CARNEGIE MELLON UNIVERSITY

MASTER'S THESIS

Patterns of Mutation in Bloom's Syndrome

Author:
Tomas MATTESON

Supervisor:
Dr. Russell SCHWARTZ

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science in Computational Biology

in the

Schwartz Lab
Computational Biology Department

December 4, 2022

Acknowledgements

First, I'd like to thank my thesis advisor Dr. Russell Schwartz, as well as my friends and mentors Dr. Dannie Durand, Dr. Luisa Hiller, and Dr. Oana Carja. Their support and advice were critical for me to continue on this journey, and their kindness and words of wisdom are always appreciated. I'd also like to thank our collaborators at the Cunniff Lab, without whom I would not have this Whole Genome Sequence data. I'd like to thank my thesis committee: Dr. Oana Carja, Dr. Luisa Hiller, and the committee chair Dr. Russell Schwartz.

CARNEGIE MELLON UNIVERSITY

Abstract

Dr. Russell Schwartz
Computational Biology Department

Master of Science in Computational Biology

Patterns of Mutation in Bloom's Syndrome

by Tomas MATTESON

Bloom's Syndrome (BS) is an autosomal recessive disorder that greatly increases the risk of early onset of multiple cancers. The putative cause of BS is loss-of-function mutations to the BLM gene, which codes for a RecQ helicase. However, 9 of 134 individuals with a BS phenotype followed by the Bloom's Syndrome Registry had no identified BLM mutation. The purpose of this research work is to uncover genomic correlations to BS from a panel of 9 Case/Control matched Whole Genome Sequenced BS patients. We found small but consistent single nucleotide polymorphisms (SNP) mutation signatures present in most of the samples, with two in particular more heavily effected. With regards to copy number variants (CNV), all samples exhibit some form of significant gain or loss, but we found no genomically consistent location where gain or loss occurs, or in the mean ploidy count. Upon analyzing the structural variants (SV), we found that almost all were translocations, suggesting significant genome disruption. The SVs seem to be concentrated around particular breakpoints with respect to both deletions and duplications. The mutation burden on BLM and other RecQ Helicases seem to suggest that damage to RecQ Helicases has high correlation to genome disruption in our Whole Genome Samples. All in all, there appears to be patterns consistent with genome instability across all samples, focusing around SNP mutation signatures and SV genomic breakpoints.

Contents

List of Figures

Chapter 1

Introduction

1.1 Bloom's Syndrome

Bloom's Syndrome (BS) is an autosomal recessive disorder that causes prenatal and postnatal growth deficiency, photosensitive skin changes, and immune deficiency. Female patients with BS have impaired but not eliminated fertility, and male patients are invariably infertile. BS causes defects in T cell and B cell lineage, and the majority of affected individuals have a deficiency in at least one of the serum immunoglobin classes. Patients with BS have a high incidence of upper respiratory and gastrointestinal infections, with no class of pathogenic organisms yet identified. BS greatly increases risk of early onset of multiple cancers, with leukemia and lymphoma being the most frequently occurring cancers in BS patients (Cunniff, 2017).

First associations to the cause of BS were uncovered when 25 of 26 affected individuals examined in a clinical study were determined to possess a homozygous polymorphic tetranucleotide locus within the FES gene in chromosome 15q26.1. The researchers observed isochromatid breaks and associated displaced acentric fragments and sister chromatid reunions, as well as transverse breakage at the centromere. This eventually helped determine the clinical diagnosis of BS, which can be confirmed by cytogenetic analysis that identifies an increased number of sister chromatid exchanges. Molecular confirmation of BS identifies biallelic mutations to the BLM gene. The current working hypothesis for BS diagnosis is that loss-of-function mutations to BLM (a RecQ helicase) are putatively causative. However in the Bloom's Syndrome Registry, 9 of 134 individuals with a BS phenotype had no identified BLM mutation (Cunniff, 2017).

No BS-causing missense mutations of BLM have been identified outside the helicase and RQC domains (Cunniff, 2017). This suggests that failure to interact to with DNA (the principle role of such domains) is suspected to be causative with loss of function. It is thought that mutations of the topoisomerase IIIα-interacting region, the strand-annealing domain, and the HRDC domain could occur, and when combined with BLM premature protein termination, mutations could be associated with a Bloom-like phenotype that is not as severe as BS itself. Absence of a functional BLM protein causes chromosome instability, excessive homologous recombination, and a greatly increased number of sister chromatid exchanges. A common founder mutation designated blmAsh is present in about 1 in 100 persons of Eastern European Jewish ancestry. The blmAsh variant is a 6-bp deletion and 7-bp insertion at position 2,281 in chromosome subband 15q26.1. A recurrent founder allele of BLM has also been identified in Slavic populations of Eastern Europe. Missense, nonsense, and frameshift mutations as well as multiexonic deletions have all been observed (Cunniff, 2017).

BLM is an abundant protein localized to the nucleus in rapidly dividing cells, such as cancer cell line models (e.g., in HeLa cells, there are 50,000 BLM molecules per cell). BLM expression is closely correlated with Ki67 expression and in dividing cells of many types. The BLM protein is not expressed in quiescent or nondividing cells, such as serum-deprived human fibroblasts or unstimulated lymphocytes. Steady-state levels of BLM are highest in late S and G2 phases of the cell cycle and lowest in early G1 phase, suggestive of genomic unwinding during replicative periods (Cunniff, 2017).

Other RecQ-related disorders include Rothmund Thomson, RAPADILINO (RA for radial (forearm bone) malformations; PA for patella (kneecap) and palate abnormalities; DI for diarrhea and dislocated joints; LI for limb abnormalities and little size). Additionally there are Baller Gerold syndrome, associated with RECQL4 mutations, and Werner syndrome, associated with mutations of WRN. Multiple genetically independent pathways have evolved that mediate the repair of DNA double-strand break (DSB), and RecQ helicases play pivotal roles in each of them. The importance of DSB repair is supported by the observations that defective DSB repair can cause chromosomal aberrations, genomic instability, senescence, or cell death, which ultimately can lead to premature aging, neurodegeneration, or tumorigenesis (H and AJ, 2021).

Bloom's Syndrome is a prototypical chromosomal instability syndrome, and the somatic mutations that occur as a result of that instability are presumed to be responsible for the increased cancer risk. In this regard we wish to utilize Cancer mutation signatures in our analysis, predominantly via COSMIC (to be further described later). By examining Cancer mutation signatures we can see if there are patterns of mutation in BS patients that are consistent with particular Cancer subtypes or known mechanisms of Cancer. In addition to looking for such patterns in the variant calls, we hoped to examine excessive homologous recombination, sister chromatid exchanges, and transverse breakage at the centromere. To determine the existence, if any, of BS subtypes, we analyzed BLM ash and the Slavic BLM variant across samples. We also searched for other RecQ mutations could be of interest, as well as any notable serum immunoglobin variants. Finally, we examined the concordance of all mutations of any variant type across all samples for a particular caller.

1.2 Genomics and Variants

Our calls were Illumina Whole Genome Sequence (WGS) calls. Illumina short read calls are a Next Generation Sequencing (NGS) technology. This is a sequence by synthesis method where a genome is digested into fragments that get sequenced and are then aligned to reference genome or de-novo assembled into a genome. We aligned our calls to the Genome Reference Consortium Human Build 38 (GRCh38) human reference genome with the BWA aligner. In general, the read depth of WGS samples is the key determinant of quality, and our samples had a read depth of around 40 across the genome, a relatively high read depth.

We first called our variants with Single Nucleotide Polymorphism (SNV) callers. SNV callers generally detect Single Base Substitutions (SBS), Multiple Base Substitutions (including double base DBS), and Insertions and Deletions of a few base pairs

in length (ID or indels). There currently exist well documented mutation signature extractors for SBS, DBS, and ID variants, particularly from COSMIC (Alexandrov, 2020). SBS, DBS, and ID mutation signatures correlate to recurrent mechanisms of mutation in Cancer. Additionally, it is useful to examine SNVs to look for missense (amino acid changing), frameshift (reading frame altering), and nonsense (premature stop codon) mutations. Such mutations are significantly more likely to cause loss of function of the BLM gene, or otherwise damage the proteins of genic regions.

Copy Number Variants (CNVs) are deletions, insertions, or duplications ranging from 50 to millions of base pairs in size. It is estimated that approximately 12% of the genome in human populations is subject to copy number change. CNVs are important to examine because gain or loss to particular genomic regions are highly correlative with certain subtypes of cancers and other human diseases (**CNV**). Gains and losses of gene copies may directly influence gene dosage within the CNV regions, which could result in a change of gene expression level. CNV detection in heterogeneous populations of cells is difficult, where purity and ploidy estimation of the sample is a key concern (Prandi D, 2019).

Structural Variants (SVs) are deletions, duplications, insertions, inversions, translocations of at least 50 bases pairs in size. CNVs are a major subtype of SVs, but many more complex combinations of SV types are known to exist including chromotripsis and chromoplexy. SVs can have a pronounced phenotypic impact by disrupting gene function and regulation or modifying gene dosage. SVs are inherently harder to detect if the event size is of a similar length to the reads, and particularly difficult if the event size size is greatly larger than the read length (Mahmoud, 2019).

1.3 COSMIC

Mutational signatures are patterns of variants throughout a genome or exome sample that are caused by particular mutational processes. Examples include smoking, UV exposure, and defects to different DNA replication machinery. The Catalogue Of Somatic Mutations In Cancer (COSMIC) is the cardinal work in this area, characterizing SBS, DBS, ID, and CNV mutations for now. The SBS, DBS, and ID derived COSMIC signatures were detected from 4,645 whole-genome and 19,184 exome sequences that encompass most types of cancer. Specifically, there are 49 single-base-substitution, 11 doublet-base-substitution, and 17 small insertion-and-deletion well characterized signatures. Given that individuals with Bloom Syndrome are at a severely inclined risk of developing cancer, cancer mutational signatures should be represented in our WGS data. In detecting mutation signatures, the nucleotide context of a particular variant can be important. This corresponds to bases immediately prior and after the SBS, DBS, or InDel event. For example SBS96 corresponds to the 96 kinds of SBS with a nucleotide context of 1 base prior and 1 base after. Other such formats extend the context, e.g. SBS1536. In the case of CNV signatures, this can signatures are generated with respect to the genomic context of heterozygosity or loss of heterozygosity and ploidy (to be elaborated further in the appropriate section).

Going slightly deeper into the methodology behind detecting such mutation signatures, we examine the algorithmic machinery of COSMIC: Non-negative matrix factorization (NMF). NMF is one instance of a greater class of algorithms called topic models. Topic models find latent structure in data. Commonly use in the

field of Natural Language Processing, this is a distribution over topics, of which each topic has a distribution over words. In COSMIC, that latent structure is mutational process contribution and the word counts are variant counts. There are many alternative approaches to topic models including Latent Dirichlet Allocation and Hierarchical Dirichlet Processes. NMF determines the signature profiles and contributions of each signature to each genome as part of its factorization of the input matrix of mutation spectra. NMF attempts to reconstruct a matrix X with low rank approximation $X \approx SH$, where S is the signature matrix COSMIC provides, and H is the importance of each signature in the sample.

1.4 Samples

We received 9 Case-Control matched WGS samples of BS patients from the Cunniff lab at Weill Cornell Medicine in raw read fastq format. We pre-processed the samples for quality with cut adapt, examined them with fastqc, then used BWA to align and index the bams. After processing the BAMs, each sample was called with 3 different variant callers for each major type of variant. For SNVs Somatic Sniper, Freebayes, and Strelka were used. For CNVs Cnvkit, Ascat, and Sequenza were used. For SVs Manta, Delly, and GRIDSS were used. Each Freebayes sample was called individually, meaning the difference in mutations in the Case versus Control did not influence the calling of each sample. The Somatic Sniper and Strelka calls were done in a joint manner, where both the Case and Control for each sample were called together in one variant file. The Control for each sample was isolated from the blood of a Bloom Syndrome patient, and the Case for each sample was isolated from a tumor of the respective patient. All CNV and SV calls were done in a joint manner as well. Furthermore, the SNV and SV variant calls were filtered for various quality metrics (to be described later), and the SNV variant calls were filtered against a panel of normals from the 1000 Genomes project.

Chapter 2

SNV Results

2.1 Freebayes

Freebayes is a haplotype-based variant caller that utilizes a Bayesian Statistical framework. Haplotype-based variant detection methods, in which short haplotypes are read directly from sequencing traces, offer a number of benefits over methods which operate on a single position at a time. Haplotype-based methods ensure semantic consistency among described variants by simultaneously evaluating all classes of alleles in the same context. The use of locally-phased genotype data can lower the computational burden of genotype imputation by reducing the possible space of haplotypes which must be considered. Locally phased genotypes can be used to improve genotyping accuracy in the context of rare variations that can be difficult

to impute due to sparse linkage information. Provided sequencing errors are independent, the use of longer haplotypes in variant detection can improve sensitivity by increasing the signal to noise ratio of the genotype likelihood space that is used in analysis. This follows from the fact that the space of possible erroneous haplotypes expands dramatically with haplotype length, while the space of true variation remains constant, with the number of true alleles less than or equal to the ploidy of the sample at a given locus (Garrison and Marth, 2012).

Freebayes assembles haplotype observations over minimal, dynamically-determined, reference-relative windows which contain multiple segregating alleles. To be used in the analysis, haplotype observations must be derived from aligned reads which are anchored by reference-matching sequence at both ends of the detection window. These haplotype observations have derived quality estimations which allow their incorporation into a general statistical model. Gradient ascent is then used to determine the maximum a posteriori estimate of a mutual genotyping over all samples under analysis and establish an estimate of the probability that the loci is polymorphic (Garrison and Marth, 2012).

In our raw Freebayes calls we observed a very high mean number of variants per sample (around 1.5 million each) and a comparatively low standard deviation (close to 0.5 million). There was a strong enrichment in C to T and T to C single base substitutions across nucleotide contexts across all the samples. This was further examined in our COSMIC SBS analysis (to be described later in this paper). Likewise we saw around 4000 Double base substitutions per sample, and a few samples had a high concentration of CC to TT substitutions, while other samples had a uniform distribution over the double substitution space. These were again further examined in our COSMIC DBS analysis. Our Freebayes calls failed to yield any viable Indel matrices and so we were unable to use COSMIC's ID signatures for analysis.

2.2 Somatic Sniper

Somatic Sniper is a likelihood based somatic SNV caller. To detect somatic mutations, the developers calculate the likelihood that a site is not somatic by using a default prior that takes into account the prior probability of a somatic mutation for a given normal genotype, accounting for interaction between normal and tumor prior and not treating the two as independent. While the uncorrected method has a predicted a maximal false Discovery Rate (FDR) of 15 percent in the absence of mapping error and in the presence of perfectly calibrated base qualities, by mitigating known issues a lower FDR was achieved. The following are known issues with Somatic Sniper. A major indicator of a false positive is strand bias, where the variant allele arises primarily from reads aligning on one strand versus the other. Another is that the Illumina pipeline uses the Read Segment Quality Control Indicator to identify 3 portions of reads that should be discarded. The author's observed that false positive bases of high quality frequently occur near these regions. Additionally, variant bases that appeared to be generated from read-through of homopolymer runs, as well as reads that appeared to map from paralogs not in the reference were issues. A suite of filtering methods were developed to address these problems, significantly decreasing the FDR but potential lower grade artefacts may remain (Larson DE, 2012).

In our raw Somatic Sniper calls we observed a highly variant amount of single base substitutions per sample, ranging from 20,000 to 2,000,000 per sample. There was a strong enrichment in C to T and T to C single base substitutions across nucleotide contexts across many samples, with other samples also having a high concentration of C to A substitutions. This was further examined in our COSMIC SBS analysis. Likewise we saw around 200 Double base substitutions per sample (with much lower variance), most samples had a high concentration of CC to AA substitutions and the other samples had a high concentration of CC to TT substitutions. These were further examined in our COSMIC DBS analysis. Our Somatic Sniper calls failed to yield any viable Indel matrices and so we were unable to use COSMIC's ID signatures for analysis.

2.3 Strelka

Strelka is a method for somatic SNV and small indel detection from sequencing data of matched tumor–normal samples (somatic SNV caller). Strelka uses a Bayesian approach which represents continuous allele frequencies for both tumor and normal samples, while leveraging the expected genotype structure of the normal. This is achieved by representing the normal sample as a mixture of germline variation with noise, and representing the tumor sample as a mixture of the normal sample with somatic variation. A natural consequence of the model structure is that sensitivity can be maintained at high tumor impurity without requiring purity estimates. There are a few complications that typically come into play when tumor impurity is high. The first is that germline variants can outnumber somatic variants by several orders of magnitude, so any tendency to mistake germline variation as somatic can substantially contaminate the somatic variant predictions. A second complicating factor is variability in the somatic allele frequencies due to the presence of normal cells in the tumor sample, copy number variation and tumor heterogeneity (Christopher T. Saunders, 2012).

In less sophisticated somatic variant callers, somatic variants are detected by independently genotyping both Case and Control samples and simply subtracting the results, an approach which can provide reasonable predictions for cell lines because the aforementioned variability in somatic allele frequency is reduced for this case. For the general case, a joint analysis of both samples should improve results by facilitating tests for candidate somatic alleles in both samples (especially important for indels) and enable better representation of sequencing noise and tumor impurity. Following realignment, Strelka uses the read alignment information from both samples to produce a somatic variant probability. The somatic caller models allele frequencies rather than diploid genotypes, representing the normal sample as a mixture of diploid germline variation with noise, and the tumor sample as a mixture of the normal sample with somatic variation. The somatic variant probability produced by this model is not used directly because it detects many variants in loss of heterzygosity (LOH) or copy number change regions. Instead, each call is reported using the joint probability of a somatic variant and a specific genotype in the normal sample, summarized as a quality score. The caller also accounts for strand bias in Illumina reads with an appropriate prior over the read distribution (Christopher T. Saunders, 2012).

In our raw Strelka calls we observed a highly variant amount of single base substitutions per sample, ranging from 1.5 million to 25 million per sample. There was a strong enrichment in C to T and C to A single base substitutions across nucleotide contexts across most of the samples. This was further examined in our COSMIC SBS analysis. Likewise we saw very high variance in the Double base substitutions per sample (from 3 to 20,000 each). Most samples had a high concentration of CC to AA substitutions and the other samples had a high concentration of CC to TT substitutions. These were further examined in our COSMIC DBS analysis. Our Strelka calls yielded viable Indel matrices and so we were able to use COSMIC's ID signatures for analysis. All the samples had a high concentration of nucleotide insertions and deletions greater than five base pairs long of Thymine.

2.4 Comparison of Callers

2.5 COSMIC Single Base Substitutions

As previously described, COSMIC is an NMF based framework that looks for patterns of mutation in variant calls and aligns such patterns to known mutational processes. For SBSs, there are 96 primary classes constituted by the 6 base substitutions C>A, C>G, C>T, T>A, T>C and T>G (in which the mutated base is represented by the pyrimidine of the base pair), plus the flanking 5′ and 3′ bases. There are additional nucleotide contexts, where the two flanking bases 5′ and 3′ to the mutated base were considered (producing 1,536 classes) or mutations within transcribed genome regions were selected and classified according to whether the mutated pyrimidine fell on the transcribed or untranscribed strand (producing 192 classes) (Alexandrov, 2020). However, for our samples all such analyzes produced similar results.

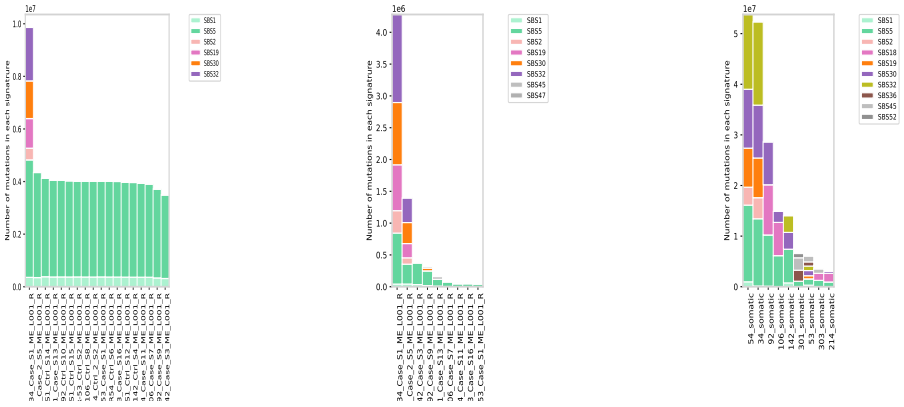

FIGURE 2.1: From left to right, the Freebayes, Somatic Sniper, and Strelka SBS96 Activity plots

Across all the Freebayes samples, SBS1 and SBS5 are significantly enriched 2.1. SBS1 corresponds to the enzymatic deamination of 5-methylcytosine to thymine, and SBS5 is unknown mutational process that has increased prevelance in bladder cancer samples with ERCC2 mutations and in many cancer types due to tobacco smoking. Sample 34 has additional enrichment in SBS2, SBS19, SBS20, and SBS32. SBS2 corresponds to loss of function in the AID/APOBEC family of cytidine deaminases. SBS19 has no known aetiology, and is very sparsely distributed in a few

cancer types. SBS30 corresponds to a deficiency in base excision repair, typically via inactivating mutations in NTHL1. SBS32 corresponds to prior treatment with azathioprine to induce immunosuppression.

Across all the Somatic Sniper samples, SBS1 and SBS5 are significantly enriched at a low level 2.1. Sample 34 and 54 have additional enrichment in SBS2, SBS19, SBS20, and SBS32. Across all the Strelka samples, the same trends from Somatic Sniper hold. Given that we saw SBS1 and SBS5 enriched across all samples, as well as minor noise, we decided to employ a panel of normals to improve the quality of our variant analysis. We also filtered the variant calls for read depth and quality metrics. These processes will be further described in the subsequent section.

2.6 Quality Control and Panel of Normals

There are a variety of best practices when it comes to filtering somatic variant calls. Similar to germline SNVs/indels, candidate somatic variants should be filtered to remove common alignment artifacts. In addition, the availability of a matched normal sample enables a direct comparison of data characteristics at the site of a candidate somatic variant call to help distinguish true variants from false positives. For example, reads supporting high-quality mutation calls should exhibit similar position and strandedness as reads supporting the wild-type allele. Other metrics, such as the difference in average mapping quality or trimmed read length, help uncover false positives due to alignment artifacts (Pedersen, 2021).

Population variant filtering is a powerful strategy for identifying and removing likely germline variants from somatic mutation callsets but should be done with caution. Simply removing all variants in dbSNP is an appealing but hazardous strategy, since that database contains a number of recurrent mutations from human tumors—such as p.(H1047R) in PIK3CA (rs121913279) and p.(R132H) in IDH1 (rs121913500)—as well as several mutations from the COSMIC somatic mutation database (Koboldt, 2020). There is a similar risk for applying a broad filter based on all variants in the gnomAD database, in which the presence of apparent somatic loss-of-function variants in hematological malignancy genes like ASXL1 has been documented. Allele frequency information can be used to safeguard against the inadvertent filtering of true somatic variants that are present in such databases. However, the 1000 Genomes project provides a more limited variant call set composed of putatively healthy samples. When filtering for a recessive disease like Bloom, an allele frequency of 0.01 is an appropriate threshold for the restricted control call set (Pedersen, 2021).

To filter our SNV calls we discarded calls with depth less than 10, as well as those with genotype quality less than 10. We additionally removed calls with allele frequency greater than 0.01 in the 1000 genomes Panel of Normals (PON). The PON is a variant call file assembled from hg38 (human genome assembly version) from both exomes and whole genomes from the 1000 Genomes Project samples and is hosted by the Broad Institute at https://console.cloud.google.com/storage/browser/gatk-best-practices/somatic-hg38. We also ran the COSMIC SBS extraction on the PON and saw significant enrichment of SBS1 and SBS5 at a lower mutation burden than in the samples 2.2.

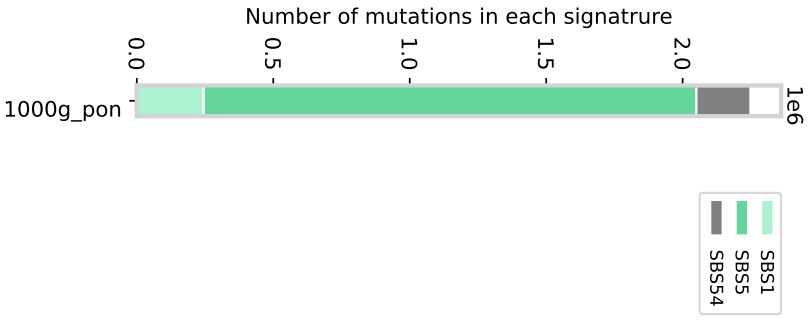

FIGURE 2.2: 1000 Genomes Reference Panel of Normals SBS96 Activity Plot

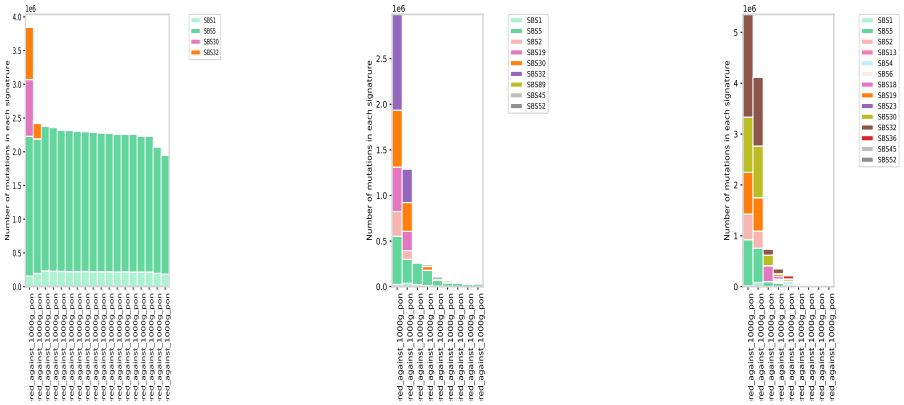

FIGURE 2.3: From left to right, the Freebayes, Somatic Sniper, and Strelka SBS96 Activity plots after 1000 Genomes Filtering

After applying these filters we noticed no significant degradation in the strength of some of the mutation process signals 2.3. All of the samples across all the callers showed very similar results to before the 1000 Genomes and quality filtering, just with slightly reduced amounts of SBS1 and SBS5. This suggests that the SBS1 and SBS5 remaining in the samples is indicative of these mutational processes contributing to the genome instability in Bloom. Likewise, the continued prevalence of SBS2, SBS19, SBS20, and SBS32 in samples 34 and 54 suggest the mutation processes continued importance in these highly degraded samples.

2.7 COSMIC Double Base Substitutions

As previously described, COSMIC is an NMF based framework that looks for patterns of mutation in variant calls and aligns such patterns to known mutational processes. For DBSs there are 78 classes constituted by the base substitutions of the form CC > AA, etc. with the flanking 5′ and 3′ bases for nucleotide context. In most cancer genomes, the number of DBSs was considerably higher than would be expected from the random adjacency of SBSs, indicating the existence of commonly occurring, single mutagenic events that cause substitutions at neighbouring bases. However, the numbers of DBSs were generally proportional to the numbers of SBSs (Alexandrov, 2020).

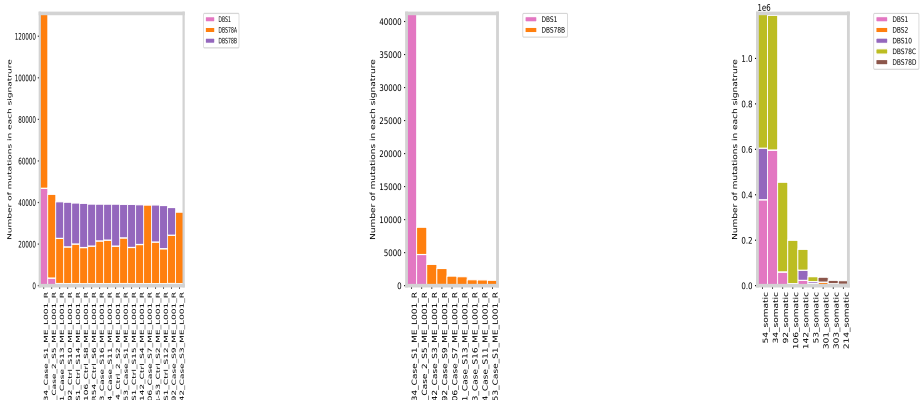

FIGURE 2.4: From left to right, the Freebayes, Somatic Sniper, and Strelka DBS78 Activity plots

For the Freebayes calls, DBS1 is enriched in Samples 34 and 54 with all other samples having only a combination of two unknown DBS mutation signatures 2.4. DBS1 corresponds to damage to cytosine transcription coupled nucleotide excision repair, which is enriched in melanoma patients. The Somatic Sniper calls follow a similar trend but with only 1 additional unknown DBS mutation signature. The Strelka calls also have DBS1 enriched in the same way with more noise.

When examining the 1000 Genomes PON, we saw no known mutational processes present 2.5. After 1000 Genomes PON and quality filtering we notice interesting results. In the Somatic Sniper and Strelka calls the DBS1 signature is not degraded but the background noise is instead eliminated, giving clear sign of the active mutation process in Samples 34 and 54 2.6. However the Freebayes calls now all display at least minor contribution from DBS11 and an additional unknown mutation signature. DBS11 has no known aetiology but may be possibly related to APOBEC mutagenesis. After careful examination of the mutation spectra, DBS11 is essentially a noiser version of DBS1 with additional mutations occurring besides the main CC to TT substitution. Given the similarity between the two mutation spectra, calling this change a minor degradation does not seem to be a stretch. With this in mind we can say that we see clear enrichment of a DBS1/DBS11 like signature in the Freebayes calls particularly for Sample 34. We can conclude that the DBS1 like signature survives filtering across all callers for Sample 34 and across the two somatic callers for Sample 54.

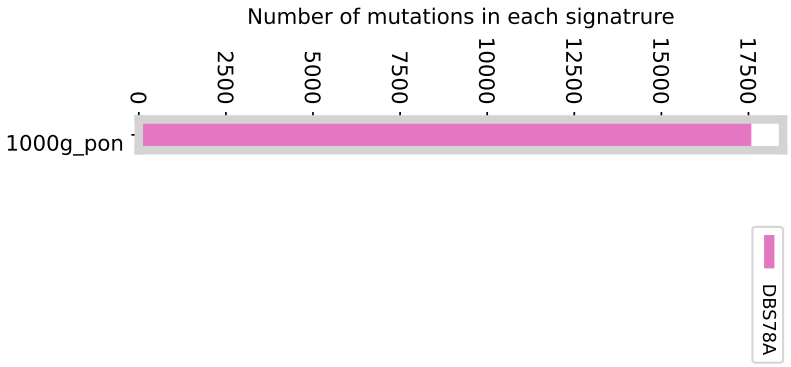

FIGURE 2.5: 1000 Genomes Reference Panel of Normals DBS78 Activity Plot

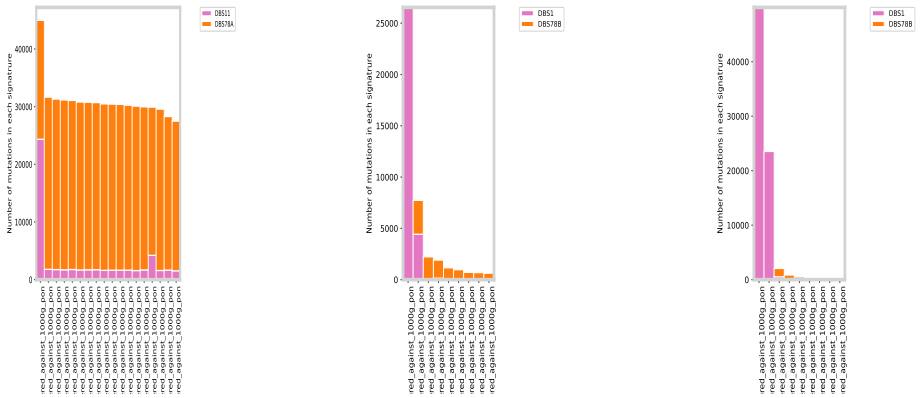

FIGURE 2.6: From left to right, the Freebayes, Somatic Sniper, and Strelka DBS78 Activity plots after 1000 Genomes filtering

2.8 COSMIC Insertions and Deletions

COSMIC applies a similar NMF based framework that looks for patterns of mutation in variant calls and aligns such patterns to known mutational processes for Insertions and Deletions. Indels are classified as deletions or insertions according to the length of the mononucleotide repeat tract in which they occurred. Longer indels are classified as occurring at repeats or with overlapping microhomology at deletion boundaries, and according to the size of indel, repeat and microhomology. There are 17 well characterized classes of Indel mutation signatures. Indels are usually present

at about 10 percent of the frequency of base substitutions (Alexandrov, 2020). In the COSMIC paper, the authors remarked that there was substantial variation between cancer genomes in the number of indels, even when cancers with evidence of defective DNA mismatch repair were excluded. Overall, the numbers of deletions and insertions were similar, but there was variation between cancer types: some cancers showed more deletions and others more insertions of various subtypes.

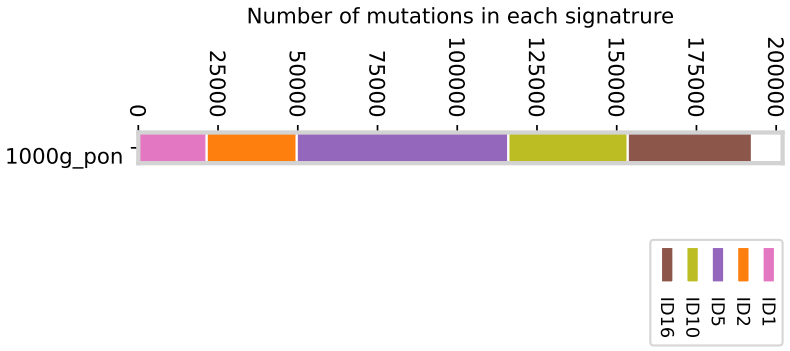

FIGURE 2.7: 1000 Genomes Reference Panel of Normals ID83 Activity Plot

Only Strelka produces non zero matrices for COSMIC ID. Prior to filtering, Samples 34 and 54 had significant enrichment in ID1, ID2, and ID9 2.8. Sample 54 also had significant enrichment for ID14. All other calls had significant enrichment for ID1, ID2, and ID12. ID1 corresponds to slippage during DNA replication of the replicated DNA strand, and ID2 corresponds to slippage during DNA replication of the template DNA strand. ID9, ID12, and ID14 all have unknown aetiologies. When examining the 1000 Genomes PON, ID1, ID2, ID5, ID10, ID16 were found to be significantly enriched 2.7. ID5, ID10, and ID16 all have unknown aetiologies. After 1000 Genome and quality filtering no significant enrichment of known ID signatures remain, meaning the mutation spectra signal was totally destroyed 2.8. With this result we can conclude that the calls have no particularly interesting mutation spectra not already present in healthy patients (or that aren't noise artefacts).

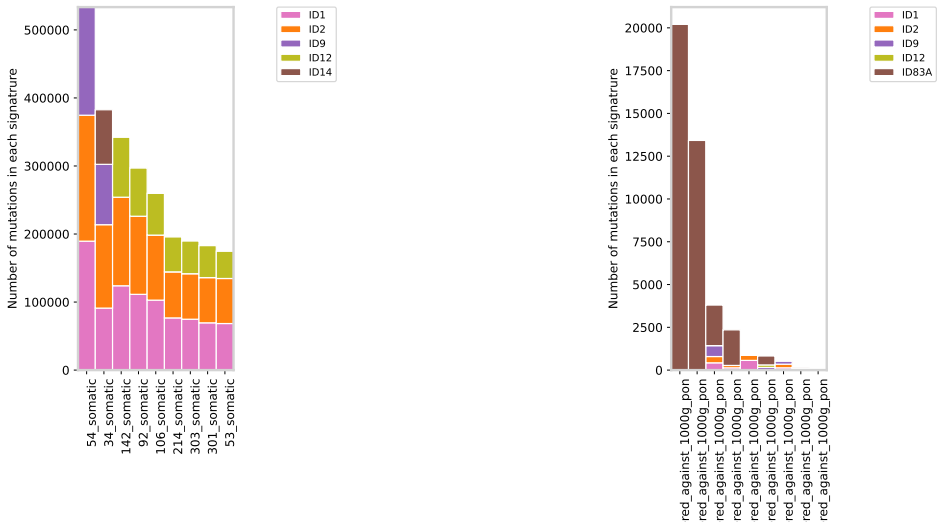

FIGURE 2.8: From left to right, the Strelka ID83 Activity plots before and after 1000 Genomes Filtering

2.9 Consensus Calls

2.10 Mutation Burden

A higher number of mutations in a tumor can also be associated with a greater probability of response to treatment (Fusco, West, and Walko, 2021). If a tumor has many different mutations, the odds are greater that the immune system will be able to recognize at least 1 of these mutations and kill the cancer. The number of mutations in a tumor cell is commonly referred to as the tumor mutation burden (TMB) of the cancer. The TMB can be helpful in predicting response to immune checkpoint inhibitor treatment across many cancer types. In our samples we use mutation burden as a proxy for genome stability, where more mutations indicate a more likely loss of function for the gene in question.

First we examined all RecQ helicases: BLM, RECQL1, WRN, RECQL4, RECQL5. RecQ helicases have distinct roles in the recovery of stalled replication fork and are broadly involved in Double Stranded Break Repair. The four currently known Double Stranded Break Repair pathways are Non Homologous End Joining (NHEJ), Homologous Recombination (HR), Strand Specific Annealing (SSA), and Microhomology Mediated End Joining (MMEJ). NHEJ is a major pathway for DSB repair during T and B cell lymphocyte development. NHEJ is a template-independent pathway that is not intrinsically accurate. Inappropriate NHEJ can lead to translocations and telomere fusion, which are hallmarks of tumor cells. HR requires a homologous DNA sequence to serve as a template, and is an accurate process. HR uses homology near the broken ends to drive repair, predominantly using the sister chromatid rather than the homologous chromosome as a template. A small percentage of DSBs are repaired through these pathways, and the factors and processes required are not well researched yet (H and AJ, 2021). MMEJ requires 2–20 nucleotide microhomology sequences for repair, and SSA needs more than 25 nucleotides of homologous sequence (typically tandem repeats). Both are intrinsically mutagenic and can cause

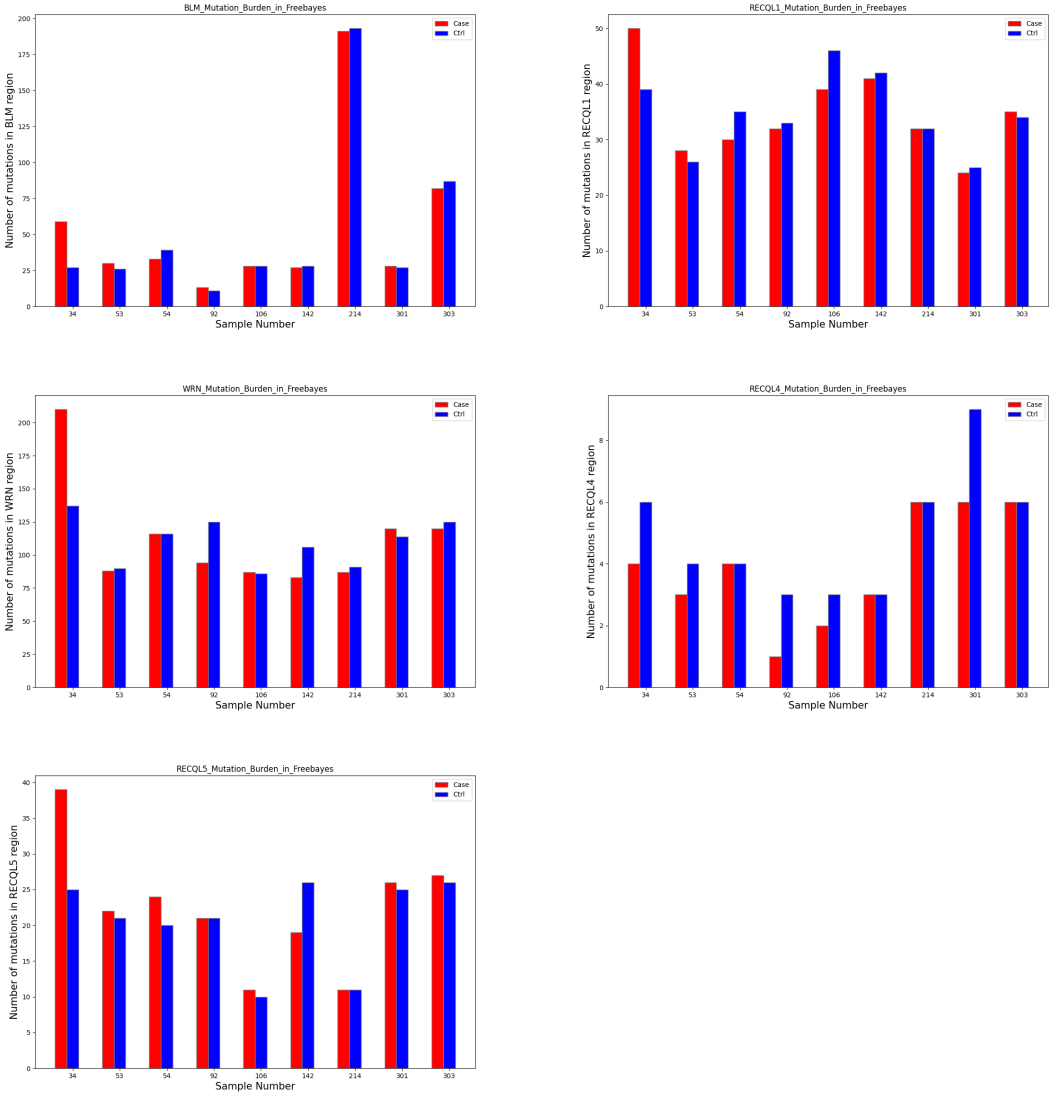

FIGURE 2.9: From left to right, the mutation burden plots for the RecQ helicases: BLM, RECQL1, WRN, RECQL4, RECQL5

deletions and rearrangements.

BLM is a major protein in HR, and is potentially implicated in SSA and MMEJ (H and AJ, 2021). RECQL1 is a major protein in HR and NHEJ, and RECQL5 is involved in HR. WRN and RECQ4 play multiple roles in NHEJ, HR, MMEJ, and SSA (H and AJ, 2021). Focusing on the BLM, we see an inconsistent pattern of mutation burden. Many samples have few BLM mutations, often with no difference in the number of mutations between Case and Control 2.9. Looking for known BLM variants, we found that for the Case samples 106, 142, 301, 303, 34, 53 were BLM ash positive, but for the Control samples 301, 303, 34, 53, 106, 142, 54 were BLM ash positive. This would suggest that the existing BLM variant in Control was mutated out in the Case for sample 54. We also observed that both Case and Control for sample 303 were positive for the slavic variant previously described, a C to T substitution at Chromosome 15, base 90761015. Looking at the other RecQ helicases, we see a moderate and consistent mutation burden on RECQL1, suggesting that damage to RECQL1 may

be correlative with Bloom Syndrome. RECQL4 and RECQL5 have no consistently high pattern of mutation burden across samples. However, each sample has at least 80 mutations in WRN, suggesting heavy damage to this critical RecQ helicase gene.

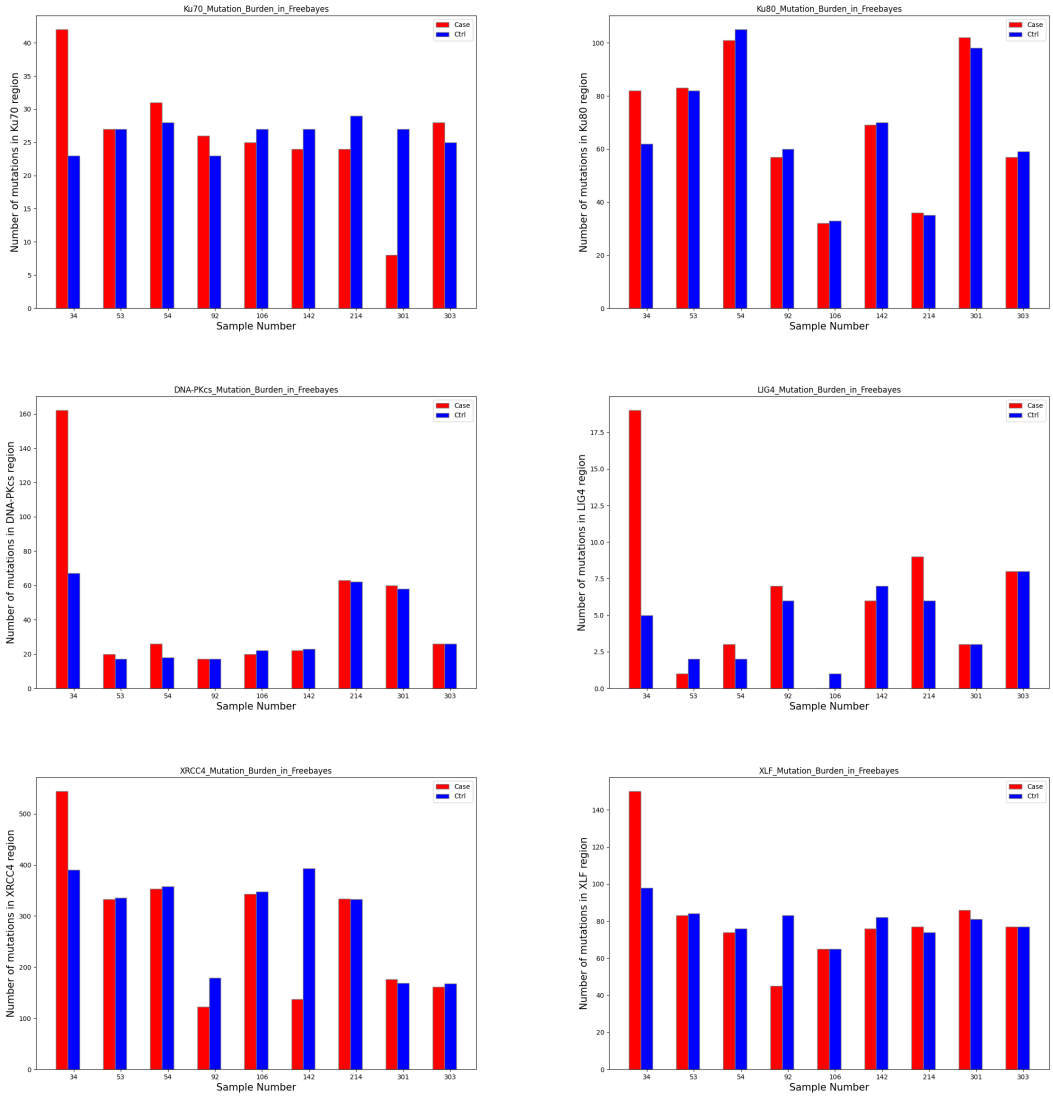

FIGURE 2.10: From left to right, the mutation burden plots for the NHEJ Pathway: Ku70, Ku80, DNA-PKcs, LIG4, XRCC4, XLF

Then we examined all proteins involved in Non Homologous End Joining (NHEJ). Initiation of NHEJ occurs when the ring-shaped Ku heterodimer, composed of the Ku70 and Ku80 proteins. Ku70/80 directly recruits DNA-dependent protein kinase catalytic subunit (DNA-PKcs). The terminal step in NHEJ is the ligation of the broken DNA ends by the DNA ligase IV(LIG4)/X-ray cross-complementing protein 4 (XRCC4) complex with the assistance of the XRCC4-like factor (XLF) (H and AJ, 2021). Looking at our Freebayes calls, If we ignore sample 301's Case we see a moderate but consistent mutation burden on Ku70 that falls broadly in line with the mutation burden on Ku80 2.10. DNA-PKcs has extremely high mutation burden in sample 34's Case, but is not consistently impacted elsewhere. LIG4 has no consistent pattern of mutation burden across samples. XRCC4 and XLF both seem to have a

consistent impact, with the mutation burden on XRCC4 being at least 100 for each sample.

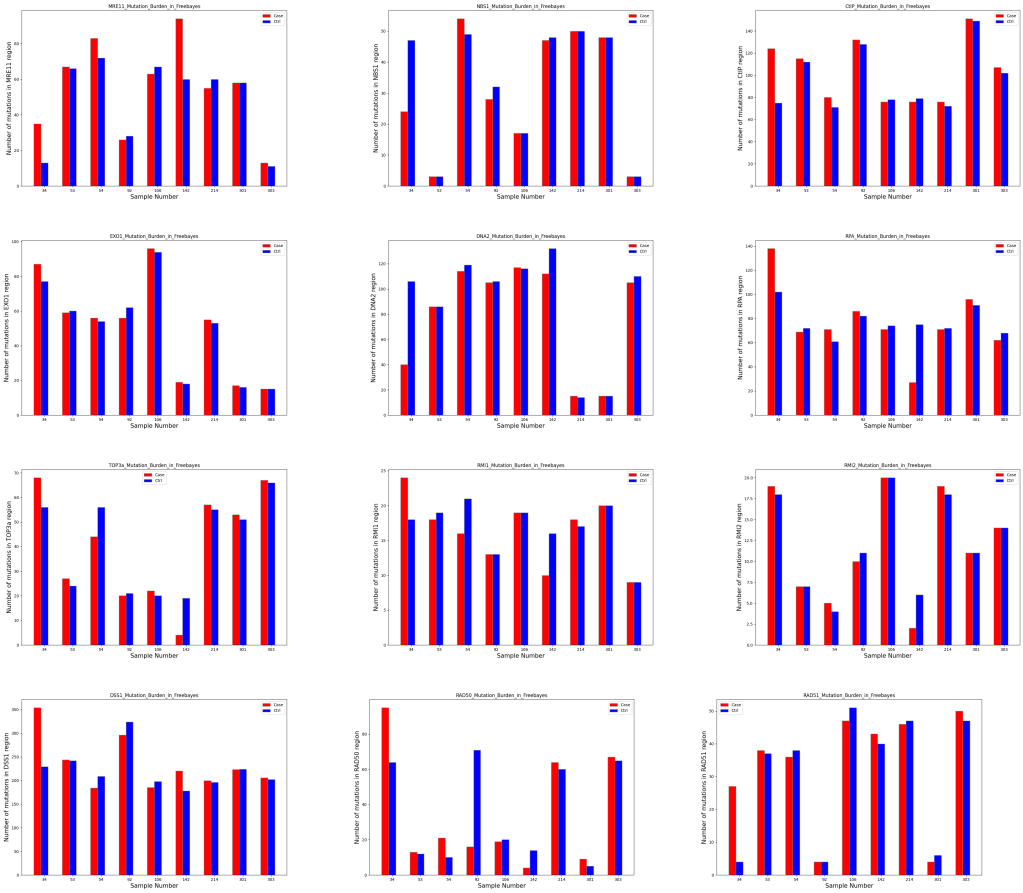

FIGURE 2.11: From left to right, the mutation burden plots for the HR pathway: MRE11, NBS1, CtIP, EXO1, DNA2, DSS1, RPA, TOP3a, RMI1, RMI2, DSS1, RAD50, RAD51

We also analyzed all proteins involved in Homologous Recombination (HR). HR is initiated by the MRE11/RAD50/NBS1 (MRN) complex in conjunction with CtIP via the endonuclease and 3′ to 5′ exonuclease activities of MRE11. The endonuclease activity of MRE11 generates a nick in the double-stranded DNA (dsDNA) near the DSB site, followed by its 3′–5′ exonuclease activity generating a short section of single-stranded DNA (ssDNA) next to the nick. This is followed by extensive resection by exonuclease 1 (EXO1) and/or the nuclease DNA2 with the RECQ helicase Bloom syndrome protein (BLM) to produce a long 3′ overhang. The resulting 3′ ssDNA is rapidly coated by the ssDNA-binding protein replication protein A (RPA), which is subsequently replaced by RAD51 via assistance by BRCA2 and DSS1. A DNA polymerase extends the 3′ end of the invasion strand past the break using the invaded homologous strand as a template, followed by dissolution via BLM-TOP3a-RMI1/2 complex, annealing, and ligation of the extended invasion strand to the other end of the DSB on the original DNA molecule (H and AJ, 2021). Looking at our samples, we see that BRCA2 has a significantly inconsistent but relatively low mutation burden across all samples 2.11. For CtIP and DSS1 we see very high and significantly inconsistent mutation burden, suggesting damage to these proteins is correlative with BLM syndrome. DNA2 and EXO1 have a moderate and consistent

pattern of mutation burden across the samples. The mutation burden on MRE11 is low across all samples, but there is significant inconsistency. NBS1, RAD50, and RAD51 have no significant pattern of mutation burden across samples. Ignoring the Case sample of 142, we see a consistently high mutation burden of at least 60 across all samples for RPA. RMI1 and RM12 have no significant pattern of mutation burden. Ignoring the Case sample of 142, we see a consistently moderate mutation burden of at least 20 across the samples for TOP3a.

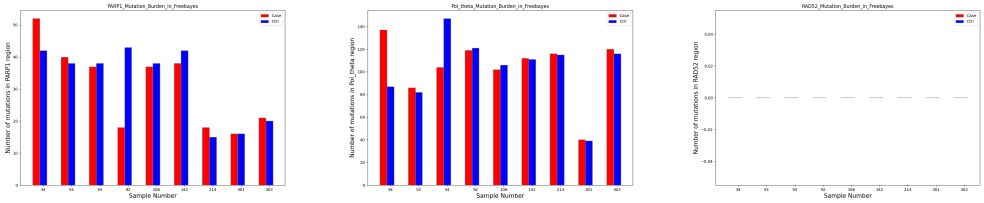

FIGURE 2.12: From left to right, the mutation burden plots for the SSA and MMEJ pathways: PARP1, Polymerase Theta, and RAD52

Finally we analyzed all proteins involved in Strand Specific Annealing (SSA) and Microhomology Mediated End Joining (MMEJ): PARP1 and Polymerase Theta. Both processes initiate with the binding of PARP1. During SSA, the generated 3 ssDNAs are annealed by Polymerase Theta and RAD52 via alignment of homologous sequences. Looking at our Freebayes calls, PARP1 has a moderate but consistent pattern of burden, with at least 10 variants per sample 2.12. Polymerase Theta is heavily burdened in all samples except 301. RAD52 has no detected variants in any of the samples, suggesting it is untouched by the genomic instability of Bloom Syndrome.

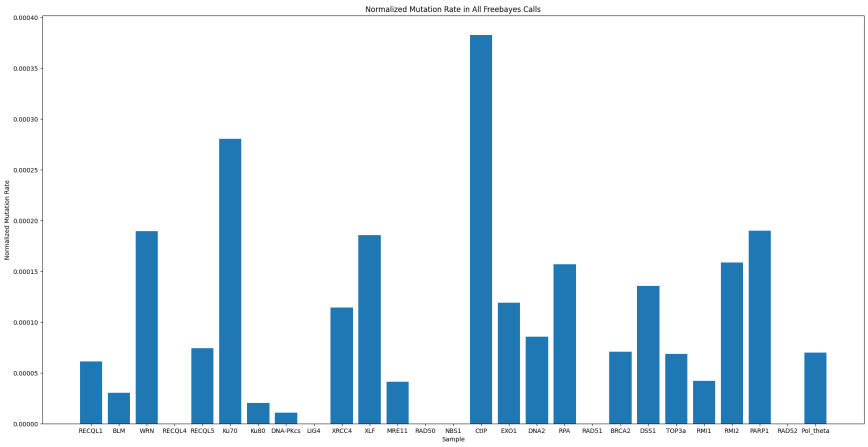

FIGURE 2.13: The genomic size normalized mutation rate for each protein previously described

Looking across samples. and normalizing for genomic size, we see that WRN, Ku70, XRCC4, XLF, CtIP, RPA , DSS1, RMI2, and PARP1 have a mutation burden of at least 100 variants per Megabase 2.13. This would suggest that there are consistent pathogenic events across samples that are caused by this high burden. We also see that sample 34's Case has a consistently high mutation burden compared

to its control and all the other Case samples. However, after screening the samples for pathogenic variants in ClinVar database we saw no known pathogenic variants for any of the proteins shared across the samples. We did see a non-genic known risk factor for colorectal cancer across all samples, as well as 3 intronic hits to APC, a tumor suppressor protein that is an antagonist of Wnt signaling that is also associated with colorectal cancer. All samples also possessed a mutation to NOD2 that is known to confer Leprosy, as well as missense variant to NAGLU which is known to cause lysosomal storage disease. The Freebayes calls also had additional intronic variants of unclear importance.

Chapter 3

CNV results

3.1 CNVkit

CNVkit uses both targeted reads and non-specifically captured off-target reads to infer copy number evenly across the genome via calculation of log2 copy ratios for each sample. This combination achieves both exon-level resolution in targeted regions and sufficient resolution in the larger intronic and intergenic regions to identify copy number changes. To briefly explain, off-target bins are assigned from the genomic positions between targeted regions, with the average off-target bin size being much larger than the average on-target bin to match their read counts. Both the on– and off-target locations are then separately used to calculate the mean read depth within each interval. However, read depth alone is an insufficient proxy for copy number because of systematic biases in coverage introduced during library preparation and sequencing. After normalizing read counts to a pooled reference, reads are evaluated and corrected for three sources of bias that explain most of the extraneous variability in the sequencing read depth: GC content, target footprint size and spacing, and repetitive sequences. A segmentation algorithm can then be run on the log2 ratio values to infer discrete copy number segments using an alignment of sequencing reads in BAM format and the positions of the on– or off-target bins in BED or interval list format.

The CNVkit calls vary massively with respect to the genomic location of gains and losses across samples. Some samples have little gain or loss at all, some have nearly all chromosomes disrupted, others have only chromosome 12 or X disrupted 3.1. All seem to occupy a narrow ploidy band of 2.5 per sample. Comparing directly across all the samples, only chromosome X is consistently lost across 8 of the samples, with no other gains or losses being anywhere near as widespread 3.2.

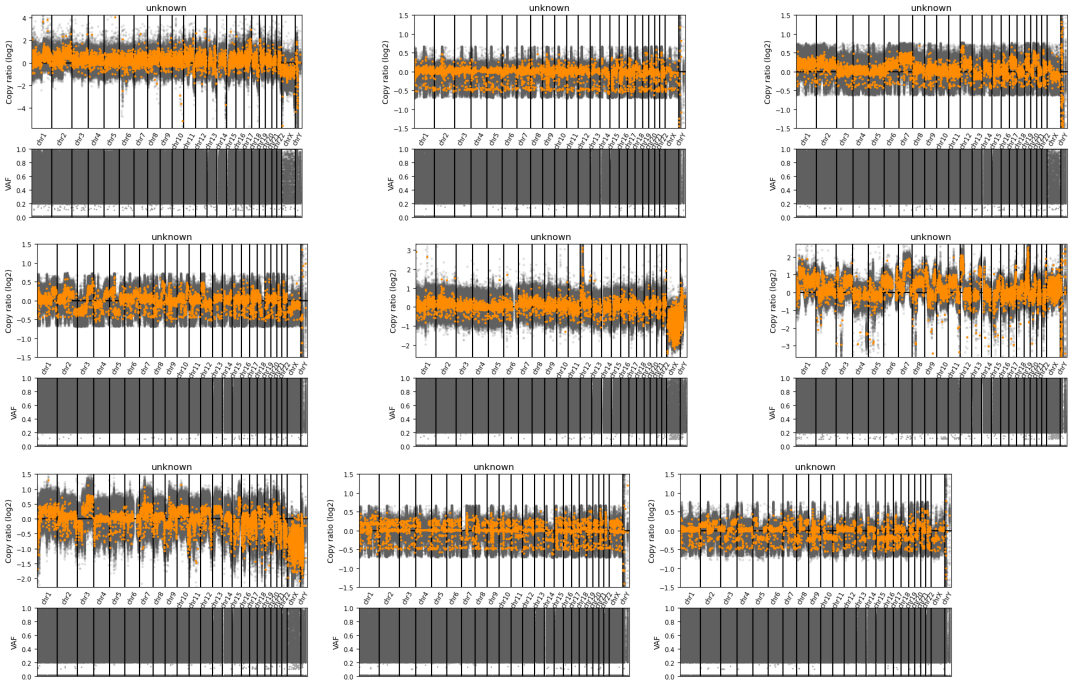

FIGURE 3.1: The LogR ratio and BAF plots for each CNVkit sample

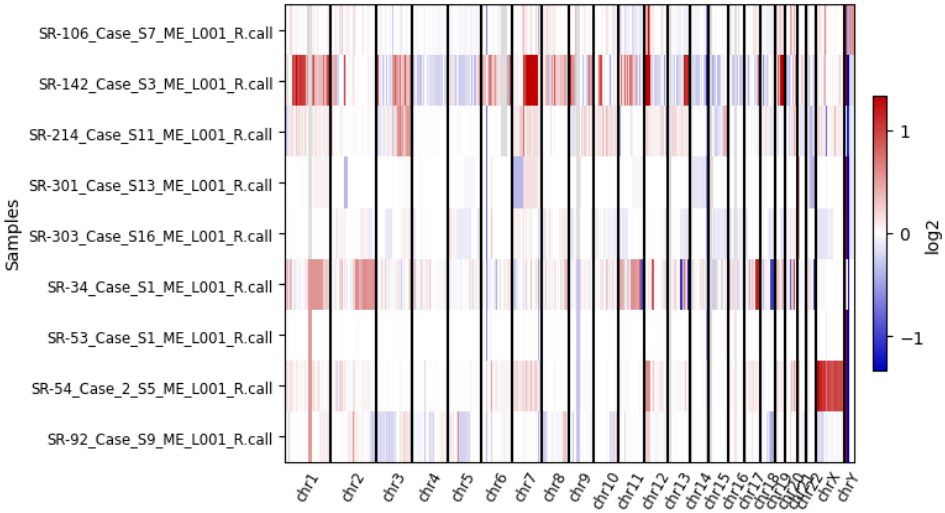

FIGURE 3.2: A heatmap of gains and losses across all CNVkit samples

3.2 ASCAT

ASCAT (allele-specific copy number analysis of tumors) is used to accurately dissect the allele-specific copy number of solid tumors, simultaneously estimating and adjusting for both tumor ploidy and nonaberrant cell admixture. This allows calculation of "ASCAT profiles" (genome-wide allele-specific copy-number profiles) from which gains, losses, copy number-neutral events, and loss of heterozygosity (LOH) can accurately be determined. By aggregation of ASCAT profiles across samples, one can obtain genomic frequency distributions of gains and losses, as well as genome-wide views of LOH and copy number-neutral events. In addition, the ASCAT profiles reveal differences in aberrant tumor cell fraction, ploidy, gains, losses, LOH, and copy number-neutral events (Peter Van Loo, 2010).

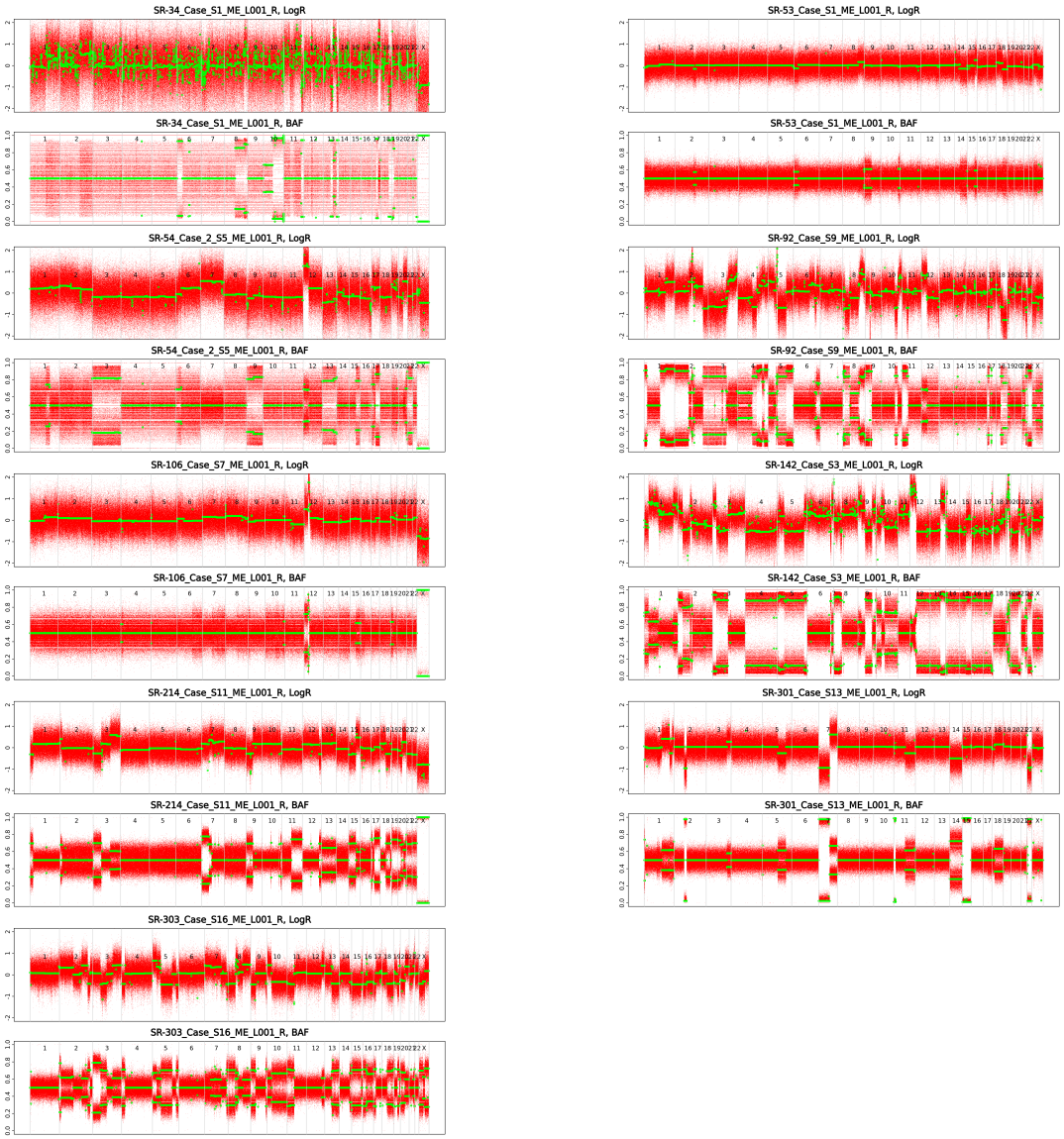

FIGURE 3.3: The LogR ratio and BAF frequency for each ASCAT sample

Illumina SNP arrays deliver two output tracks: Log R, a measure of total signal intensity, and B allele frequency (BAF), a measure of allelic contrast. The Log R track quantifies the (total) copy number of each genomic locus. The BAF track shows the relative presence of each of the two alternative nucleotides (called "A" and "B") at each SNP locus profiled. To make the method less sensitive to noise in the input data, both Log R and BAF are preprocessed by a specially designed segmentation and filtering algorithm, Allele-Specific Piecewise Constant Fitting (ASPCF). As a result, a segmentation of the genome is obtained, each segment corresponding to a genomic region between two adjacent change points (or between a change point and the start/end of a chromosome arm). For Log R, a single fitted value is obtained for each segment, whereas for BAF the output from ASPCF may consist of either one or two values per segment. These ASPCF-smoothed data are subsequently used as input to the ASCAT algorithm, to estimate the aberrant cell fraction and tumor ploidy parameters, as well as the absolute allele-specific copy number calls. The

method optimizes for parameter values such that the allele-specific copy number estimates are as close as possible to nonnegative whole numbers for germline heterozygous SNPs. Optimal parameter values are estimated as a grid search over a narrowly defined space. For each parameter value combination, the total distance to a nonnegative whole-number solution for the genome-wide allele-specific copy number profiles was calculated and summed over all SNPs (Peter Van Loo, 2010).

The ASCAT calls have a mean ploidy of 3 and standard deviation in ploidy of 1. These also have have a mean purity of 57.75 percent and standard deviation of 25 percent. The mean goodness of fit of the samples is 92 percent and standard deviation of 5 percent. Samples 34, 303, and 142 have significant gain or loss disruption across the entire genome, while the other samples have very few events from a changing and small set of hotspots 3.3. Many samples have a gain at the beginning of chromosome of 12, some samples have a loss in chromosome X, others have a gain at the end of chromosome 1. Unfortunately there does not seem to be a clear pattern of gain or loss across the samples.

3.3 Sequenza

Sequenza uses paired tumor-normal DNA sequencing data to estimate tumor cellularity and ploidy, and to calculate allele-specific copy number profiles and mutation profiles. Sequenza detects the correct ploidy in samples with tumor content as low as 30 percent. Tumor tissue specimens comprise a mixture of cancer cells and normal cells; therefore, analysis of tumor data must take the specimen cellularity into consideration. However, it is currently not possible to make a histological estimate of tumor cellularity and extract high-quality DNA from the very same specimen; therefore, cellularity estimates based on histology are commonly made from an adjacent tumor section which often does not reflect the cellularity of the section used for DNA sequencing. However, using the DNA itself to make cellularity estimates is an emerging approach. Sequenza is based on a probabilistic model applied to segmented data. The observations include the average depth ratio between tumor and normal, as well as the B allele frequency. The B allele frequency is the lesser of the two allelic fractions as measured at germline heterozygous positions for each segment. The model parameters include overall tumor ploidy and cellularity, and segment-specific copy number and minor allele copy number. The location of the segments and the segment-level dispersion are taken as known constants. Model parameters are estimated using a maximum a posteriori approach in which prior probabilities are defined for the copy number such that two copies (by default) are preferred over other values. Under this model, given values for cellularity and ploidy, the segment-level parameters can be quickly estimated. The the overall estimation problem is solved using a grid search over reasonable values of cellularity and ploidy (Favero F, 2015).

Sequenza performs GC-content normalization of the tumor versus normal depth ratio, and performs allele-specific segmentation. It is possible with Sequenza to explicitly specify the ploidy rather than determine it by model fitting, but we were unable to do so for our samples. In cases with substantial disagreement in copy number profile the ploidy and cellularity were often the problem. This may be the case with our samples. Finally it is worth noting that Sequenza does not account for possible heterogeneity of mutations within a tumor specimen or uncertainty in the

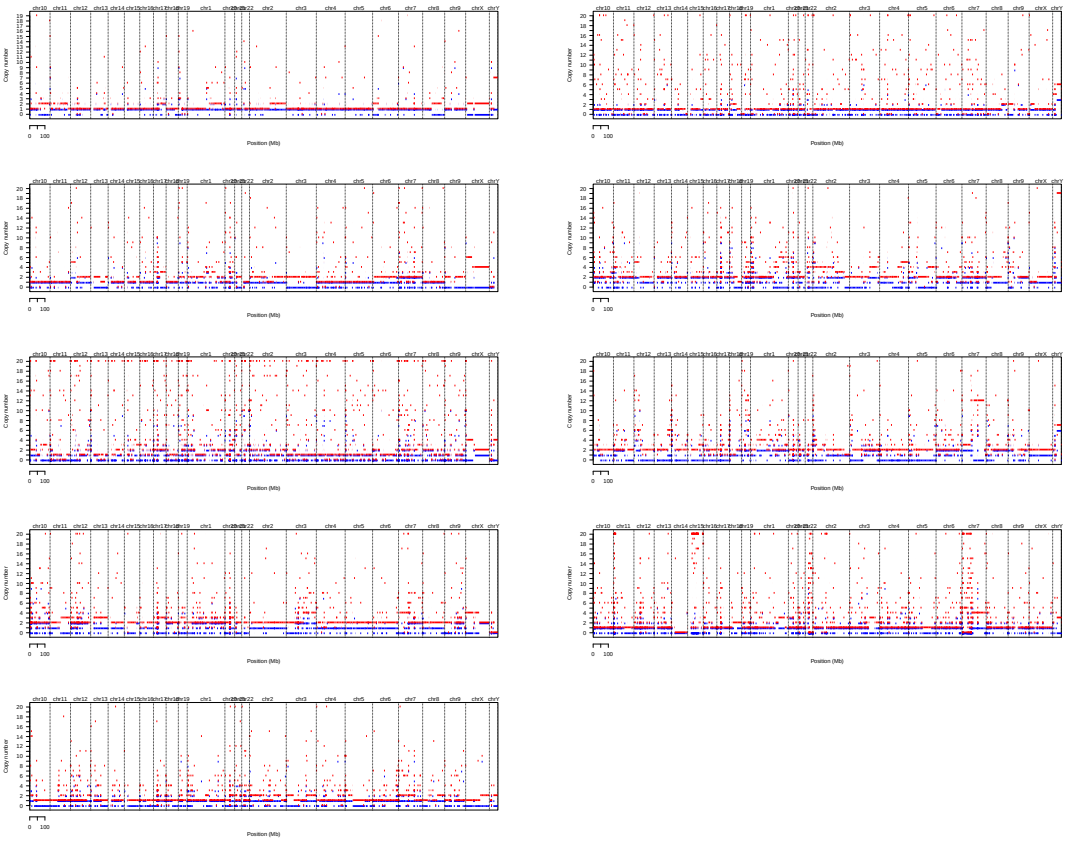

FIGURE 3.4: The allele specific copy number segmentation of each Sequenza sample

assignment of segment boundaries (Favero F, 2015).

The Sequenza calls vary massively with respect to the genomic location of gains and losses across samples. Some samples have little gain or loss at all, others have only chromosome 12 or X disrupted, other samples have random chromosomes disrupted with no clear pattern 3.4. All seem to occupy a narrow ploidy band of 3.25 per sample. Comparing directly across all the samples, there are no consistent gains or losses.

3.4 Comparison of Callers

3.5 COSMIC CNV

The extent of genomic instability can be measured through the number of copy number segments, the proportion of the genome displaying loss of heterozygosity (LOH) and the detection of genome doubling—varied. To distil copy number heterogeneity and to capture biologically relevant copy number features, Alexandrov and others developed a classification framework that encodes the copy number profile of a sample by summarizing the counts of copy number segments into a 48-dimensional vector on the basis of the total copy number (TCN), the heterozygosity status and the segment size. Using SNP6 microarray data, copy number profiles were generated for 9,873 cancers and matching germline DNA of 33 different types from TCGA

using allele-specific copy number analysis of tumours (ASCAT). In addition, a set of whole-genome sequences from 512 cancers of the International Cancer Genome Consortium that overlapped with tumor profiles in TCGA were analysed to generate WGS-derived copy number profiles. A set of 3,175 allele-specific copy number profiles called using the ABSOLUTE algorithm on SNP6 microarray data were obtained, to validate against the corresponding ASCAT profiles for the same samples (Steele, 2022).

Copy number segments were classified into three heterozygosity states: heterozygous segments with copy number of ($A > 0, B > 0$ where A and B are alleles), segments with LOH with copy number of ($A > 0, B = 0$); and segments with homozygous deletions ($A = 0, B = 0$). Segments were further subclassified into five classes on the basis of the sum of major and minor alleles (TCN). The biological relevance is as follows: $TCN = 0$ is homozygous deletion; $TCN = 1$ is deletion leading to LOH; $TCN = 2$ is wild type, including copy-neutral LOH; $TCN = 3$ or 4 is minor gain; $5 \leq TCN \leq 8$ is moderate gain; and TCN$\geq$ 9 is high-level amplification. Each of the heterozygous and LOH TCN states were then subclassified into five classes on basis of the size of their segments: 0–100kb, 100kb–1Mb, 1Mb–10Mb, 10Mb–40Mb and >40Mb. This subclassification was used to capture focal, large-scale and chromosomal-scale copy number changes. In this way, copy number profiles were summarized as counts of 48 combined copy number categories defined by heterozygosity, copy number and size (Steele, 2022).

The CNV mutation spectra for each sample is generally highly discordant between the three callers. An extreme example of this is sample 301, where ASCAT reports whole genome duplication, a major mutation signature enriched in many kinds of cancer, and CNVkit reports a normal ploidy for the sample. For sample 142 there is a similar issue, where ASCAT reports a ploidy close to 5 and CNVkit reports a ploidy close to 3. Generally the ploidy between callers for the same sample is not consistent. The broad genomic locations of the events do line up better between callers though. The mutation spectra for each sample are very inconsistent due to the ploidy problem, with major Loss of Heterozygosity and Heterozygous gain and loss events being highly discordant between callers for the same sample. Strangely there does not seem to be a consistent shift of distribution that one can attribute to only a ploidy difference between the callers. Rather it seems the issue is more complicated than that, with each caller making different shifts in ploidy in either direction for each sample.

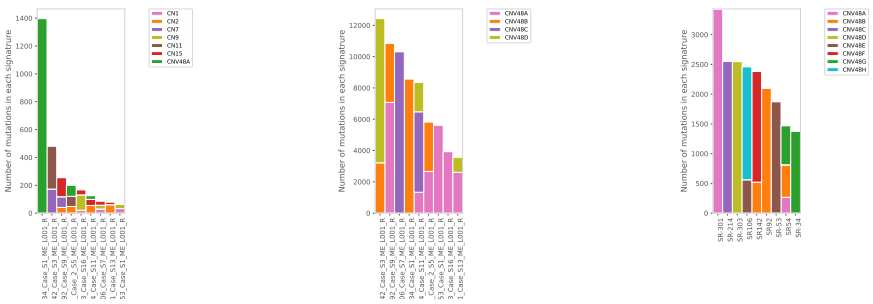

FIGURE 3.5: From left to right, the ASCAT, Sequenza, and CNVkit CNV48 Activity plots

Looking at mutation spectra across all samples for each caller individually, we find no consistent mutation spectra patterns, and very few known COSMIC CNV spectra 3.5. Examining the ASCAT calls, several known mutation spectra are observed in the samples in seemingly random proportions. Sample 34 is highly enriched for an unknown mutation spectra. Looking at the CNVkit and Sequenza calls, we find absolutely no known copy number spectra. This is concerning, but understandable given the samples used to contruct the known COSMIC CNV mutation signatures. Almost all samples were ASCAT calls, with a fraction of the ASCAT SNP6 array calls validated by additional calling with ABSOLUTE. The authors did not validate their results on other copy number callers, despite Sequenza and other callers being mentioned directly on the Alexandrov Lab's github. Taking all this is into account, we can only conclude that we have observed no pattern of known copy number mutation signatures, but rather a series of gains and losses across the entire genome of some samples that is indicative of widespread genome instability.

Chapter 4

SV results

4.1 Delly

DELLY integrates short insert paired-ends, long-range mate-pairs and split-read alignments to accurately delineate genomic rearrangements at single-nucleotide resolution. DELLY is suitable for detecting copy-number variable deletion and tandem duplication events as well as balanced rearrangements such as inversions or reciprocal translocations. DELLY has been specifically geared towards enabling SV calling in the presence of different paired-end sequencing libraries with distinct insert sizes. For each input BAM file, DELLY computes the default read-pair orientation and the paired-end insert size distribution characterized by the median and standard deviation of the library. Based on these parameters, DELLY then identifies all discordantly mapped read-pairs that either have an abnormal orientation or an insert size greater than the expected range (Rausch T, 2012).

DELLY focuses on uniquely mapping paired-ends with a default insert size cutoff is three standard deviations from the median insert size. Any discordantly mapped paired-ends are binned by chromosome and sorted according to the left-most alignment position. For translocations, DELLY sorts all paired ends according to the lexicographically smaller chromosome. This sorted vector of discordantly mapped paired ends is subsequently used to build an undirected, weighted graph that indicates which paired-ends support the same structural rearrangement. For each paired-end, the graph contains one node, and an edge that indicates both paired ends support the same SV. The weight of edge is the absolute difference between the predicted SV sizes induced by the mapping locations of the paired-ends. Since this is not possible for translocations, DELLY takes in this case the sum of the absolute differences in the left-most alignment position of both read-pairs. To achieve maximum specificity, the developers of DELLY only cluster paired-ends that show

the same mapping pattern. In particular, left- and right-spanning paired-ends are clustered separately for inversions, as well as for four types of translocations. The paired-end clusters identified in the previous mapping analysis are interpreted as breakpoint-containing genomic intervals, which are subsequently screened for split-read support to fine map the genomic rearrangements at single-nucleotide resolution and to investigate the breakpoints for potential microhomologies and microinsertions (Rausch T, 2012). The output is a VCF file in a format that is amenable to BED paired end file conversion.

In our Delly calls, all of the samples were significantly enriched for translocations and no other SV events. The Delly calls were around 10,000 events per sample with low variance. There is no consistency across samples of the clustered or unclustered nature of the translocations.

4.2 GRIDSS

The authors of GRIDSS2 showed that breakpoint detection alone was insufficient for the comprehensive characterisation of somatic genomic rearrangements that occur in cancer. As such the author's tool focuses on a new genomic rearrangement primitive: single breakends. GRIDSS2 is the first structural variant caller to explicitly report single breakends—breakpoints in which only one side can be unambiguously determined. By treating single breakends as a fundamental genomic rearrangement signal on par with breakpoints, GRIDSS2 can explain 47 percent of somatic centromere copy number changes using single breakends to non-centromere sequence. On a cohort of 3782 deeply sequenced metastatic cancers, GRIDSS2 achieved an unprecedented 3.1 percent false negative rate and 3.3 percent false discovery rate. GRIDSS2 simplifies complex rearrangement interpretation through phasing of structural variants (Cameron, 2021).

The variant call format (VCF) defines a single breakend as a breakpoint in which only one side can be unambiguously placed. This can occur due to one of two reasons. Firstly, the sequence on one side of the breakpoint could be absent from the reference. Either non-reference sequence could be present due to the integration of foreign DNA (e.g. provirus) or the reference could lack sequence present in the sample. Secondly, breakpoints into highly repetitive regions cannot be unambiguously placed. Single breakends allow the representation of such breakpoints. Such rearrangements are common in cancer and by reporting single breakends the rearrangement landscape of regions previously considered inaccessible to short read sequence can be explored (Cameron, 2021).

Short read-based SV detection algorithms identify breakpoints by finding clusters of reads that do not support the reference allele. Typically, these use discordant read pairs, or split reads, with some callers also considering reads with unmapped mates and soft-clipped reads. More sophisticated callers incorporate assembly either through de novo assembly, targeted breakpoint assembly, or breakend assembly. These callers report breakpoints, that is, novel adjacencies. When reads cannot be unambiguously mapped on either side, a breakpoint call cannot be made and information is lost. Some callers have attempted to address this by considering multiple

alignment locations for each read, but this only works for regions with a small number of potential alignment locations and has proven impractical for general use. Single breakend calling has the potential to improve short read caller sensitivity above the 50 percent reported in recent benchmarking (Cameron, 2021).

GRIDSS2 assembles all reads that potentially support a structural variant using a positional de Bruijn graph breakend assembly algorithm. Breakend contigs are then realigned back to the reference to identify breakpoints and probabilistic structural variant calling is performed based on both the aligned reads and assembled contigs. Single breakend variant calling uses the same probabilistic variant calling approach as breakpoint calling, but instead of split reads, discordant read pairs, and assembly contigs with chimeric alignments support, single breakends are called based on soft-clipped reads, reads with unmapped or ambiguously mapping mates, and assemblies with unmapped or ambiguously mapping breakend sequence. SV phasing is performed based on assembly contigs and the presence of transitive calls. SVs are phased cis if an assembly spans both breaks or a transitive call is found and phased trans if an assembly involves one SV but supports the reference at the other. Since assembly contig length is limited by the library fragment size, only nearby SVs can be phased (Cameron, 2021).

Breakpoints are called using a probabilistic model based on the empirical distribution of CIGAR operators, the library fragment size distribution, and mapping rate. Each read/read pair is given a phred-scaled quality score based on the mapping quality and the probability of encountering that read/read pair given the library empirical distribution. Split reads and soft clipped reads use the distribution of soft clipping CIGAR operators. Discordant read pairs use the discordant mapping mate if distal or the library fragment size distribution if falling within the range reported by Picard tools CollectInsertSizeMetrics. Reads with unmapped mates use the unmapped mate fragment mapping rate and indels based on rate of alignments with insertion/deletion CIGAR elements of matching lengths. As with GRIDSS, split reads and breakpoint-supporting assemblies incorporate the mapping quality scores on both sides of the supported break (Cameron, 2021). The output is a VCF file in a format that is amenable to BED paired end file conversion.

Due to the unique calling convention of GRIDSS the variant call files are massive and poorly annotated. Each of the files has millions of breakpoint events, with very high standard deviation between samples. As such, the files are too large to run significant analysis on, and the individual events are poorly annotated. To be clear, each event is simply a breakpoint rather than something like a "translocation" or "inversion". We attempted to filter the calls based on quality and were able to achieve a 10x reduction in file size, but the files still had the same issues described above. Perhaps in the future we will have a better way to analyze and filter these calls.

4.3 Manta

Manta is optimized for rapid germline and somatic analysis, calling structural variants, medium-sized indels and large insertions on standard compute hardware in less than a tenth of the time that comparable methods require. Manta can discover and score variants based on supporting paired and split-read evidence, with scoring models optimized for germline analysis of diploid individuals and somatic analysis

of tumor-normal sample pairs. Call quality is similar to or better than comparable methods, as determined by pedigree consistency of germline calls and comparison of somatic calls to COSMIC database variants (Xiaoyu Chen, 2016).

Manta operates in two phases: first a graph of all breakend associations within the genome is built, then the components of this graph are processed for variant hypothesis generation, assembly, scoring and VCF reporting. The breakend graph contains edges between any genomic regions where evidence of a long range adjacency exists, indel assembly regions are denoted in this scheme as self-edges. The graph does not express specific variant hypotheses so it is very compact, and can be constructed from segments of the genome in parallel. Following graph construction, individual edges (or larger subgraphs) are analyzed for variants in parallel. Each edge is analyzed to find imprecise variant hypotheses, for which variant reads are assembled and aligned back to the genome. Assembly is attempted for all cases, but is not required to report a variant. All paired and split-read evidence is consolidated to a quality score under either a germline or somatic variant model, and filtration metrics complement this quality score to improve call precision. For ease of use, Manta automates estimation of insert size distribution and exclusion of high depth reference compression regions (Xiaoyu Chen, 2016). The output is a VCF file in a format that is amenable to BED paired end file conversion.

All of our Manta called samples were significantly enriched for translocations and no other SV events. The Manta calls are around 500,000 events per sample with higher variance. All samples had a significant enrichment in unclustered translocations.

4.4 SV Annotation and Pathogenicity

AnnotSV is an annotation pipeline that finds support for the pathogenicity of variants based on a variety of databases. Starting with SV's genomic co-ordinates called from NGS data and available in a standard VCF or BED file, AnnotSV performs the annotation process first by identifying the genomic overlaps between the input and the annotation features. The overlapping criteria can be either a reciprocal or non-reciprocal overlap between the SV and the annotation. AnnotSV generates for each SV (i) one annotation based on the full length SV and (ii) one annotation for each gene within the SV. AnnotSV provides annotations for the sample SVS via overlapping features of the genes/transcripts from RefSeq (i.e. ID, Coding DNA Sequence (CDS), transcript length, SV co-ordinates within the gene), DGV, DECIPHER, 1000 Genomes project (phase 3), OMIM, ExAC, ClinVar, dbVar, and ClinGen (Véronique Geoffroy, 2018). This screens not only for genic pathogenicity but also haploinsufficiency, promoters, disruption of Topologically Associating Domain, deviations in GC content, and known repeated sequences in non-genic space. An ACMG/ClinGen compliant prioritization module then allows the scoring and the ranking of SV into 5 SV classes from pathogenic to benign (Véronique Geoffroy, 2021).

To determine the pathogenicity of our structural variant we ran each of our Delly calls through the AnnotSV pipeline. The Delly calls were chosen as they were the smallest files and were in a parsible format for AnnotSV to use. In the future we

can also run the Manta SV calls through the AnnotSV pipeline. As the resulting annotated files were a few gigabytes each, we subset the annotated VCF files based on pathogenicity score. As previously described, AnnotSV ranks the potential risk of each SV event based on a complex formula that takes into account the known pathogenicity of known SV events. These are output as an integer score from 1 to 5, where 5 is the most pathogenic score possible. To streamline the analysis, we subset all the annotated VCFs based on only those events that scored a 5. With these aggressively subset calls, we then joined all calls between samples that had an overlap of the genomic coordinates of at least 80 percent. This overlap was used as it was easy to observe that there were many events of roughly the same size (100,00 to 200 million base pairs) that were only a few dozen base pairs at the start and end coordinates of the same chromosome. To investigate any common patterns of pathogenic SV events across all the samples, we implemented the previously described analysis as a snowballing join between each of the the level 5 subset annotated SV samples.

The results were very surprising, as there are several massive events that are roughly shared between all calls. At Chromosome 1 we saw 200 million bp duplication around the same breakpoint in all samples, as well as 20 million bp and 60 million bp deletion. The 200 million bp duplication alone affected more than 2000 genes. At Chromosome 2 we saw more than 200 big SVs with definite overlap between dozens of the SVs. This many large SV events suggests possible chromothripsis at this chromosome, but more work is needed. At Chromosome 3 we saw a 100 million bp duplication event across all samples. At Chromosome 5 we saw a 40 million bp duplication and a 5 million bp deletion across all samples. At Chromosome 7 we saw 3 deletions and 2 duplications, all greater than 500,000 bp. At Chromosome 9 we saw one 180 million deletion, and one 30 million deletion. At Chromosome 11 we saw 1, maybe 2 duplications (overlap), each of 50 million bp. At Chromosome 12 we saw 1 duplication and 1 deletion each 200 million bp. At Chromosome 15 we saw 600,000 duplication. At Chromosome 16 we saw one deletion and duplication of 60 million bp, same break points. At Chromosome 17 we saw one duplication and deletion at same breakpoint of 5 million, and another 60 million bp deletion. At Chromosome 18 we saw one deletion of 40 million across all samples. At Chromosome 19 we saw one deletion of 300,000. At Chromosome X we saw one deletion and duplication of 3 million bp at same breakpoint. At Chromosome Y we saw one (40 million) duplication and one deletion at around the same breakpoint coordinates. With event sizes in the tens of millions of base pairs, it is no surprise that these events were marked as maximally pathogenic. While this line of analysis seems promising, further investigation must be left to the future due to the time constraints of my graduation.

Chapter 5

Conclusion

In conclusion, we see complicated patterns of mutation in our samples that are difficult to directly tie to distinct aetiologies. The mutation burden on BLM and other

RecQ Helicases seem to suggest that damage to RecQ Helicases has high correlation to genome disruption in our Whole Genome Samples. Additional damage to other DNA repair proteins could suggest some correlative effect, but such substitutions are not conserved across all samples. Across our SNV calls, we see SBS1 and SBS5 consistently across all samples even after filtering, and see additional signatures in the most disrupted samples 34 and 54. We also see a less consistent but likely pattern of DBS1 across all samples. We see no known mutation signatures across our ID samples after filtering, suggesting the BS patients have no notable ID signatures above the general populace genomic background. We also see no consistent CNV signatures across our callers, even for ASCAT. This is particularly surprisingly, as known signatures were predominantly extracted from ASCAT CNV calls. This suggests that the pattern of CNV mutation is seemingly too complex, rare, or random for COSMIC CNV to meaningfully contribute information. Looking at our SV calls, we see consistently high patterns of translocations across all samples and all callers. This is potentially promising, but additional analysis through complex calling is necessary. Likewise, we see potentially promising signs of consistent and large SV events conserved between all samples, but more work is needed. We believe we have only found weak patterns of mutation across the samples with this existing analysis, but it is likely that with more in depth analysis of SV calls we may come to more useful conclusions.

In the future, we would like to extend our analysis in several directions. First, with the existing data we would like to get more information from our structural variant calls. The Marcin Imieliński lab has produced a pair of tools called Jabba and GGnome that would allow us to re-call each of our structural variant calls to look for complex events. These complex events include chromoplexy, chromothripsis, pyrgo, rigma, and tyfonas. Chromoplexy is when large chains of rearrangements that affect multiple chromosomes occur. Chromotripsis is chromosome shattering, when thousands of clustered chromosomal rearrangements occur in a single event in a confined genomic region. Pyrgo are towers of low junction copy number duplications that are associated with early-replicating regions and superenhancers. Rigma are low junction copy number deletions that are enriched in late-replicating fragile sites and gastrointestinal carcinomas. Tyfonas are high junction copy number regions and fold-back inversions associated with expressed protein-coding fusions and breakend hypermutation (Hadi et al., 2020). Through these complex calls we may be able to better trace the mutational processes and thus aetiologies beyond loss of function of BLM that shape the genome instability of BS.

We would also like to extend the analysis of the clinical significance of our structural variant calls. In particular we would like to run the AnnotSV pipeline on our Manta calls, and if possible, additionally on our GRIDSS calls. This would give us several points of reference for each sample to see if there are consistent patterns of pathogenic structural variants. While the snowballing join with overlap procedure could be repeated without AnnotSV annotations, the script has $O(n^3)$ runtime complexity and so scales poorly to the smallest files with only 10,000 events. By aggressively subsetting the structural variant calls to only those with the highest pathogenicity, and running the same script across all callers for each sample, we could see if there were consistent patterns of pathogenic SVs between callers for each sample. Similarly, we could run the same script across all samples and all callers and see if there are any pathogenic SVs conserved across all samples and all callers. Additionally, further investigation into the common diseases shared by the

joint pathogenic variants and their possible relation to Bloom's Syndrome could be very useful. Generally speaking, a much more thorough analysis of the pathogenically annotated SVs is in order.

If we had more time and more resources, we would like to do things a little differently. If possible, a larger cohort with a higher average read depth per sample would greatly increase the power of our study and better let us gauge the consistency between callers. Additionally it would be nice to have good histological estimates of the purity and ploidy of each sample, so we could eliminate any difference in estimation between copy number callers. This way we would statically fix the hyper parameters and hopefully have more concordance between our callers.

In any situation, our analysis has seemingly made clear that there are few tools to properly understand the significance of our structural variant calls in a scalable way. Genome graphs and other simple visualizations fall apart for even our smallest files as there simply are too many SV events in each sample. What is needed is a tool akin to COSMIC's signature extraction but for SV calls. However the featurization of the mutations (what to count) and the presumed latent space (the mutational processes contributing different mutation signatures) for SVs are unclear. Currently there is a hierarchical dirichlet process model that extracts signatures from SV calls (Li, 2020). However it is not a tool, rather it is a commented R notebook with little documentation. Similarly, the featurization of the mutations is not complex and the tool lacks any known aetiologies like those that COSMIC provide. These aetiologies are crucial to making sense of what could be causing the mutations we see in the samples. In this way we could hopefully find concordant mutational process(s) to correlate with Bloom's Syndrome given our variant calls. Thankfully others in the Schwartz Lab are working on new tools that can better account for and understand SVs. Hopefully some day soon we will have a better understanding of SVs through a more sophisticated analysis of the raw reads, aligned BAM files, or SV called VCFs.

Bibliography

Alexandrov L.B., Kim J. Haradhvala N.J. et al. (2020). "The repertoire of mutational signatures in human cancer". In: *Nature*, pp. 94–101. URL: https://doi.org/10.1038/s41586-020-1943-3.

Cameron D.L., Baber J. Shale C. et al. (2021). "GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing". In: *Genome Biology*. URL: https://doi.org/10.1186/s13059-021-02423-x.

Christopher T. Saunders Wendy S. W. Wong, Sajani Swamy Jennifer Becq Lisa J. Murray R. Keira Cheetham (2012). "Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs". In: *Bioinformatics*, pp. 1811–1817. URL: https://doi.org/10.1093/bioinformatics/bts271.

Cunniff, Christopher et al. (2017). "Bloom's Syndrome: Clinical Spectrum, Molecular Pathogenesis, and Cancer Predisposition". In: *Molecular syndromology*, pp. 4–23. URL: doi:10.1159/000452082.

Favero F Joshi T, Marquard AM et al. (2015). "Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data". In: *Ann Oncol*, pp. 64–70. URL: doi:10.1093/annonc/mdu479.

Fusco, Michael J., Howard (Jack) West, and Christine M. Walko (Feb. 2021). "Tumor Mutation Burden and Cancer Treatment". In: *JAMA Oncology* 7.2, pp. 316–316. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2020.6371. eprint: https://jamanetwork.com/journals/jamaoncology/articlepdf/2773840/jamaoncology_fusco_2020_pg_200004_1613590913.70503.pdf. URL: https://doi.org/10.1001/jamaoncol.2020.6371.

Garrison, Erik and Gabor Marth (2012). "Haplotype-based variant detection from short-read sequencing". In: DOI: 10.48550/ARXIV.1207.3907. URL: https://arxiv.org/abs/1207.3907.

H, Lu and Davis AJ (2021). "Human RecQ Helicases in DNA Double-Strand Break Repair". In: *Front. Cell Dev. Biol.* URL: doi:10.3389/fcell.2021.640755.

Hadi, Kevin et al. (2020). "Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs". In: *Cell* 183.1, 197–210.e32. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2020.08.006. URL: https://www.sciencedirect.com/science/article/pii/S0092867420309971.

Koboldt, D.C. (2020). "Best practices for variant calling in clinical sequencing". In: *Genome Medicine*. URL: https://doi.org/10.1186/s13073-020-00791-w.

Larson DE Harris CC, Chen K et al (2012). "SomaticSniper: identification of somatic point mutations in whole genome sequencing data". In: *Bioinformatics*, pp. 311–317. URL: doi:10.1093/bioinformatics/btr665.

Li Y., Roberts N.D. Wala J.A. et al. (2020). "Patterns of somatic structural variation in human cancer genomes". In: *Nature*. URL: https://doi.org/10.1038/s41586-019-1913-9.

Mahmoud M., Gobet N. Cruz-Dávalos D.I. et al. (2019). "Structural variant calling: the long and the short of it". In: *Genome Biology*. URL: https://doi.org/10.1186/s13059-019-1828-7.

Pedersen B.S., Brown J.M. Dashnow-H. et al. (2021). "Effective variant filtering and expected candidate variant yield in studies of rare human disease." In: *npj Genom. Med.* URL: https://doi.org/10.1038/s41525-021-00227-3.

Peter Van Loo Silje H. Nordgard, Ole Christian Lingjærde et. al. (2010). "Allele-specific copy number analysis of tumors". In: *PNAS*, pp. 16910–16915. URL: https://doi.org/10.1073/pnas.1009843107.

Prandi D, Demichelis F. (2019). "Ploidy- and Purity-Adjusted Allele-Specific DNA Analysis Using CLONETv2". In: *Curr Protoc Bioinformatics*.

Rausch T Zichner T, Schlattl A Stütz AM Benes V Korbel JO (2012). "DELLY: structural variant discovery by integrated paired-end and split-read analysis". In: *Bioinformatics*, pp. 333–339. URL: doi:10.1093/bioinformatics/bts378.

Steele C.D., Abbasi A. Islam S.M.A. et al. (2022). "Signatures of copy number alterations in human cancer". In: *Nature*, pp. 984–991. URL: https://doi.org/10.1038/s41586-022-04738-6.

Véronique Geoffroy Thomas Guignard, Arnaud Kress et al. (2021). "AnnotSV and knotAnnotSV: a web server for human structural variations annotations, ranking and analysis". In: *Nucleic Acids Research*, W21–W28. URL: https://doi.org/10.1093/nar/gkab402.

Véronique Geoffroy Yvan Herenger, Arnaud Kress et al. (2018). "AnnotSV: an integrated tool for structural variations annotation". In: *Bioinformatics*, pp. 3572–3574. URL: https://doi.org/10.1093/bioinformatics/bty304.

Xiaoyu Chen Ole Schulz-Trieglaff, Richard Shaw et al. (2016). "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications". In: *Bioinformatics*, pp. 1220–1222. URL: https://doi.org/10.1093/bioinformatics/btv710.