list of this project. This section will display a brief example of ASCAT output followed by summary statistics and a box and whisker plot as in the previous section.

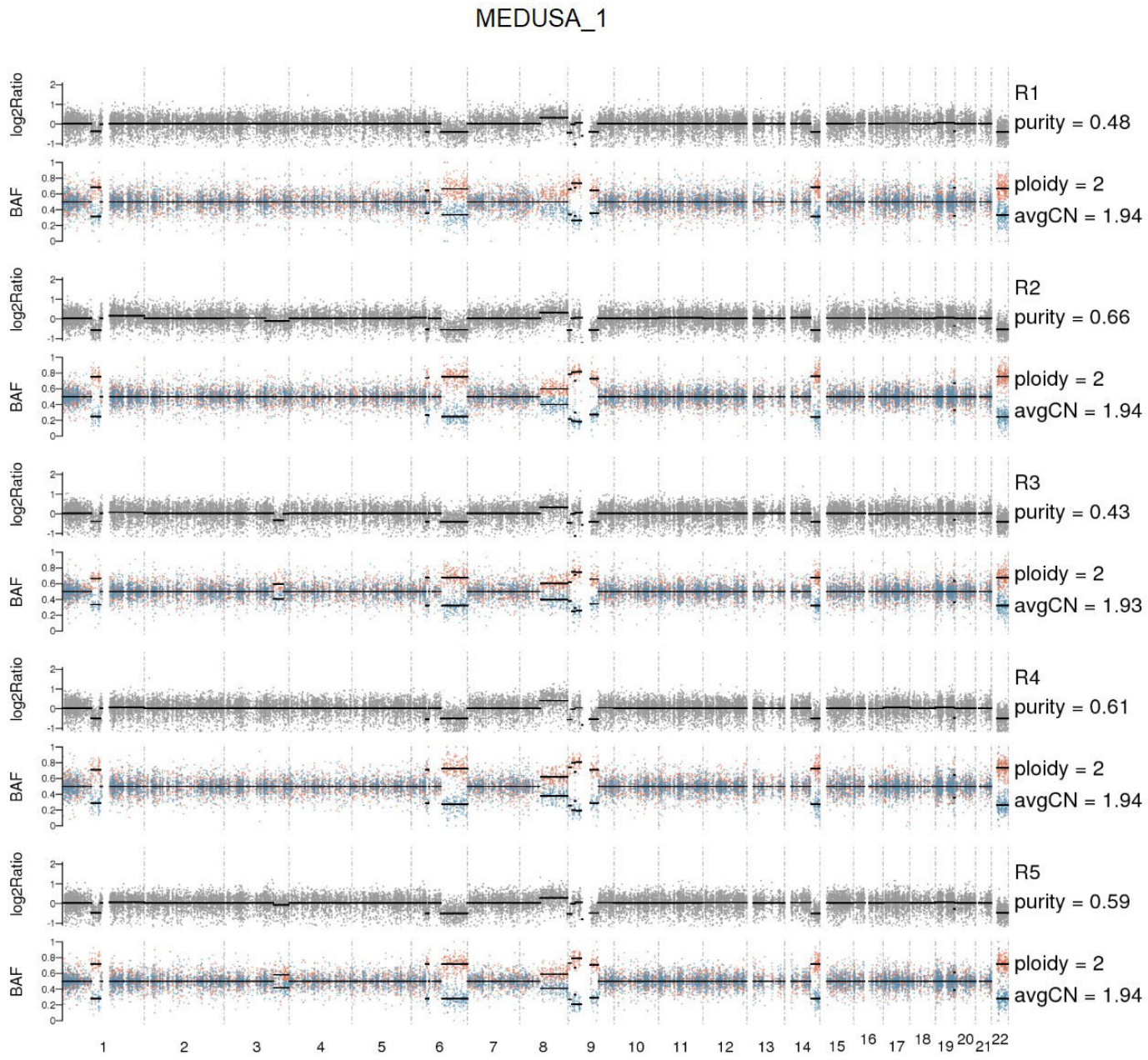


Figure 3.13: The profiles graph produced by Novogene using ASCAT for Patient 1. All 5 samples of Patient 1 are displayed across twin graphs per sample, the top graph displaying the log ratio and the lower graph displaying the BAF. Purity (cellularity), ploidy and the average copy number for each patient are also displayed.

Figure 3.13 is a profile graph produced by Novogene using the ASCAT software for Patient 1 from the MPM cohort. All 5 samples are displayed in the figure, with each set of twin graphs assigned to one sample, the higher graph in each pair showing the log ratio calculated by ASCAT (essentially the depth ratio) and

the lower graph in each pair displaying the BAF. Each sample also has the predicted ploidy and cellularity estimates displayed as well as the average copy number.

	Segments		Mean Cellularity		Mean Ploidy	
Patient Number	Sequenza	ASCAT	Sequenza	ASCAT	Sequenza	ASCAT
1	166	41	0.572	0.554	2	2
6	219	50	0.525	0.523	2	2
12	319	57	0.354	0.484	2.4	2
16	256	42	0.293	0.21	1.9	2
18	280	64	0.503	0.49	1.8	2
23	214	56	0.304	0.314	1.9	2
24	302	83	0.73	0.688	1.9	2
27	272	79	0.58	0.568	1.9	2
33	225	56	0.413	0.383	1.9	2
34	316	97	0.64	0.632	1.8	2
37	200	58	0.505	0.513	1.8	2
64	372	54	0.198	0.253	2.1	2
75	261	54	0.605	0.598	1.9	2
78	215	56	0.428	0.41	1.8	2
84	264	73	0.475	0.46	2	2
85	239	47	0.44	0.43	1.8	2
91	332	46	0.405	0.413	1.6	2

Table 3.7: A table showing comparative statistics from the Sequenza and ASCAT copy number calls, as well as mean average estimations of cellularity and ploidy.

Table 3.7 showcases the main difference in the output of Sequenza and ASCAT, which is in the number of segments called for each patient, with the values for ASCAT being far lower than for Sequenza. The amount called for Sequenza ranges from 166 in Patient 1 to 372 in Patient 64 and a mean number of 262 segments reported across all patients. However, for ASCAT the lower range sits at 41 in Patient 1 and the upper range at 97 in Patient 34 and a mean of 60 segments reported across all patients. It is clear that the sensitivity for detecting copy number events in ASCAT is much lower than in Sequenza, calling approximately 6 times fewer segments. One possibility into what has caused this is that ASCAT was primarily designed with whole genome sequencing in mind, not whole exome sequencing, and that its use with whole exome sequencing data may have resulted in more regions of the data being called with a “normal” copy number, meaning they did not contribute to any segments. Conversely, it may be that Sequenza generates results with more noise, and divides large segments called by ASCAT into smaller segments which incorporate the same region.

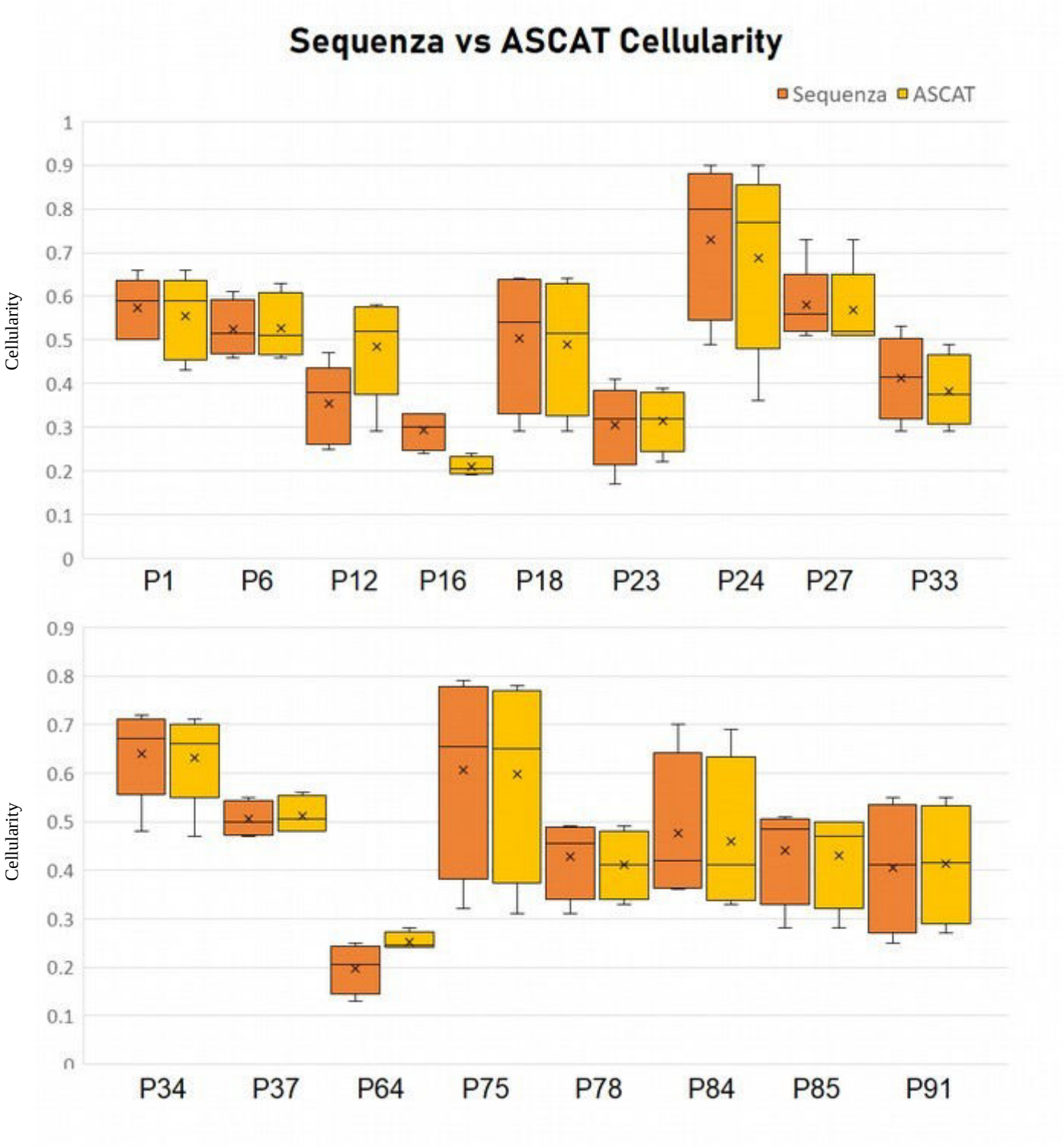


Figure 3.14: Box and Whisker plot of cellularity covering the 17 patients in the reduced cohort, with values provided by Sequenza (orange) and ASCAT (yellow). The plots are separated into 2 graphs due to convenience in displaying the figure and not for any analytical reasons. The plots are grouped into pairs for each patient so they can be directly compared. The upper and lower values of each plot, the “whiskers”, represent the max and min range of each set of values. The top and bottom of the “box” display the upper and lower interquartile range respectively, with the inner line showing the median value. The x in the box shows the mean value of the data.

As shown by Figure 3.14, in terms of mean cellularity, the two methods have a good level of agreement, with only Patients 12, 16 and 64 having a difference greater than 0.05 estimated. The Sequenza estimates display a greater range with an upper value of 0.73 and a lower value of 0.198, whereas ASCAT has an upper value of 0.688 and a lower value of 0.21. Patient 12 has the biggest difference in estimated average cellularity, of 0.13. The general agreement of these values does imply that the ASCAT software was able to work to a reasonably accurate degree, even with the use of whole exome sequencing data, assuming that the Sequenza estimates are also accurate. The difference in average values between the two softwares for Patient 12 would have made it the only patient not to pass this brief quality control step, however, for the following reasons it was not straightforward to eliminate Patient 12 because of this.

Curiously, all samples from all patients have an estimated ploidy value of exactly 2 in the ASCAT results, including samples 1 and 4 in Patient 12 which were thought to have had a genome doubling event present, occurring some time after the initial divergence of the cancer cell populations. In order to check further on whether the genome doubling event was in fact present in the BAM files of samples 1 and 4 (or at least the implication that they could be there), the Integrative Genomics Viewer (IGV) was used, which allows for the visualisation of BAM files with a display of read counts across the regions within the file (Robinson, Thorvaldsdóttir et al. 2011; Robinson, Thorvaldsdóttir et al. 2017). Loading either of the sample files to the IGV alongside the normal sample from Patient 12 clearly showed a large increase in read counts across large portions of the exome, so samples 3 and 5 were also checked, both displaying read

counts more similar to that of the normal tissue, though still slightly increased. This implies that genome doubling event is not merely an artifact and is in fact present in both samples 1 and 4. This raises the question as to why the ASCAT analysis by Novogene either missed or obscured the event from its work flow so that it was absent from the results.

The best explanation is that Novogene used an option present in the ASCAT software to enforce a particular ploidy in the model, that ploidy value being 2 (which is why all sample ploidy estimates are exactly 2), which forced all copy number calls to be made as if that was the case. This is done is place of letting ASCAT estimate the ploidy itself and is actually also an option available to use with Sequenza. The reason ASCAT does this is to control for cell admixture between separate tumour tissue and normal tissue, leaving these parameters of the model to the user. This is likely the reason for the higher average cellularity value for Patient 12 reported by ASCAT compared to Sequenza, as can be seen in Figures 3.6 and 3.7, when the estimated ploidy of a given sample is reduced, it causes an increase in the cellularity estimate to compensate for the change. This is also likely the reason why Patient 12 has the lowest mean average copy number across all samples, with a value of 1.72. As such, it is likely ASCAT would miss a genome doubling event if one was present, and as one is suspected to be present in Patient 12, it was decided that this was likely the reason it didn't meet the comparison thresholds. As such, it was decided to keep Patient 12 as part of the cohort as there was no evidence of low data quality in this particular case.

3.3.5) Sequenza Summary Results

The following table (Table 3.8) displays the summary results for Sequenza across the 17 patients carried forward after comparison.

Patient Number	Total segments reported	Proportion of exome effected %	Losses %	Gains %
1	166	12.5	8.9	3.6
6	219	32	23.5	8.5
12	319	93.7	27.2	66.5
16	256	27.9	23.4	4.5
18	280	34.8	31.9	2.9
23	214	25	15.1	9.9
24	302	44.9	40.5	4.4
27	272	33.4	29.8	3.6
33	225	22.5	21.4	1.1
34	316	38	36.3	1.7
37	200	25.6	25.1	0.5
64	372	33	23.8	9.2
75	261	36.6	26.7	9.9
78	215	34.3	33.4	0.9
84	264	26.5	19.3	7.2
85	239	27.8	24.4	3.4
91	332	34.8	31.4	3.4

Table 3.8: A table displaying the total number of segments called by Sequenza per patient. The proportion of exome effected refers to how much of the sequence of te exome was effected by copy-number changes in total, expressed as a percentage. The losses and gains columns refer to the proportion of the copy-number change that was either a decrease or increase in copy-number respectively.

The total segments reported in Table 3.8 were extracted from the segments file produced by Sequenza as its main output, which contains the genomic positions and copy-number of each distinct segment as reported by Sequenza. The total number of segments has a range of 166 in Patient 1 up to 372 in Patient 64 and the mean number of segments reported across the cohort was 262 per patient. There doesn't appear to be any correlation between either ploidy or cellularity and the number of segments that were detected in a patient, though it should be noted that Patient 64 did have the highest amount of segments by a large margin. It is possible that this is due to the lower cellularity of Patient 64 resulting in larger segments being broken into separate smaller ones due to regions of lower quality, though this is conjecture. It should be

noted that the minimum number of possible segments that can be called by Sequenza on this dataset is 22, as that is the number of chromosomes included for each patient, and as stated previously, a chromosome with no reported copy-number change events would be designated as a single unbroken segment.

The proportion of the exome effected was determined using the positional data included in the Sequenza segments file, by comparing the length of all segments with a copy number bigger or smaller than 2, to the total length of all reported segments. This value had a range of 12.5% in Patient 1 up to 93.7% in Patient 12, and the mean proportion of the exome effected was 34.3%. However, Patient 12 is clearly an outlier here, as the only patient with an exome coverage of over 50%, and only one of two with an exome coverage of over 40% (the other being Patient 24). The possible genome doubling in Patient 12 is what will have caused this phenomenon, resulting in large sections of the exome data to report with higher copy number. The reason that Patient 12 could still be reporting an exome effected coverage of less than 100% even if genome doubling has occurred is due to the timeline of the mutation events in the tumour. If the genome doubling event was the first to occur, then it is likely that the reported number would be even closer to 100%, as single copy gains or losses would be much more difficult for Sequenza to identify. It is also possible that in a patient where genome doubling had occurred as one of the very first copy-number events, it would be difficult to identify that it had occurred, though the use of depth ratio in the Sequenza software would help to solve this issue. The mean average when Patient 12 is excluded is 30.6%.

This is an important finding, as it provides further supporting evidence that MPM is a cancer that involves large-scale copy number events, though as discussed in previous chapters, this has been known for at least 14 years. A similar observation can be made when comparing the proportion of decreases or increases in copy number change in the cohort, or losses and gains respectively. The Sequenza findings agree with previous reports that it is losses that are more prevalent in MPM than gains, with all patients showing a greater amount of loss than gain, except for Patient 12, with the genome doubling event accounting for the huge amount of gains reported.

3.4) TuMult Results

3.4.1) TuMult Patient Results

All 17 remaining patients in the cohort were successfully processed by the Perl scripts and subsequently the TuMult R script, which generates two phylogenetic trees and a table per patient. The first tree is the most important visualisation and will be referred to as the cytobands tree, as it posts the events in cytoband format along the branches of the tree. This is demonstrated in Figures 3.15 and 3.16. The second tree is identical to the first except it posts the events using unique “segment IDs” instead of as cytobands (referred to as the segments tree). These IDs refer directly to the segments file which TuMult produces that lists each event posted on the tree, giving the probes that the event covers, the start and end position, and assigns a unique segment ID to each event. The purpose of the segments tree is simply to allow for mapping of results from the cytobands tree to the segments table without having to manually do it using the cytoband information. An example of a segments tree is provided in Figure 3.17.

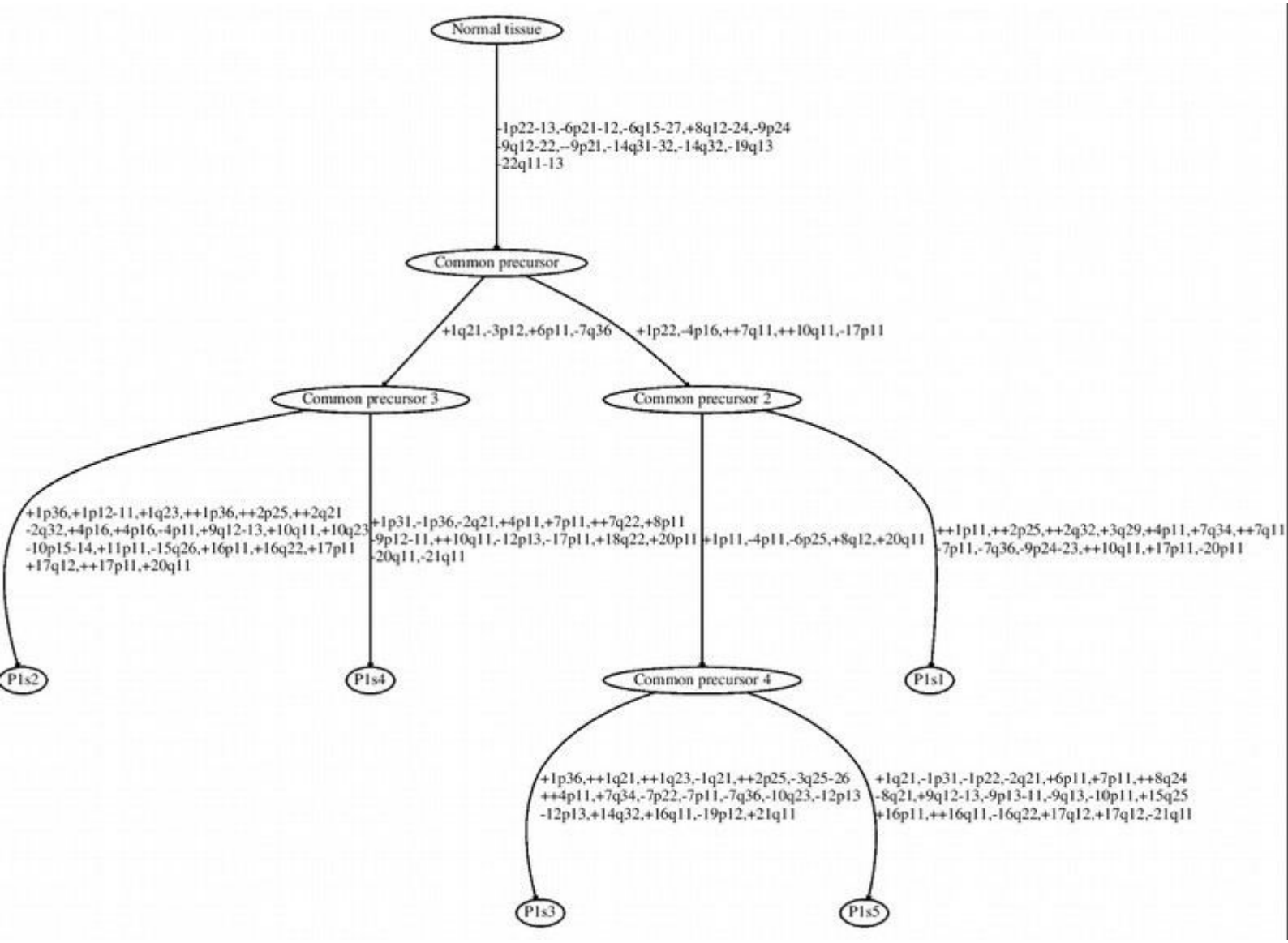


Figure 3.15: The TuMult cytobands tree for Patient 1. Events are displayed as cytoband label and separated from each other via a comma. A '+' sign indicates it is a heterozygous gain in the region of the label, '++' indicates a copy number gain of 2 or more copies, '-' indicates a heterozygous loss, and '–' indicates a homozygous loss. The nodes of the trees have labels regarding the patient number and sample number, for example, P1s1 is the far right node and refers to the patient 1 sample 1 data.

Each terminal node on a TuMult phylogenetic tree represents one of the samples included in the input, and TuMult enforces that these must be unique nodes. This results in the internal “common precursor” nodes which represent hypothetical evolutionary states TuMult predicts were present at an earlier time. The length of the branches which connect nodes to precursors is not representative of evolutionary time and is simply established by TuMult when formatting the tree. Due to TuMult making the assumption that all events are

equal in terms of evolutionary time, a basic method of determining the true length of a tree edge is to count all events which occur between any given node and the normal tissue at the root of the tree. For example, in Figure 3.15, sample 1 (P1s1) is located on the far right of the tree, branching from common precursor 2. Between sample 1 and common precursor 2 there are 1 3 events, between common precursor 2 and common precursor 1 there are 5 events and between common precursor 1 and the normal tissue there are 11 events, giving a total edge length of 29 for sample 1.

Samples which have more events in common are assumed to have diverged later in time than those with fewer events in common. In Figure 3.15, samples 3 and 5 are located at the bottom of the image, and are shown to have the 5 events between common precursor 4 and comm on precursor 2 in common, the 5 events between common precursor 2 and common precursor 1 in common, and the 11 events between common precursor 1 and the normal tissue in common. With a total of 21 common events they have the highest number of common events on the tree and so could be said to be the two most related samples in Patient 1, or rather, the two samples which diverged most recently compared to all other samples.

The most important region of any phylogenetic tree to consider in this project is the branch between the normal tissue and the first common precursor. This is the branch known as the truncal region and represents events which are common to all samples in any given patient and so can be said to have occurred the earliest, and importantly, before the samples were able to significantly genetically diverge. This is the region where driver mutations would

be present, as they would cause the rapid genetic divergence as well as the physical spread of the cancer across the pleura. As such, reported truncal events which recur across the cohort are the key events to consider when searching for possible genetic drivers, and subsequently, identify regions which may contain druggable targets for the treatment of MPM.

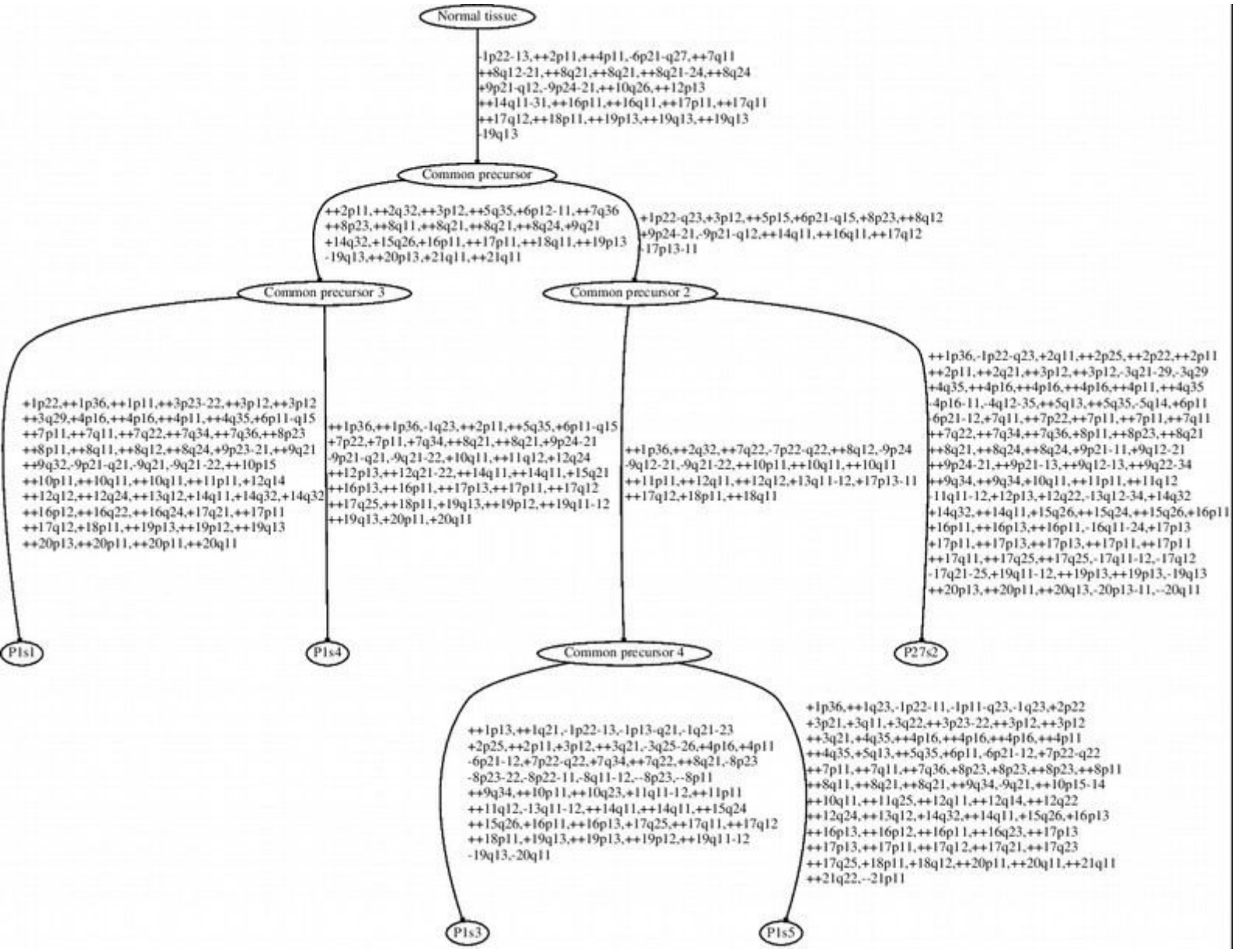


Figure 3.16: The TuMult cytobands tree for Patient 1 generated using a bin size of 10kb. Events are displayed as cytoband label and separated from each other via a comma. A '+' sign indicates it is a heterozygous gain in the region of the label, '++' indicates a copy number gain of 2 or more copies, '-' indicates a heterozygous loss, and '–' indicates a homozygous loss. The nodes of the trees have labels regarding the patient number and sample number, for example, P1s1 is the far left node and refers to the patient 1 sample 1 data.

Figure 3.16 acts as an example of what a lower 10kb bin size results in when compared to the tree in Figure 3.15 with a 100kb bin size. Generally, the trees appear the be quite similar, with the same topology of samples, though it is clear that there is a much greater number of events reported in the 10kb tree. This is because many copy number events exist with sizes between 10kb and

100kb which would be excluded when binned in the latter size but not the former. It can also be seen that these events are far more likely to be larger scale copy number gains, often reporting copy numbers above the exome average but only for relatively small regions. There is a strong possibility that these events are artifacts generated by Sequenza due to lower qualities in these small regions which can affect the depth ratio calls made.

It can also be seen that the truncal region is very different between the two trees, with a much larger number of gain events in the 10kb bin size tree (Figure 3.16), whereas in the 100kb bin size tree (Figure 3.15) there was only a single gain event in the truncal region. The overall topology of the tree is also altered between the two versions, with samples 1 and 4 being grouped in Figure 3.16 rather than samples 2 and 4 in Figure 3.15. It was decided that the 100kb trees would reduce the noise present in the 10kb trees whilst still maintaining enough sensitivity to include the vast majority of eligible segments produced by Sequenza. This was due to MPM being characterised by large-scale copy number loss events, with an influx of small gain events unlikely to be contributing to the disease if they are even present in the first place.

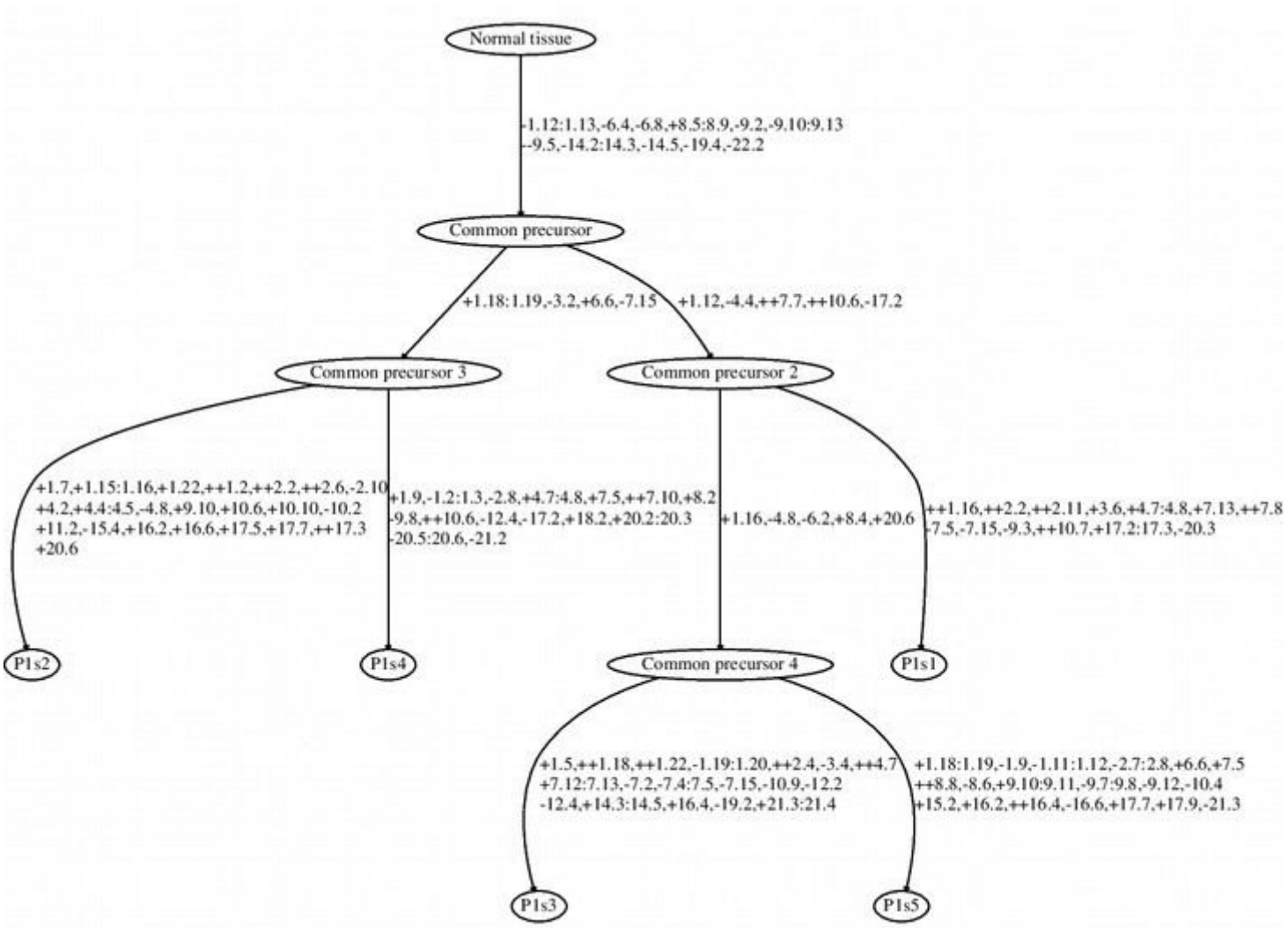


Figure 3.17: The Tumult segments tree for Patient 1. This tree is the sister tree to the one presented in Figure 3.15, and both trees are actually identical in structure. Rather than showing the cytobands as events, the segments tree shows the unique ID of the segment or segments used to create specific events. These IDs correspond to the segments file generated by TuMult. The nodes of the trees have labels regarding the patient number and sample number, for example, P1s1 is the far right node and refers to the patient 1 sample 1 data.

Figure 3.17 is the sister tree of the cytobands displayed in Figure 3.15 and it can be seen that structurally both trees are identical, with the only difference being the labels on each branch. These labels actually refer to segment IDs that are established by TuMult and displayed in a segments file produced as additional output. These IDs link the information present in the file with the position on the cytobands tree (Figure 3.15), with the segments tree acting as a

key to link the two together. The reason for producing this figure is simply as an example to showcase what TuMult produces and how specific genomic positions were identified based on the cytoband given in the cytobands tree. All 3 outputs are required in order to properly identify which regions are effected by the loss or gain of copy number and where they were then placed on the phylogenetic tree, as just having the cytoband displayed on the cytobands tree does not necessarily mean the entire cytoband is actually involved in the event, just that it incorporates some portion of it. This avoids the accidental association with genes that may exist within a certain cytoband, but not actually exist within the copy number event (defined by the segment) which is present on the tree.

One particular tree of interest is that of Patient 12, as the increased ploidy estimates and very large proportion of exome predicted to be effected by copy number change both suggest that a genome doubling event may have taken place.

3.4.2) Genome Doubling in Patient 12

Patient 12 samples 1 and 4 have been mentioned multiple times throughout this report up to this point, in relation to the much higher ploidy value estimated by Sequenza and ABSOLUTE in comparison with the other samples and other patients, which in turn resulted in a large amount of copy number gain events being reported by Sequenza for these samples. Besides genome doubling being responsible, two alternative explanations were provided in the form of chromothripsis, and admixture of multiple tumour populations.

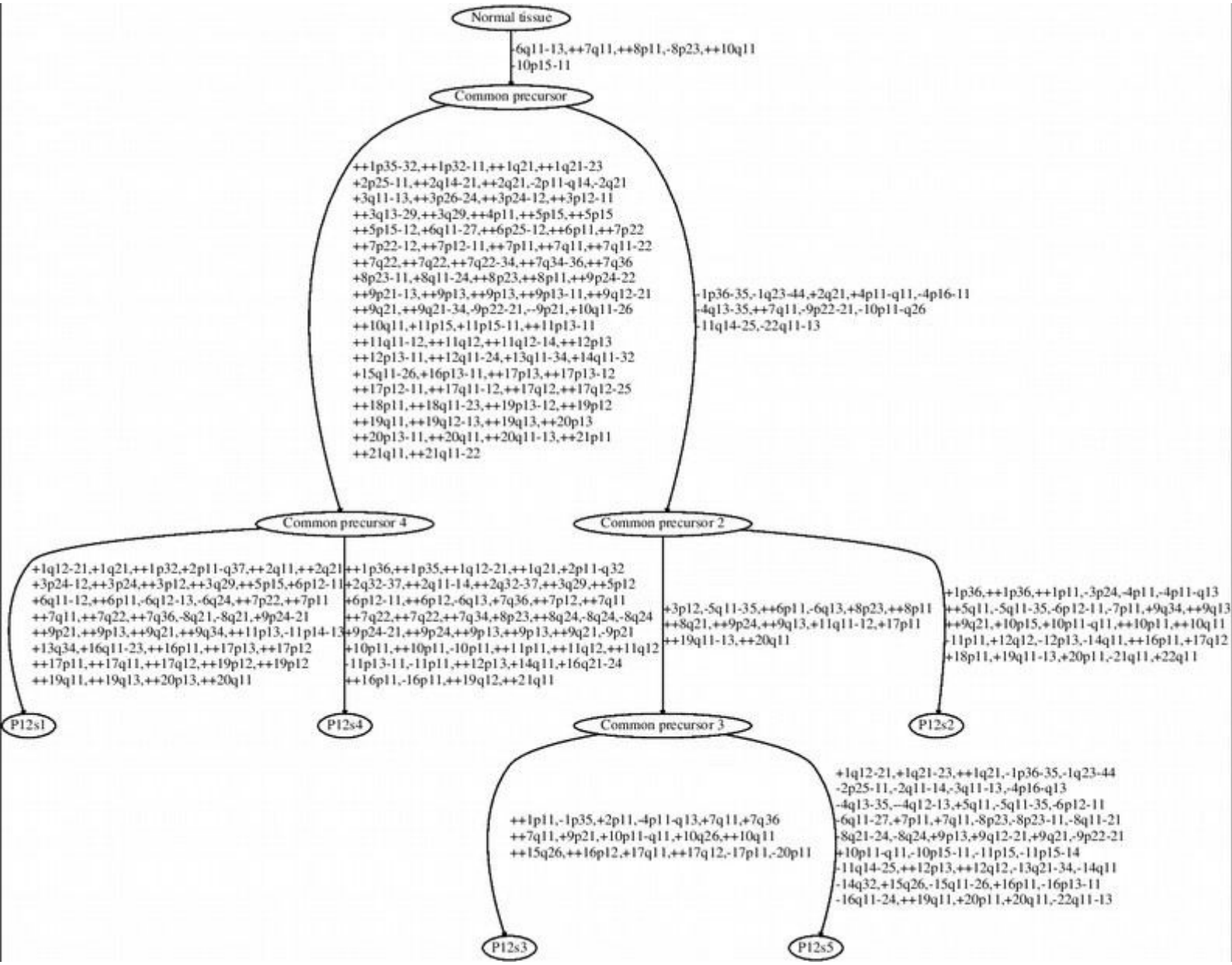


Figure 3.18: The cytobands tree for Patient 12 produced by TuMult. The large number of gain events seen between common precursor 1 and common precursor 4, as well as between the sample 1 and sample 4 nodes is indicative of a genome doubling event occuring.

Figure 3.18 allows for a visualisation of what might have happened in these Patient 12 samples, and it suggests that genome doubling may in fact be the best explanation for the large amount of copy number gain, even though the initial spread of copy number values shown in Figure 3.11 suggested that this may be unlikely.

The truncal region between the normal tissue and the first common precursor represent the point in evolutionary time where the cancer had not yet diverged at all. During the period represented by the truncal branch, TuMult has calculated that 6 large scale copy number events occurred. It is after this point, it can be seen that the cancer diverged into at least two distinct cancer cell populations (other cancer cell populations will also have formed but no longer be detectable due to either being too small in number or no longer being present, both caused by selection within the populations), which are represented by common precursor 4 and common precursor 2. Examining the patient data in Table 3.1 indicates that patient 12 had epithelioid MPM, with this phenomenon possibly being explained if they had biphasic MPM.

This is a good opportunity to mention a limitation with the use of TuMult as a phylogenetic software, though one that can be overcome through more careful analysis of the results. As stated previously, TuMult assumes all events are equal in terms of evolutionary time, and it is instead the number of events which is used to determine how recently two samples may have diverged. In this scenario, as all the branch lengths for samples 2, 3 and 5 are shorter than the branch length between common precursor 1 and common precursor 4, it would be assumed that these populations were already established before common precursor 4 had become fully established. However, if it is a genome doubling event responsible for the huge amount of gain events seen, then it is likely that all of these events took place simultaneously, due to endoreduplication (where a genome has duplicated but no cell mitosis has taken place) or a failure in cytokinesis (Nik-Zainal, Van Loo et al. 2012). There is no way for TuMult to list such an event as a single event, due to the technical constraints of the

software, with no way to account for events which take place over multiple chromosomes simultaneously. In this specific example, the event is easy to spot and so does not convolute analysis, though it should be noted that although the branch lengths for samples 1 and 4 would imply that they diverged from the common precursor less recently, this may not be the case.

Therefore, all that can be determined about the genome doubling event is that it happened after the initial divergence of the samples, and so probably occurred after the cancer had already entered its more aggressive stage and begun to spread across physical distance.

The genome doubling may have eclipsed events that had taken place prior to it happening, with loss events then regaining copy number in regions where it was decreased. This phenomenon can explain why the copy number distribution seen in Figure 3.11 was so dispersed. Unfortunately, the genome doubling event may have also resulted in limiting how many truncal events could be found. Earlier it was mentioned that the Patient 12 TuMult tree had 6 reported truncal events, however this could be an underestimation of how many events actually took place prior to the initial divergence of the samples. Clearly, this is an issue as the events which occurred before the initial divergence are the events with the most clinical significance, as well as the events that this pipeline was built to identify. Therefore, it could be the case that amongst the events assigned to the branch between common precursor 1 and common precursor 2, there are actually truncal events, masked in samples 1 and 4 by the genome doubling, which were unable to be accurately identified by Sequenza. Although not recorded as truncal events, this branch will be discussed again later in the

chapter.

The presence of genome doubling in cancer cells is not a novel finding, and they have been previously reported both in MPM (Hmeljak, Sanchez-Vega et al. 2018) and in other cancers (Bielski, Zehir et al. 2018; López, Lim et al. 2020; Nik-Zainal, Van Loo et al. 2012). It has been established that these genome doubling events tend to take place during the later stages of cancer evolution (Hmeljak, Sanchez-Vega et al. 2018) and this could be due to a specific type of selection which results in the proliferation of cancer cells which have had a genome doubling event. Cancer cells are able to mutate at a much greater scale than normal cells and still remain viable, however this does have a limit and at a certain point, a cancer cell can have picked up so many deleterious mutations that it can no longer maintain replication. It has been suggested that it is this scenario, which results in the establishment of a genome doubling cancer phenotype, as the cells may be able to recover some of the lost sequence in order to maintain viability. This results in positive selection towards cells which have genome doubling as other cells in the population are unable to maintain their own replication (López, Lim et al. 2020). This hypothesis makes sense here, as we have already established that MPM is a cancer which displays a large amount of genomic loss, and provided evidence which supports that claim, and so could be providing the perfect selection pressure to cause genome doubling events to occur.

Overall, only 1 of the 17 patients displayed any sign of genome doubling, and it is likely that the event occurred later on in the evolution of the cancer (or at least, occurred after the cancer had already diverged and began to spread).

This makes it unlikely that genome doubling is a significant driver in the spread of MPM, though its presence can cause problems with genome analysis.

The complete set of TuMult trees for the cohort can be seen in Appendix A.

3.4.3) TuMult Summary Results

This section covers the summary results from the TuMult analysis and includes the following figures: the total proportion of the exome sequence effected by copy number alteration events, as well as the proportion that were clonal (present in the truncal region) or subclonal (present in shared or node branches) (Table 3.9); the total number of events reported for each patient tree and the total branch length for each sample per patient (Table 3.10); a table showing all clonal regions which showed recurrence between patients (Table 3.11); and a table showing in which patients the recurrent clonal regions were detected (Table 3.12).

Patient Number	Proportion of exome effected %	Clonal proportion %	Subclonal proportion %	Heterogeneity index
1	12.5	10.4	2.1	4.95
6	32	26.1	5.9	4.44
12	93.7	1.8	91.9	0.02
16	27.9	0.1	27.8	0
18	34.8	14.4	20.4	0.71
23	25	12.3	12.7	0.97
24	44.9	11.8	33.1	0.36
27	33.4	20.7	12.7	1.64
33	22.5	14.8	7.7	1.93
34	38	13.1	24.9	0.53
37	25.6	16.8	8.8	1.89
64	33	6.8	26.2	0.26
75	36.6	21.8	14.8	1.47
78	34.3	21.6	12.7	1.7
84	26.5	16.7	9.8	1.69
85	27.8	16.8	11	1.53
91	34.8	12.6	22.2	0.56

Table 3.9: A table displaying the proportion of the exome effected by copy number changes per patient. Also displays the proportion of the change which was assigned as clonal and subclonal, with a heterogeneity index showcasing the relationship between the two.

In Table 3.9 the proportion of exome effected by copy number changes is the same values that can be seen in Table 3.8, though now, instead of showcasing the proportion of losses versus the proportion of gains, this table displays the proportion of events which were assigned as clonal or subclonal by TuMult. In this context clonal means events which were present in the trunk due to being present in all samples within the patient, and subclonal refers to all other events on the tree not assigned to the trunk, i.e. events in the shared branches and branch nodes. The clonality of these events has not only be defined by their presence in each sample though, as the breakpoints at either end of the event had to be similar enough for TuMult to classify it as a single event. This is essentially the key aspect of the pipeline, and is the biggest strength of this analysis, as the likelihood of the same breakpoints occurring in parallel after the initial divergence of samples should be extremely low. The heterogeneity index is calculated by dividing the clonal proportion by the subclonal proportion, with a value of 1 indicating an identical amount of both clonal and subclonal, a value

higher than 1 indicating that the clonal proportion is higher and a value lower than 1 indicating that the subclonal proportion is higher.

The range of the proportion of the exome which has been classified as clonal ranges from 0.1 in Patient 16, up to 26.1 in Patient 6. The Patient 16 tree only has two events in the truncal region, one gain and one amplification (a gain with a copy number increase of more than 2), both of which are small in size, being 300kb and 400kb respectively. Patient 16 has not had any previously abnormal results, so the very small number of truncal events is difficult to explain. It could be the case that there was just an extremely early divergence in this patient, with the samples then deriving separately from these distinct populations. Another explanation is that there was no divergence and that multiple tumours formed independently in this patient, leaving no clonal events and the samples were never all part of a single tumour cell population, but there was no mention in the patient characteristics data that this may be the case. However, in order for the second explanation to fit, it would mean the two events that are listed as truncal would have to have occurred by chance, which has already been stated as being extremely unlikely. Or, it would mean that the two truncal events are in some way artificial, however, besides the events being fairly small there is no other evidence that would suggest that this is the case.

The range of the proportion of the exome which has been classified as subclonal ranges from 2.1 in Patient 1 and 91.9 in Patient 12. The reasons for this high number in Patient 12 have been discussed at length already, with the genome doubling event occurring after the first divergence being responsible. Taking the next highest value puts the top range at 33.1 in Patient 24. The

heterogeneity index in the final column of Table 3.9 basically represents whether a patient has more copy number change events in the truncal region or on the branches of the tree. Overall, 9 patients display a heterogeneity index value of above 1 and 8 patients display a value of below 1, giving a fairly even split of patients that have a majority of change in the trunk and patients which have a majority of change in the branches.

Logically, the longer an MPM tumour has been present in a patient in its aggressive form, the lower its heterogeneity index value will be. This is due to there being no change to the number of clonal events once the populations have already diverged from the original tumour, but an increase in the number of subclonal events as the cancer continues to mutate and accumulate more copy number changes. Whilst it would be very interesting to compare the heterogeneity index values calculated here to the progression of the cancer in each of the patients at the time of surgery and sample extraction, unfortunately, this data is not available.

		Branch Lengths				
Patient Number	Total number of events	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
1	175	29	37	39	30	40
6	196	37	####	60	47	52
12	431	132	45	47	132	75
16	185	31	####	51	40	63
18	285	66	####	56	77	86
23	232	52	43	69	29	39
24	302	74	32	79	47	70
27	349	67	94	54	68	66
33	198	38	####	58	47	55
34	364	73	79	64	63	85
37	197	43	####	46	60	48
64	426	112	####	81	121	112
75	232	59	####	46	76	51
78	198	47	####	49	54	48
84	306	73	####	90	78	65
85	258	62	####	62	71	63
91	316	77	####	66	80	93
	Average Branch Length	63.1	55	59.8	65.9	65.3

Table 3.10: A table showing the total number of events reported on the TuMult tree for each patient. It also displays the branch length of each sample for each patient, and then an average branch length across all samples. The #### values present in the sample 2 column indicate that the given patient didn't have sample 2 data available, as discussed in Methods.

In Table 3.10 the total number of events on the TuMult tree for each patient is listed in the second column. This is where a noticeable divergence from the Sequenza results can be seen, in Table 3.4. The numbers of total events calculated by TuMult can actually exceed the number of segments reported by Sequenza for the same patient as the same segment can be used to account for multiple different events that have occurred in separate samples. The range of the total number of events has a low of 175 for Patient 1 (which also had the lowest proportion of the exome effected) and a high of 431 for Patient 12 due to genome doubling and having 5 samples instead of 4. The next highest would be Patient 64 which has a total of 426, possibly due to the lower cellularity values estimated in this patient, resulting in a larger amount of smaller segments being reported by Sequenza.

Recurrent Clonal Region	Number of Patients
-22q	13
-3p21	10
-9p21	9
-14q	8
-6q	5
-13q	5
-1p36	4
-4	3
-17p	3
-10q23	2

Table 3.11: A table displaying recurrent clonal regions found in the patient cohort. For an event to be included, it had to be present in at least 2 patients.

Table 3.11 is a table showing all of the recurrent clonal regions calculated by TuMult. In order to be classified as recurrent and clonal, a region simply had to be present in full, in the truncal region of at least 2 patients. These results are essentially the key findings of the Sequenza/TuMult pipeline, as these regions are where the genes which have the potential to trigger the proliferation of MPM and cause it to shift from its latent phase into its aggressive phase (i.e. driver genes) will be found. The next section of text will discuss the findings presented in the table, with candidate genes being mentioned throughout but not fully discussed. However, detailed discussion on the possible biological pathways that these genes involve will be done later in the report, in the final section of Chapter 5.

Overall, there were 10 recurrent clonal regions calculated by TuMult across the

patient cohort, with the most common region being 22q, which was present in 13 of the 17 patients. The first obvious thing that can be noticed with the regions, is that they are all caused by copy number loss events, supporting the notion that MPM is a cancer mainly driven by genomic loss in the cancer cells. Whilst all of these regions are shown as heterozygous losses, where one copy of the chromosome covering that section is sequence is lost, it should be noted that in the case of the 9p21 region, homozygous losses were detected in 2 patients, Patient 1 and Patient 37. Homozygous loss is when both copies of the sequence for a given region are completely deleted, resulting in a total loss of function for any protein-coding sequence present there. The heterozygous loss of one copy of a gene region can be enough to cause functional change in a cell due to the lower amount of protein expression, mainly through a lower efficiency at a specific stage of a cellular pathway or the inability to meet a protein threshold in the case of regulation. Homozygous losses are a step up from this, with the protein expression reduced to zero. Although only detected in 2 patients out of 17, and only in a single recurrent region, this finding indicates that the 9p21 region may be especially important in the acceleration of MPM progression.

Recurrent Clonal Region	P1	P6	P12	P16	P18	P23	P24	P27	P33	P34	P37	P64	P75	P78	P84	P85	P91
-22q																	
-3p21																	
-9p21																	
-14q																	
-6q																	
-13q																	
-4																	
-1p36																	
-17p																	
-10q23																	

Table 3.12: A table displaying the patients where each clonal region was detected, coloured in blue. In order for a clonal region to be counted, it had to be present in at least 2 of the patients in the cohort.

Table 3.12 displays the recurrent clonal regions in association with the patients they were detected in, ordered by frequency of occurrence of each specific recurrent event. These recurrent deletion events are a key finding of the project and are discussed at length at the end of the chapter.

3.4.4) ASCAT/TuMult Results

The ASCAT results produced by Novogene were also run through TuMult to evaluate whether the findings were significantly different. An example ASCAT tree is displayed as well as 2 summary tables representing the TuMult/ASCAT analysis.

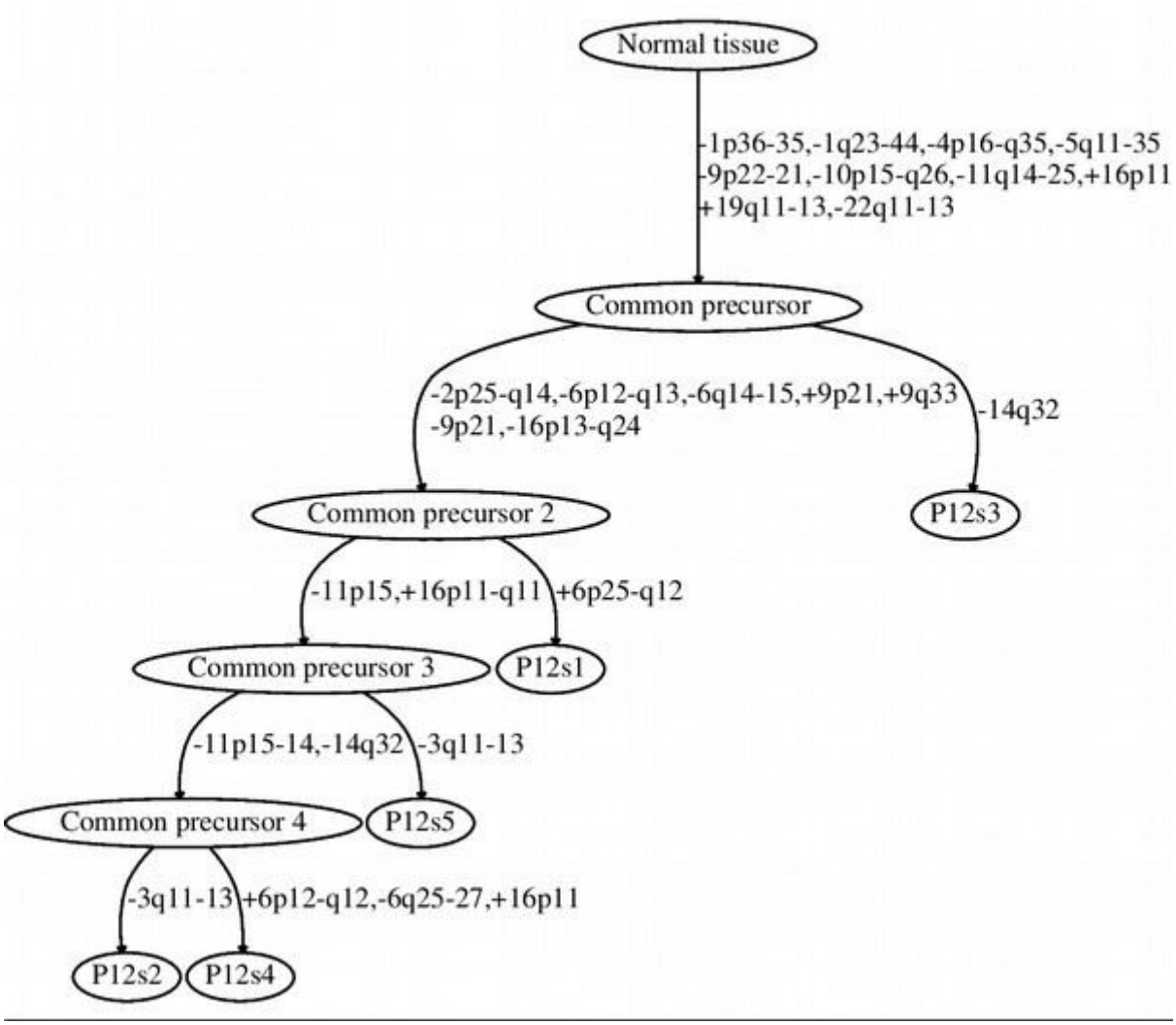


Figure 3.19: TuMult tree produced using copy number segments called by ASCAT (generated by Novogene) for Patient 12. Events with a – are heterozygous deletions, + are single copy number gains, -- is a homozygous deletion and ++ is a copy number gain of 2 or more. The labels in each node refer to the patient number and sample number represented by that particular branch of the tree, for example, P12s2 in the far right node indicates that it is data from Patient 12 sample 2.

The ASCAT/TuMult cytobands tree is shown in Figure 3.19, appearing very differently to how it looked when generated using the Sequenza/TuMult pipeline. The topology of the tree has completely changed, as the lack of the genome doubling event results in the samples being placed into different clades due to the lower number of overall events. In terms of topology, this tree is far weaker than the Sequenza/TuMult tree, as it is highly likely that samples 1 and 4 should be placed together, though the elimination of the genome doubling may excuse this. Interestingly there are now far more events present in the truncal region of

the tree compared to the other pipeline, with this being a common theme throughout most of the ASCAT trees. Many of these events also coincide with the recurrent clonal events reported in Sequenza/TuMult.

Recurrent Clonal Region	Number of Patients
-22q	14
-3p21	11
-9p21	15
-14q	10
-6q	8
-13q	5
-1p36	9
-4	10
-17p	4
-10q23	3

Table 3.13: A table displaying recurrent clonal regions found in the patient cohort using ASCAT/TuMult. For the sake of comparison, only events that were recurrent in Sequenza/TuMult are considered.

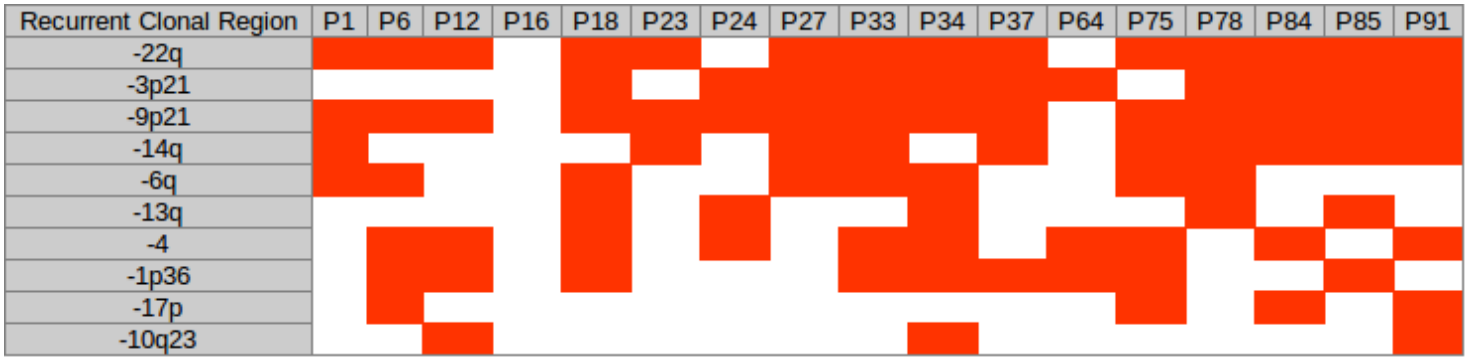


Table 3.14: A table displaying the patients where each clonal region was detected in the ASCAT/TuMult trees, coloured in red. For the sake of comparison, only events that were recurrent in Sequenza/TuMult are considered.

Table 3.13 and 3.14 display the increased amount of recurrent clonal regions which are predicted using ASCAT rather than Sequenza, with all events except 13q loss being reported more frequently. Losses in 9p21, 1p36 and complete loss of chromosome 4 are all greatly enriched, appearing at least 5 times more in the ASCAT/TuMult trees. All patients also saw either an increase in the number of clonal events reported or the same amount, with all Sequenza/TuMult clonal events also being reported in ASCAT/TuMult, with the exception of 1p36 loss in Patient 24, and 10q23 loss in Patient 85. The

increased frequency of 9p21 loss is especially interesting as it is the most recurrent loss in the ASCAT/TuMult trees. In order to find out whether it was the Sequenza calls or the TuMult process which was preventing these 9p21 loss events from being called as clonal, the original Sequenza segmentation files were browsed to find the specific break points in chromosome 9. For Patients 6, 23, 24, 34 and 78, there were 9p21 loss events present in all samples, however, at least one of the samples always had different chromosome breakpoints encapsulating the region, which prevented TuMult from placing the events on the truncal branch. However, there were never more than two sets of distinct breakpoints in any of these 5 patients. This implies, that at least for these patients, the 9p21 loss event was not clonal, but in fact an event that occurred shortly after divergence of the first cancer cell populations.

Clearly, the difference in ASCAT and Sequenza when calling copy number segments results in variable events being associated with the truncal region in TuMult. However, in the opinion of the author, Sequenza is the superior tool for calling copy number segments when trying to determine accurate breakpoints. Sequenza considers the read counts and therefore, depth ratio, of all of the sequence provided to it, whereas ASCAT uses an SNP based process, meaning it does not consider every single base when establishing copy number predictions. This could lead to the missing of specific breakpoints and result in the algorithm assigning whole regions to certain copy numbers between SNPs, essentially creating artificial breakpoints where an SNP is present. This would explain why there is an increase of clonally called events in TuMult, as it is not in fact a higher frequency of events, but instead a higher frequency of breakpoints being detected.

Despite this limitation with ASCAT in relation to generating TuMult trees, it is still useful to see that most of the events called by Sequenza are also confirmed by ASCAT, increasing the confidence that these events have actually occurred clonally. The comparison of different methodologies via practical means such as this is a key part of the scientific process, especially in the field of bioinformatics, where software is often taken at face value. ASCAT has shown that it has the ability to call copy number events accurately, but is ill suited to TuMult because it uses an SNP-based method to estimate copy number, resulting in a surplus of false positive calls in relation to specific breakpoints. However, for phylogenetic methods where specific breakpoints are not taken into account, the segments produced by ASCAT would be more likely to accurately call clonal regions.

The full ASCAT/TuMult results figures are available in Appendix B.

3.5) Discussion of Copy Number Pipeline Results

3.5.1) Recurrent Events and Patient Exceptions

Examining Table 3.11, it has been the top three most common recurrent regions in the table which have been most extensively associated with MPM, 22q, 3p21 and 9p21, with copy number losses in these regions appearing frequently throughout the literature (Bott, Brevet et al. 2011; Bueno, Stawiski et al. 2016; Hylebos, Van Camp et al. 2016; Hylebos, Van Camp et al. 2017; Lindholm, Salmenkivi et al. 2007; Prins, Williamson et al. 1998; Sekido, Pass et al. 1995). Table 3.12 reveals, that in the case of the patient cohort used here, almost every patient experienced a clonal loss event which effected at least one of these 3 regions, with the exception of 3 patients: Patient 12, Patient 16 and Patient 64. This is especially interesting as these are 3 patients which have all been mentioned previously in the report due to various problems during analysis or abnormalities in the results compared to the rest of the cohort.

Earlier it was shown that there was likely a genome doubling even in Patient 12, which may have had a convoluting effect on the Sequenza/TuMult analytical pipeline, in that it may have obscured certain clonal events from being detected. It was suggested that the shared branch between common precursors 1 and 2 could actually hint at events that in reality, are clonal in Patient 12, but were undetectable by Sequenza in samples 1 and 4. Interestingly, when observing the Patient 12 tree (Figure 3.18) it can be seen that on the shared branch between common precursors 1 and 2, losses in 1p36, chromosome 4, 9p21, 10q23 and 22q are present. One explanation for this could be that these events did in fact occur clonally, prior to the initial divergence of cancer cell populations

and subsequent genome doubling event in one of those populations. Further evidence towards this point is that no events associated with these regions can be seen in the shared branch between common precursor 1 and 4, or on either of the node branches for samples 1 and 4. Should these regions have been unaffected in the samples, then it would logically make sense for them to appear as gain events, along with most of the rest of the genome. These loss events occurring, causing a loss in copy number, and then having that copy number recovered during the genome doubling event would make sense. In theory, Sequenza should be able to detect this kind of phenomenon as the depth ratio of the regions in question should still be half of that of the majority of the genome which was not initially affected by any sort of copy number change. Though there is a strong possibility that these events were missed due to the complexity introduced by the genome doubling event subsequent to their initial loss.

Patient 16 only had a very small proportion of the exome effected by events which were calculated as truncal, with only two smaller events reported in that branch of the tree. In the node branches of the tree, where events specific to each sample are listed, samples 1 and 5 both displayed losses in 22q, samples 4 and 5 shared a 3p21 loss event, and sample 5 also showed an almost complete loss of chromosome 4 as well as a heterozygous loss event in 9p21. Though these events are present lower in the tree, it is still confusing as to why no significant events were reported in the truncal branch, and as to why sample 3 contained no loss events in common with the other samples. Overall, the defining feature of Patient 16 is the accumulation of many small gain events, each only effecting small sections of the genome. It is possible that the two gain

events calculated to be in the truncal region do play a role in the proliferation of MPM, being in the region of 10q11 and 11q14. There are studies implication both of these regions in cancer, but only in the case of prostrate cancer for which they have both been associated (Pomerantz, Shrestha et al. 2010; Schleutker, Baffoe-Bonnie et al. 2003). This does seem unlikely however, with both association being in non-coding regions of the genome and the study that implicates 11q14 reporting a heterozygous loss in the region rather than a gain. Neither region has been previously implicated with MPM in any previous literature. The results for Patient 16 remain a mystery, though the most robust explanation is that the samples derived from two distinct cancer cell populations which never shared a single population, or, that the slightly lower cellularity estimate of the patient (0.29) meant that copy number could not be accurately called.

In the case of Patient 64, it is probable that it is simply the quality of the data which resulted in the the lack of any clonal loss of 22q, 3p21 or 9p21. As mentioned in previous sections of this chapter when discussing the selection of patients to move forward with in the pipeline, Patient 64 was very close to being excluded, and actually fell below the minimum threshold for selection in relation to the Sequenza results (with a value 0.198). In actuality, every sample did have a 3p21 heterozygous loss called by Sequenza, but the separation of chromosome breakpoints meant that TuMult did not calculate this to be caused by a single event, but instead as two separate events, one incorporating samples 1, 4 and 5 and one incorporating sample 3. It is unclear whether this is the truth though based purely on the output of TuMult, this can not be called as a clonal event and so must be excluded from the figure, though it is still

interesting to note the involvement of 3p21 in every sample. This example, though it may be caused by low quality data, actually represents an essential argument in the use of phylogenetics to analyse cancer cell data in order to find key driver mutations. A simple association study using the Sequenza output would have found that the 3p21 event in all of the Patient 64 samples was clonal, simply because it was present across all physical regions of the cancer. However, as mentioned in Chapter 1, simply finding events which have occurred in all samples is not enough, as it neglects that parallel evolution may have resulted in the events happening subsequent to the cancer already entering its aggressive state. When trying to find candidate gene targets to arrest the spread of MPM (or any cancer), it is imperative that the mutation events must occur prior to divergence of the cancer cell populations.

Whilst copy number loss events in the three most recurrent clonal regions (22q, 3p21, 9p21) are clearly very important in the proliferation of MPM, as displayed here and reported in the literature, it is interesting that no patient exhibits a loss of the 9p21 region without also showing a loss in one of the other two regions. This may imply that in order for 9p21 to influence the transformation of the cancer into its aggressive state, the genome must also experience a copy number loss event in one of the other two regions. (Bueno, Stawiski et al. 2016; Hylebos, Van Camp et al. 2016; Hylebos, Van Camp et al. 2017; Lindholm, Salmenkivi et al. 2007). There are tumour-suppressor genes in each of these regions, which are also well established, *NF2* in the 22q region (Sekido, Pass et al. 1995), *CDKN2A* in the 9p21 region (Kettunen, Savukoski et al. 2019; Prins, Williamson et al. 1998), and *BAP1* in the 3p21 region(Bott, Brevet et al. 2011), though *SETD2,MTAP* and *CRI1 (EP300)* have also been associated with cancer

(Fahey, Davis 2017; Fukazawa, Matsuoka et al. 2008; Hida, Hamasaki et al. 2017) which are also present in the 3p21 and 9p21 regions respectively. The loss of function in tumour suppression from these genes may be what allow the cancer to begin to proliferate. It should be noted that the reason the top three are grouped like this for the purpose of the discussion of this analysis is not only that they are the three most recurrent clonal copy number events, but also because they are so prevalent in the literature, and so can be seen to be almost standard findings in the genetic analysis of MPM.

The 14q loss event is reported on a near similar scale as the top three most recurrent, but interestingly has less of a presence in the MPM genetics literature, appearing in studies far less commonly than 22q, 3p21 and 9p21, though it is still associated with the cancer in a few studies and so cannot be considered a novel finding (Borczuk, Pei et al. 2016; Lindholm, Salmenkivi et al. 2007). This could be due to the fact that there are no reported genes within the region that have been associated with MPM, and so it is neglected from gene studies looking to find associations with already established cancer genes. Despite this, it has been proposed that both the *HECTD1* and *NFKBIA* genes have regulatory effects in pathways that can lead to tumorigenesis (Bredel, Scholtens et al. 2011; Sarkar, Zohn 2012). The fact that a loss event in 14q is reported as being a clonal event in 8 of the 17 patients by the Sequenza/TuMult pipeline indicates that it is likely to have some kind of importance in regulating the suppression of tumours, and that loss of function in this region can result in the proliferation of the cancer. However, a loss in 14q is never independent from the top three events and actually is only found in patients with at least 2 of the top three also present. This implies that the effect of a heterozygous loss in this

region alone may not be significant enough to result in the transformation of the MPM tumour, and that the genome requires additional events to occur in conjunction with 14q in order for the process to take place.

This effect can also be seen with the rest of the 6 recurrent events, as well as with 9p21, and has some interesting implications into explaining what could be happening prior to the spread of the cancer. A broad explanation could be that whilst the cancer is still in its latency period, and the various cancer cells within the original population are accumulating various mutations, certain specific copy number change events are increasing the ability of a given cancer cell to begin rapid replication and expansion, but that there is always a specific trigger event required, regardless of how many of these other events build up. A good candidate for this specific copy number change event would be the loss of 22q and the *NF2* gene due to its prevalence in most patients and the fact that it is present in 2 patients where neither of the other two most prevalent events are found. That is to say, that the point at which a cancer cell obtains a 22q loss event, it undergoes positive selection in the cancer cell population and proliferates until that phenotype dominates, whilst also causing the tumour to undergo rapid expansion in physical space, culminating in metastasis and spread across the tissue of the pleural membrane. This is not to say that the other recurrent clonal regions determined by TuMult are not undergoing positive selection, on the contrary, they will be and will resulting a larger proportion of the cancer cell population harbouring those events, but they will not be undergoing strong enough positive selection to result in the kind of event seen in MPM, with mass rapid replication of cancer cells over a short period of time. This explanation would mean that -22q would be the ideal region to search for a

candidate gene, such as *NF2*, which could then be targeted in drug development studies to arrest the spread of the cancer in patients with known exposure to asbestos, before the cancer was able to transform.

In most of the patients, the results of the Sequenza/TuMult pipeline would support this hypothesis, with the presence of clonal -22q in all but 4 patients (Patients 12, 16, 24 and 64). However, 2 of these patients do have recurrent clonal regions suggesting a possible alternative path in the proliferation of MPM (Patients 12 and 16 were excluded from this part of the discussion as neither had any recurrent clonal regions). Patient 24 has a copy number loss event in 3p21 that is assigned by TuMult as clonal and also appears in other patients making it recurrent, though it is the only patient which does not also harbour a clonal 22q event alongside 3p21. Patient 24 also has a copy number loss event in 1p36 assigned as clonal, an event shared with Patient 64, and although stated previously, Patient 64 also harbours loss events in 3p21 in every sample, though this was not called as clonal by TuMult due to difficulties with the data quality. Copy number losses in 1p36 have been associated with multiple human cancers and are one of the most well established chromosome aberrations in the field (Bagchi, Mills 2008; Henrich, Schwab et al. 2012; White, Maris et al. 1995).

In terms of tumour suppression coding regions in the short arm of chromosome 1, it has been a constant challenge in the genetic analysis of cancers to locate specific genes which are responsible for the loss of function in tumour suppression when the genomic region has suffered a copy number loss. But rather than there being a shortage of eligible genes, there is actually a surplus,

with tumour suppressors in the region including *CHD5, CAMTA1, CASZ1* and *KIF1B* (Henrich, Schwab et al. 2012). However, due to the often large size of the copy number losses which occur in the region, it is difficult for specific genes to be identified as causative factors in the cancer proliferation. This also applies to both Patient 24 and 64, with the entire 1p36 region being lost in both patients. There have been suggestions that the region is inherently unstable, and so more prone to copy number loss, which results in the whole cytoband often undergoing loss (Bagchi, Mills 2008).

The results shown for Patient 24 indicate that the 22q loss pathway may not be the only one which can cause the cancer to progress, and that an alternative pathway may exist by way of 3p21 loss, possibly supported by a loss of 1p36. Based on the results shown here, if this alternative pathway does exist, it is much rarer than the 22q loss pathway proposed earlier, and also may not be dependent on the involvement of a 1p36 loss. A copy number loss of 1p36 is only present clonally in 4 of the patients, but subclonally it is present in an additional 10 (Patients 1, 12, 18, 27, 33, 34, 75, 78, 85 and 91) implying that it is actually a change that takes place most commonly after the cancer has already diverged, and is more likely to evolve in parallel in the samples (i.e. it is unlikely to have common chromosome breakpoints which would be picked up by TuMult). This is backed up by the notion that the 1p36 region of the genome may be unstable and so is likely to accumulate copy number losses and allow for the loss in function of tumour suppressor genes in the region to provide further positive selection for a given cancer cell population. Heterozygous loss of 1p36 is also present clonally in Patients 33 and 37, but this is in conjunction with losses in 22q, 3p21, 9p21 and 14q in both patients.

The rest of the recurrent clonal regions can now be discussed briefly, as they are present in a smaller proportion of the patients and only in conjunction with regions that have already been discussed. Heterozygous loss in the 6q region is present clonally in 5 of the patients, and is a region which encompasses the tumour suppressor gene *LATS1* which has been associated with MPM in previous studies (Miyanaga, Masuda et al. 2015; Zhang, Dai et al. 2017). The region only had a copy number loss in conjunction with a loss in 22q. A heterozygous loss of the 13q region was also present in 5 patients, though it was only ever lost in conjunction with the 6q region in Patient 78. The region encompasses two previously reported tumour suppressor genes associated with MPM, *LATS2* and *BRCA2* (Betti, Casalone et al. 2017; Hassan, Morrow et al. 2019; Murakami, Mizuno et al. 2011). 13q only had a copy number loss in conjunction with a loss in 3p21. It is unclear whether there is a link between copy number loss in 13q and 3p21, as opposed to a loss in 22q, as the 13q loss is also seen in conjunction with a 22q loss in all but one patient where it is present (Patient 24).

In the 3 patients where there was a clonal loss of chromosome 4, the event incorporated the entire chromosome (as near to this as the TuMult input would allow, listed as a loss from 4p16-4q35) and so caused heterozygous loss to all coding regions. However, the *FBXW7* gene is present of chromosome 4 and has been previously associated with MPM (Kato, Tomson et al. 2016; Yeh, Bellon et al. 2018). The entire loss of one copy of a chromosome is actually quite common in cancers, and is referred to as aneuploidy (Duijf, Benezra 2013; Thomas, Marks et al. 2018). The whole loss of one copy of chromosome 4

actually introduces an interesting limitation with the TuMult software, as regardless of when a total chromosome loss had occurred in different samples, it would always be reported in the truncal region by TuMult. This is because the chromosome breakpoints of a total chromosome loss will always be the same, simply either end of the chromosome. However, it could be argued that it is unlikely that the entire loss of one copy of a chromosome would occur twice, though that is assuming no genetic changes have already occurred which make chromosome loss more likely. Overall, it is a minor limitation that is unlikely to have a major effect on analysis, and the fact that the loss of chromosome 4 was recurrent in 3 patients, when no other chromosomes suffered the loss of an entire copy, suggests that this event may well be clonal in the patients where it is reported.

Heterozygous loss in the 17p region was clonal in 3 of the patients, with the MPM associated gene *TP53* (de Assis, Isoldi 2014; Sementino, Menges et al. 2018) present in this region. Heterozygous loss in the 10q23 region were clonal in only 2 patients, with MPM associated gene *PTEN* (de Assis, Isoldi 2014; Sementino, Menges et al. 2018) present here. Interestingly, these two genes have often been reported to be lost in conjunction with one another, although they are only shared by Patient 91 in this cohort. Looking at subclonal branches, samples 3, 4 and 5 of Patient 85 do have losses in the 17p region, though it is not detected in sample 1 and so is not clonal to the patient, though it is possible that the event is selected for if a 10q23 loss is already present in the genome.

A final point to address on the Sequenza/TuMult results presented here, is the

lack of any recurrent clonal gain events across the patient cohort, which is unusual as copy number gain events have been reported several times throughout the literature (Furukawa, Toyooka et al. 2015; Lindholm, Salmenkivi et al. 2007). This report has already established that losses account for a greater proportion of change than do gains (Figure 3.18) and that although a large number of gain events can be observed on the TuMult trees, they are mostly very small and certainly much smaller than the loss events. An explanation for the lack of any clonally reported gain events is that they simply do not recurrently occur prior to the transformation of an MPM to its aggressive state and only begin to appear frequently once divergence of cancer cell populations has taken place. This assumes that any clonal gain events that are reported on the TuMult trees are simply passenger mutations, i.e. they do not undergo positive selection and do not offer any additional function to promote the replication of the cancer cells. It is also possible that these events are just rare and so were just less likely to be detected in a cohort size of just 17 patients.

Whilst these results do not offer any novel findings in the form of new regions of interest associated with MPM, they are novel in that the phylogenetic analysis of MPM sequencing data has allowed the copy number loss events in these regions to be mapped based on the time that they occurred in the evolution of the cancer.

3.5.2) Use of Branch Length to Determine Anatomical Spread of MPM

The average branch lengths were calculated for each sample in an attempt to try and determine if there was a clear order of divergence which could then be mapped to the physical space of the pleura, using the data presented in Table 3.10. One of the project aims was to see if a pattern could be observed with how MPM spreads once it has entered its aggressive state and began to metastasise across the pleura. Previous observations in the field have hypothesised that the cancer initially spreads upwards, towards the apex of the lung (sample 1 region) before traversing down the lung pleura (Collins, Sundar et al. 2020; Tertemiz, Ozgen Alpaydin et al. 2014). In order for the results to support that hypothesis, it would be expected for the average branch length for sample 1 to be higher than the others, though in the case of the Sequenza/TuMult pipeline, it has a value higher than the average for samples 2 and 3 and lower than the average for samples 4 and 5. It would seem unlikely that the apex of the lung is the first site of metastasis by this metric, but there are a few considerations to take which suggest that this is not the best method for trying to determine this pattern.

Firstly, this method would need to assume that all instances of this cancer would need to spread the same way, with subsequent populations always spreading to the next region after the apex. This is unlikely to be the case and is also heavily influenced by the initial site of cancer development in the pleura, which is also variable. The second assumption would be that all sites are equally as easy for the cancer to metastasise too, which has been reported to be a false assumption (Collins, Sundar et al. 2020) with the pericardium region

(sample 2) being described as a fairly rare occurrence. This could also explain why the sample 2 region is missing from the majority of patients, as often the patient will have passed before the pericardium was metastasised. However, the biggest concern with this is also one of TuMults biggest limitations, and that is the creation of artificial events in order for the tree topology to exist.

The phenomenon of creating artificial events is quite complicated though it does happen regularly in TuMult trees in order to support the topology which has been determined by the neighbour-joining method that operates at the core of the TuMult software. When constructing the trees, TuMult has to decide which patients will be grouped together based on the shared chromosome breakpoints between them by using the copy number profiles it is provided to determine regions of consensus between samples. It does this incrementally, starting with the pair of samples which have the lowest distance between them and then adding on further samples until there are none left to add. At the end of this process, the topology of the tree will be determined, with the sample nodes arranged at the end of each branch and linked to the closest other sample in terms of shared chromosome breakpoints. However, in some cases, two of the samples (or groups) it has linked will have contradicting events despite having the most common chromosome breakpoints at that point in the process. After the grouping stage is done, TuMult then populates the tree with the events which resulted in the software arranging the samples in that particular topology. In order to resolve any contradictions, TuMult will then add an artificial event on a shared branch. This is a confusing concept to grasp, though it is essentially the software resolving the tree in the best way it can via the incremental process. It is not a hugely common occurrence across the cohort reported on in

this thesis, however, every patient has at least one of these events, with Patient 12 having a substantial amount in particular due to the uncertainty caused by the subclonal genome doubling.

This means that using the number of events listed by TuMult, as determined by common breakpoints in the input profile, will not always be entirely accurate. As such, TuMult is not an ideal software when trying to determine the pattern of spreading a cancer has undergone. It is actually unlikely that any phylogenetics software would be able to resolve this due to the issues described previously. Whereas TuMult assumes a constant rate of evolution in the cancer cell populations, this is not the case for all methods, and the possibility of accelerated mutation as the cancer evolves adds an additional confounding variable to the problem.

Even though TuMult does have limitations such as this, it is still a demonstrably good method when considering what this project is aiming to do. In terms of the main goal of identifying recurrent clonal regions which may harbour druggable targets, TuMults use of chromosome breakpoints to determine the origin point of copy number events is a very robust tool.

3.5.3) Copy Number Based Phylogenetic Pipeline Final Comments

The aims of this chapter were to identify recurrent clonal regions via copy number based phylogenetic methods in order to identify possible driver regions and subsequently, actionable genes which they may contain. This was to be done whilst critically evaluating the software used to ensure that the final output was of sufficient quality for this level of research. Furthermore, an additional goal was to attempt to use the branch lengths provided by phylogenetic trees produced as part of the analysis, to try and determine whether anatomical spread of MPM over the lung could be tracked via phylogenetic methods.

The two initial aims were achieved via the use of the Sequenza/TuMult pipeline, which identified multiple possible regions that contain known MPM-associated genes, as truncal regions. Unfortunately, the last aim was not met, though the understanding of the limitations of the methodology in that process can allow for more robust analysis in future attempts. The critical appraisal of why the aim could not be met is sufficient in regards to this project.

4) Single Nucleotide Based Pipeline and Results

4.1) Introduction to Single Nucleotide Pipeline

Chapter 4 contains the results and discussion for the single nucleotide based pipeline which was described in Section 2.4 on the thesis. The VarScan2 results are displayed first followed by results from PhyloWGS and a discussion of the findings.

The importance of CNV events in MPM has been mentioned throughout the thesis, with a vast amount of literature focused on deciphering these events. However, the potential of SNVs to affect the outcome and progression of MPM needs to be thoroughly explored alongside the CNVs, and largely has been as discussed in Chapter 1. This chapter aims to; identify recurrent clonal SNV events in the patient cohort and evaluate their possible impact; evaluate the structure of the pipeline used; and cross-reference any SNV-related findings to see if they affect CNV events reported in the previous chapter.

4.2) VarScan2 Results

Unlike Sequenza, ABSOLUTE and ASCAT, VarScan2 does not generate any visual output and so the only figure to be discussed in this section is a table displaying the total amount of variants detected for each sample (Table 4.1).

	Number of Variants called by VarScan2 after quality control					
Patient Number	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average per Patient
1	83	105	97	92	93	94
6	69	####	77	66	75	57.4
12	92	84	64	86	95	84.2
16	12	####	15	47	99	34.6
18	94	####	79	78	83	66.8
23	68	61	57	68	75	65.8
24	88	85	78	74	68	78.6
27	139	110	136	123	122	126
33	114	####	109	121	88	86.4
34	110	115	120	111	121	115.4
37	75	####	81	82	73	62.2
64	56	####	70	65	62	50.6
75	103	####	84	106	82	75
78	63	####	76	74	70	56.6
84	61	####	58	59	56	46.8
85	83	####	104	94	89	74
91	160	####	138	125	133	111.2
Average per Sample	86.47	93.33	84.88	86.53	87.29	

Table 4.1: A table displaying the total amount of variants called by VarScan2 for each patient and sample, after quality control. The average values for each sample and each patient are also displayed. The #### values in the sample 2 column represent lack of data due to unavailability in those specific patients, as discussed in Methods.

Table 4.1 is a table displaying the amount of variants in each sample for each patient that were called by VarScan2, after the quality control steps in the previous section had been performed. It also includes a mean average amount of variants called per patient and a mean average amount of variants called in each sample region across all patients. Overall, an immediate point of consideration is that the total number of variants detected in all samples is quite low, and this is not simply a result of the quality control steps, with each sample retaining approximately 30% of variant calls. The range of total variants in a

sample ranges from a low of 12 variants in Patient 16 sample 1 and a high of 160 in Patient 91 sample 1. The range of mean average variants called per patient ranges from a low of 34.6 in Patient 16 and a high of 126 in Patient 27.

4.3) PhyloWGS Results

As discussed in the Methods section, visual output in the form of trees could not be obtained from PhyloWGS due to the immense amount of computational power that would be required to display them. However, due to the workaround employed, results were still able to be extracted and are displayed here across 2 tables.

Patient Number	Number of clonal SNVs
1	3
6	2
12	32
16	4
18	8
23	14
24	3
27	2
33	42
34	2
37	19
64	3
75	6
78	1
84	1
85	1
91	68

Table 4.2: A table displaying the number of clonal SNVs reported by PhyloWGS in each of the patients.

Table 4.2 displays the total number of clonal SNVs reported by PhyloWGS in each of the patients, with the overall number for each being quite low. This is not an unexpected result when considering that the total number called by VarScan2 for each patient was also low.

Patient Number	Cytoband	Gene	Genomic Position	Genotype	Synonymous = 0 Non-Synonymous = 1
18	22q	NF2	30067929	G → T	1
33	22q	NF2	30035172	G → T	1
75	22q	NF2	30050645	G → A	1
27	22q	NF2	30051598	C → T	1
27	9p21	CDKN2A	21974746	C → G	0
84	3p21	BAP1	52443760	C → T	1
91	3p21	SETD2	47162471	G → A	1
6	4	FBXW7	153249543	T → A	0

Table 4.3: Clonal SNVs as reported by PhyloWGS, present in suspected driver regions. Synoymous refers to a single-base change in the DNA that does not result in a different amino acid being produced, indicating that the change is very unlikely to have any causative effect. Non-synonymous is the opposite, where the base change has caused a change in amino acid resulting in a change in the protein structure.

Table 4.3 displays the SNVs that were classified as clonal through PhyloWGS whilst also being present in suspected driver genes identified via the Sequenza/TuMult pipeline, with the patient they were found in, the cytoband and gene in which the SNV is located, the specific genomic position that the nucleotide change has occurred, and whether the change is synonymous or non-synonymous. The VarScan2/PhyloWGS pipeline was built in order to produce findings such as those displayed in the table, as they present a route by which SNVs found in MPM cancer tissue cell populations can influence the pathophysiology of the cancer, and possibly confer positive selection in cells where they occur (dN/ds analysis cannot be performed on such a small amount of variants, and so positive selection cannot be validated but a method by which it could occur can be remarked upon). Though not recurrent when considering exact genomic position, the SNVs found in Patients 18, 33, 75 and 27 within the

22q region are recurrent in that they all occur within the *NF2* gene, and also all result in non-synonymous mutation in the gene. Synonymous and non-synonymous mutations refer to the codon code, in which several combinations of nucleotide triplets all produce the same amino acid in a protein. If a mutation is synonymous, that means the nucleotide change which has happened, has not caused the amino acid in the protein to change, and so should not effect the function of the protein. Non-synonymous mutations refer to a nucleotide change which has caused the amino acid to change (as with the 4 *NF2* SNVs shown in Table 4.3), and so may confer a functional change in the produced protein. There is also the possibility that a non-synonymous mutation may occur to produce a stop codon, which would then result in a truncated protein, or disrupt a stop codon, resulting in a long amino acid chain unable to fold into the correct shape to retain function of the protein.

4.4) Discussion of Single Nucleotide Pipeline Results

4.4.1) Discussion of VarScan2 Findings

The low number of variants could be due to the nature of the cancer, in that it undergoes large amounts of copy number loss and so also has a high chance of losing variants when these loss events occur. It could also simply be due to the exome data only representing regions of coding DNA, so variants outside of these regions are not counted, though it should be noted that this is not an issue in terms of the analysis as SNVs in these regions, though not detectable, are also extremely unlikely to play a role in the pathophysiology of MPM. Another explanation is that the the vast majority of SNVs don't play any role in the proliferation of the cancer cells, and so offer no positive (or negative) selection pressure and instead simply act as passenger mutations. As such, throughout the early cancer cell population, a huge variety of SNVs would be found across the individual cancer cells in the population, as none would be driving the cells to replicate and thus increase their presence in said population. That being said of course, it would still be expected for these mutations to be retained if a copy number event (or other event) allowed for a cancer cell to be positively selected for, as with the case with copy number events that also confer no positive selection value (which is why having recurrent events is so important when trying to identify regions of interest in relation to driver genes). The key difference is, that copy number events, especially on a large scale, are very difficult to reverse compared to SNV events, and the latency time of MPM is very long, giving a long period of time where these SNV events could be lost without effecting the proliferation of the cancer cell population and causing it to be diluted in the population.

This may suggest that SNV events do only have low impact, or no impact, on the proliferation of MPM and its transformation from the latent to aggressive stage, though it is not possible to confirm this without analysing the data set to find SNVs that were present prior to the divergence of the cancer in physical space, and then attempting to find recurrent SNVs which may hint at some role played in MPM. The use of phylogenetics is perfect for this.

There is no correlation between average ploidy or cellularity, as both reported by Sequenza and ABSOLUTE, to the average amount of variants detected in each patient, though Patient 16 and 64 both display the lowest amount of variants as well as the lowest average cellularity. This is possibly due to lower quality data in some of the samples of these patients, with samples 1 and 3 in Patient 16 showing particularly low numbers of variants (12 and 15 respectively).

In terms of differences between sample regions, the average number of variants per region tends to be quite similar, with a range of 81.88 being the lowest for region 3 and an upper limit of 93.33 for region 2. The reason for the increased average in region 2 is likely due to an increased average number of variants called in patients with 5 samples, with 3 of the top 5 average values being in 5 sample patients (Patients 27, 34 and 1).

4.4.2) Discussion of PhyloWGS Findings

The range of the number of clonal SNVs that is detected in the cohort by PhyloWGS is unusually large, as can be seen in Table 4.2, with 3 patients hosting only a single clonal SNV, up to 68 clonal events hosted by Patient 91. The presence of such a high amount of clonal SNVs in Patient 91 does seem unusual when compared to the rest of the cohort, as most of the patients have 3 or less clonal SNVs. There could be a few reasons for this, but the first thing to consider is that Patient 91 was reported as having the lowest ploidy by Sequenza after analysis in the Sequenza/TuMult pipeline. This may have resulted in large numbers of SNVs to be grouped together based on a common dip in the read depth, which also may have resulted in the ploidy estimates of the patient samples to be lower, though the cause of such an event is a mystery. It is also interesting to note that Patient 91 has clonal SNVs present in every chromosome, again indicating that this phenomenon has been caused by some kind of whole genome occurrence. Of the other patients with a high number of SNVs called clonal, Patient 12 has obvious ploidy abnormalities resulting from the genome doubling event in samples 1 and 4, though this does not extend to all samples and so should not be resulting in a clonal effect. Patient 33 did not have a lower than average ploidy estimate or cellularity estimate, though Patient 37 was second lowest for ploidy. In terms of biological explanation, it could be that these patients had the initial aggressive spread of MPM more recently, and so more SNVs that were present in the initial population are still harboured by the new population which diverged from it. Patients 37 and 33 both have lower than average exome coverage which also hints towards this possibility, though Patients 91 and 12 do not.

In terms of recurrent clonal SNVs across the cohort, there were zero found that were present at the same nucleotide position, implying that, at least in this 17 patient cohort, there is no individual SNV which contributes to MPM as a driver. However, the case could also be that there is an SNV mutation in MPM which can confer positive selection to the cancer cell population, but that it is rare and so undetectable in a cohort of only 17 patients. However, among the clonal SNVs there are cases where the SNV lies within candidate driver genes proposed in Chapter 3, as can be seen in Table 4.3.

What is particularly interesting with the SNVs reported in Table 4.3 is that they all occur in patients where the gene region in question has already undergone a heterozygous loss in copy number, meaning an entire copy of the gene region has already been lost. That means, at least in the cases of non-synonymous mutation seen in the table, for the particular loci where the SNVs are located, the gene is homozygously hit, possibly resulting in a complete loss of function. This concept is paramount when considering the possible effects of SNVs when they occur alongside copy number changes, as the complete loss of function of a gene will produce a stronger effect in terms of selection when the gene itself operates as a tumour suppressor. That being said, it does not necessarily double the effect, as the loss of one copy of a gene can be enough to reduce overall functionality to almost nothing. For example, in the case of regulation, there are often threshold amounts of protein which need to be reached in order for a biological pathway to proceed to a further step. If one copy of a regulatory gene is lost, the ability to reach that threshold may be greatly impacted, and make it practically impossible for it to be reached, meaning the heterozygous loss essential acts as a homozygous loss in terms of selective pressure (when

considering cancer cells). It is unlikely that these SNVs are themselves acting as drivers however, as if this were the case, it would be expected to see a higher frequency of these mutations in MPM cancer tissues. Further to this, in this cohort these SNVs have only been seen to be clonal when in a gene region that has already undergone copy number loss. Were the SNVs acting as drivers, it would be expected to see them clonally in copy number normal regions, as in that scenario, a non-synonymous nucleotide mutation would be equal to a heterozygous copy number loss in that it prevents the full expression of two gene copies.

That these SNVs have occurred in the 22q and 3p21 regions, in genes already associated with MPM (*NF2, BAP1* and *SETD2*), as well as having been identified as clonal in the Sequenza/TuMult pipeline is strong evidence that the genes play a role in the development of the cancer and act as drivers in the characteristic accelerated replication and physical spread of MPM. It also validates the ability of the PhyloWGS method to determine clonal SNVs, though it should be noted that the false-positive rate in the findings is not established. The clonal SNVs found in Patient 27 in the region of 9p21 and within the *CDKN2A* gene, and in Patient 6 within the *FBXW7* gene are non-synonymous events, meaning they do not alter the amino acid code in either of the genes, and so therefore are unlikely to be conferring any sort of positive selection for the cancer cell populations where they are present. Even though no deleterious effect is being caused by either SNV, they are still useful for this analysis, as they act as a secondary marker to infer clonality on the regions where they are present. For example, in Patient 27, the *CDKN2A* gene has shown to have undergone a copy number loss due to an alteration event during the history of

the cancer, and this event has also been shown to have occurred clonally by the Sequenza/TuMult pipeline, before the cancer cell populations diverged and physically spread across the pleural membrane. The presence of the SNV in this region, which has also been called as clonal, establishes further evidence that this region was altered prior to cancer divergence. The logic to this is that, just as it is highly unlikely that a copy number change event would occur twice at the same breakpoints through parallel evolution in separate cancer cell populations, it is also highly unlikely that the same nucleotide base would be mutated in the same way in. As such, it acts as a biomarker of clonality. This same logic can be applied to all of the SNV events reported to happen clonally that are present within a clonal copy number change event called by the previous pipeline, including those that are non-synonymous.

4.4.3) Single Nucleotide Based Pipeline Final Comments

In conclusion, this pipeline failed to identify any novel clonal SNV events that may act as potential drivers, though this could be considered the expected result based on the lack of single nucleotide mutations identified in the literature (except those which are thought to confer susceptibility). As well as previous research providing strong evidence that MPM is a cancer mainly driven by copy number alterations (and losses in particular), which, due to their large scale are more likely to cause deleterious effects in normal cells and drive the proliferation of cancer through widespread loss of function. However, the identification of clonal SNV events within known associated MPM genes (*NF2, BAP1, SETD2*) causing non-synonymous mutations, alongside copy number loss, is a novel finding not previously documented in the literature. The use of phylogenetics to identify these homozygous loci, calculated to have occurred prior to the major divergence of cancer populations reinforces the evidence that these genes are important drivers in the proliferation of MPM. Furthermore, it provides evidence that these SNVs could themselves be conferring positive selection on the cancer cells where they are present, though with less selective power than a copy number loss.

5) Revolver Analysis

5.1) Introduction to the Revolver Analysis

Chapter 5 covers the results and discussion for the Revolver Analysis, performed as described in Section 2.5 of Chapter 2. It is a short chapter with only a single result and some discussion, though it establishes one of the most important findings of the thesis.

Revolver was chosen as it was a freshly released software which suggested that it could calculate trajectories of recurrent events in phylogenies, which was an attractive prospect when considering the structure of the project up to this point. This chapter displays the single result obtained from Revolver, which is then followed by a short discussion on the impact of that result. The sole aim of this chapter is to try and establish trajectories for recurrent CNV events that were reported by the copy number pipeline in the Chapter 3, and evaluate the importance of any findings.

5.2) Revolver Results

As binary values were used rather than CCF values (explained in Methods), the amount of plots which can be generated using Revolver were more limited. One of the plots which could be generated was the 'drivers occurrence' plot, though it was essential identical in the data that it displayed to Table 3.12, and so was excluded from this report. One plot that was generated and included can be seen in Figure 5.1, and is the drivers graph displaying the driver trajectories across the entire patient cohort.

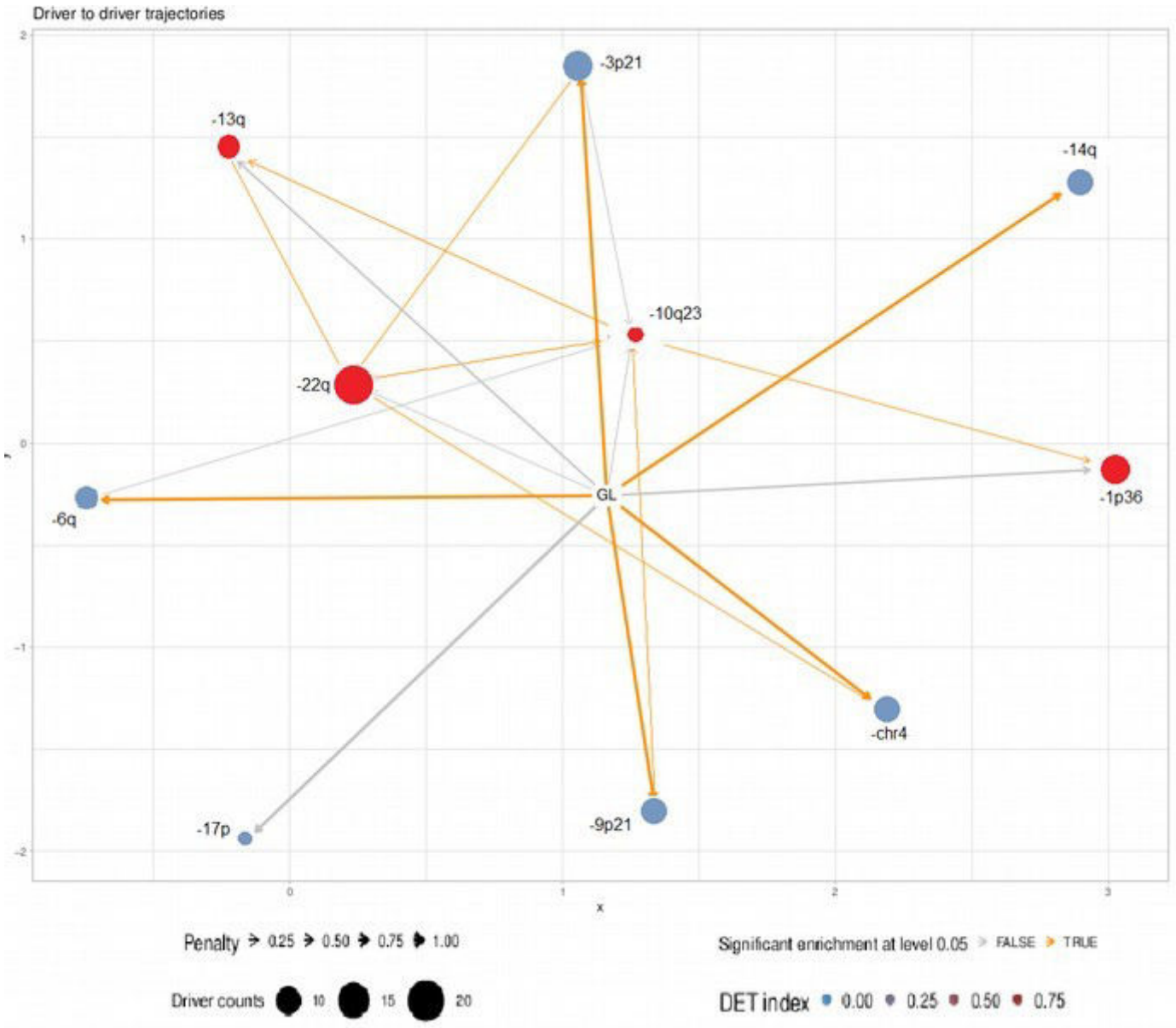


Figure 5.1: The drivers graph produced for the 17 patient cohort using Revolver. The graph displays all drivers that were part of the Revolver input and the evolutionary distance calculated between them (x and y access). The germline is represented in the middle of the graph. Yellow arrows indicate a direction of trajectory that is significant (p-value < 0.05). The colour of each node represents whether it had multiple or single trajectories going to it, calculated via the DET index. The size of the arrow head indicates the penalty that was applied to the driver during the fit command process. The size of the node indicates how many of that driver were present in the patient cohort.

The drivers graph is a visualisation of the trajectories calculated by Revolver during the revolver_fit step of the work flow. In order to fully appreciate what the graph is displaying, each part of the figure needs to be explained in greater detail. Firstly, the GL displayed in the middle part of the figure represents the germline tissue, which in the patient cohort here, all drivers have to derive from.

This can be seen as there is an arrow to every node from the germline here. This is due to drivers being defined as events which have appeared in the truncal region of TuMult trees more than once across all patients in the cohort, meaning that all drivers have been directly derived from the germline at least twice.

The next thing to be aware of is the x and y axis, which display the evolutionary distance calculated between the drivers. The physical space on the graph represents evolutionary distance between different nodes, this can best be seen when comparing nodes to the germline. For example, the 10q23 loss event node is the closest to the germline on the graph, which is representative of 10q23 loss appearing most frequently in the truncal region compared to elsewhere in the TuMult trees. It should be noted that this is only in relation to itself, as it is in fact the least abundant driver, only appearing twice, however, both appearances are in the truncal region, whereas other drivers appear more frequently in subclonal regions.

The direction of an arrow on the graph indicates the direction of the trajectory, meaning that Revolver has calculated that the order in which the events occur in the cancer cell population follow the direction of the arrows. Whether an arrow is yellow or grey indicates the statistical significance, with arrows that are yellow showing enrichment at p-values < 0.05. The thickness of an arrow indicates the penalty that was applied to the likelihood during the Revolver algorithm, and is directly reflective of how many individual trajectories each node has. The penalty starts with a value of 1 and shrinks every time a individual trajectory between two nodes is established during the computational

step of Revolver, meaning that the penalty of all incoming arrows to a node must be equal to 1 (i.e. a node with two incoming arrows will have a penalty of 0.5 on each one).

Lastly, the nodes themselves represent each driver reported in the input file, with the size of the node corresponding to the total number of that driver event in the cohort, including subclonally. The colour of the node represents the value of the DET index (Divergent Evolutionary Trajectories), which is a measure of heterogeneity. Essentially, if the DET index value is above 0, then there is a heterogeneous trajectory path to the node. On this graph, blue nodes are ones which only have a single trajectory, whereas red nodes have multiple incoming trajectories.

5.3) Revolver Discussion

In terms of discussing the results displayed on the drivers graph, only significant trajectories will be discussed. However, one interesting point in relation to this is that the 17p loss event, which can be seen in the lower left, only has a single trajectory, which is directly from the germline, and yet it is not significantly enriched. In a practical sense, this implies that the 17p loss event is not a true driver as far as the Revolver algorithm can determine. This is unusual, as it appears in the truncal region of three different patients in the Sequenza/TuMult pipeline, and does harbour the *TP53* gene, which has been historically associated to MPM and is part of the same biological pathway as *NF2* .

Loss events in 9p21, 3p21, 6q, 14q and chr4 are all linked directly from the germline with a single trajectory. This finding does not imply that these events are not drivers, just that there is less evidence for them to be drivers compared to alternative events in the dataset. The frequency of 9p21 and 3p21 alone indicate that they are both important loss events in the positive selection of cancer cells. However, it could be the case that neither event results in the progression of MPM into its hyper-aggressive state. Clear trajectories can be seen going into loss events of 10q23, 1p36 and 13q, though of the two events going into both 1p36 and 13q, only one is statistically significant. The trajectories calculated for these regions displayed by Revolver here implies these events occur later on in the development of the cancer, possibly showing that they have potential to be drivers in the progression. What is particularly interesting about the 1p36 event, is that it does not show any outbound trajectories. This implies that 1p36 may be an alternative driver, either that it operates via a different pathway or that it is present as a major driver in only a

subset of patients.

The key event on the graph is clearly the loss event which occurs in 22q, with 4 alternative and significant trajectories leading to it. This provides strong evidence that 22q loss is a major driver event in the progression of MPM, showing good concordance with the previous pipelines utilised during this project. Interestingly, the lack of a significant trajectory implies that 22q loss alone is not always enough to cause rapid progression of the cancer, and instead, what is being seen is that 22q loss alongside another driver is required for this to happen. This implies that it may not always be a 22q loss which is the last event to occur, but that it is 22q loss paired with another event (though 22q is still the catalyst for the change in cancer state).

In conclusion, although this analysis was shorter than the other two, it still provided new evidence that a 22q loss event is a key driver in the proliferation of MPM. This type of methodology has never been used before with MPM data, making the identification of 22q loss as a driver a novel finding in the cancer research field.

6) General Discussion, Conclusions and Further Work

The identification of the 22q region, and more specifically, the *NF2* gene as a driver event in the progression of malignant pleural mesothelioma was established during this report, via three different phylogenetic based pipelines incorporating both copy number and single nucleotide somatic mutation calls. As discussed in Chapter 1, *NF2* has long been established as an important gene in the development of MPM, with evidence of its association with the disease presented in a great deal of literature. As such, identification of *NF2* alone is not sufficient to provide a novel and useful contribution to the field. However, it is the methodology by which *NF2* was identified which is the true strength to the body of work in this thesis.

Many studies which have identified *NF2* have simply examined the prevalence of mutations in the gene in order to establish it as an important factor (Borczuk, Pei et al. 2016; Sato, Sekido 2018; Sekido, Pass et al. 1995). Whilst this does provide evidence towards association with MPM, it is not enough to label *NF2* as one of the main driver events in the disease. The prevalence of mutations detected could be due to positive selection of *NF2* in the later stages of malignancy, and mean that the gene itself is not what is responsible for the initial proliferation of the cancer from a latent state. The reason why this distinction is important is due to ideal drug design.

Developing a drug that can successfully target an MPM associated gene (or a

gene in a linked pathway) which is present in high numbers long after the initial boost may help to treat symptoms or reduce the rate of proliferation, but it does not address the root problem. If a drug can be developed which can inhibit the key driver event, then recurrence of the cancer in a patient post treatment via chemotherapy or surgery for example, may be reduced or even completely stopped. It should be noted that inhibiting the key driver event in the case of *NF2* in MPM will not target *NF2* or merlin, but rather another target with connected function. Whilst this seems idealist, it is a core reason as to why phylogenetic inference is far superior in identifying driver events in MPM (and other cancers) when considering druggable targets.

At least two previous studies have used the prevalence of *NF2* to identify protein targets in order to induce synthetic lethality in cancer cells where merlin is deficient. One targeting the focal adhesion kinase (FAK) protein (Shapiro, Irina M et al. 2014) and one targeting the mammalian target of rapamycin (mTOR) protein (Rodrik-Outmezguine, Vanessa et al. 2016). Synthetic lethality in this case refers to targeting another protein, which alongside the lack of merlin, will result in the death of any cells where neither protein is present. As *NF2* is a tumour-suppressor gene characterised by loss-of-function mutations in MPM, it cannot be directly targeted itself, as no protein is being expressed. FAK was selected as a target because the researchers had discovered that merlin depleted cells were more sensitive to FAK-inhibition and realised that the cells were more dependent on FAK for cell-to-cell adhesion. The decisions to target the specific proteins in these studies was done due to the recurrence of *NF2* deletion (or LOF). Incorporation of phylogenetic inference as in this body of work could result in increased priority for searching for appropriate drug targets

linked to *NF2*, due to its nature as a driver event.

A further issue in targeting specific gene pathways without identifying driver events, is that of drug resistance. Phylogenetic inference ensures that you are targeting the pathway of the key gene, or genes, involved in a given cancer. Targeting non-drivers which occur in greater number once proliferation has already started to occur, will lead to selection pressure in the cancer cell population, possibly making further treatment more difficult or even impossible. In regards to the FAK-inhibition study, clinical trials were subsequently approved but stopped shortly after due to unspecified reasons. This could have been due to drug resistance or possibly even to *NF2*-pathway resistance, i.e. the loss of merlin may be making it harder for the drugs to be effectively delivered. This idea is also suggested in a review paper for merlin associated druggable targets (Sato, Tatsuhiro 2018).

An additional reason as to why a phylogenetic inference study to identify driver events is superior to a simple prevalence study, is to allow identified clonal events to be used as markers for relapse in patients after treatment. It is common practice for patients to receive significant after-care in treatment of MPM and if regular screening of blood samples around the site of the tumour were taken, the expression levels of proteins in the blood could be compared to matched normals to see if they were reflective of possible driver events taking place. In the case of this study this would be seen as a decrease in expression of *NF2*, i.e. there would be less merlin and *NF2* RNA in the sample than expected.

In terms of comparing the body of work in the thesis to another phylogenetic study analysing MPM, as of the time of writing only one exists, and is the Zhang paper discussed previously. The thesis work and the Zhang paper use overlapping patient cohorts and employ similar methodologies, as well as similar results, though the Zhang paper reports higher frequency of the clonal events across the samples. If critically comparing the methodologies employed, it could be argued that the phylogenetic inference performed in the thesis is probably more accurate than the one in the Zhang paper, though the overall methodology of the Zhang paper was superior. This is due to the software choices, in particular the use of ASCAT as the copy number estimation software in the Zhang paper compared to Sequenza in the thesis. The control of cellularity and ploidy values used by ASCAT causes a bias in the copy number reported, and will report copy numbers as closer in value to one another than will Sequenza. When this is paired with a phylogenetic inference software, it will then lead to a greater number of clonal events being called. This phenomenon can be seen in Chapter 3.4.4, indicating that ASCAT will produce a higher rate of false positive calls when compared to Sequenza, specifically when producing copy number calls to be used for phylogenetic inference. However, as stated, the quality control steps and post-phylogenetic analysis in the Zhang paper was far more extensive than in the thesis.

There are several limitations to the thesis which will be discussed now. The most minor limitation is the small patient cohort, only 17 patients were left after quality control steps were taken, covering 74 samples. Due to MPM being a rare cancer it's difficult to obtain greater patient cohort sizes, though there are several online databases which could be utilised for testing purposes. The small

cohort size may have led to inaccuracies in the final results due to selection bias.

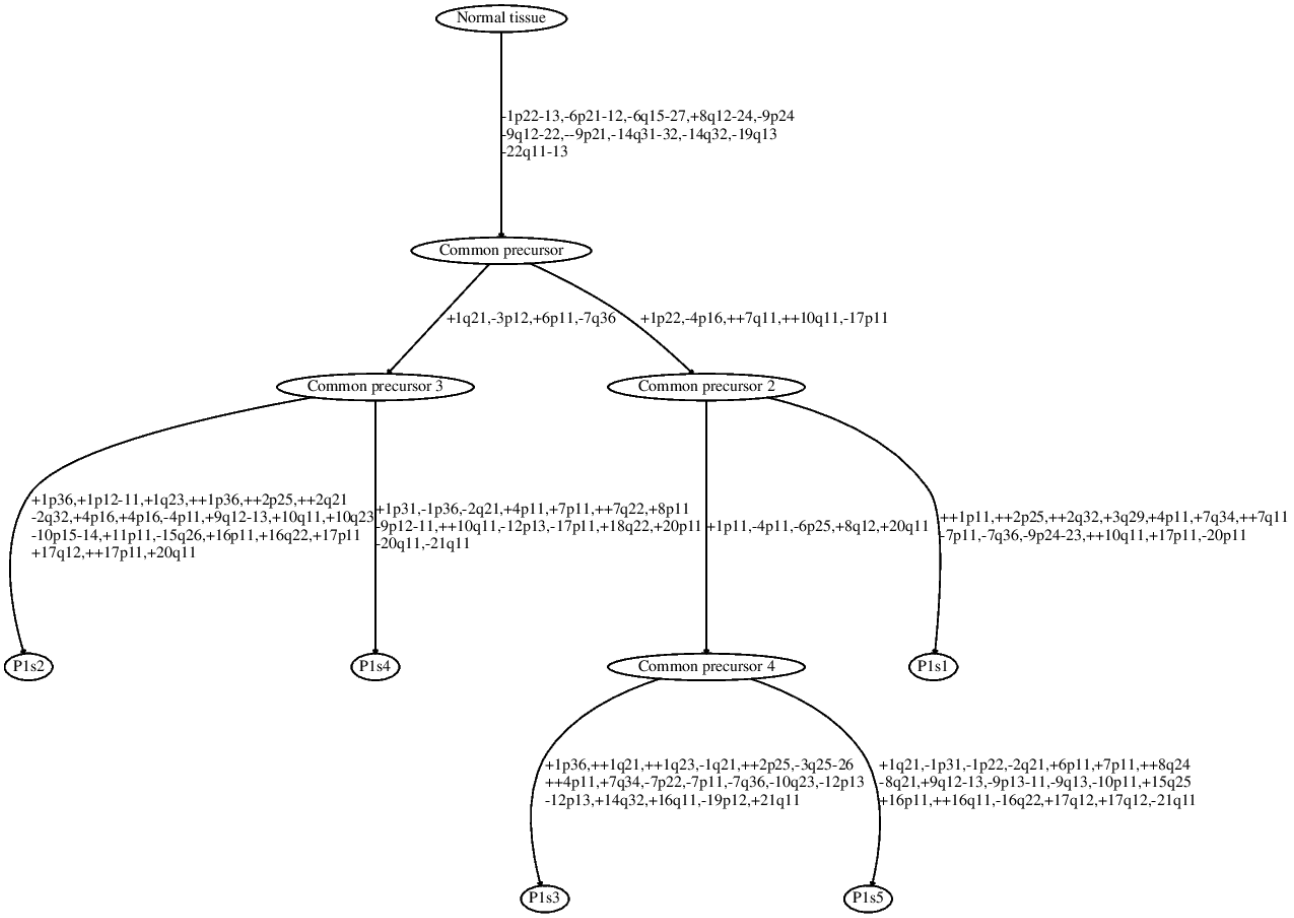
The lack of greater statistical analysis is a major limitation in the project, especially in the case of the PhyloWGS and Revolver pipelines. The range of additional results which could have been obtained, especially from Revolver, was great, and the pipeline should be modified in future to ensure CCF values are calculated so that full Revolver analysis can be performed.

The biggest limitation in the pipeline is probably the required computational resources it needs to complete a full throughput. Sequenza and PhyloWGS in particular are computationally heavy and take a long time to produce output, resulting in a robust but slow pipeline. Further to this, the PhyloWGS tree output was inaccessible due to lack of computational power, though this could also be attributed to the strange design choice by the creators of the software to display the tree output in such a way.

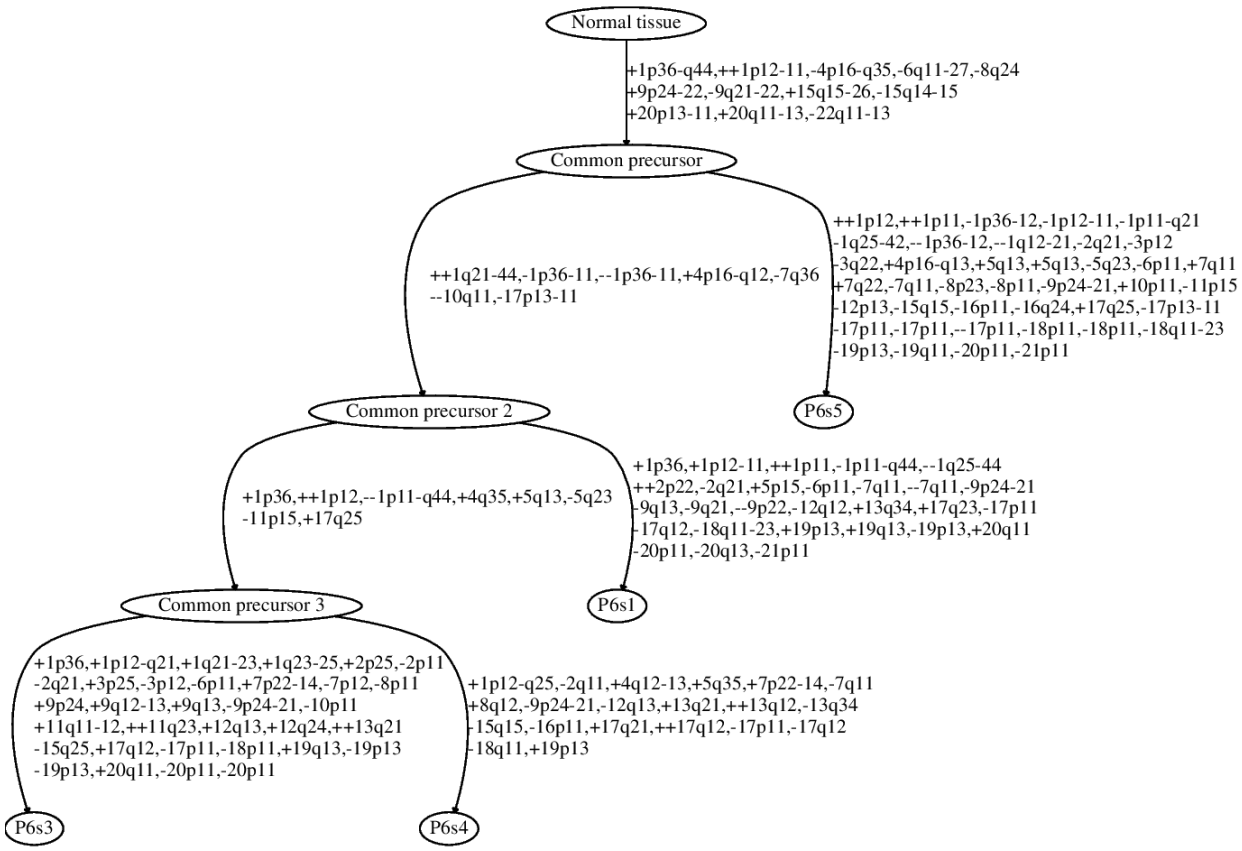
Further work that could be undertaken as part of this project would be incorporation of several parallel softwares at each step of the pipeline in a variety of combinations, so that the final results from each could be compared to form a combined results output. It would be particularly interesting to put such a result set into Revolver to see what trajectories in would present. Furthermore, in order to test the robustness of the pipeline, datasets from other cancer types could be used to see if it is able to handle them in much the same way as MPM data. An eventual final future objective of the project would be the specification of a druggable target in MPM, based on the results from a greater cohort.

In conclusion, the overall aims of the project were each fulfilled by the body of work in the thesis. Potential driver events were identified in *NF2, BAP1 and CDKN2A*, using a combination of both copy number and single-nucleotide derived data. The pipeline used several softwares at different stages to establish consensus results and perform quality control, and post-phylogenetic inference was used to generate trajectories to establish a “true driver” event in *NF2*.

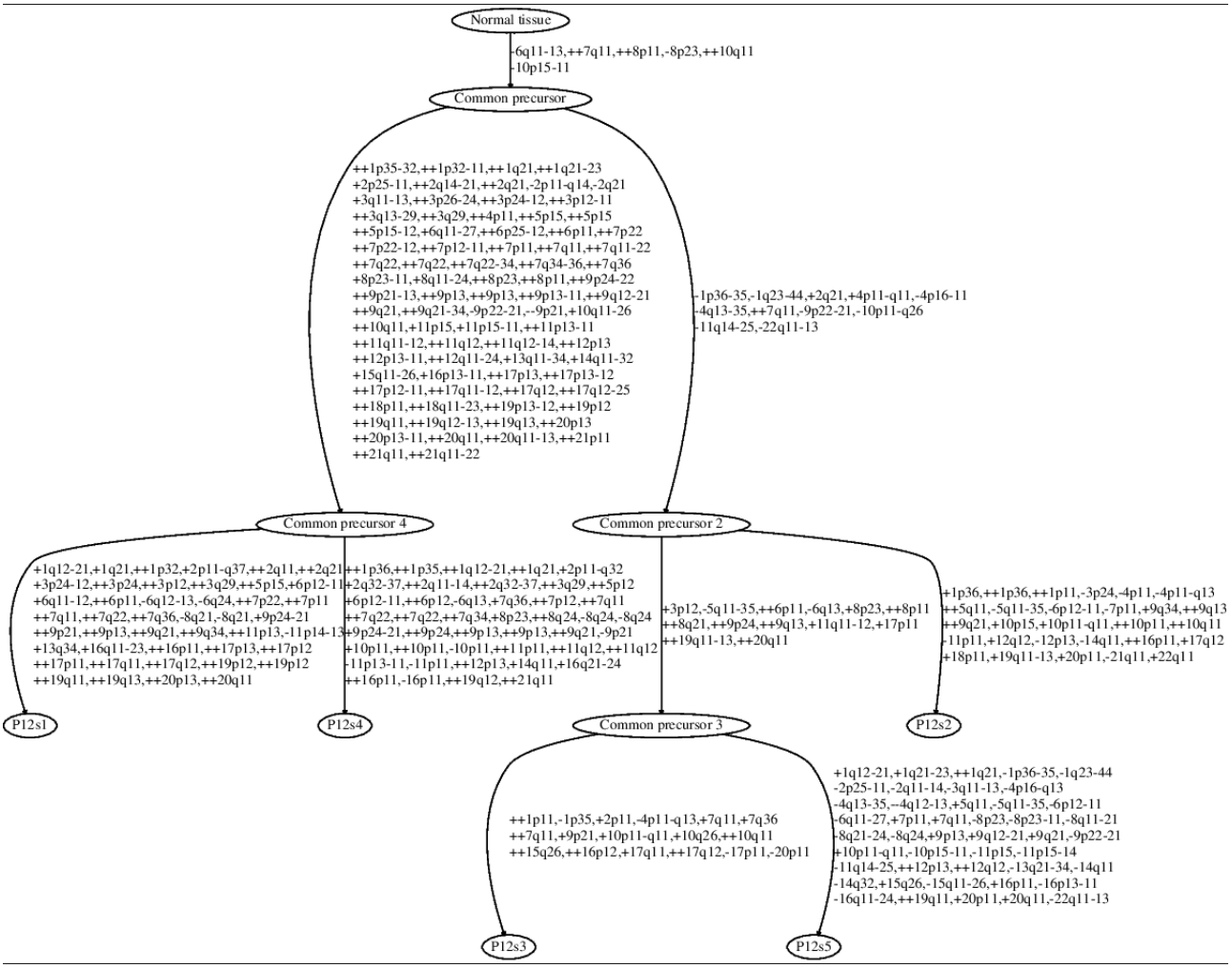
Appendix A - Sequenza/TuMult Cytobands Trees



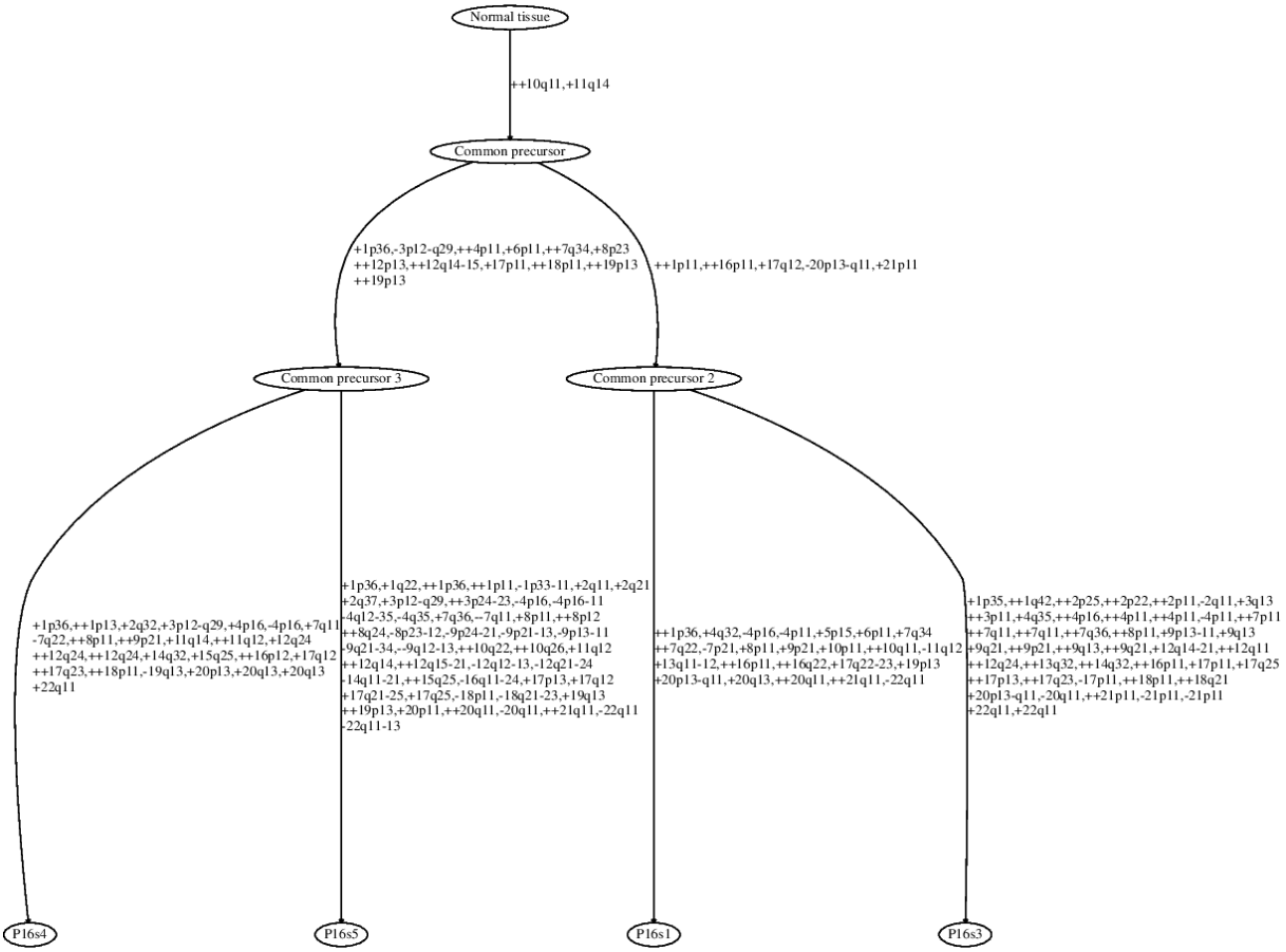
Sequenza/TuMult cytobands tree Patient 1



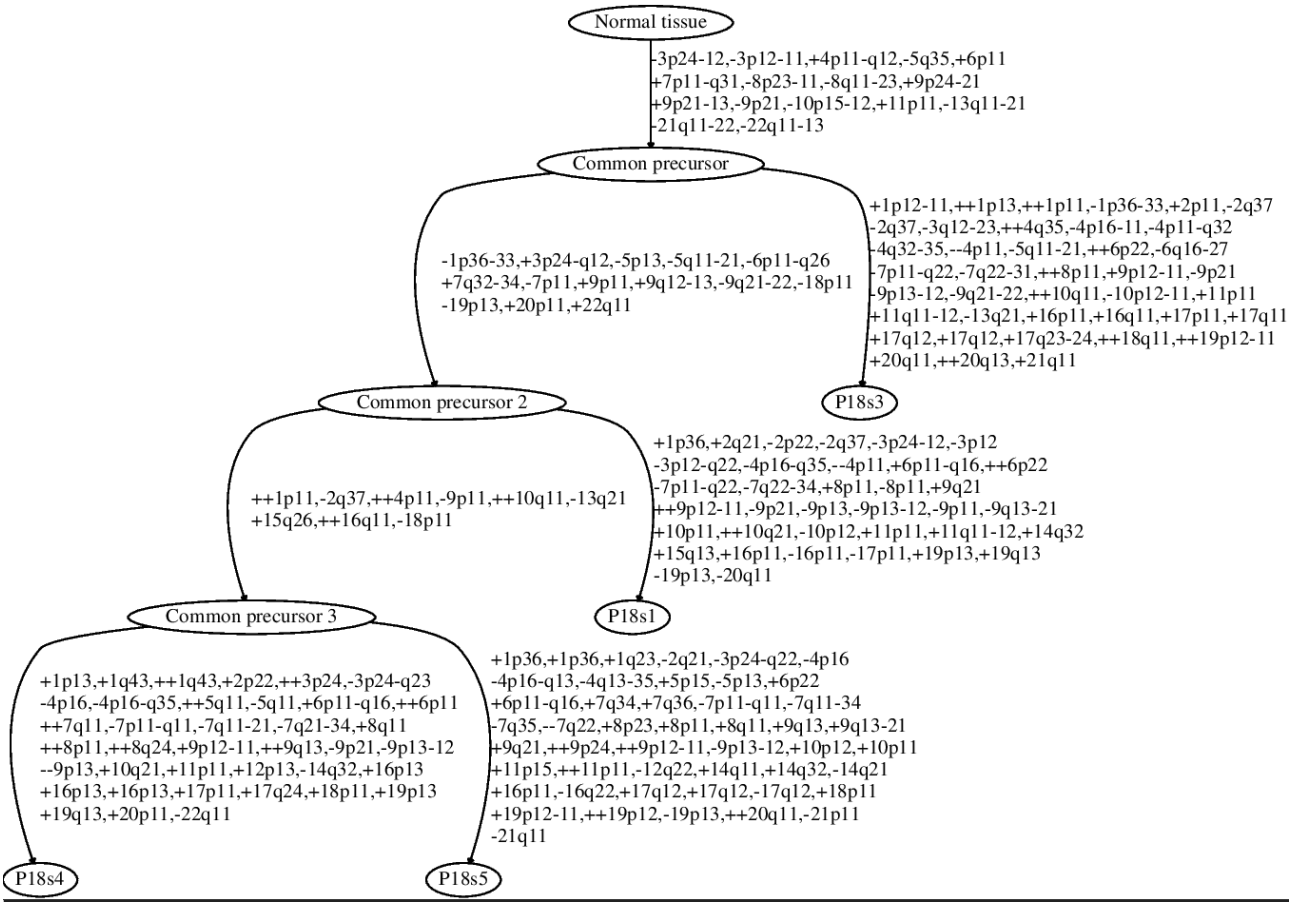
Sequenza/TuMult cytobands tree Patient 6



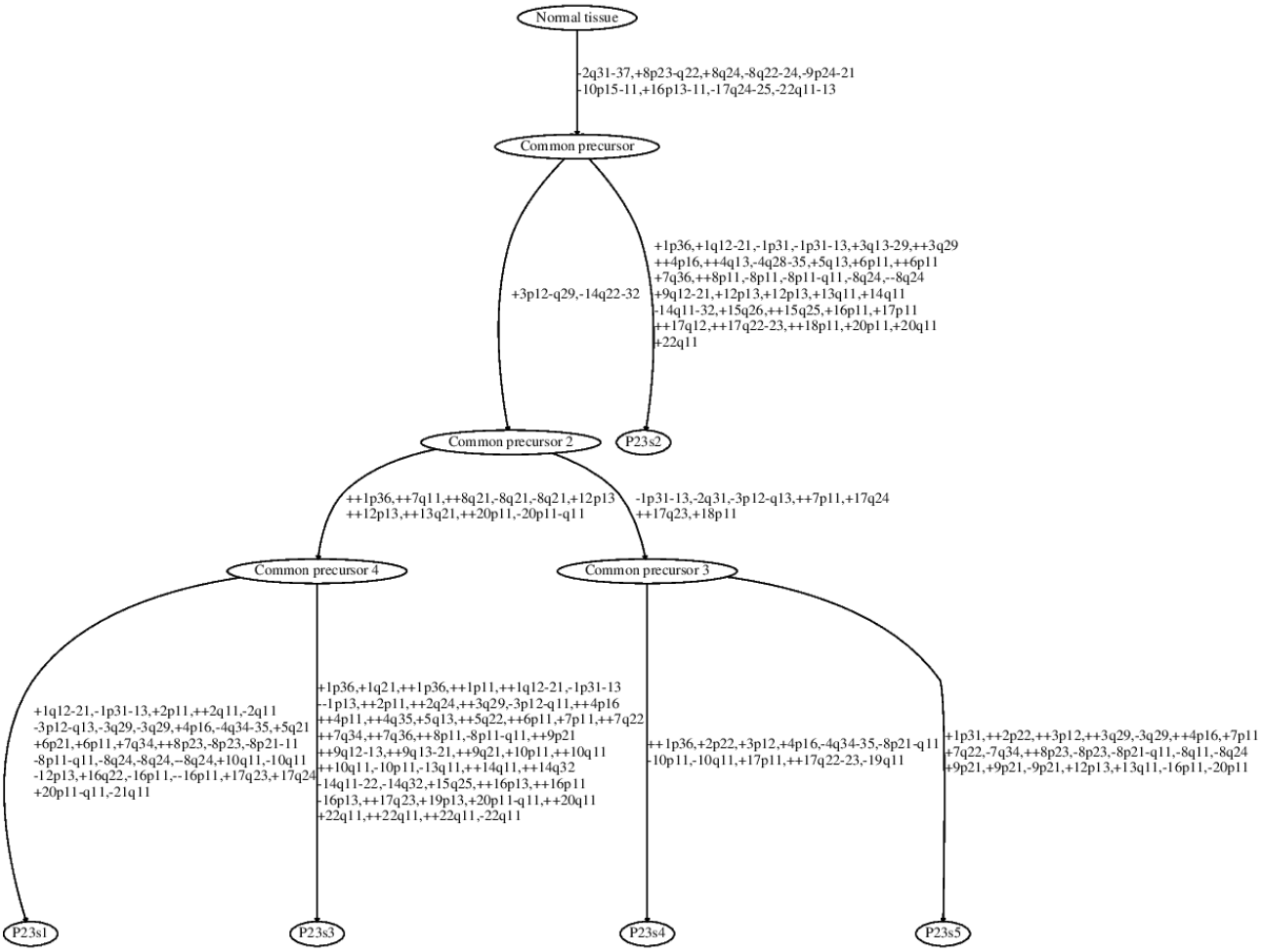
Sequenza/TuMult cytobands tree Patient 12



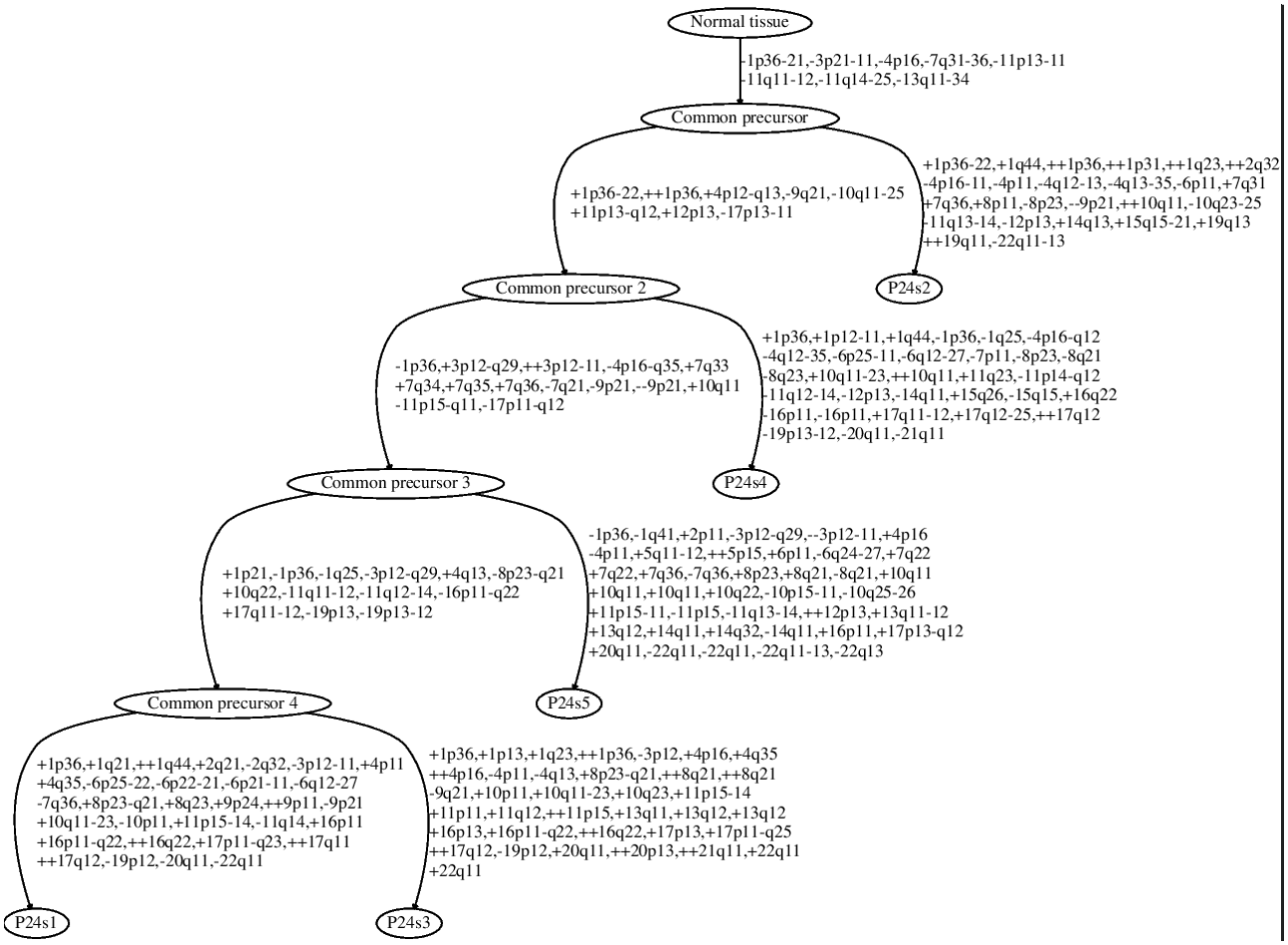
Sequenza/TuMult cytobands tree Patient 16



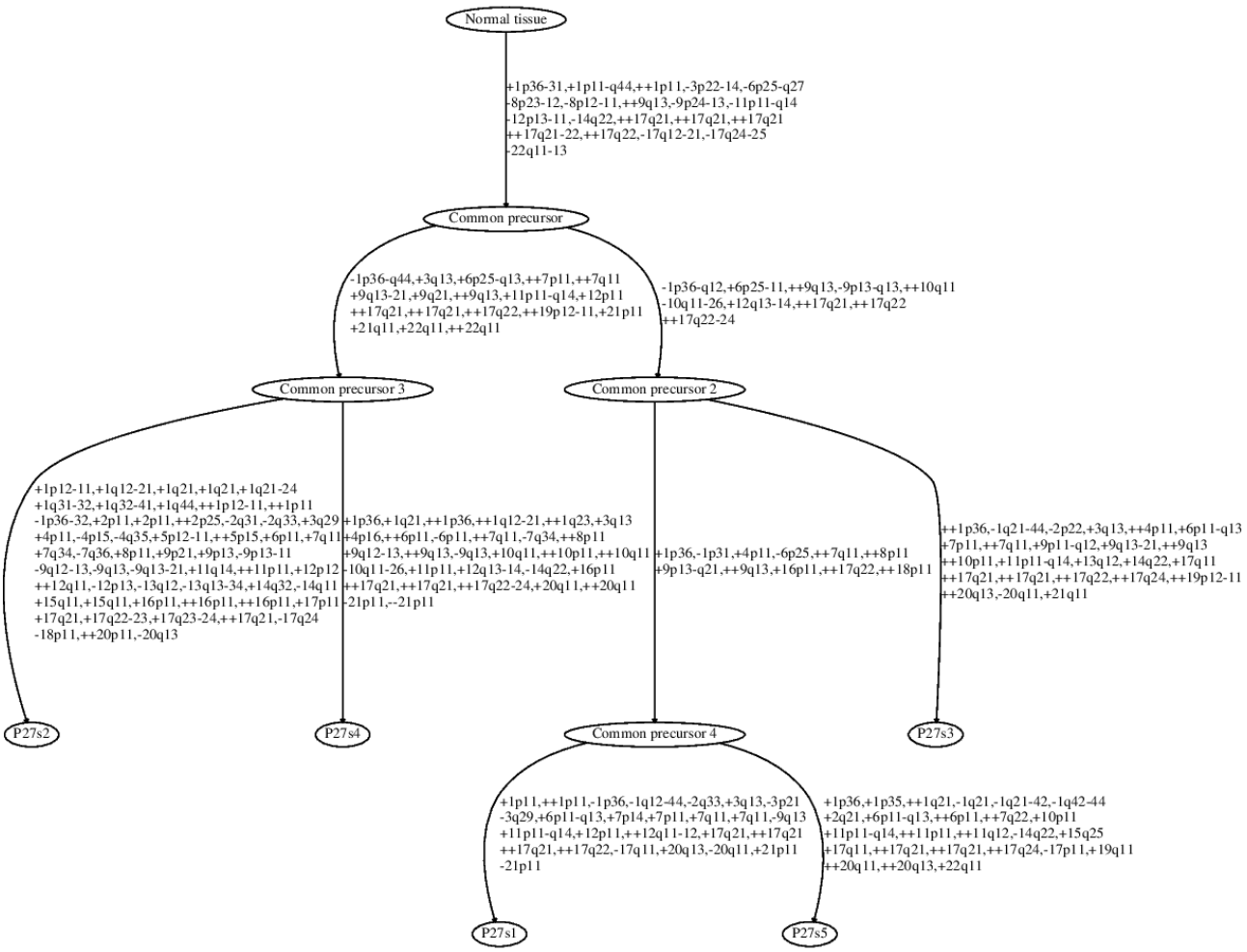
Sequenza/TuMult cytobands tree Patient 18



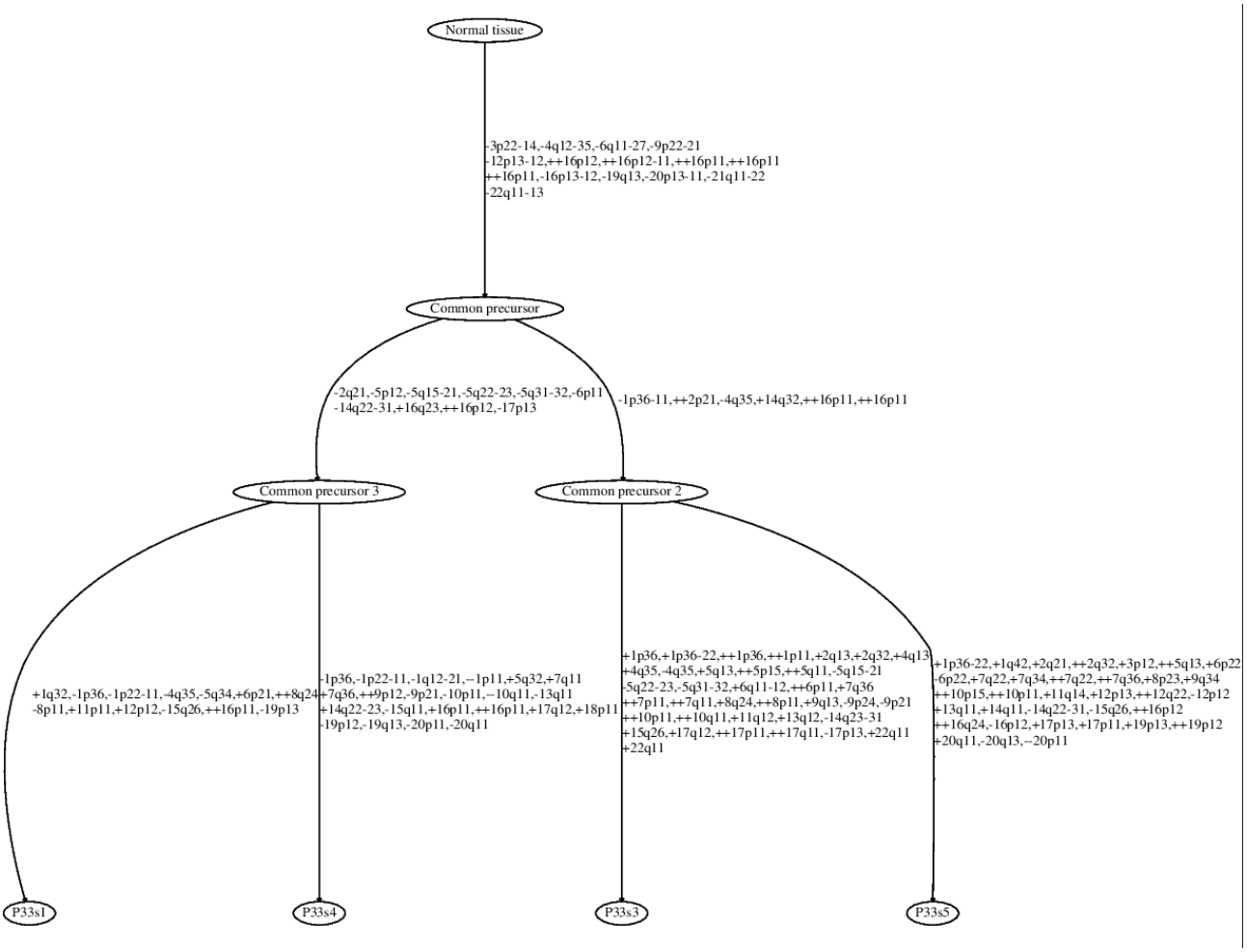
Sequenza/TuMult cytobands tree Patient 23



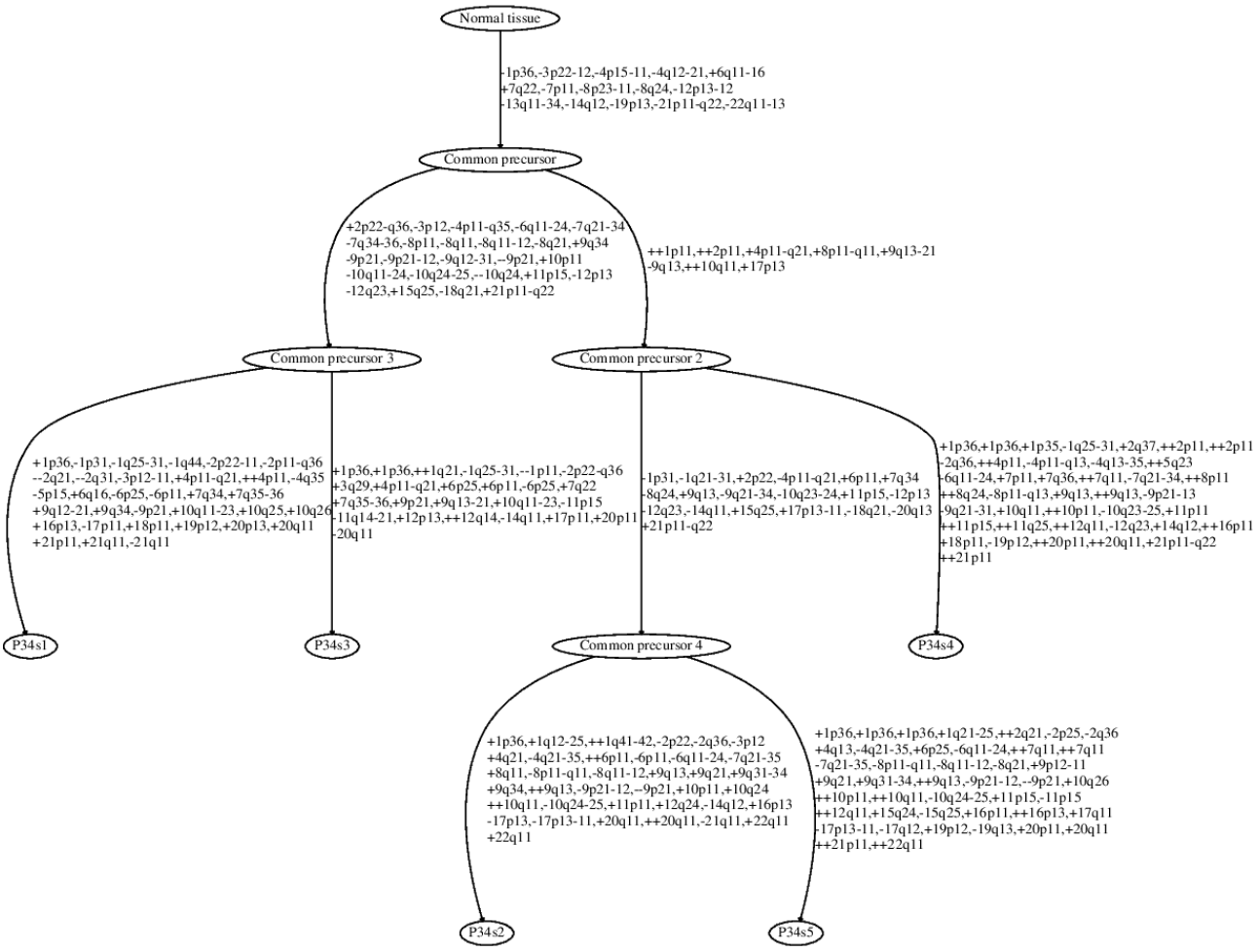
Sequenza/TuMult cytobands tree Patient 24



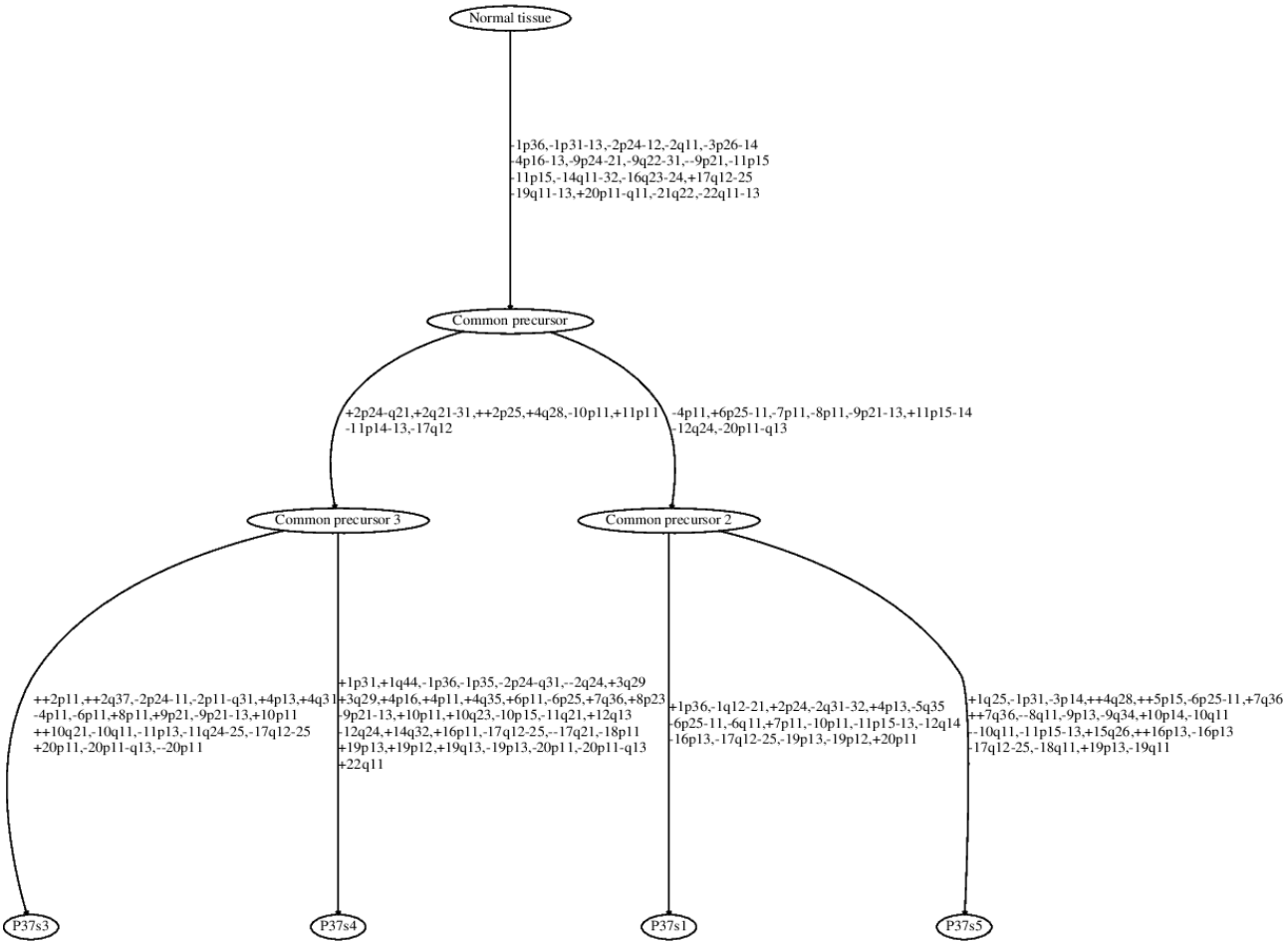
Sequenza/TuMult cytobands tree Patient 27



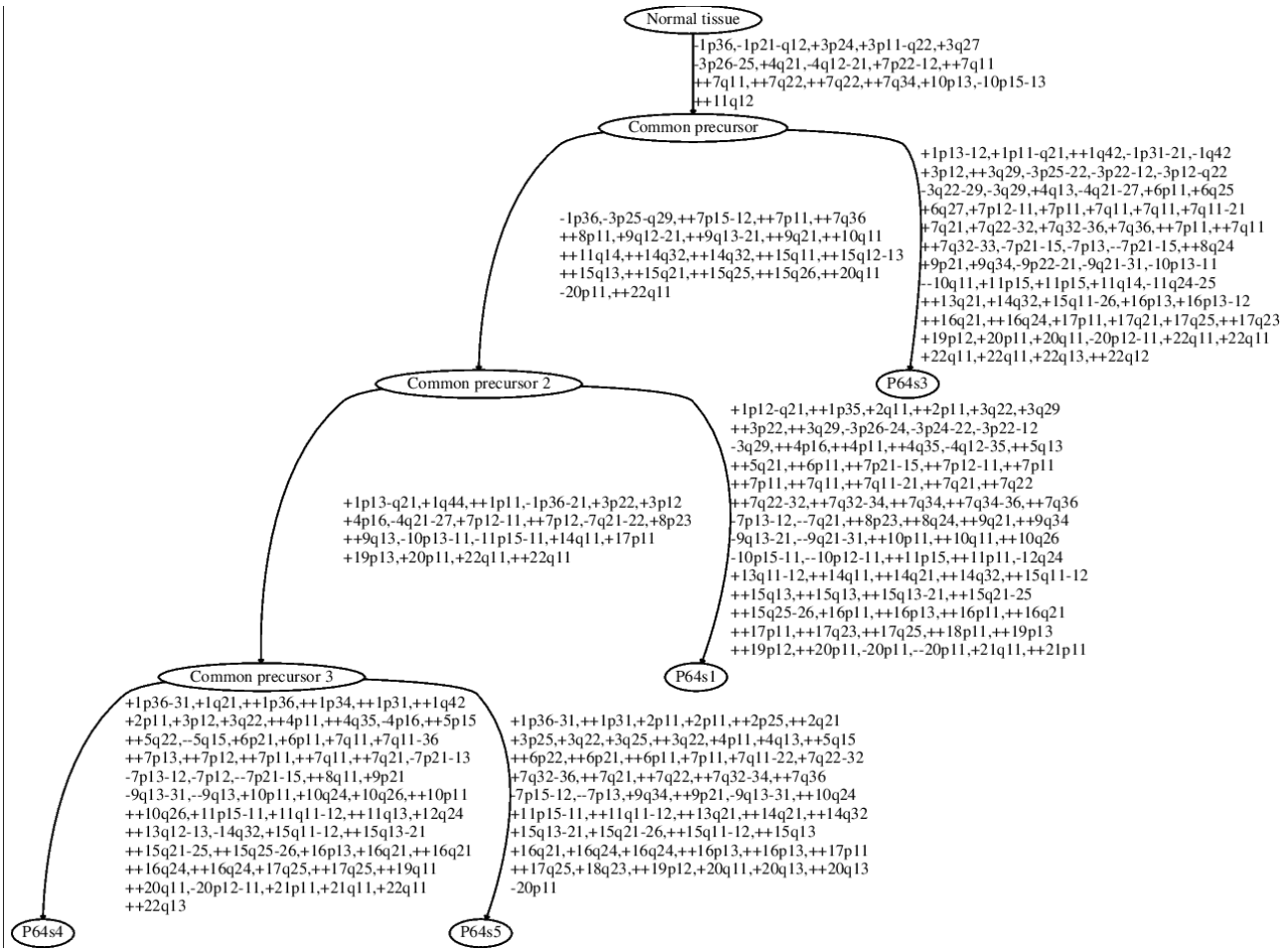
Sequenza/TuMult cytobands tree Patient 33



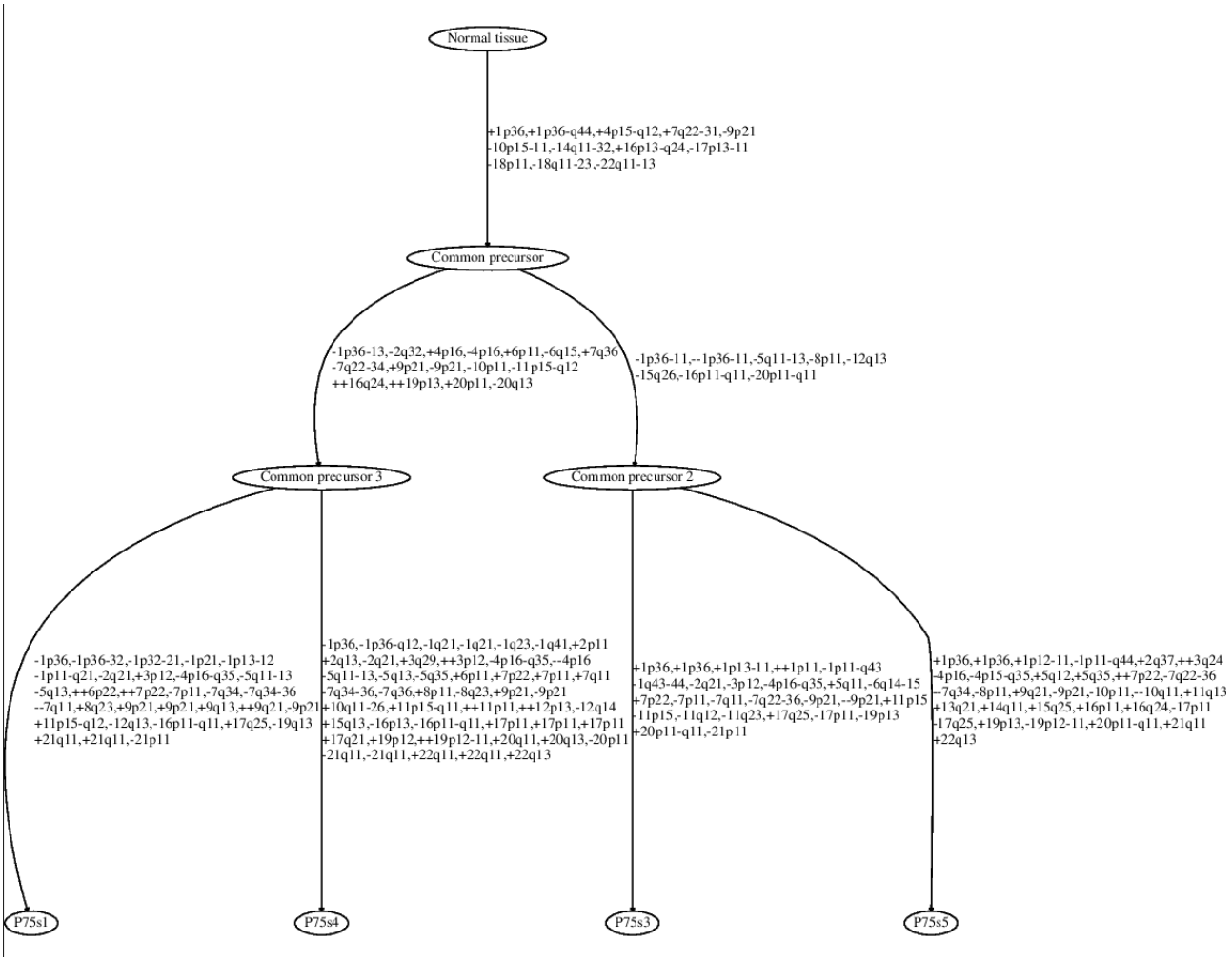
Sequenza/TuMult cytobands tree Patient 34



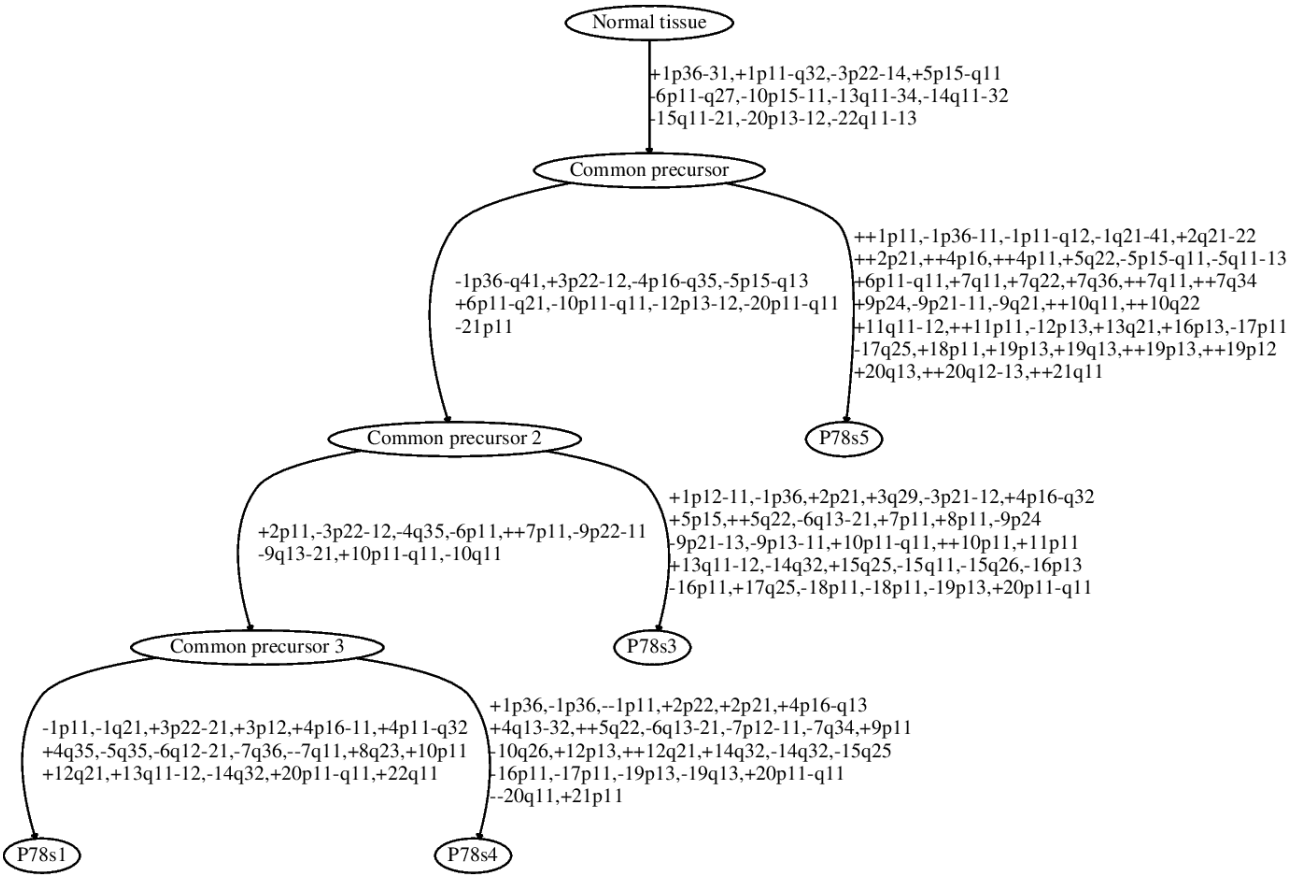
Sequenza/TuMult cytobands tree Patient 37



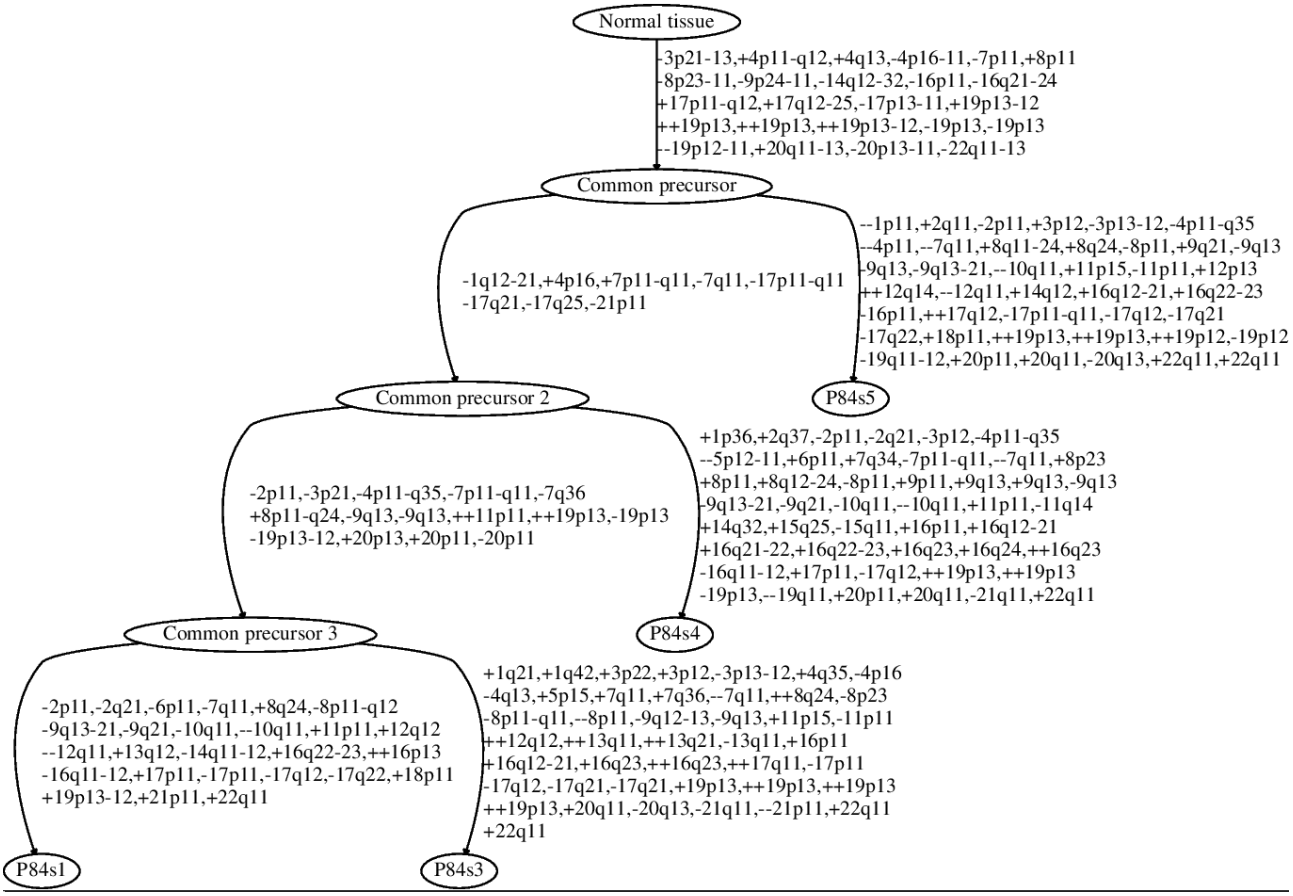
Sequenza/TuMult cytobands tree Patient 64



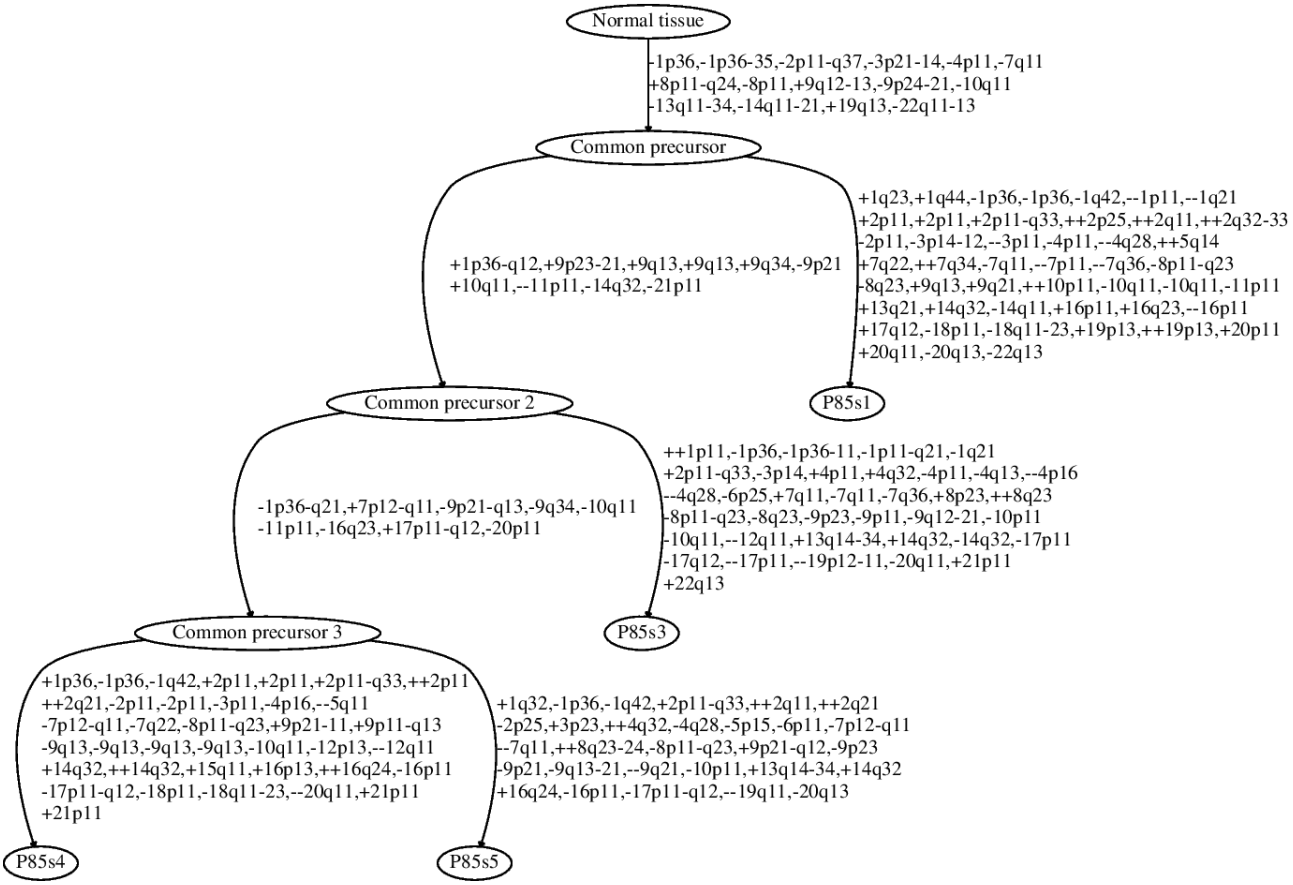
Sequenza/TuMult cytobands tree Patient 75



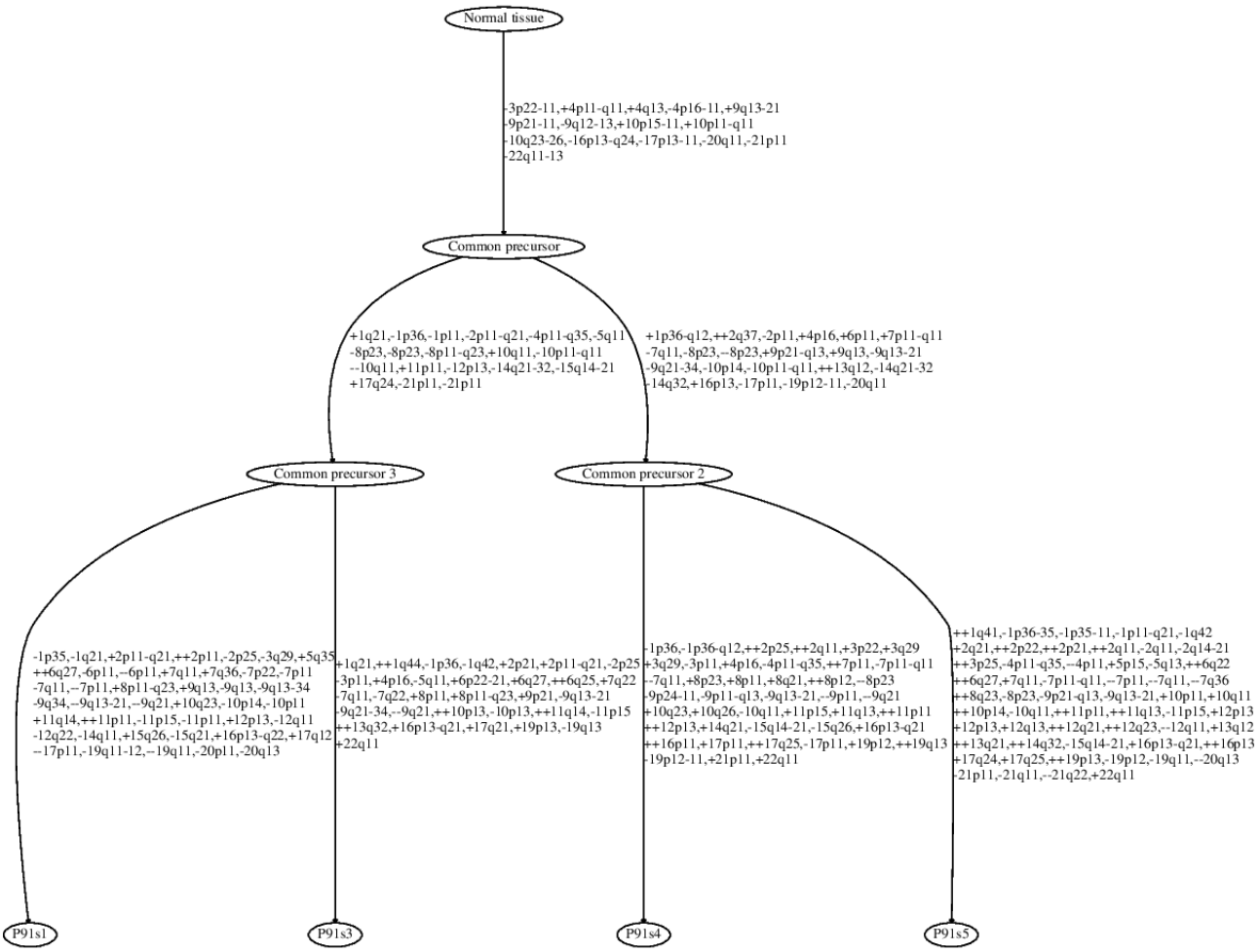
Sequenza/TuMult cytobands tree Patient 78



Sequenza/TuMult cytobands tree Patient 84

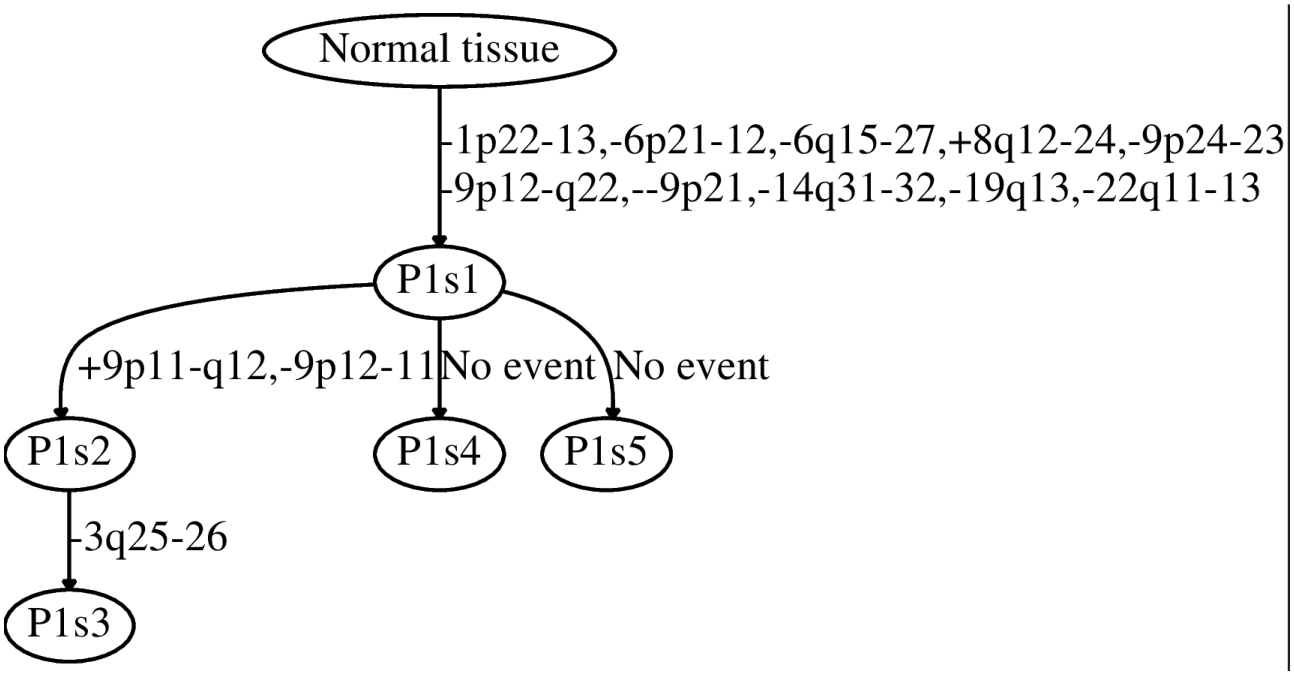


Sequenza/TuMult cytobands tree Patient 85

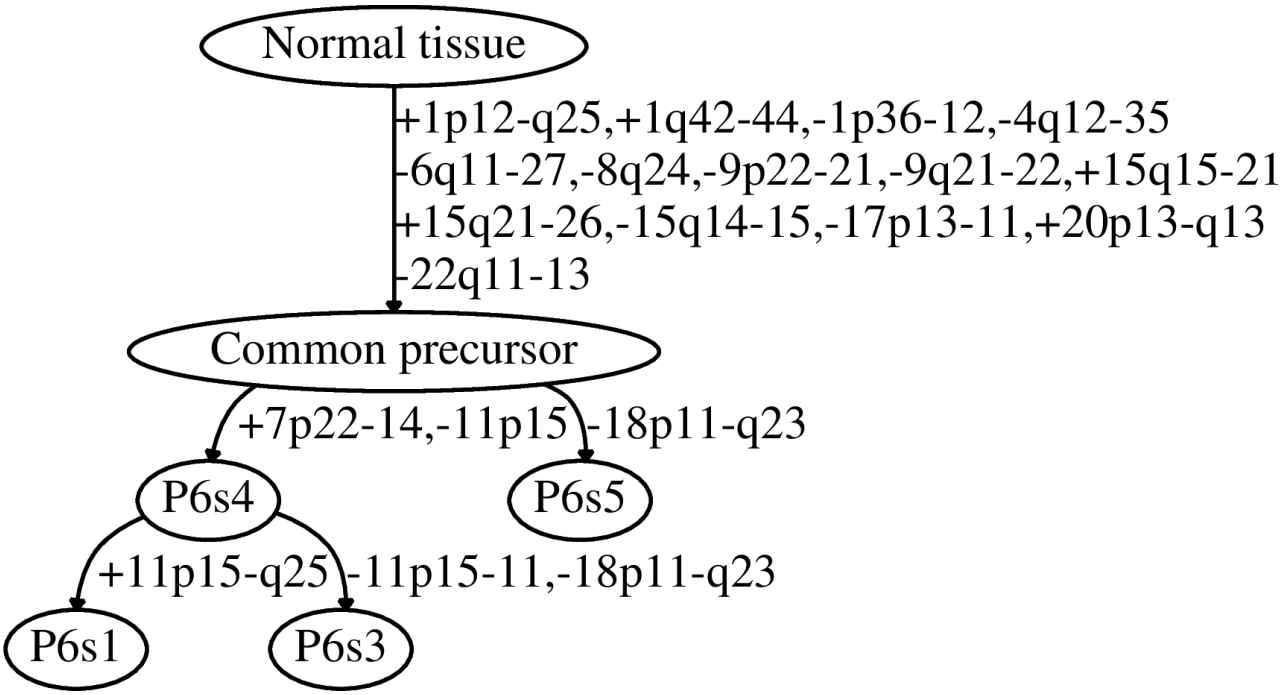


Sequenza/TuMult cytobands tree Patient 91

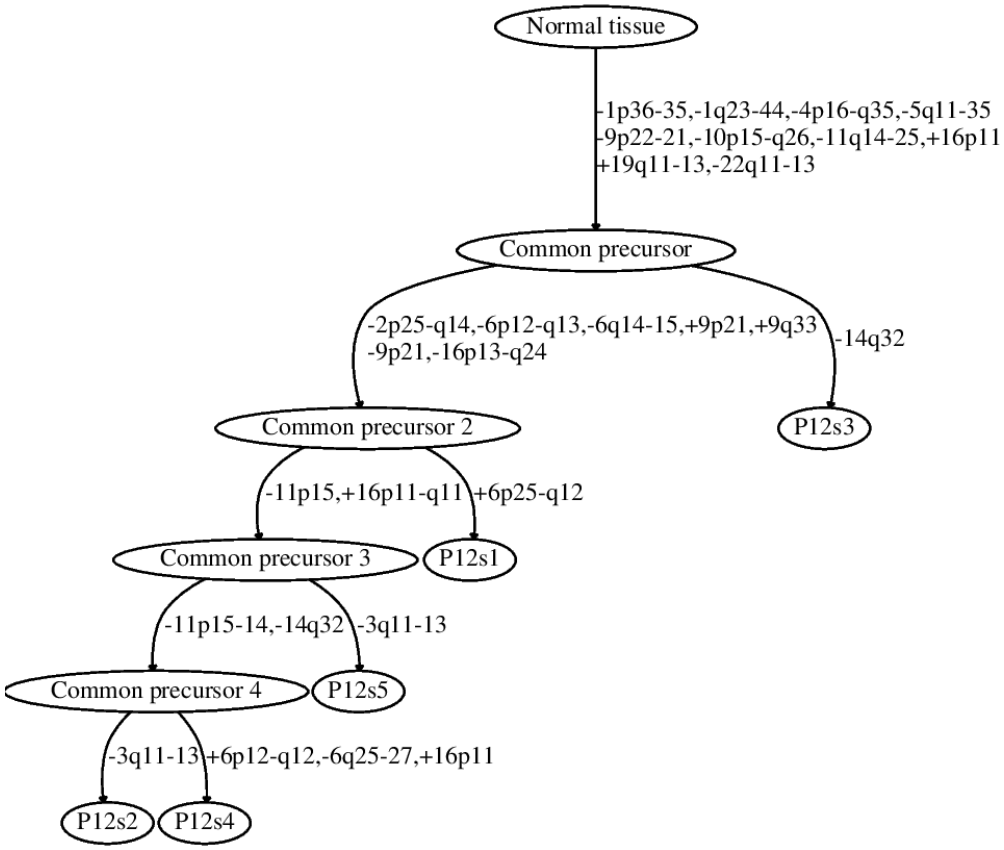
Appendix B - ASCAT/TuMult Cytobands Trees



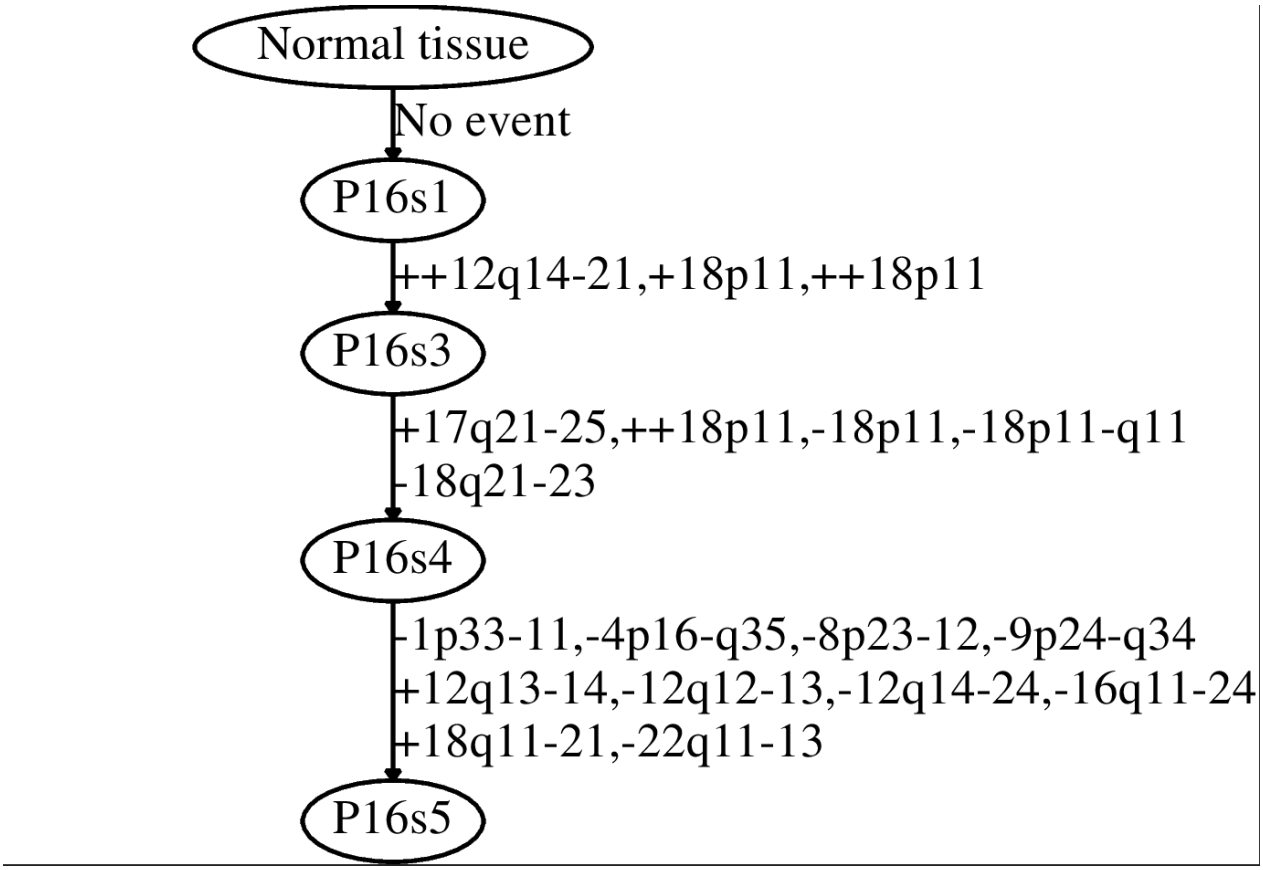
ACSCAT/TuMult cytobands tree Patient 1



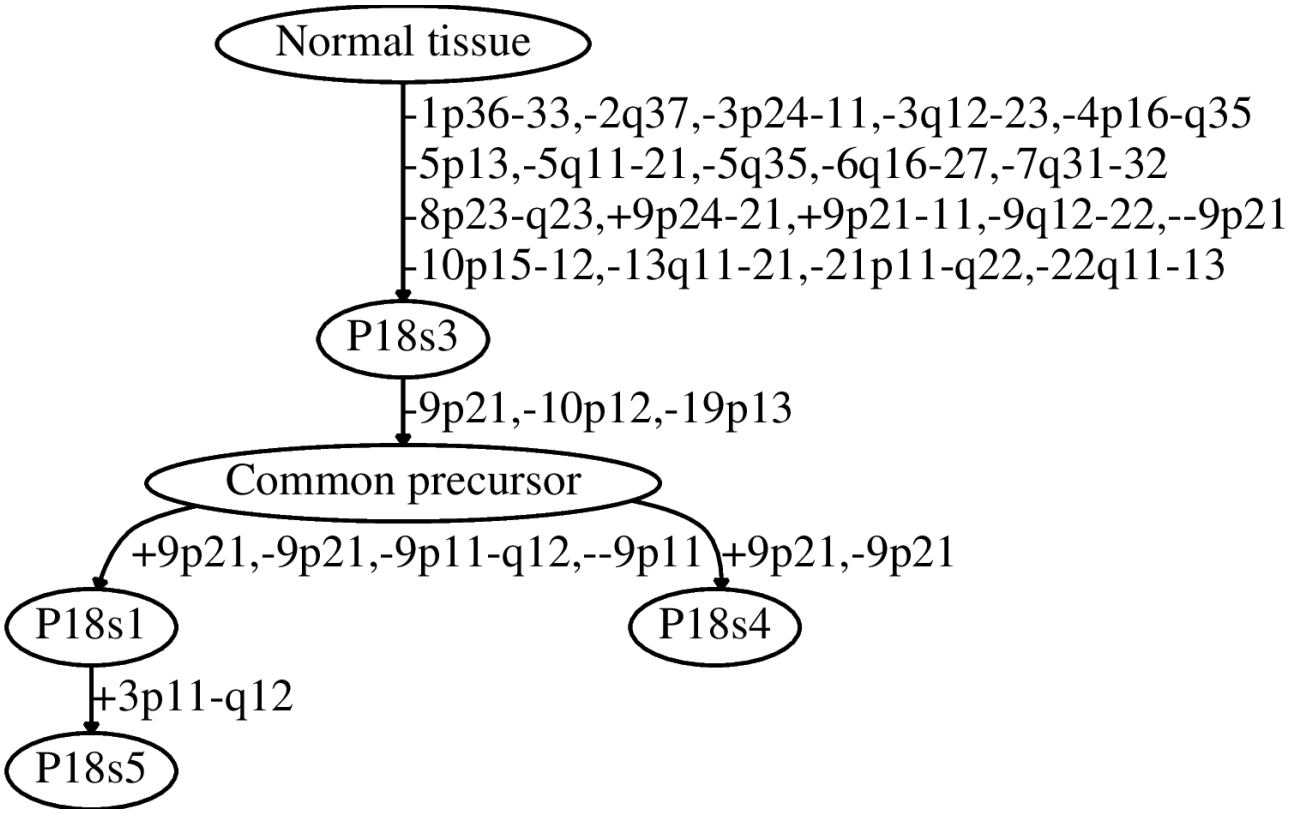
ACSCAT/TuMult cytobands tree Patient 6



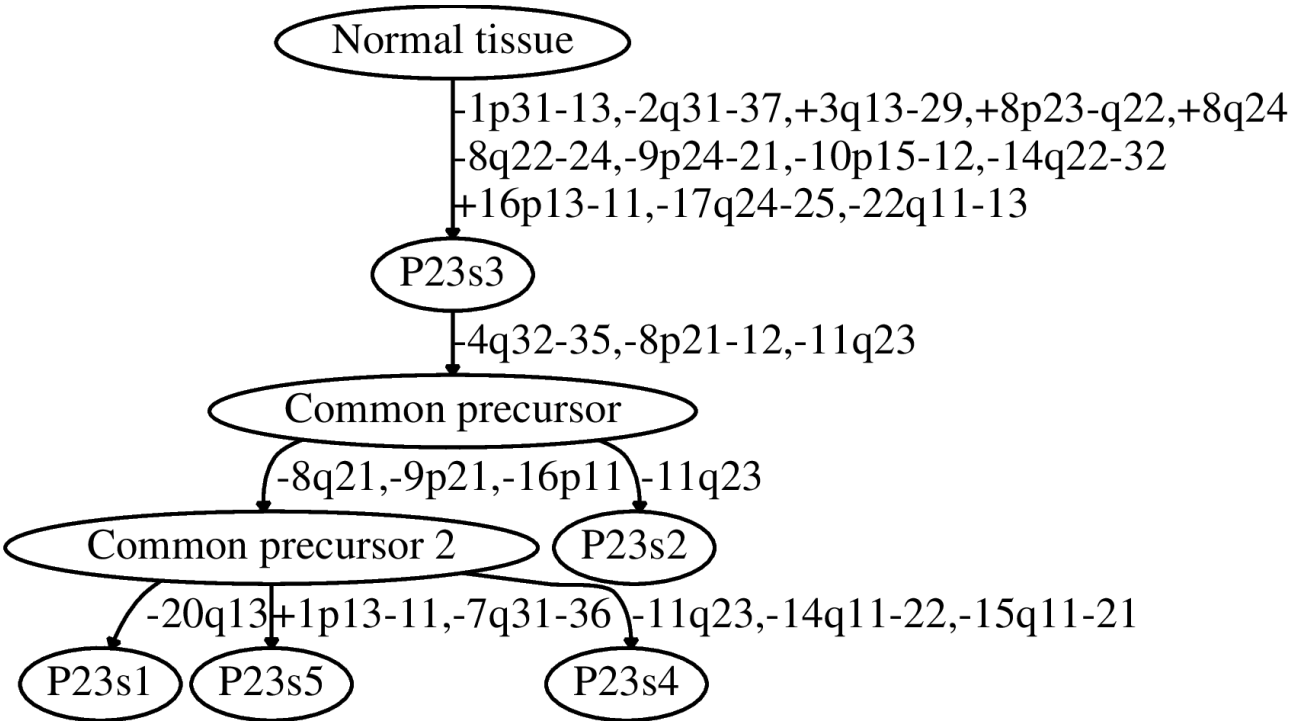
ACSCAT/TuMult cytobands tree Patient 12



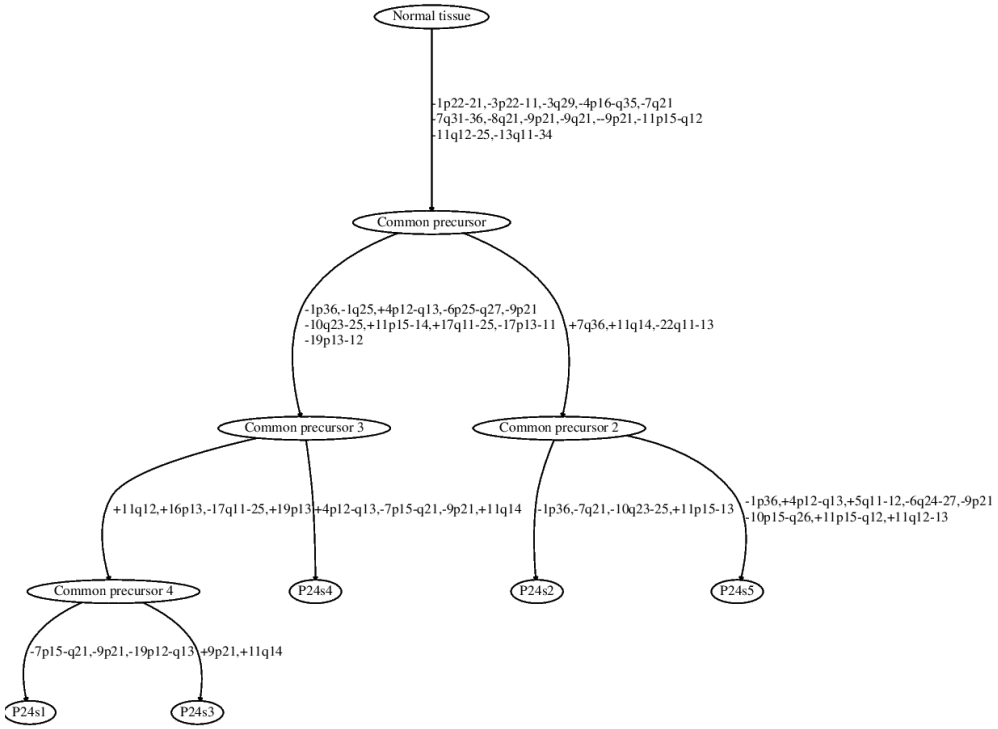
ACSCAT/TuMult cytobands tree Patient 16



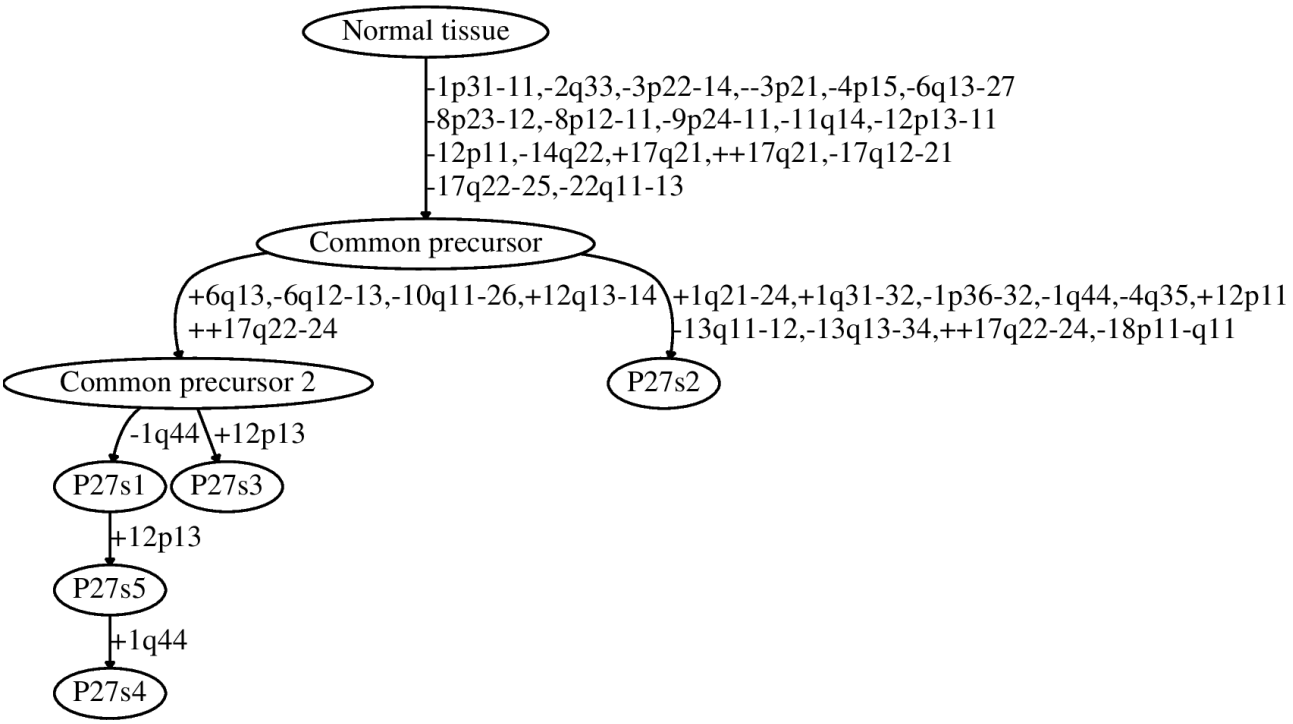
ACSCAT/TuMult cytobands tree Patient 18



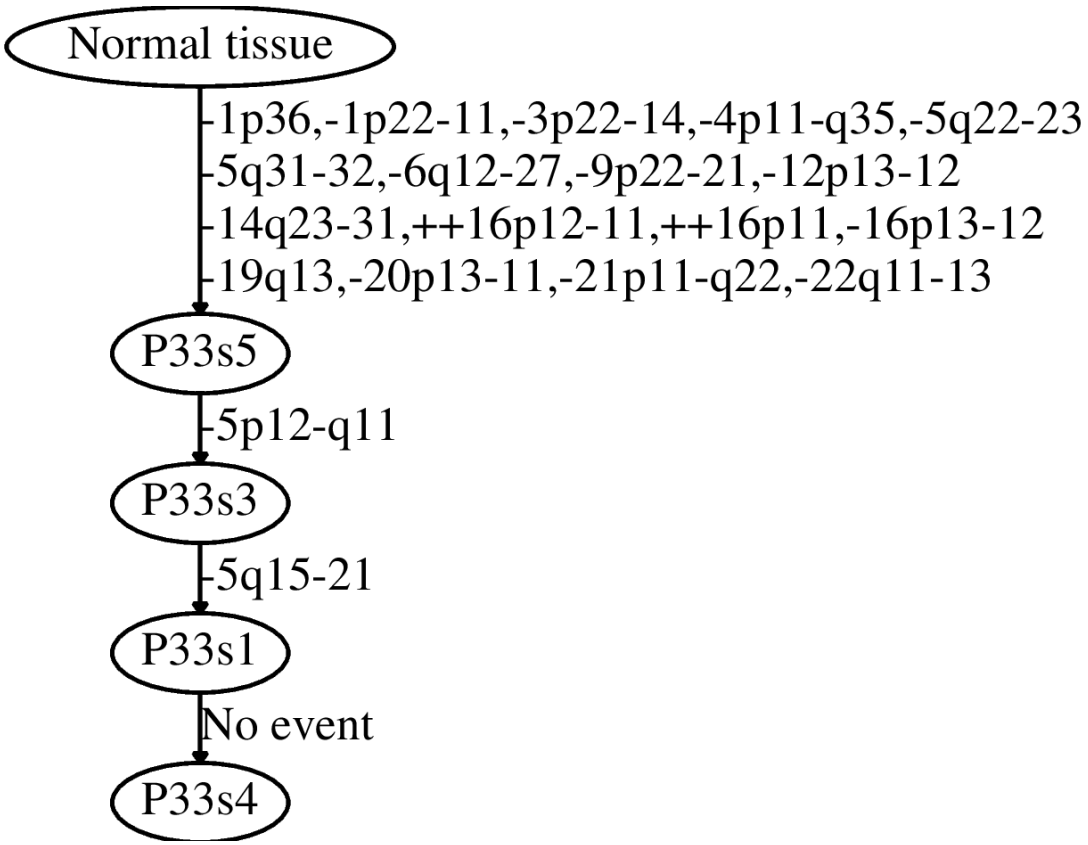
ACSCAT/TuMult cytobands tree Patient 23



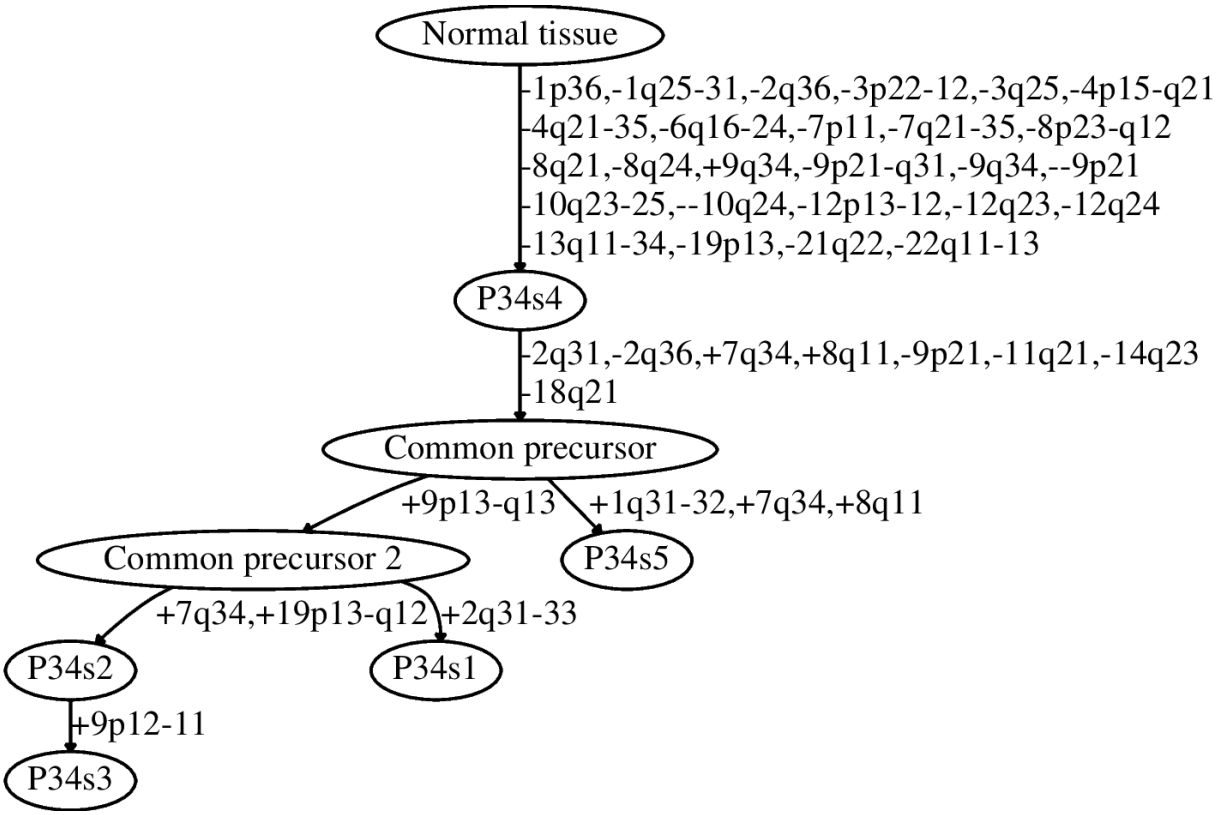
ACSCAT/TuMult cytobands tree Patient 24



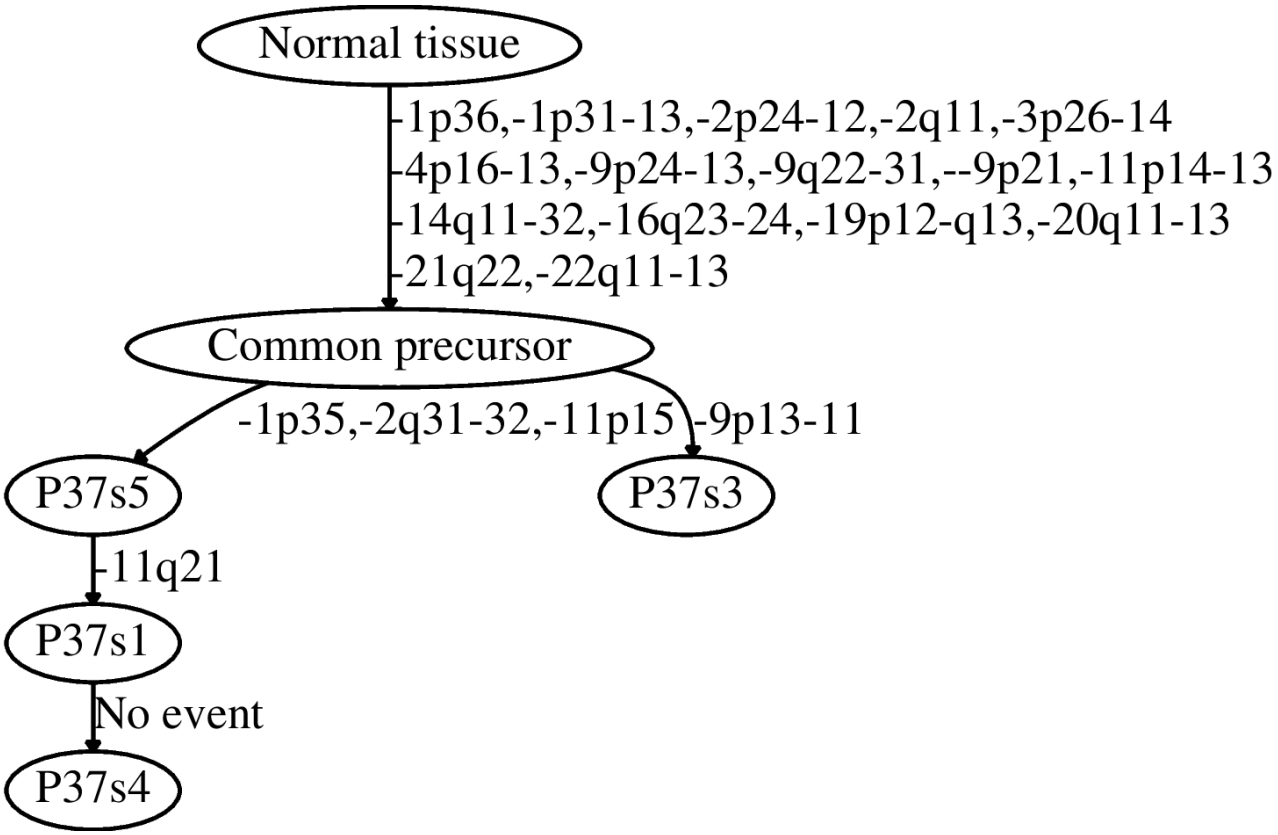
ACSCAT/TuMult cytobands tree Patient 27



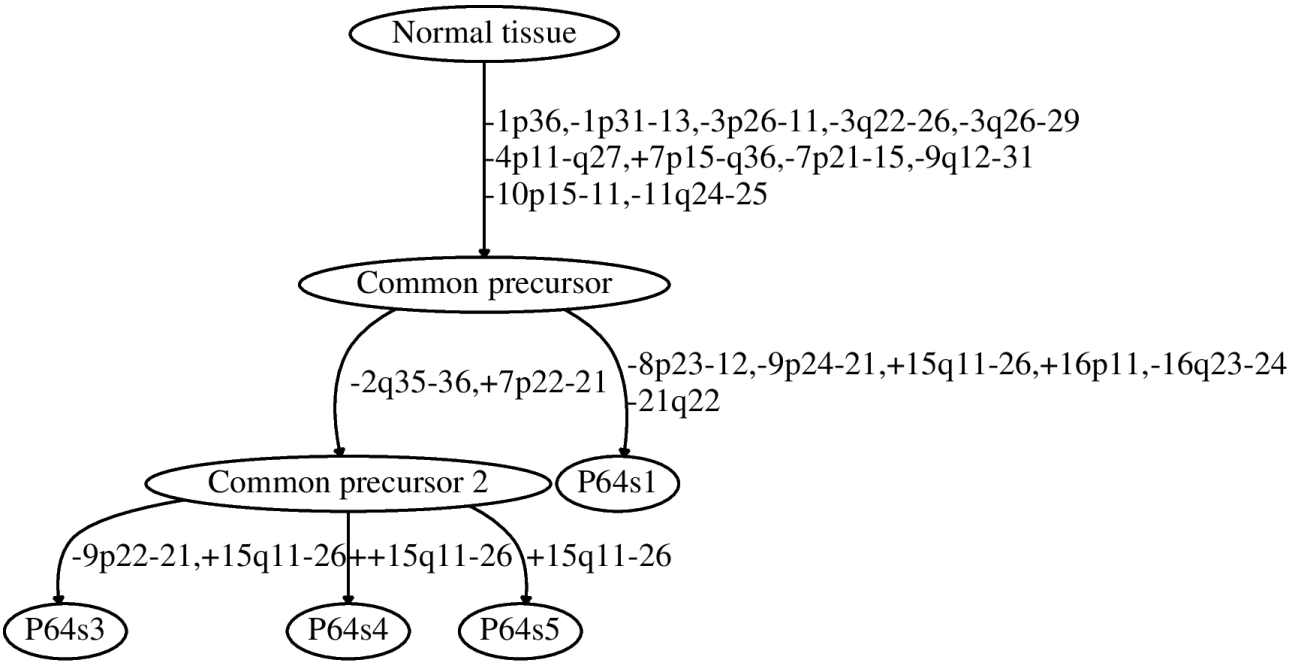
ACSCAT/TuMult cytobands tree Patient 33



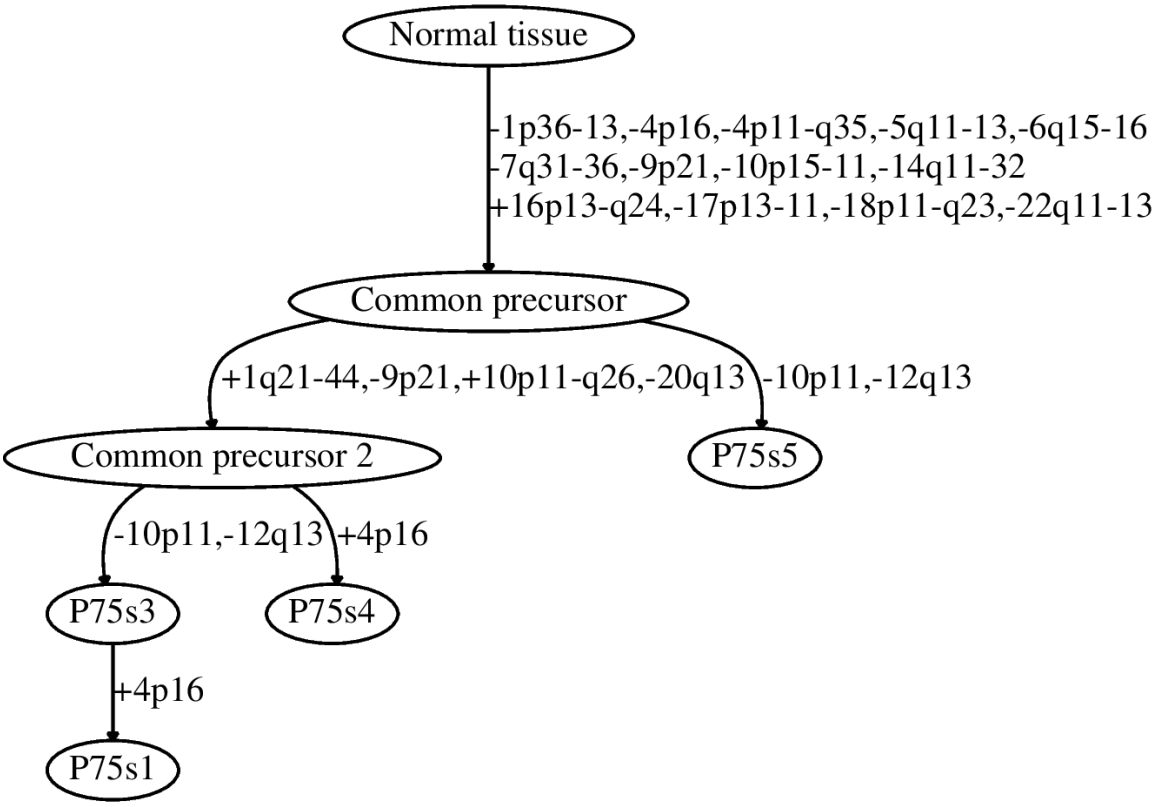
ACSCAT/TuMult cytobands tree Patient 34



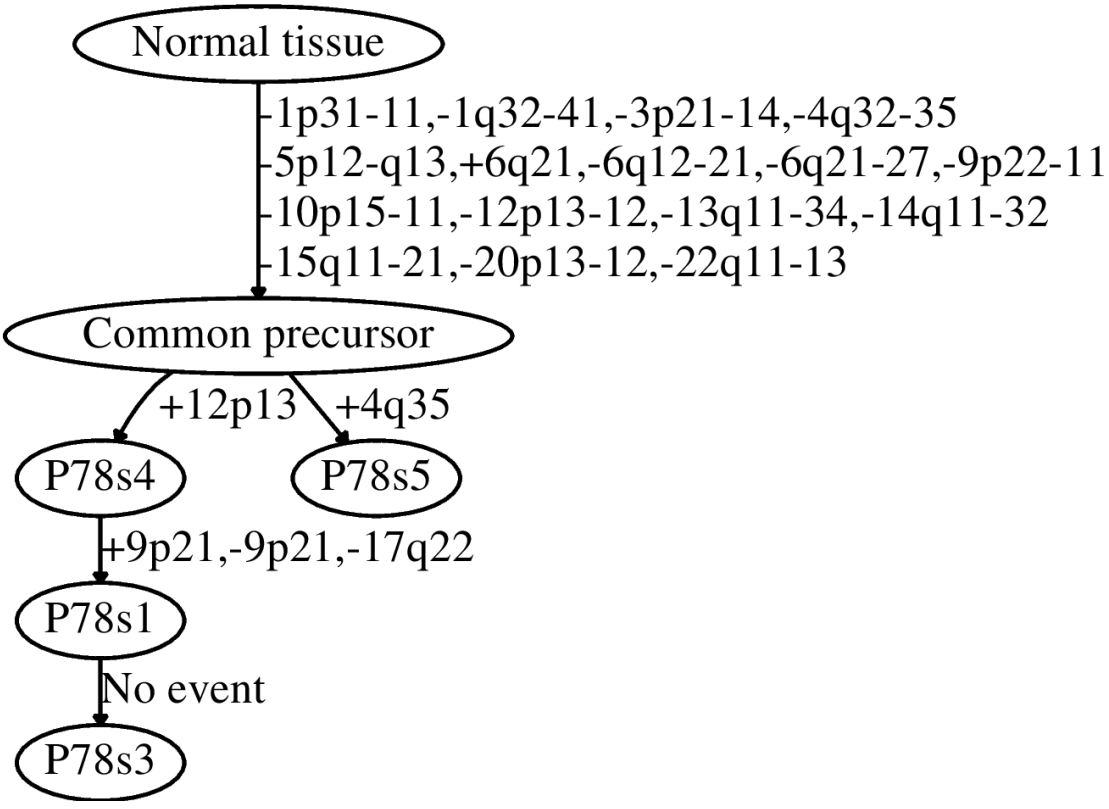
ACSCAT/TuMult cytobands tree Patient 37



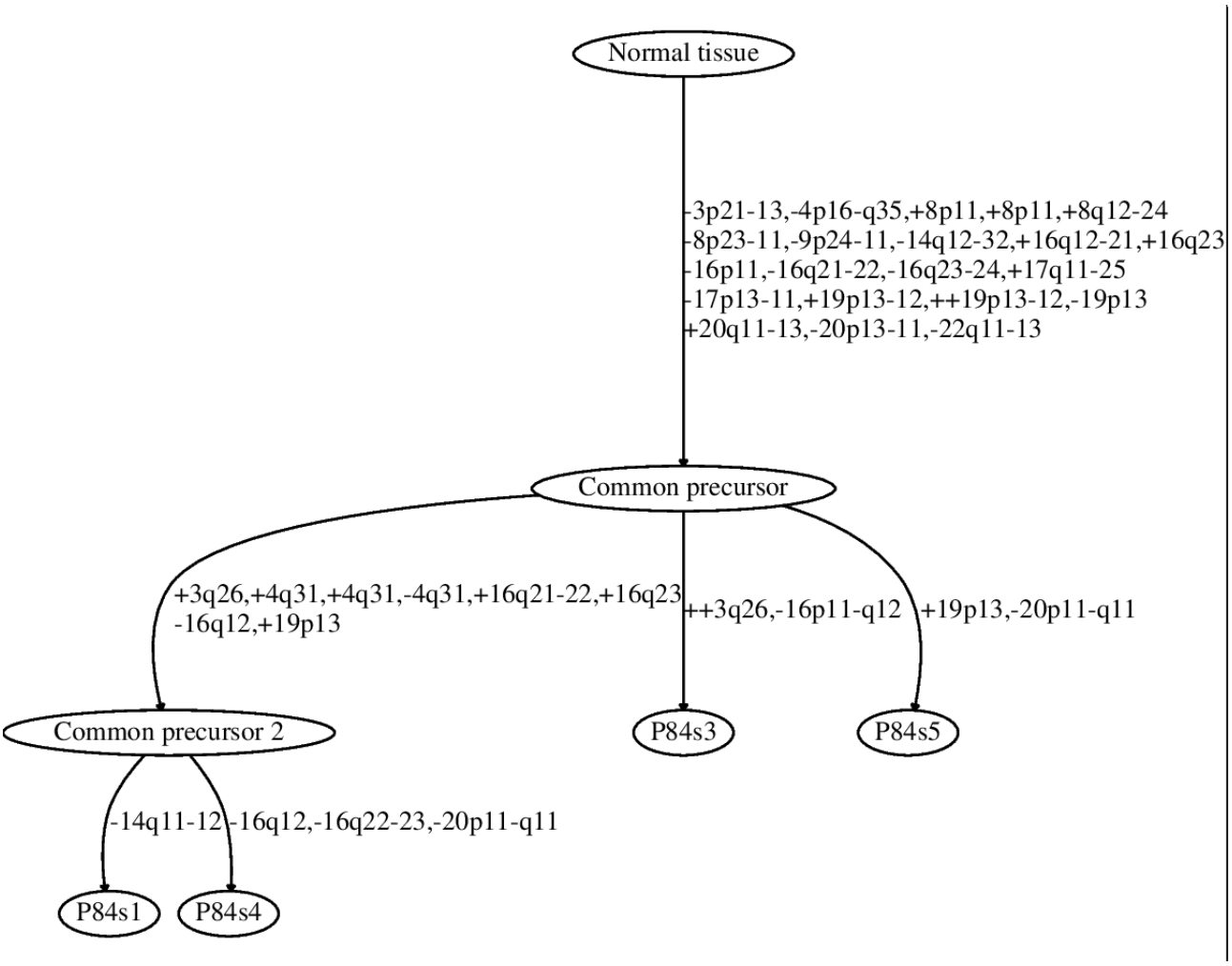
ACSCAT/TuMult cytobands tree Patient 64



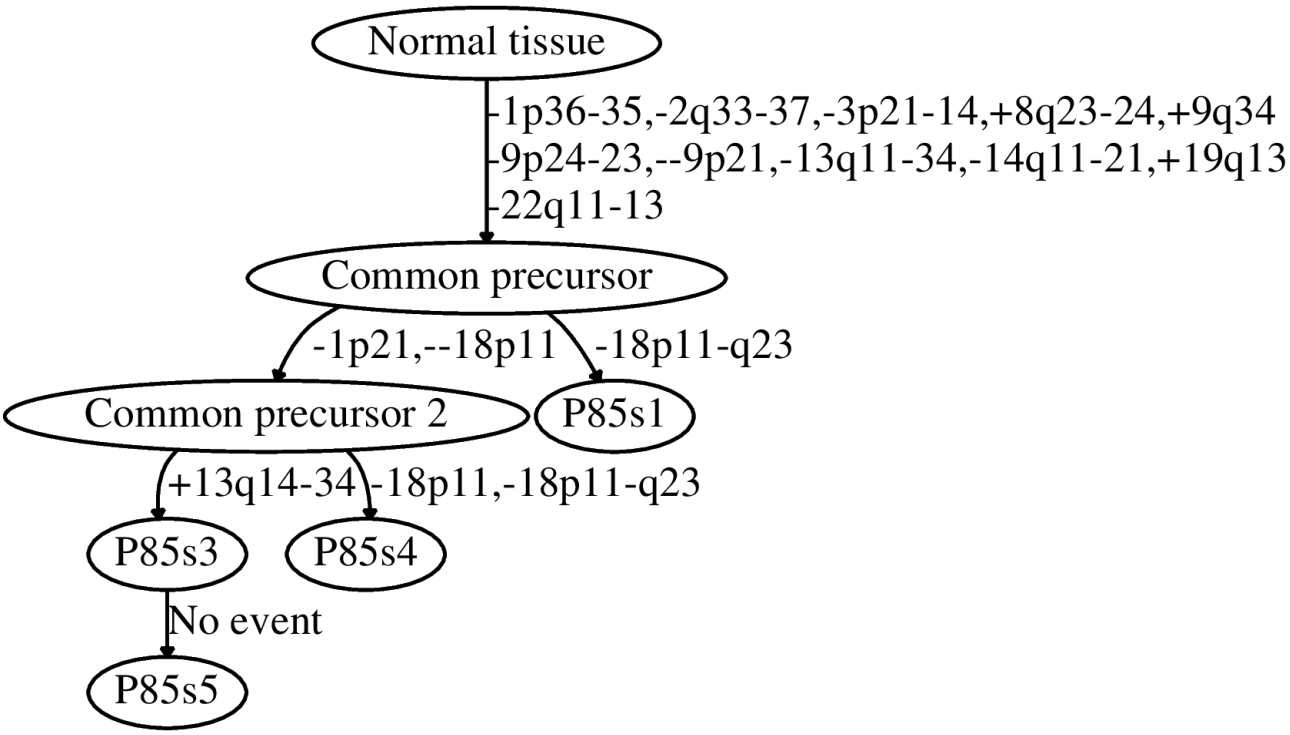
ACSCAT/TuMult cytobands tree Patient 75



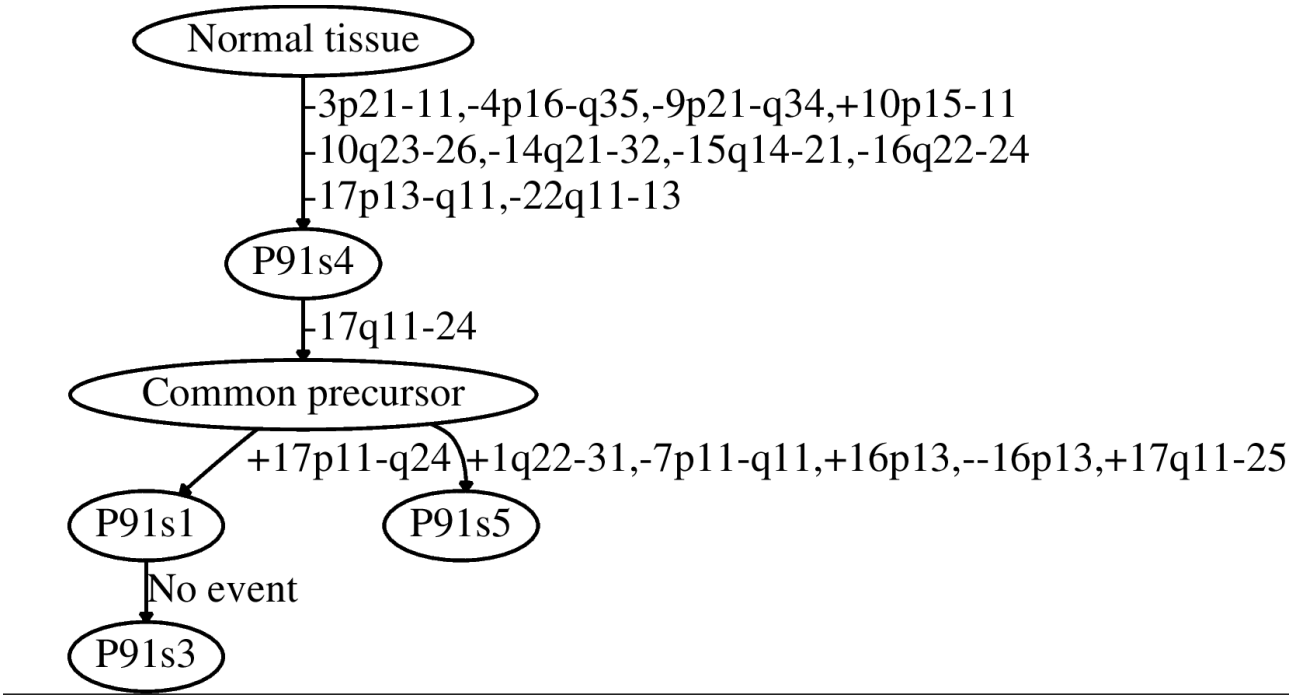
ACSCAT/TuMult cytobands tree Patient 78



ACSCAT/TuMult cytobands tree Patient 84



ACSCAT/TuMult cytobands tree Patient 85



ACSCAT/TuMult cytobands tree Patient 91

Bibliography

1000 GENOMES PROJECT CONSORTIUM, AUTON, A., BROOKS, L.D., DURBIN, R.M., GARRISON, E.P., KANG, H.M., KORBEL, J.O., MARCHINI, J.L., MCCARTHY, S., MCVEAN, G.A. and ABECASIS, G.R., 2015. A global reference for human genetic variation. *Nature*, **526**(7571), pp. 68-74.

ABBOSH, C., BIRKBAK, N.J., WILSON, G.A., JAMAL-HANJANI, M., CONSTANTIN, T., SALARI, R., LE QUESNE, J., MOORE, D.A., VEERIAH, S., ROSENTHAL, R., MARAFIOTI, T., KIRKIZLAR, E., WATKINS, T.B.K., MCGRANAHAN, N., WARD, S., MARTINSON, L., RILEY, J., FRAIOLI, F., AL BAKIR, M., GRÖNROOS, E., ZAMBRANA, F., ENDOZO, R., BI, W.L., FENNESSY, F.M., SPONER, N., JOHNSON, D., LAYCOCK, J., SHAFI, S., CZYZEWSKA-KHAN, J., ROWAN, A., CHAMBERS, T., MATTHEWS, N., TURAJLIC, S., HILEY, C., LEE, S.M., FORSTER, M.D., AHMAD, T., FALZON, M., BORG, E., LAWRENCE, D., HAYWARD, M., KOLVEKAR, S., PANAGIOTOPOULOS, N., JANES, S.M., THAKRAR, R., AHMED, A., BLACKHALL, F., SUMMERS, Y., HAFEZ, D., NAIK, A., GANGULY, A., KAREHT, S., SHAH, R., JOSEPH, L., MARIE QUINN, A., CROSBIE, P.A., NAIDU, B., MIDDLETON, G., LANGMAN, G., TROTTER, S., NICOLSON, M., REMMEN, H., KERR, K., CHETTY, M., GOMERSALL, L., FENNELL, D.A., NAKAS, A., RATHINAM, S., ANAND, G., KHAN, S., RUSSELL, P., EZHIL, V., ISMAIL, B., IRVIN-SELLERS, M., PRAKASH, V., LESTER, J.F., KORNASZEWSKA, M., ATTANOOS, R., ADAMS, H., DAVIES, H., OUKRIF, D., AKARCA, A.U., HARTLEY, J.A., LOWE, H.L., LOCK, S., ILES, N., BELL, H., NGAI, Y., ELGAR, G., SZALLASI, Z., SCHWARZ, R.F., HERRERO, J., STEWART, A., QUEZADA, S.A., PEGGS, K.S., VAN LOO, P., DIVE, C., LIN, C.J., RABINOWITZ, M., AERTS, HUGO J W L, HACKSHAW, A., SHAW, J.A., ZIMMERMANN, B.G., TRACERX CONSORTIUM, PEACE CONSORTIUM and SWANTON, C., 2017. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*, **545**(7655), pp. 446-451.

AHUJA, D., SAENZ-ROBLES, M.T. and PIPAS, J.M., 2005. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. *Oncogene*, **24**(52), pp. 7729-7745.

ALBERTSON, D.G., COLLINS, C., MCCORMICK, F. and GRAY, J.W., 2003. Chromosome aberrations in solid tumors. *Nature genetics*, **34**(4), pp. 369-376.

ALEXANDROV, L.B., NIK-ZAINAL, S., WEDGE, D.C., APARICIO, S.A., BEHJATI, S., BIANKIN, A.V., BIGNELL, G.R., BOLLI, N., BORG, A., BØRRESEN-DALE, A.L., BOYAULT, S., BURKHARDT, B., BUTLER, A.P., CALDAS, C., DAVIES, H.R., DESMEDT, C., EILS, R., EYFJÖRD, J.E., FOEKENS, J.A., GREAVES, M., HOSODA, F., HUTTER, B., ILICIC, T., IMBEAUD, S., IMIELINSKI, M., JÄGER, N., JONES, D.T., JONES, D., KNAPPSKOG, S., KOOL, M., LAKHANI, S.R., LÓPEZ-OTÍN, C., MARTIN, S., MUNSHI, N.C., NAKAMURA, H., NORTHCOTT, P.A., PAJIC, M., PAPAEMMANUIL, E., PARADISO, A., PEARSON, J.V., PUENTE, X.S., RAINE, K., RAMAKRISHNA, M., RICHARDSON, A.L., RICHTER, J., ROSENSTIEL, P., SCHLESNER, M., SCHUMACHER, T.N., SPAN, P.N., TEAGUE, J.W., TOTOKI, Y., TUTT, A.N., VALDÉS-MAS, R., VAN BUUREN, M.M., VAN 'T VEER, L., VINCENT-SALOMON, A., WADDELL, N., YATES, L.R., AUSTRALIAN PANCREATIC CANCER GENOME INITIATIVE, ICGC BREAST CANCER CONSORTIUM, ICGC MMML-SEQ CONSORTIUM, ICGC PEDBRAIN, ZUCMAN-

ROSSI, J., FUTREAL, P.A., MCDERMOTT, U., LICHTER, P., MEYERSON, M., GRIMMOND, S.M., SIEBERT, R., CAMPO, E., SHIBATA, T., PFISTER, S.M., CAMPBELL, P.J. and STRATTON, M.R., 2013. Signatures of mutational processes in human cancer. *Nature*, **500**(7463), pp. 415-421.

BAGCHI, A. and MILLS, A.A., 2008. The quest for the 1p36 tumor suppressor. *Cancer research*, **68**(8), pp. 2551-2556.

BAO, L., PU, M. and MESSER, K., 2014. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics (Oxford, England)*, **30**(8), pp. 1056-1063.

BENEDETTI, S., NUVOLI, B., CATALANI, S. and GALATI, R., 2015. Reactive oxygen species a double-edged sword for mesothelioma. *Oncotarget*, **6**(19), pp. 16848-16865.

BEROUKHIM, R., MERMEL, C.H., PORTER, D., WEI, G., RAYCHAUDHURI, S., DONOVAN, J., BARRETINA, J., BOEHM, J.S., DOBSON, J., URASHIMA, M., MC HENRY, K.T., PINCHBACK, R.M., LIGON, A.H., CHO, Y.J., HAERY, L., GREULICH, H., REICH, M., WINCKLER, W., LAWRENCE, M.S., WEIR, B.A., TANAKA, K.E., CHIANG, D.Y., BASS, A.J., LOO, A., HOFFMAN, C., PRENSNER, J., LIEFELD, T., GAO, Q., YECIES, D., SIGNORETTI, S., MAHER, E., KAYE, F.J., SASAKI, H., TEPPER, J.E., FLETCHER, J.A., TABERNERO, J., BASELGA, J., TSAO, M.S., DEMICHELIS, F., RUBIN, M.A., JANNE, P.A., DALY, M.J., NUCERA, C., LEVINE, R.L., EBERT, B.L., GABRIEL, S., RUSTGI, A.K., ANTONESCU, C.R., LADANYI, M., LETAI, A., GARRAWAY, L.A., LODA, M., BEER, D.G., TRUE, L.D., OKAMOTO, A., POMEROY, S.L., SINGER, S., GOLUB, T.R., LANDER, E.S., GETZ, G., SELLERS, W.R. and MEYERSON, M., 2010. The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**(7283), pp. 899-905.

BETTI, M., CASALONE, E., FERRANTE, D., ASPESI, A., MORLEO, G., BIASI, A., SCULCO, M., MANCUSO, G., GUARRERA, S., RIGHI, L., GROSSO, F., LIBENER, R., PAVESI, M., MARIANI, N., CASADIO, C., BOLDORINI, R., MIRABELLI, D., PASINI, B., MAGNANI, C., MATULLO, G. and DIANZANI, I., 2017. Germline mutations in DNA repair genes predispose asbestos-exposed patients to malignant pleural mesothelioma. *Cancer letters*, **405**, pp. 38-45.

BIANCHI, C. and BIANCHI, T., 2007. Malignant mesothelioma: global incidence and relationship with asbestos. *Industrial health*, **45**(3), pp. 379-387.

BIBBY, A.C., TSIM, S., KANELLAKIS, N., BALL, H., TALBOT, D.C., BLYTH, K.G., MASKELL, N.A. and PSALLIDAS, I., 2016. Malignant pleural mesothelioma: an update on investigation, diagnosis and treatment. *European respiratory review : an official journal of the European Respiratory Society*, **25**(142), pp. 472-486.

BIELSKI, C.,M., ZEHIR, ,AHMET, PENSON, A.,V., DONOGHUE, M.T.,A., CHATILA, ,WALID, ARMENIA, ,JOSHUA, CHANG, M.,T., SCHRAM, A.,M., JONSSON, ,PHILIP, BANDLAMUDI, ,CHAITANYA, RAZAVI, ,PEDRAM, IYER, ,GOPA, ROBSON, M.,E., STADLER, Z.,K., SCHULTZ, ,NIKOLAUS, BASELGA, ,JOSE, SOLIT, D.,B., HYMAN, D.,M., BERGER, M.,F. and TAYLOR, B.,S., 2018. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature Genetics*, 50(8), pp. 1189–1195.

BONOMI, M., DE FILIPPIS, C., LOPCI, E., GIANONCELLI, L., RIZZARDI, G., CERCHIARO, E., BORTOLOTTI, L., ZANELLO, A. and CERESOLI, G.L., 2017. Clinical staging of malignant pleural mesothelioma: current perspectives. *Lung Cancer (Auckland, N.Z.),* **8**, pp. 127-139.

BORCZUK, A.C., PEI, J., TAUB, R.N., LEVY, B., NAHUM, O., CHEN, J., CHEN, K. and TESTA, J.R., 2016. Genome-wide analysis of abdominal and pleural malignant mesothelioma with DNA arrays reveals both common and distinct regions of copy number alteration. *Cancer biology & therapy,* **17**(3), pp. 328-335.

BOTT, M., BREVET, M., TAYLOR, B.S., SHIMIZU, S., ITO, T., WANG, L., CREANEY, J., LAKE, R.A., ZAKOWSKI, M.F., REVA, B., SANDER, C., DELSITE, R., POWELL, S., ZHOU, Q., SHEN, R., OLSHEN, A., RUSCH, V. and LADANYI, M., 2011. The nuclear deubiquitinase BAP1 is commonly inactivated by somatic mutations and 3p21.1 losses in malignant pleural mesothelioma. *Nature genetics,* **43**(7), pp. 668-672.

BREDEL, M., SCHOLTENS, D.M., YADAV, A.K., ALVAREZ, A.A., RENFROW, J.J., CHANDLER, J.P., YU, I.L., CARRO, M.S., DAI, F., TAGGE, M.J., FERRARESE, R., BREDEL, C., PHILLIPS, H.S., LUKAC, P.J., ROBE, P.A., WEYERBROCK, A., VOGEL, H., DUBNER, S., MOBLEY, B., HE, X., SCHECK, A.C., SIKIC, B.I., ALDAPE, K.D., CHAKRAVARTI, A. and HARSH, G.R.,4TH, 2011. NFKBIA deletion in glioblastomas. *The New England journal of medicine,* **364**(7), pp. 627-637.

BUENO, R., STAWISKI, E.W., GOLDSTEIN, L.D., DURINCK, S., DE RIENZO, A., MODRUSAN, Z., GNAD, F., NGUYEN, T.T., JAISWAL, B.S., CHIRIEAC, L.R., SCIARANGHELLA, D., DAO, N., GUSTAFSON, C.E., MUNIR, K.J., HACKNEY, J.A., CHAUDHURI, A., GUPTA, R., GUILLORY, J., TOY, K., HA, C., CHEN, Y., STINSON, J., CHAUDHURI, S., ZHANG, N., WU, T.D., SUGARBAKER, D.J., DE SAUVAGE, F.J., RICHARDS, W.G. and SESHAGIRI, S., 2016. Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nature genetics,* **48**(4), pp. 407-416.

BUENO, R., STAWISKI, E.W., GOLDSTEIN, L.D., DURINCK, S., DE RIENZO, A., MODRUSAN, Z., GNAD, F., NGUYEN, T.T., JAISWAL, B.S., CHIRIEAC, L.R., SCIARANGHELLA, D., DAO, N., GUSTAFSON, C.E., MUNIR, K.J., HACKNEY, J.A., CHAUDHURI, A., GUPTA, R., GUILLORY, J., TOY, K., HA, C., CHEN, Y.J., STINSON, J., CHAUDHURI, S., ZHANG, N., WU, T.D., SUGARBAKER, D.J., DE SAUVAGE, F.J., RICHARDS, W.G. and SESHAGIRI, S., 2016. Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nature genetics,* **48**(4), pp. 407-416.

CAO, C., CROCE, B. and HARRIS, R., 2012. MPM: Malignant Pleural Mesothelioma. *Annals of cardiothoracic surgery,* **1**(4), pp. 544-319X.2012.11.03.

CARAVAGNA, G., GIARRATANO, Y., RAMAZZOTTI, D., TOMLINSON, I., GRAHAM, T.A., SANGUINETTI, G. and SOTTORIVA, A., 2018. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature methods,* **15**(9), pp. 707-714.

CARBONE, M., HARBOUR, J.W., BRUGAROLAS, J., BONONI, A., PAGANO, I., DEY, A., KRAUSZ, T., PASS, H.I., YANG, H. and GAUDINO, G., 2020. Biological Mechanisms and Clinical Significance of BAP1 Mutations in Human Cancer. *Cancer*

discovery, **10**(8), pp. 1103-1120.

CARBONE, M., PASS, H.I., RIZZO, P., MARINETTI, M., DI MUZIO, M., MEW, D.J., LEVINE, A.S. and PROCOPIO, A., 1994. Simian virus 40-like DNA sequences in human pleural mesothelioma. *Oncogene,* **9**(6), pp. 1781-1790.

CARTER, S.L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P.W., ONOFRIO, R.C., WINCKLER, W., WEIR, B.A., BEROUKHIM, R., PELLMAN, D., LEVINE, D.A., LANDER, E.S., MEYERSON, M. and GETZ, G., 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology,* **30**(5), pp. 413-421.

CASULA, M., PALIOGIANNIS, P., AYALA, F., DE GIORGI, V., STANGANELLI, I., MANDALÀ, M., COLOMBINO, M., MANCA, A., SINI, M.C., CARACÒ, C., ASCIERTO, P.A., SATTA, R.R., MELANOMA UNIT OF SASSARI (MUS), LISSIA, A., COSSU, A., PALMIERI, G. and ITALIAN MELANOMA INTERGROUP (IMI), 2019. Germline and somatic mutations in patients with multiple primary melanomas: a next generation sequencing study. *BMC cancer,* **19**(1), pp. 772-019-5984-7.

CERAMI, E., GAO, J., DOGRUSOZ, U., GROSS, B.E., SUMER, S.O., AKSOY, B.A., JACOBSEN, A., BYRNE, C.J., HEUER, M.L., LARSSON, E., ANTIPIN, Y., REVA, B., GOLDBERG, A.P., SANDER, C. and SCHULTZ, N., 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery,* **2**(5), pp. 401-404.

CHEN, F., ZHANG, Y. and CREIGHTON, C.J., 2021. Systematic identification of non-coding somatic single nucleotide variants associated with altered transcription and DNA methylation in adult and pediatric cancers. *NAR cancer,* **3**(1), pp. zcab001.

CHEN, R., LEE, W.C., FUJIMOTO, J., LI, J., HU, X., MEHRAN, R., RICE, D., SWISHER, S.G., SEPESI, B., TRAN, H.T., CHOW, C.W., LITTLE, L.D., GUMBS, C., HAYMAKER, C., HEYMACH, J.V., WISTUBA, I.I., LEE, J.J., FUTREAL, P.A., ZHANG, J., REUBEN, A., TSAO, A.S. and ZHANG, J., 2020. Evolution of Genomic and T-cell Repertoire Heterogeneity of Malignant Pleural Mesothelioma Under Dasatinib Treatment. *Clinical cancer research : an official journal of the American Association for Cancer Research,* **26**(20), pp. 5477-5486.

CHOI, Y., LIM, S. and PAEK, D., 2013. Trades of dangers: a study of asbestos industry transfer cases in Asia. *American Journal of Industrial Medicine,* **56**(3), pp. 335-346.

CHURCH, D.M., SCHNEIDER, V.A., GRAVES, T., AUGER, K., CUNNINGHAM, F., BOUK, N., CHEN, H.C., AGARWALA, R., MCLAREN, W.M., RITCHIE, G.R., ALBRACHT, D., KREMITZKI, M., ROCK, S., KOTKIEWICZ, H., KREMITZKI, C., WOLLAM, A., TRANI, L., FULTON, L., FULTON, R., MATTHEWS, L., WHITEHEAD, S., CHOW, W., TORRANCE, J., DUNN, M., HARDEN, G., THREADGOLD, G., WOOD, J., COLLINS, J., HEATH, P., GRIFFITHS, G., PELAN, S., GRAFHAM, D., EICHLER, E.E., WEINSTOCK, G., MARDIS, E.R., WILSON, R.K., HOWE, K., FLICEK, P. and HUBBARD, T., 2011. Modernizing reference genome assemblies. *PLoS biology,* **9**(7), pp. e1001091.

CIBULSKIS, K., LAWRENCE, M.S., CARTER, S.L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E.S. and GETZ, G., 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer

samples. *Nature biotechnology,* **31**(3), pp. 213-219.

CICALA, C., POMPETTI, F. and CARBONE, M., 1993. SV40 induces mesotheliomas in hamsters. *The American Journal of Pathology,* **142**(5), pp. 1524-1533.

COLLINS, D.,CATHERINE, SUNDAR, ,RAGHAV, CONSTANTINIDOU, ,ANASTASIA, DOLLING, ,DAVID, YAP, T.,ANTHONY, POPAT, ,SANJAY, O'BRIEN, M.,E., BANERJI, ,UDAI, DE BONO, J.,SEBASTIAN, LOPEZ, J.,SUZANNE, TUNARIU, ,NINA and MINCHOM, ,ANNA, 2020. Radiological evaluation of malignant pleural mesothelioma - defining distant metastatic disease. *BMC Cancer,* 20(1).

CURTIS, C., SHAH, S.P., CHIN, S.F., TURASHVILI, G., RUEDA, O.M., DUNNING, M.J., SPEED, D., LYNCH, A.G., SAMARAJIWA, S., YUAN, Y., GRAF, S., HA, G., HAFFARI, G., BASHASHATI, A., RUSSELL, R., MCKINNEY, S., METABRIC GROUP, LANGEROD, A., GREEN, A., PROVENZANO, E., WISHART, G., PINDER, S., WATSON, P., MARKOWETZ, F., MURPHY, L., ELLIS, I., PURUSHOTHAM, A., BORRESEN-DALE, A.L., BRENTON, J.D., TAVARE, S., CALDAS, C. and APARICIO, S., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature,* **486**(7403), pp. 346-352.

DE ASSIS, L.V. and ISOLDI, M.C., 2014. The function, mechanisms, and role of the genes PTEN and TP53 and the effects of asbestos in the development of malignant mesothelioma: a review focused on the genes' molecular mechanisms. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine,* **35**(2), pp. 889-901.

DENG, N., ZHOU, H., FAN, H. and YUAN, Y., 2017. Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget,* **8**(66), pp. 110635-110649.

DESHWAR, A.G., VEMBU, S., YUNG, C.K., JANG, G.H., STEIN, L. and MORRIS, Q., 2015. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology,* **16**(1), pp. 35-015-0602-8.

DESPER, R., JIANG, F., KALLIONIEMI, O.P., MOCH, H., PAPADIMITRIOU, C.H. and SCHÄFFER, A.A., 1999. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of computational biology : a journal of computational molecular cell biology,* **6**(1), pp. 37-51.

DUIJF, ,P.H.G. and BENEZRA, ,R., 2013. The cancer biology of whole-chromosome instability. *Oncogene,* 32, pp. 4727-4736.

FAHEY, C.C. and DAVIS, I.J., 2017. SETting the Stage for Cancer Development: SETD2 and the Consequences of Lost Methylation. *Cold Spring Harbor perspectives in medicine,* **7**(5), pp. a026468. doi: 10.1101/cshperspect.a026468.

FAVERO, F., JOSHI, T., MARQUARD, A.M., BIRKBAK, N.J., KRZYSTANEK, M., LI, Q., SZALLASI, Z. and EKLUND, A.C., 2015. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology,* **26**(1), pp. 64-70.

FORBES, S., BHAMRA, G., BAMFORD, S., DAWSON, E., KOK, C., CLEMENTS, J., MENZIES, A., TEAGUE, J., FUTREAL, P. and STRATTON, M., 2008. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human*

genetics / editorial board, Jonathan L.Haines ...[et al.], **CHAPTER**, pp. Unit-10.11.

FRANK, A.L. and JOSHI, T.K., 2014. The global spread of asbestos. *Annals of global health,* **80**(4), pp. 257-262.

FRANKO, A., KOTNIK, N., GORICAR, K., KOVAC, V., DODIC-FIKFAK, M. and DOLZAN, V., 2018. The Influence of Genetic Variability on the Risk of Developing Malignant Mesothelioma. *Radiology and Oncology,* **52**(1), pp. 105-111.

FUKAZAWA, T., MATSUOKA, J., NAOMOTO, Y., MAEDA, Y., DURBIN, M.L. and TANAKA, N., 2008. Malignant pleural mesothelioma-targeted CREBBP/EP300 inhibitory protein 1 promoter system for gene therapy and virotherapy. *Cancer research,* **68**(17), pp. 7120-7129.

FURUKAWA, M., TOYOOKA, S., HAYASHI, T., YAMAMOTO, H., FUJIMOTO, N., SOH, J., HASHIDA, S., SHIEN, K., ASANO, H., AOE, K., OKABE, K., PASS, H.I., TSUKUDA, K., KISHIMOTO, T. and MIYOSHI, S., 2015. DNA copy number gains in malignant pleural mesothelioma. *Oncology Letters,* **10**(5), pp. 3274-3278.

GAN, K.A., CARRASCO PRO, S., SEWELL, J.A. and FUXMAN BASS, J.I., 2018. Identification of Single Nucleotide Non-coding Driver Mutations in Cancer. *Frontiers in genetics,* **9**, pp. 16.

GANSNER, E. and NORTH, S., 2000. An open graph visualization system and its applications to software engineering. *Software- Practice and Experience,* **30**(11), pp. 1203-1233.

GAO, J., AKSOY, B.A., DOGRUSOZ, U., DRESDNER, G., GROSS, B., SUMER, S.O., SUN, Y., JACOBSEN, A., SINHA, R., LARSSON, E., CERAMI, E., SANDER, C. and SCHULTZ, N., 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling,* **6**(269), pp. pl1.

GERLINGER, M., HORSWELL, S., LARKIN, J., ROWAN, A.J., SALM, M.P., VARELA, I., FISHER, R., MCGRANAHAN, N., MATTHEWS, N., SANTOS, C.R., MARTINEZ, P., PHILLIMORE, B., BEGUM, S., RABINOWITZ, A., SPENCER-DENE, B., GULATI, S., BATES, P.A., STAMP, G., PICKERING, L., GORE, M., NICOL, D.L., HAZELL, S., FUTREAL, P.A., STEWART, A. and SWANTON, C., 2014. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature genetics,* **46**(3), pp. 225-233.

GOLDSTEIN, A.M., CHAN, M., HARLAND, M., GILLANDERS, E.M., HAYWARD, N.K., AVRIL, M., AZIZI, E., BIANCHI-SCARRA, G., BISHOP, D.T., BRESSAC-DE PAILLERETS, B., BRUNO, W., CALISTA, D., CANNON ALBRIGHT, L.A., DEMENAIS, F., ELDER, D.E., GHIORZO, P., GRUIS, N.A., HANSSON, J., HOGG, D., HOLLAND, E.A., KANETSKY, P.A., KEFFORD, R.F., LANDI, M.T., LANG, J., LEACHMAN, S.A., MACKIE, R.M., MAGNUSSON, V., MANN, G.J., NIENDORF, K., NEWTON BISHOP, J., PALMER, J.M., PUIG, S., PUIG-BUTILLE, J.A., DE SNOO, F.A., STARK, M., TSAO, H., TUCKER, M.A., WHITAKER, L., YAKOBSON, E. and MELANOMA GENETICS CONSORTIUM (GENOMEL), 2006. High-risk melanoma susceptibility genes and pancreatic cancer, neural system tumors, and uveal melanoma across GenoMEL. *Cancer research,* **66**(20), pp. 9818-9828.

GREENMAN, C., STEPHENS, P., SMITH, R., DALGLIESH, G.L., HUNTER, C., BIGNELL, G., DAVIES, H., TEAGUE, J., BUTLER, A., STEVENS, C., EDKINS, S.,

O'MEARA, S., VASTRIK, I., SCHMIDT, E.E., AVIS, T., BARTHORPE, S., BHAMRA, G., BUCK, G., CHOUDHURY, B., CLEMENTS, J., COLE, J., DICKS, E., FORBES, S., GRAY, K., HALLIDAY, K., HARRISON, R., HILLS, K., HINTON, J., JENKINSON, A., JONES, D., MENZIES, A., MIRONENKO, T., PERRY, J., RAINE, K., RICHARDSON, D., SHEPHERD, R., SMALL, A., TOFTS, C., VARIAN, J., WEBB, T., WEST, S., WIDAA, S., YATES, A., CAHILL, D.P., LOUIS, D.N., GOLDSTRAW, P., NICHOLSON, A.G., BRASSEUR, F., LOOIJENGA, L., WEBER, B.L., CHIEW, Y.E., DEFAZIO, A., GREAVES, M.F., GREEN, A.R., CAMPBELL, P., BIRNEY, E., EASTON, D.F., CHENEVIX-TRENCH, G., TAN, M.H., KHOO, S.K., TEH, B.T., YUEN, S.T., LEUNG, S.Y., WOOSTER, R., FUTREAL, P.A. and STRATTON, M.R., 2007. Patterns of somatic mutation in human cancer genomes. *Nature,* **446**(7132), pp. 153-158.

GUALTIERI, A.F., 2021. Bridging the gap between toxicity and carcinogenicity of mineral fibres by connecting the fibre crystal-chemical and physical parameters to the key characteristics of cancer. *Current research in toxicology,* **2**, pp. 42-52.

HANSEN, N.F., GARTNER, J.J., MEI, L., SAMUELS, Y. and MULLIKIN, J.C., 2013. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics (Oxford, England),* **29**(12), pp. 1498-1503.

HASSAN, R., MORROW, B., THOMAS, A., WALSH, T., LEE, M.K., GULSUNER, S., GADIRAJU, M., PANOU, V., GAO, S., MIAN, I., KHAN, J., RAFFELD, M., PATEL, S., XI, L., WEI, J.S., HESDORFFER, M., ZHANG, J., CALZONE, K., DESAI, A., PADIERNOS, E., ALEWINE, C., SCHRUMP, D.S., STEINBERG, S.M., KINDLER, H.L., KING, M.C. and CHURPEK, J.E., 2019. Inherited predisposition to malignant mesothelioma and overall survival following platinum chemotherapy. *Proceedings of the National Academy of Sciences of the United States of America,* **116**(18), pp. 9008-9013.

HE, S. and SHARPLESS, N.E., 2017. Senescence in Health and Disease. *Cell,* **169**(6), pp. 1000-1011.

HENRICH, K.O., SCHWAB, M. and WESTERMANN, F., 2012. 1p36 Tumor Suppression--a Matter of Dosage? *Cancer research,* **72**(23), pp. 6079-6088.

HIDA, T., HAMASAKI, M., MATSUMOTO, S., SATO, A., TSUJIMURA, T., KAWAHARA, K., IWASAKI, A., OKAMOTO, T., ODA, Y., HONDA, H. and NABESHIMA, K., 2017. Immunohistochemical detection of MTAP and BAP1 protein loss for mesothelioma diagnosis: Comparison with 9p21 FISH and BAP1 immunohistochemistry. *Lung cancer (Amsterdam, Netherlands),* **104**, pp. 98-105.

HMELJAK, J., SANCHEZ-VEGA, F., HOADLEY, K.A., SHIH, J., STEWART, C., HEIMAN, D., TARPEY, P., DANILOVA, L., DRILL, E., GIBB, E.A., BOWLBY, R., KANCHI, R., OSMANBEYOGLU, H.U., SEKIDO, Y., TAKESHITA, J., NEWTON, Y., GRAIM, K., GUPTA, M., GAY, C.M., DIAO, L., GIBBS, D.L., THORSSON, V., IYPE, L., KANTHETI, H., SEVERSON, D.T., RAVEGNINI, G., DESMEULES, P., JUNGBLUTH, A.A., TRAVIS, W.D., DACIC, S., CHIRIEAC, L.R., GALATEAU-SALLÉ, F., FUJIMOTO, J., HUSAIN, A.N., SILVEIRA, H.C., RUSCH, V.W., RINTOUL, R.C., PASS, H., KINDLER, H., ZAUDERER, M.G., KWIATKOWSKI, D.J., BUENO, R., TSAO, A.S., CREANEY, J., LICHTENBERG, T., LERAAS, K., BOWEN, J., TCGA RESEARCH NETWORK, FELAU, I., ZENKLUSEN, J.C., AKBANI, R., CHERNIACK, A.D., BYERS, L.A., NOBLE, M.S., FLETCHER, J.A.,

ROBERTSON, A.G., SHEN, R., ABURATANI, H., ROBINSON, B.W., CAMPBELL, P. and LADANYI, M., 2018. Integrative Molecular Characterization of Malignant Pleural Mesothelioma. *Cancer discovery,* **8**(12), pp. 1548-1565.

HMELJAK, J., SANCHEZ-VEGA, F., HOADLEY, K.A., SHIH, J., STEWART, C., HEIMAN, D., TARPEY, P., DANILOVA, L., DRILL, E., GIBB, E.A., BOWLBY, R., KANCHI, R., OSMANBEYOGLU, H.U., SEKIDO, Y., TAKESHITA, J., NEWTON, Y., GRAIM, K., GUPTA, M., GAY, C.M., DIAO, L., GIBBS, D.L., THORSSON, V., IYPE, L., KANTHETI, H., SEVERSON, D.T., RAVEGNINI, G., DESMEULES, P., JUNGBLUTH, A.A., TRAVIS, W.D., DACIC, S., CHIRIEAC, L.R., GALATEAU-SALLÉ, F., FUJIMOTO, J., HUSAIN, A.N., SILVEIRA, H.C., RUSCH, V.W., RINTOUL, R.C., PASS, H., KINDLER, H., ZAUDERER, M.G., KWIATKOWSKI, D.J., BUENO, R., TSAO, A.S., CREANEY, J., LICHTENBERG, T., LERAAS, K., BOWEN, J., TCGA RESEARCH NETWORK, FELAU, I., ZENKLUSEN, J.C., AKBANI, R., CHERNIACK, A.D., BYERS, L.A., NOBLE, M.S., FLETCHER, J.A., ROBERTSON, A.G., SHEN, R., ABURATANI, H., ROBINSON, B.W., CAMPBELL, P. and LADANYI, M., 2018. Integrative Molecular Characterization of Malignant Pleural Mesothelioma. *Cancer discovery,* **8**(12), pp. 1548-1565.

HU, X., ESTECIO, M.R., CHEN, R., REUBEN, A., WANG, L., FUJIMOTO, J., CARROT-ZHANG, J., MCGRANAHAN, N., YING, L., FUKUOKA, J., CHOW, C.W., PHAM, H.H.N., GODOY, M.C.B., CARTER, B.W., BEHRENS, C., ZHANG, J., ANTONOFF, M.B., SEPESI, B., LU, Y., PASS, H.I., KADARA, H., SCHEET, P., VAPORCIYAN, A.A., HEYMACH, J.V., WISTUBA, I.I., LEE, J.J., FUTREAL, P.A., SU, D., ISSA, J.J. and ZHANG, J., 2021. Evolution of DNA methylome from precancerous lesions to invasive lung adenocarcinomas. *Nature communications,* **12**(1), pp. 687-021-20907-z.

HYLAND, R.A., WARE, S., JOHNSON, A.R. and YATES, D.H., 2007. Incidence trends and gender differences in malignant mesothelioma in New South Wales, Australia. *Scandinavian journal of work, environment & health,* **33**(4), pp. 286-292.

HYLEBOS, M., VAN CAMP, G., VAN MEERBEECK, J.P. and OP DE BEECK, K., 2016. The Genetic Landscape of Malignant Pleural Mesothelioma: Results from Massively Parallel Sequencing. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer,* **11**(10), pp. 1615-1626.

HYLEBOS, M., VAN CAMP, G., VANDEWEYER, G., FRANSEN, E., BEYENS, M., CORNELISSEN, R., SULS, A., PAUWELS, P., VAN MEERBEECK, J.P. and OP DE BEECK, K., 2017. Large-scale copy number analysis reveals variations in genes not previously associated with malignant pleural mesothelioma. *Oncotarget,* **8**(69), pp. 113673-113686.

INAI, K., 2008. Pathology of mesothelioma. *Environmental health and preventive medicine,* **13**(2), pp. 60-64.

JAMAL-HANJANI, M., WILSON, G.A., MCGRANAHAN, N., BIRKBAK, N.J., WATKINS, T.B.K., VEERIAH, S., SHAFI, S., JOHNSON, D.H., MITTER, R., ROSENTHAL, R., SALM, M., HORSWELL, S., ESCUDERO, M., MATTHEWS, N., ROWAN, A., CHAMBERS, T., MOORE, D.A., TURAJLIC, S., XU, H., LEE, S.M., FORSTER, M.D., AHMAD, T., HILEY, C.T., ABBOSH, C., FALZON, M., BORG, E., MARAFIOTI, T., LAWRENCE, D., HAYWARD, M., KOLVEKAR, S., PANAGIOTOPOULOS, N., JANES, S.M., THAKRAR, R., AHMED, A.,

BLACKHALL, F., SUMMERS, Y., SHAH, R., JOSEPH, L., QUINN, A.M., CROSBIE, P.A., NAIDU, B., MIDDLETON, G., LANGMAN, G., TROTTER, S., NICOLSON, M., REMMEN, H., KERR, K., CHETTY, M., GOMERSALL, L., FENNELL, D.A., NAKAS, A., RATHINAM, S., ANAND, G., KHAN, S., RUSSELL, P., EZHIL, V., ISMAIL, B., IRVIN-SELLERS, M., PRAKASH, V., LESTER, J.F., KORNASZEWSKA, M., ATTANOOS, R., ADAMS, H., DAVIES, H., DENTRO, S., TANIERE, P., O'SULLIVAN, B., LOWE, H.L., HARTLEY, J.A., ILES, N., BELL, H., NGAI, Y., SHAW, J.A., HERRERO, J., SZALLASI, Z., SCHWARZ, R.F., STEWART, A., QUEZADA, S.A., LE QUESNE, J., VAN LOO, P., DIVE, C., HACKSHAW, A., SWANTON, C. and TRACERX CONSORTIUM, 2017. Tracking the Evolution of Non-Small-Cell Lung Cancer. *The New England journal of medicine*, **376**(22), pp. 2109-2121.

JAURAND MC and FLEURY-FEITH J, Pathogenesis of malignant pleural mesothelioma. - *Respirology.2005 Jan;10(1):2-8.doi: 10.1111/j.1440-1843.2005.00694.x.*, (1323-7799 (Print); 1323-7799 (Linking)),.

JIAO, W., VEMBU, S., DESHWAR, A.G., STEIN, L. and MORRIS, Q., 2014. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, **15**, pp. 35-2105-15-35.

KANNAN, L. and WHEELER, W.C., 2012. Maximum Parsimony on Phylogenetic networks. *Algorithms for Molecular Biology : AMB*, **7**, pp. 9-7188-7-9.

KATO, S., TOMSON, B.N., BUYS, T.P., ELKIN, S.K., CARTER, J.L. and KURZROCK, R., 2016. Genomic Landscape of Malignant Mesotheliomas. *Molecular cancer therapeutics*, **15**(10), pp. 2498-2507.

KENT, W.J., ZWEIG, A.S., BARBER, G., HINRICHS, A.S. and KAROLCHIK, D., 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics (Oxford, England)*, **26**(17), pp. 2204-2207.

KETTUNEN, ,EEVA, SAVUKOSKI, ,SAULI, SALMENKIVI, ,KAISA, BÖHLING, ,TOM, VANHALA, ,ESA, KUOSMA, ,EEVA, ANTTILA, ,SISKO and WOLFF, ,HENRIK, 2019, CDKN2A copy number and p16 expression in malignant pleural mesothelioma in relation to asbestos exposure. *BMC Cancer*, 19(1).

KIDD, J.M., COOPER, G.M., DONAHUE, W.F., HAYDEN, H.S., SAMPAS, N., GRAVES, T., HANSEN, N., TEAGUE, B., ALKAN, C., ANTONACCI, F., HAUGEN, E., ZERR, T., YAMADA, N.A., TSANG, P., NEWMAN, T.L., TÜZÜN, E., CHENG, Z., EBLING, H.M., TUSNEEM, N., DAVID, R., GILLETT, W., PHELPS, K.A., WEAVER, M., SARANGA, D., BRAND, A., TAO, W., GUSTAFSON, E., MCKERNAN, K., CHEN, L., MALIG, M., SMITH, J.D., KORN, J.M., MCCARROLL, S.A., ALTSHULER, D.A., PEIFFER, D.A., DORSCHNER, M., STAMATOYANNOPOULOS, J., SCHWARTZ, D., NICKERSON, D.A., MULLIKIN, J.C., WILSON, R.K., BRUHN, L., OLSON, M.V., KAUL, R., SMITH, D.R. and EICHLER, E.E., 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**(7191), pp. 56-64.

KOBOLDT, D.C., ZHANG, Q., LARSON, D.E., SHEN, D., MCLELLAN, M.D., LIN, L., MILLER, C.A., MARDIS, E.R., DING, L. and WILSON, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**(3), pp. 568-576.

KRAUSE, A., ROMA, L., LORBER, T., DIETSCHE, T., PERRINA, V., MÜLLER, D.C., LARDINOIS, D., RUIZ, C., SAVIC PRINCE, S., PISCUOGLIO, S., NG, C.K.Y. and BUBENDORF, L., 2021. Genomic evolutionary trajectory of metastatic squamous cell carcinoma of the lung. *Translational lung cancer research,* **10**(4), pp. 1792-1803.

KROCZYNSKA, B., CUTRONE, R., BOCCHETTA, M., YANG, H., ELMISHAD, A.G., VACEK, P., RAMOS-NINO, M., MOSSMAN, B.T., PASS, H.I. and CARBONE, M., 2006. Crocidolite asbestos and SV40 are cocarcinogens in human mesothelial cells and in causing mesothelioma in hamsters. *Proceedings of the National Academy of Sciences of the United States of America,* **103**(38), pp. 14128-14133.

KUHNER, M.K. and FELSENSTEIN, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution,* **11**(3), pp. 459-468.

LAI, Z., MARKOVETS, A., AHDESMAKI, M., CHAPMAN, B., HOFMANN, O., MCEWEN, R., JOHNSON, J., DOUGHERTY, B., BARRETT, J.C. and DRY, J.R., 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic acids research,* **44**(11), pp. e108.

LARSON, N.B. and FRIDLEY, B.L., 2013. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics (Oxford, England),* **29**(15), pp. 1888-1889.

LAWRENCE, M.S., STOJANOV, P., MERMEL, C.H., ROBINSON, J.T., GARRAWAY, L.A., GOLUB, T.R., MEYERSON, M., GABRIEL, S.B., LANDER, E.S. and GETZ, G., 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature,* **505**(7484), pp. 495-501.

LEONG, S.L., ZAINUDIN, R., KAZAN-ALLEN, L. and ROBINSON, B.W., 2015. Asbestos in Asia. *Respirology (Carlton, Vic.),* **20**(4), pp. 548-555.

LETOUZE, E., ALLORY, Y., BOLLET, M.A., RADVANYI, F. and GUYON, F., 2010. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome biology,* **11**(7), pp. R76-2010-11-7-r76. Epub 2010 Jul 22.

LI, H. and DURBIN, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England),* **25**(14), pp. 1754-1760.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. and 1000 GENOME PROJECT DATA PROCESSING SUBGROUP, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England),* **25**(16), pp. 2078-2079.

LI, Y. and XIE, X., 2014. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. *Bioinformatics (Oxford, England),* **30**(15), pp. 2121-2129.

LINDHOLM, P.M., SALMENKIVI, K., VAUHKONEN, H., NICHOLSON, A.G., ANTTILA, S., KINNULA, V.L. and KNUUTILA, S., 2007. Gene copy number analysis in malignant pleural mesothelioma using oligonucleotide array CGH. *Cytogenetic and genome research,* **119**(1-2), pp. 46-52.

LIU, G., CHERESH, P. and KAMP, D.W., 2013. Molecular basis of asbestos-induced lung disease. *Annual review of pathology,* **8**, pp. 161-187.

LÓPEZ, S., LIM, E.L., HORSWELL, S., HAASE, K., HUEBNER, A., DIETZEN, M., MOURIKIS, T.P., WATKINS, T.B.K., ROWAN, A., DEWHURST, S.M., BIRKBAK, N.J., WILSON, G.A., VAN LOO, P., JAMAL-HANJANI, M., TRACERX CONSORTIUM, SWANTON, C. and MCGRANAHAN, N., 2020. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nature genetics,* **52**(3), pp. 283-293.

LUPSKI, J.R., 2007. Genomic rearrangements and sporadic disease. *Nature genetics,* **39**(7 Suppl), pp. S43-7.

MARINACCIO, A., BINAZZI, A., CAUZILLO, G., CAVONE, D., ZOTTI, R.D., FERRANTE, P., GENNARO, V., GORINI, G., MENEGOZZO, M., MENSI, C., MERLER, E., MIRABELLI, D., MONTANARO, F., MUSTI, M., PANNELLI, F., ROMANELLI, A., SCARSELLI, A., TUMINO, R. and ITALIAN MESOTHELIOMA REGISTER (RENAM) WORKING GROUP, 2007. Analysis of latency time and its determinants in asbestos related malignant mesothelioma cases of the Italian register. *European journal of cancer (Oxford, England : 1990),* **43**(18), pp. 2722-2728.

MARTINCORENA, I. and CAMPBELL, P.J., 2015. Somatic mutation in cancer and normal cells. *Science (New York, N.Y.),* **349**(6255), pp. 1483-1489.

MELAIU, O., GEMIGNANI, F. and LANDI, S., 2018. The genetic susceptibility in the development of malignant pleural mesothelioma. *Journal of thoracic disease,* **10**(Suppl 2), pp. S246-S252.

MENSI, C., DE MATTEIS, S., DALLARI, B., RIBOLDI, L., BERTAZZI, P.A. and CONSONNI, D., 2016. Incidence of mesothelioma in Lombardy, Italy: exposure to asbestos, time patterns and future projections. *Occupational and environmental medicine,* **73**(9), pp. 607-613.

MERLO, L.M., PEPPER, J.W., REID, B.J. and MALEY, C.C., 2006. Cancer as an evolutionary and ecological process. *Nature reviews.Cancer,* **6**(12), pp. 924-935.

MEYERSON, M., GABRIEL, S. and GETZ, G., 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature reviews.Genetics,* **11**(10), pp. 685-696.

MIYANAGA, A., MASUDA, M., TSUTA, K., KAWASAKI, K., NAKAMURA, Y., SAKUMA, T., ASAMURA, H., GEMMA, A. and YAMADA, T., 2015. Hippo pathway gene mutations in malignant mesothelioma: revealed by RNA and targeted exon sequencing. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer,* **10**(5), pp. 844-851.

MOTT, F.E., 2012. Mesothelioma: A Review. *The Ochsner Journal,* **12**(1), pp. 70-79.

MURAKAMI, H., MIZUNO, T., TANIGUCHI, T., FUJII, M., ISHIGURO, F., FUKUI, T., AKATSUKA, S., HORIO, Y., HIDA, T., KONDO, Y., TOYOKUNI, S., OSADA, H. and SEKIDO, Y., 2011. LATS2 is a tumor suppressor gene of malignant mesothelioma. *Cancer research,* **71**(3), pp. 873-883.

MURTHY, S.S. and TESTA, J.R., 1999. Asbestos, chromosomal deletions, and tumor

suppressor gene alterations in human malignant mesothelioma. *Journal of cellular physiology,* **180**(2), pp. 150-157.

NIK-ZAINAL, S., VAN LOO, P., WEDGE, D.C., ALEXANDROV, L.B., GREENMAN, C.D., LAU, K.W., RAINE, K., JONES, D., MARSHALL, J., RAMAKRISHNA, M., SHLIEN, A., COOKE, S.L., HINTON, J., MENZIES, A., STEBBINGS, L.A., LEROY, C., JIA, M., RANCE, R., MUDIE, L.J., GAMBLE, S.J., STEPHENS, P.J., MCLAREN, S., TARPEY, P.S., PAPAEMMANUIL, E., DAVIES, H.R., VARELA, I., MCBRIDE, D.J., BIGNELL, G.R., LEUNG, K., BUTLER, A.P., TEAGUE, J.W., MARTIN, S., JÖNSSON, G., MARIANI, O., BOYAULT, S., MIRON, P., FATIMA, A., LANGERØD, A., APARICIO, S.A., TUTT, A., SIEUWERTS, A.M., BORG, Å., THOMAS, G., SALOMON, A.V., RICHARDSON, A.L., BØRRESEN-DALE, A.L., FUTREAL, P.A., STRATTON, M.R., CAMPBELL, P.J. and BREAST CANCER WORKING GROUP OF THE INTERNATIONAL CANCER GENOME CONSORTIUM, 2012. The life history of 21 breast cancers. *Cell,* **149**(5), pp. 994-1007.

PANOU, V., GADIRAJU, M., WOLIN, A., WEIPERT, C.M., SKARDA, E., HUSAIN, A.N., PATEL, J.D., ROSE, B., ZHANG, S.R., WEATHERLY, M., NELAKUDITI, V., KNIGHT JOHNSON, A., HELGESON, M., FISCHER, D., DESAI, A., SULAI, N., RITTERHOUSE, L., RØE, O.D., TURAGA, K.K., HUO, D., SEGAL, J., KADRI, S., LI, Z., KINDLER, H.L. and CHURPEK, J.E., 2018. Frequency of Germline Mutations in Cancer Susceptibility Genes in Malignant Mesothelioma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* **36**(28), pp. 2863-2871.

PERSHOUSE, M.A., HEIVLY, S. and GIRTSMAN, T., 2006. The role of SV40 in malignant mesothelioma and other human malignancies. *Inhalation toxicology,* **18**(12), pp. 995-1000.

PETRILLI, A.M. and FERNÁNDEZ-VALLE, C., 2016. Role of Merlin/NF2 inactivation in tumor biology. *Oncogene,* **35**(5), pp. 537-548.

PIETROFESA, R.A., VELALOPOULOU, A., ALBELDA, S.M. and CHRISTOFIDOU-SOLOMIDOU, M., 2016. Asbestos Induces Oxidative Stress and Activation of Nrf2 Signaling in Murine Macrophages: Chemopreventive Role of the Synthetic Lignan Secoisolariciresinol Diglucoside (LGM2605). *International journal of molecular sciences,* **17**(3), pp. 322.

PIRA, E., DONATO, F., MAIDA, L. and DISCALZI, G., 2018. Exposure to asbestos: past, present and future. *Journal of thoracic disease,* **10**(Suppl 2), pp. S237-S245.

PLATO, N., MARTINSEN, J.I., SPAREN, P., HILLERDAL, G. and WEIDERPASS, E., 2016. Occupation and mesothelioma in Sweden: updated incidence in men and women in the 27 years after the asbestos ban. *Epidemiology and health,* **38**, pp. e2016039.

POMERANTZ, M.M., SHRESTHA, Y., FLAVIN, R.J., REGAN, M.M., PENNEY, K.L., MUCCI, L.A., STAMPFER, M.J., HUNTER, D.J., CHANOCK, S.J., SCHAFER, E.J., CHAN, J.A., TABERNERO, J., BASELGA, J., RICHARDSON, A.L., LODA, M., OH, W.K., KANTOFF, P.W., HAHN, W.C. and FREEDMAN, M.L., 2010. Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. *PLoS genetics,* **6**(11), pp. e1001204.

PRINS, J.B., WILLIAMSON, K.A., KAMP, M.M., VAN HEZIK, E.J., VAN DER

KWAST, T.H., HAGEMEIJER, A. and VERSNEL, M.A., 1998. The gene for the cyclin-dependent-kinase-4 inhibitor, CDKN2A, is preferentially deleted in malignant mesothelioma. *International journal of cancer,* **75**(4), pp. 649-653.

PRINS, J.B., WILLIAMSON, K.A., KAMP, M.M., VAN HEZIK, E.J., VAN DER KWAST, T.H., HAGEMEIJER, A. and VERSNEL, M.A., 1998. The gene for the cyclin-dependent-kinase-4 inhibitor, CDKN2A, is preferentially deleted in malignant mesothelioma. *International journal of cancer,* **75**(4), pp. 649-653.

QUETEL, L., MEILLER, C., ASSIÉ, J.B., BLUM, Y., IMBEAUD, S., MONTAGNE, F., TRANCHANT, R., DE WOLF, J., CARUSO, S., COPIN, M.C., HOFMAN, V., GIBAULT, L., BADOUAL, C., PINTILIE, E., HOFMAN, P., MONNET, I., SCHERPEREEL, A., LE PIMPEC-BARTHES, F., ZUCMAN-ROSSI, J., JAURAND, M.C. and JEAN, D., 2020. Genetic alterations of malignant pleural mesothelioma: association with tumor heterogeneity and overall survival. *Molecular oncology,* **14**(6), pp. 1207-1223.

RABBANI, B., TEKIN, M. and MAHDIEH, N., 2014. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics,* **59**(1), pp. 5-15.

RAINE, K.M., VAN LOO, P., WEDGE, D.C., JONES, D., MENZIES, A., BUTLER, A.P., TEAGUE, J.W., TARPEY, P., NIK-ZAINAL, S. and CAMPBELL, P.J., 2016. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current protocols in bioinformatics,* **56**, pp. 15.9.1-15.9.17.

ROBINSON, B.M., 2012. Malignant pleural mesothelioma: an epidemiological perspective. *Annals of Cardiothoracic Surgery,* **1**(4), pp. 491-496.

ROBINSON, J.T., THORVALDSDÓTTIR, H., WENGER, A.M., ZEHIR, A. and MESIROV, J.P., 2017. Variant Review with the Integrative Genomics Viewer. *Cancer research,* **77**(21), pp. e31-e34.

ROBINSON, J.T., THORVALDSDÓTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E.S., GETZ, G. and MESIROV, J.P., 2011. Integrative genomics viewer. *Nature biotechnology,* **29**(1), pp. 24-26.

RODRIK-OUTMEZGUINE, V.S., OKANIWA, M., YAO, Z., NOVOTNY, C.J., MCWHIRTER, C., BANAJI, A., WON, H., WONG, W., BERGER, M., DE STANCHINA, E., BARRATT, D.G., COSULICH, S., KLINOWSKA, T., ROSEN, N. and SHOKAT, K.M., 2016. Overcoming mTOR resistance mutations with a new-generation mTOR inhibitor. *Nature,* **534**(7606), pp. 272-276.

ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. and SHAH, S.P., 2014. PyClone: statistical inference of clonal population structure in cancer. *Nature methods,* **11**(4), pp. 396-398.

SARKAR, A.A. and ZOHN, I.E., 2012. Hectd1 regulates intracellular localization and secretion of Hsp90 to control cellular behavior of the cranial mesenchyme. *The Journal of cell biology,* **196**(6), pp. 789-800.

SATHIRAPONGSASUTI, J.F., LEE, H., HORST, B.A., BRUNNER, G., COCHRAN, A.J., BINDER, S., QUACKENBUSH, J. and NELSON, S.F., 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics (Oxford, England),* **27**(19), pp. 2648-2654.

SATO, T. and SEKIDO, Y., 2018. NF2/Merlin Inactivation and Potential Therapeutic Targets in Mesothelioma. *International journal of molecular sciences,* **19**(4), pp. 988. doi: 10.3390/ijms19040988.

SATO, T. and SEKIDO, Y., 2018. NF2/Merlin Inactivation and Potential Therapeutic Targets in Mesothelioma. *International journal of molecular sciences,* **19**(4), pp. 988. doi: 10.3390/ijms19040988.

SCHLEUTKER, J., BAFFOE-BONNIE, A.B., GILLANDERS, E., KAINU, T., JONES, M.P., FREAS-LUTZ, D., MARKEY, C., GILDEA, D., RIEDESEL, E., ALBERTUS, J., GIBBS, K.D.,JR, MATIKAINEN, M., KOIVISTO, P.A., TAMMELA, T., BAILEY-WILSON, J.E., TRENT, J.M. and KALLIONIEMI, O.P., 2003. Genome-wide scan for linkage in finnish hereditary prostate cancer (HPC) families identifies novel susceptibility loci at 11q14 and 3p25-26. *The Prostate,* **57**(4), pp. 280-289.

SCHWARTZ, R. and SCHÄFFER, A.A., 2017. The evolution of tumour phylogenetics: principles and practice. *Nature reviews.Genetics,* **18**(4), pp. 213-229.

SEKIDO, Y., PASS, H.I., BADER, S., MEW, D.J., CHRISTMAN, M.F., GAZDAR, A.F. and MINNA, J.D., 1995. Neurofibromatosis type 2 (NF2) gene is somatically mutated in mesothelioma but not in lung cancer. *Cancer research,* **55**(6), pp. 1227-1231.

SEMENTINO, E., MENGES, C.W., KADARIYA, Y., PERI, S., XU, J., LIU, Z., WILKES, R.G., CAI, K.Q., RAUSCHER, F.J.,3RD, KLEIN-SZANTO, A.J. and TESTA, J.R., 2018. Inactivation of Tp53 and Pten drives rapid development of pleural and peritoneal malignant mesotheliomas. *Journal of cellular physiology,* **233**(11), pp. 8952-8961.

SEMENTINO, E., MENGES, C.W., KADARIYA, Y., PERI, S., XU, J., LIU, Z., WILKES, R.G., CAI, K.Q., RAUSCHER, F.J.,3RD, KLEIN-SZANTO, A.J. and TESTA, J.R., 2018. Inactivation of Tp53 and Pten drives rapid development of pleural and peritoneal malignant mesotheliomas. *Journal of cellular physiology,* **233**(11), pp. 8952-8961.

SHAH, K.V., GALLOWAY, D.A., KNOWLES, W.A. and VISCIDI, R.P., 2004. Simian virus 40 (SV40) and human cancer: a review of the serological data. *Reviews in medical virology,* **14**(4), pp. 231-239.

SHAPIRO, I.M., KOLEV, V.N., VIDAL, C.M., KADARIYA, Y., RING, J.E., WRIGHT, Q., WEAVER, D.T., MENGES, C., PADVAL, M., MCCLATCHEY, A.I., XU, Q., TESTA, J.R. and PACHTER, J.A., 2014. Merlin deficiency predicts FAK inhibitor sensitivity: a synthetic lethal relationship. *Science translational medicine,* **6**(237), pp. 237ra68.

SHAVELLE, R., VAVRA-MUSSER, K., LEE, J. and BROOKS, J., 2017. Life Expectancy in Pleural and Peritoneal Mesothelioma. *Lung Cancer International,* **2017**, pp. 10.1155/2017/2782590.

SHLIEN, A. and MALKIN, D., 2009. Copy number variations and cancer. *Genome medicine,* **1**(6), pp. 62.

SHUKLA, A., GULUMIAN, M., HEI, T.K., KAMP, D., RAHMAN, Q. and MOSSMAN, B.T., 2003. Multiple roles of oxidants in the pathogenesis of asbestos-induced diseases. *Free radical biology & medicine,* **34**(9), pp. 1117-1129.

SOMARELLI, J.A., WARE, K.E., KOSTADINOV, R., ROBINSON, J.M., AMRI, H., ABU-ASAB, M., FOURIE, N., DIOGO, R., SWOFFORD, D. and TOWNSEND, J.P., 2017. PhyloOncology: Understanding cancer through phylogenetic analysis. *Biochimica et biophysica acta,* **1867**(2), pp. 101-108.

SONG, S., NONES, K., MILLER, D., HARLIWONG, I., KASSAHN, K.S., PINESE, M., PAJIC, M., GILL, A.J., JOHNS, A.L., ANDERSON, M., HOLMES, O., LEONARD, C., TAYLOR, D., WOOD, S., XU, Q., NEWELL, F., COWLEY, M.J., WU, J., WILSON, P., FINK, L., BIANKIN, A.V., WADDELL, N., GRIMMOND, S.M. and PEARSON, J.V., 2012. qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PloS one,* **7**(9), pp. e45835.

SPELEMAN, F., KUMPS, C., BUYSSE, K., POPPE, B., MENTEN, B. and DE PRETER, K., 2008. Copy number alterations and copy number variation in cancer: close encounters of the bad kind. *Cytogenetic and genome research,* **123**(1-4), pp. 176-182.

STAYNER, L., WELCH, L.S. and LEMEN, R., 2013. The worldwide pandemic of asbestos-related diseases. *Annual Review of Public Health,* **34**, pp. 205-216.

STEAD, L.F., SUTTON, K.M., TAYLOR, G.R., QUIRKE, P. and RABBITTS, P., 2013. Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: applications in tumor subclone resolution. *Human mutation,* **34**(10), pp. 1432-1438.

TAIOLI, E., VAN GERWEN, M., MIHALOPOULOS, M., MOSKOWITZ, G., LIU, B. and FLORES, R., 2017. Review of malignant pleural mesothelioma survival after talc pleurodesis or surgery. *Journal of Thoracic Disease,* **9**(12), pp. 5423-5433.

TALEVICH, E., SHAIN, A.H., BOTTON, T. and BASTIAN, B.C., 2016. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS computational biology,* **12**(4), pp. e1004873.

TAN, E., WARREN, N., DARNTON, A.J. and HODGSON, J.T., 2010. Projection of mesothelioma mortality in Britain using Bayesian methods. *British journal of cancer,* **103**(3), pp. 430-436.

TARASOV, A., VILELLA, A.J., CUPPEN, E., NIJMAN, I.J. and PRINS, P., 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics (Oxford, England),* **31**(12), pp. 2032-2034.

TERTEMIZ, K.C., OZGEN ALPAYDIN, A., GUREL, D., SAVAS, R., GULCU, A. and AKKOCLU, A., 2014. Multiple distant metastases in a case of malignant pleural mesothelioma. *Respiratory medicine case reports,* **13**, pp. 16-18.

THOMAS, R., MARKS, D.H., CHIN, Y. and BENEZRA, R., 2018. Whole chromosome loss and associated breakage-fusion-bridge cycles transform mouse tetraploid cells. *The EMBO journal,* **37**(2), pp. 201-218.

VAN LOO, P., NORDGARD, S.H., LINGJÆRDE, O.C., RUSSNES, H.G., RYE, I.H., SUN, W., WEIGMAN, V.J., MARYNEN, P., ZETTERBERG, A., NAUME, B., PEROU, C.M., BØRRESEN-DALE, A.L. and KRISTENSEN, V.N., 2010. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America,* **107**(39), pp. 16910-16915.

VISONÀ, S.D., VILLANI, S., MANZONI, F., CHEN, Y., ARDISSINO, G., RUSSO, F., MORETTI, M., JAVAN, G.T. and OSCULATI, A., 2018. Impact of asbestos on public health: a retrospective study on a series of subjects with occupational and non-occupational exposure to asbestos during the activity of Fibronit plant (Broni, Italy). *Journal of public health research,* **7**(3), pp. 1519.

WALL, L., CHRISTIANSEN, T. and ORWANT, J., 2000. *Programming Perl.* 3rd Edition edn. O'Reilly Media.

WHITE, P.S., MARIS, J.M., BELTINGER, C., SULMAN, E., MARSHALL, H.N., FUJIMORI, M., KAUFMAN, B.A., BIEGEL, J.A., ALLEN, C., HILLIARD, C., VALENTINE, M.B., LOOK, A.T., ENOMOTO, H., SAKIYAMA, S. and BRODEUR, G.M., 1995. A region of consistent deletion in neuroblastoma maps within human chromosome 1p36.2-36.3. *Proceedings of the National Academy of Sciences of the United States of America,* **92**(12), pp. 5520-5524.

XU, C., 2018. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal,* **16**, pp. 15-24.

YEH, ,CHIEN-HUNG, BELLON, ,MARCIA and NICOT, ,CHRISTOPHE, 2018. FBXW7: a critical tumor suppressor of human cancers. *Molecular Cancer,* 17(115).

YU, K., CHEN, B., ARAN, D., CHARALEL, J., YAU, C., WOLF, D.M., VAN 'T VEER, L.J., BUTTE, A.J., GOLDSTEIN, T. and SIROTA, M., 2019. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature communications,* **10**(1), pp. 3574-019-11415-2.

ZARE, F., DOW, M., MONTELEONE, N., HOSNY, A. and NABAVI, S., 2017. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC bioinformatics,* **18**(1), pp. 286-017-1705-x.

ZHANG, F., GU, W., HURLES, M.E. and LUPSKI, J.R., 2009. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics,* **10**, pp. 451-481.

ZHANG, L., BAI, W., YUAN, N. and DU, Z., 2019. Comprehensively benchmarking applications for detecting copy number variation. *PLoS computational biology,* **15**(5), pp. e1007069.

ZHANG, M., LUO, J., SUN, Q., HARBER, J., DAWSON, A.G., NAKAS, A., BUSACCA, S., SHARKEY, A.J., WALLER, D., SHEAFF, M.T., RICHARDS, C., WELLS-JORDAN, P., GABA, A., POILE, C., BAITEI, E.Y., BZURA, A., DZIALO, J., JAMA, M., LE QUESNE, J., BAJAJ, A., MARTINSON, L., SHAW, J.A., PRITCHARD, C., KAMATA, T., KUSE, N., BRANNAN, L., DE PHILIP ZHANG, P., YANG, H., GRIFFITHS, G., WILSON, G., SWANTON, C., DUDBRIDGE, F., HOLLOX, E.J. and FENNELL, D.A., 2021. Clonal architecture in mesothelioma is prognostic and shapes the tumour microenvironment. *Nature communications,* **12**(1), pp. 1751-021-21798-w.

ZHANG, W.Q., DAI, Y.Y., HSU, P.C., WANG, H., CHENG, L., YANG, Y.L., WANG, Y.C., XU, Z.D., LIU, S., CHAN, G., HU, B., LI, H., JABLONS, D.M. and YOU, L., 2017. Targeting YAP in malignant pleural mesothelioma. *Journal of*

Cellular and Molecular Medicine, 21(11), pp. 2663-2676.