CHAPTER THREE: OPTIMISATION OF MUTATION-CALLING IN ORDER TO OBTAIN A CURATED CATALOGUE OF SOMATIC SUBSTITUTIONS FOR DOWNSTREAM ANALYSES

3.1 INTRODUCTION

Obtaining raw sequence data for twenty-one breast cancer genomes was only the beginning of a complex process that required multiple iterations of computational processing in order to translate raw sequence data into a comprehensive list of somatic variants. In order to excavate the cancer genome for patterns in somatic substitutions, insertions/deletions and rearrangements, it was critical to obtain a set of high-confidence mutations with high specificity i.e. a low false positive rate.

Calling single nucleotide substitution and insertion/deletion variants from short-read sequencing data can be problematic in general but is particularly so in cancer genome sequences. A general issue associated with sequencing of short read data includes decline in sequencing qualities at lattermost cycles of the sequencing-by-synthesis process. In addition, certain sequencing motifs (for example strings of G bases (–GGGG)) have been known to cause an increase in polymerase errors resulting in sequencing errors immediately following such motifs (Abnizova et al., 2012). Inaccuracies in base assignment following photo-laser capture can also occur. The confidence in a base call made during the sequencing process is simply the probability estimate of that base call being a true nucleotide. The likelihood of the accuracy of a base call is reflected in the base quality score or Phred score. Base qualities can therefore be taken into account when considering variant calls. Finally, accurate mapping of short-read data to the reference genome can be hampered by the large proportion of repetitive sequence in the human genome. Errors in mapping can be seen, for example, as excessive coverage in certain regions of the genome due to inaccurately assembled reference genome sequence given by sites of low complexity. Non-unique mapping is reflected in mapping qualities and like base qualities, can be taken into consideration when curating catalogues of variants.

In whole-genome sequencing family-based studies of the germline, relatives enable the efficient elimination of errors based on Mendelian inheritance patterns and knowledge of parental haplotype blocks. This has, in fact, permitted successful identification of genes underlying a host of inherited disorders despite sequencing very few individuals, for example Miller syndrome and Freeman-Sheldon Syndrome (Ng et al., 2009; Roach et al., 2010). Furthermore, the digital nature of next-generation sequencing technology provides additional means of supporting variant-calling in the germline. For every base in the genome, coverage of 40-fold would mean that sequencing information from 40 DNA molecules is available at that particular genomic coordinate. A heterozygous mutation in the germline would be expected to be present in approximately 50% of

reads for a diploid genome and a homozygous mutation should be present in 100% of reads (Figure 3.2). Using this reasoning, sequencing artefacts that arise in just a small proportion of reads, for example, could be filtered from the variant dataset.

Figure 3.1

(a) Q = -10 * log(E)

(b) mE = 10 ^ (-mQ / 10.0)

Figure 3.1 Phred score and mapping qualities. (a) A base quality score or Phred score is a score of an estimate of a base call being the true nucleotide. The probability that a base call is wrong is called an error probability. If the error probability of a base call is E, then the Phred base quality score is ***Q*** where is as seen in the figure. If the quality of a base call is 30, the probability that it is wrong is 0.001. Therefore, on average 1 in every 1000 base calls with Q=30 is erroneous. (b) A similar principle applies for mapping qualities. Each read alignment is a probabilistic estimate of the true alignment. If the mapping quality of a read alignment is ***mQ***, the probability ***mE*** that the alignment is wrong is as above. Once again, one in every 1000 read alignments with mapping quality of 30 will be wrong on average.

The approach of using Mendelian-based elimination of errors cannot be applied directly to cancer genome sequencing. On top of the general problems associated with the next-generation sequencing process and mapping of short-read sequencing data, digital calling of variants in cancers is plagued further by issues such as intra-tumour heterogeneity, contamination by normal cells and marked abnormalities of ploidy. Unlike calling mutations in the diploid human germline genome, calling of variants in cancer requires consideration of these additional parameters in order to maximize the likelihood of detection (Figure 3.2). This however, may come at a cost on the specificity or the false positive rate of variant-calling.

In the last few years, multiple substitution-calling algorithms have been published although many of these result in an extremely large numbers of variants which turn out to be errors or false positive calls. Given the high false positive rate in studies utilizing short-read sequencing technology for the detection of somatic single-nucleotide variants, independent mid- to large-scale validation experiments has been obligatory, (preferably) on an orthogonal platform in order to avoid reproducing systematic sequencing artifacts. For instance, more than 500 somatic substitutions in a lung cancer were validated using mass spectrometry (Lee et al., 2010) whereas other studies re-sequenced hundreds of substitution variants using Sanger sequencing (Pleasance et al., 2010a; Pleasance et al., 2010b). These validation experiments rapidly become as costly as the initial discovery experiment and are labour-intensive.

Across the cancer genomics community, filters have been developed and applied to raw variant-called datasets in order to reduce the false positive rate. However, there is little consensus on what

filters should be used and at what threshold applied. Additionally, the extent to which filters discard true variants has not been formally assessed.

To the best of my knowledge, there is only one example of a study which has documented the challenges of variant-calling from short-read data and provided some detail on additional processing of raw data in order to obtain a set of high-confidence substitutions (Reumers et al., 2012). In that study, post-processing filter optimisation was performed on whole genome sequences in the germline obtained from a pair of identical twins. The authors reasoned that shared variants were more likely to be real whereas discordant variants were more likely to be false positive calls. Filters were optimised to remove as many discordant single nucleotide variants and as few shared ones as possible. There were drawbacks with this analysis. First, systematic sequencing artifacts with a predilection for certain sequence motifs were precisely the sort of systematic false positives that could be shared between twin genomes. Their metric for measuring the effectiveness of filters, which was based on the ratio of shared versus discordant variants, was therefore systematically overestimated. Second, aggregation of the fraction of the genome removed across the filters meant that up to 32% of the genome could be removed. Third, and acknowledged by the authors, calling of variants in tumour-normal pairs of ovarian cancer was attempted, and although generally was able to call variants, was plagued by difficulties in over-calling mutations at regions of extremes of ploidy (zones of amplification and loss-of-heterozygosity). Furthermore, validation of their method concentrated on coding regions of these cancer genomes. Coding sequences are generally more unique, show more sequence complexity and are less troubled by false positives than intronic/intergenic regions, and again this validation step is likely to have overestimated the effectiveness of their filters.

In this chapter, the challenge of distinguishing true mutations from errors in whole genome sequences is deliberated using substitution-calling as a foremost example, and the solutions that have been created, in the form of post-processing filters, are described.

Figure 3.2

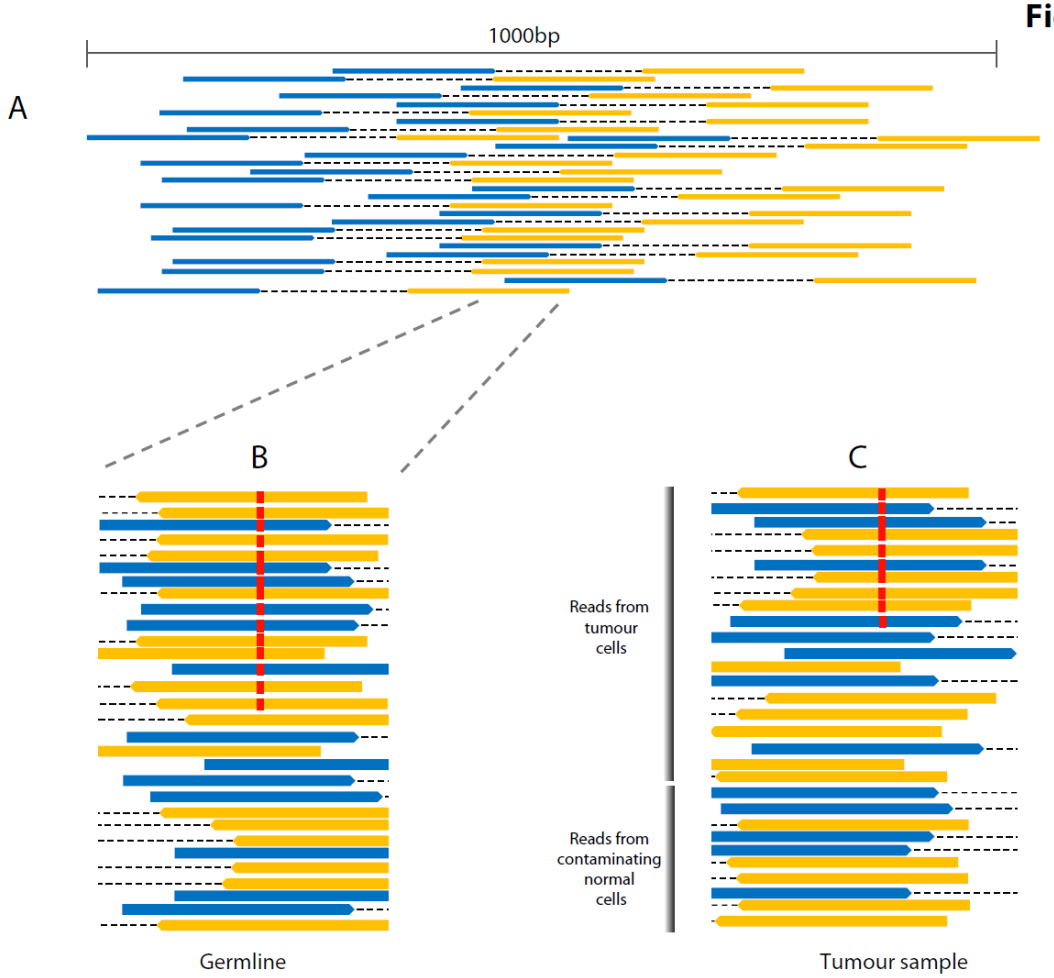


Figure 3.2: Differences in calling variants in a germline genome and a tumour genome. (A) Blue and yellow reads joined by a dotted line represent forward and reverse reads respectively of a 500bp fragment. (B) A higher resolution depiction of (A), which is a germline sample, showing 30-fold coverage of reads in the region of interest. The red marks represent a variant allele which is different to the reference genome. This heterozygous SNP in the diploid germline genome is seen in approximately 50% of reads or has a variant allele fraction of 0.5. (C) This higher resolution schematic of a tumour sample also has 30-fold coverage but has 1/3 of reads originating from *contaminating normal cells*. In this region which is diploid in the tumour, the somatic variant is a heterozygous mutation and is present at a lower variant allele fraction (when compared to the germline genome) of 0.33. However, if the variant allele fraction of a true variant is lower than expected for the level of ploidy and normal contamination, then this may be taken as evidence of a somatic mutation in a subclonal population (*intra-tumoural heterogeneity*). In contrast, a *polyploid* region where a somatic variant is present on only 1 of multiple alleles will be present at a much lower variant allele fraction.

3.2 THE METRICS USED FOR THIS ANALYSIS

In order to track the improvements in the performance of the mutation-calling and post-processing procedure, some statistical measures of the performance of a binary classification test (where a mutation is called as somatic or not) was required. Sensitivity, or the recall rate, measures the proportion of true positives which are correctly identified (e.g. the percentage of affected people who are correctly identified as having the condition). Specificity measures the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition). An alternative metric which is easier to calculate for the purposes of this analysis is the positive predictive value (PPV). This metric measures the proportion of positives which are correctly identified. Specificity and the positive predictive value are sometimes used interchangeably although in theory reflect subtly different concepts.

The two measures of sensitivity and specificity are closely related to the concepts of type I and type II errors. The perfect algorithm would have 100% sensitivity and specificity. However, for any test, there is usually a trade-off between the measures. In this thesis, it was in theory impossible to measure the sensitivity given that a priori knowledge of mutations in any given cancer was not known. However, an attempt was made to infer sensitivity from a cross-comparison with a high-confidence set of mutations produced by an alternative substitution-calling algorithm produced by Illumina© as well as a cross-comparison with whole exome sequences for 3 samples. The positive predictive value (PPV) was the metric that was used to track the progress and improvements in mutation-calling and post-processing.

3.3 CaVEMan IS A BESPOKE SUBSTITUTION-CALLING ALGORITHM

An in-house bespoke substitution-calling algorithm, CaVEMan (Cancer Variants Through Expectation Maximization) was used for calling somatic substitutions. CaVEMan is a naïve Bayesian probabilistic classifier which utilizes the expectation maximization (EM) algorithm and is designed for calling substitution variants in new sequencing technology reads. Given prior information regarding reference and variant alleles, copy number status or ploidy, fraction of aberrant tumour cells present in each cancer sample and quality scores relating to sequencing and mapping, CaVEMan generates a probability score for potential genotypes at each genomic position. CaVEMan requires mapped, paired-end reads in the form of a sorted and indexed BAM file for the tumour and matched normal samples. An indexed reference sequence in FASTA format is also a prerequisite.

There are two main steps in the core CaVEMan algorithm. The first *maximization* or *M*-step generates a prior depiction of each genomic position by gathering data from all valid reads (reads that are properly paired and not marked as duplicates) that are available at that coordinate. These data or covariates include read information (1^{st} or 2^{nd} read of a pair), mapping qualities of the reads, lane information, base qualities, the expected reference allele (A, C, G or T), the variant allele (A, C, G or T) and the position of the variant in the read. CaVEMan iterates through each genomic position generating a multi-dimensional array of information in order to build an "error profile" for each coordinate.

The second *expectation* or *E*-step uses this profile to generate a probability for each possible genotype at this position, again iterating through each position in the genome. A number of parameters can be set to enhance the accuracy of the probability estimates in cancer. The degree of contamination from normal cells as well as the ploidy of each section of the genome (both obtained from SNP6 copy number analysis) can be provided to CaVEMan in order to enhance mutation-calling. In order to produce a set of raw variants, other parameters that are factored into this step include mutation rate (6e-6), SNP rate (1e-3), reference bias (0.95), a SNP probability cutoff (0.95) and a mutation probability cut-off (0.8). At the end of the *E*-step, a list of potential genotypes at each base is obtained. Three output files are generated following this process; a "raw substitutions" output file for those variants in which the sum of the genotype probabilities exceeds the mutation probability cut-off, a "raw SNPs" file if the sum of the SNP genotype probabilities exceeds the SNP probability cut-off and an "uncategorised" file for variants which meet neither of these criteria.

On average, tens of thousands of variants per raw output of breast cancer sample were obtained (Table 3.2). However, these variants were unlikely to all be true somatic variants. In the following section, the development of filters in order to remove false positive calls (post-processing) will be described.

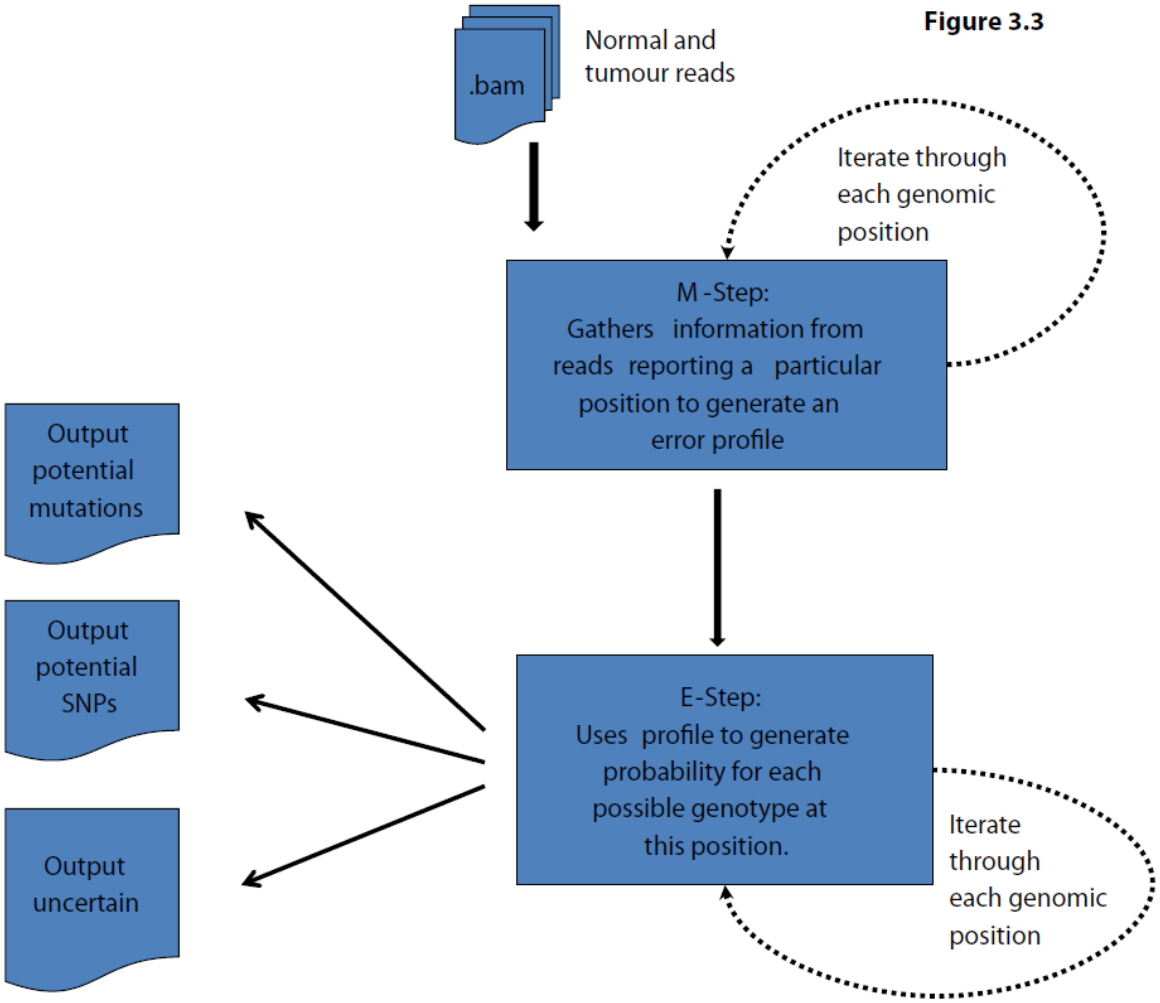


Figure 3.3: The CaVEMan workflow. CaVEMan takes a BAM file as an input file and performs two main steps, the M-step and E-step before generating three output files, a file of potential somatic substitutions, a file of possible SNPs and a file for variants that meet neither of the criteria for the other two files.

3.4 A FIRST COMPARISON BETWEEN DATA FROM CaVEMan AND THE ILLUMINA SUBSTITUTION-CALLING ALGORITHM REVEALED GOOD SENSITIVITY AND ALLOWED IDENTIFICATION OF FALSE POSITIVES FOR THE DEVELOPMENT OF EARLY FILTERS

The first and only breast cancer sample to be sequenced at Illumina© was PD3890a. 4836 highly-filtered high-confidence substitution variants were identified using the Illumina© substitution-calling algorithm. 201 variants were selected for validation by Sanger sequencing, comprising all coding variants and a random selection of non-coding variants. 168 were confirmed as somatic (83.6%) and 33 were found to be false positive calls (16.4%). The PPV of the Illumina substitution-calling process was 83.6%. This PPV was, however, possibly an overestimate of the true PPV of the Illumina © substitution-calling algorithm. Variant selection for validation was targeted to the coding exons where genomic sequence shows higher complexity. These variants were more likely to be called correctly and to therefore be true somatic variants, given the favourable mapping characteristics of the coding sequence.

On the first iteration of CaVEMan, 76235 raw variants were called in PD3890a. 100% of the 4836 variants identified by Illumina were present in this raw list of CaVEMan variants. All of the 168 confirmed somatic variants were identified demonstrating that the sensitivity or the ability to recall true variants was high. However, the total number of variants called by CaVEMan was vastly more than Illumina, likely to be overwhelmed by a variety of mis-calls and unlikely to reflect the true mutation burden in the cancer. Therefore, some early intrinsic filters were used to remove potential false positive variants whilst maintaining the number of true somatic variants.

3.5 EARLY POST-PROCESSING FILTERS

The earliest thresholds used were relatively simple. Firstly, only variants with a high likelihood (of 0.95 and above) were retained (*Mutation Probability Threshold*). Secondly, it was reasoned that a variant reported in the tumour had to be appropriately represented in the tumour. Substitution variants were identified as mismatches relative to the reference genome (Figure 3.4). However, true substitution variants were usually of a good base quality. In contrast, false positive calls arising from sequencing artifacts could also present as mismatches but were frequently at lower base qualities. With this knowledge, putative somatic variants were required to be appropriately represented in the tumour with at least a third of the reads carrying the variant allele showing a base quality score of more than or equal to 25 (*Read Depth*). Thirdly, it was considered that any putative somatic variant should not be present in the matched normal sample as well. A variant present in 5% of reads or more in the matched normal sample at base qualities of 15 or more would fail this filter and be excluded from further analysis (*Matched normal*). Using these three main criteria, the total number of variants fell to 21659 from 76235 variants.

Figure 3.4

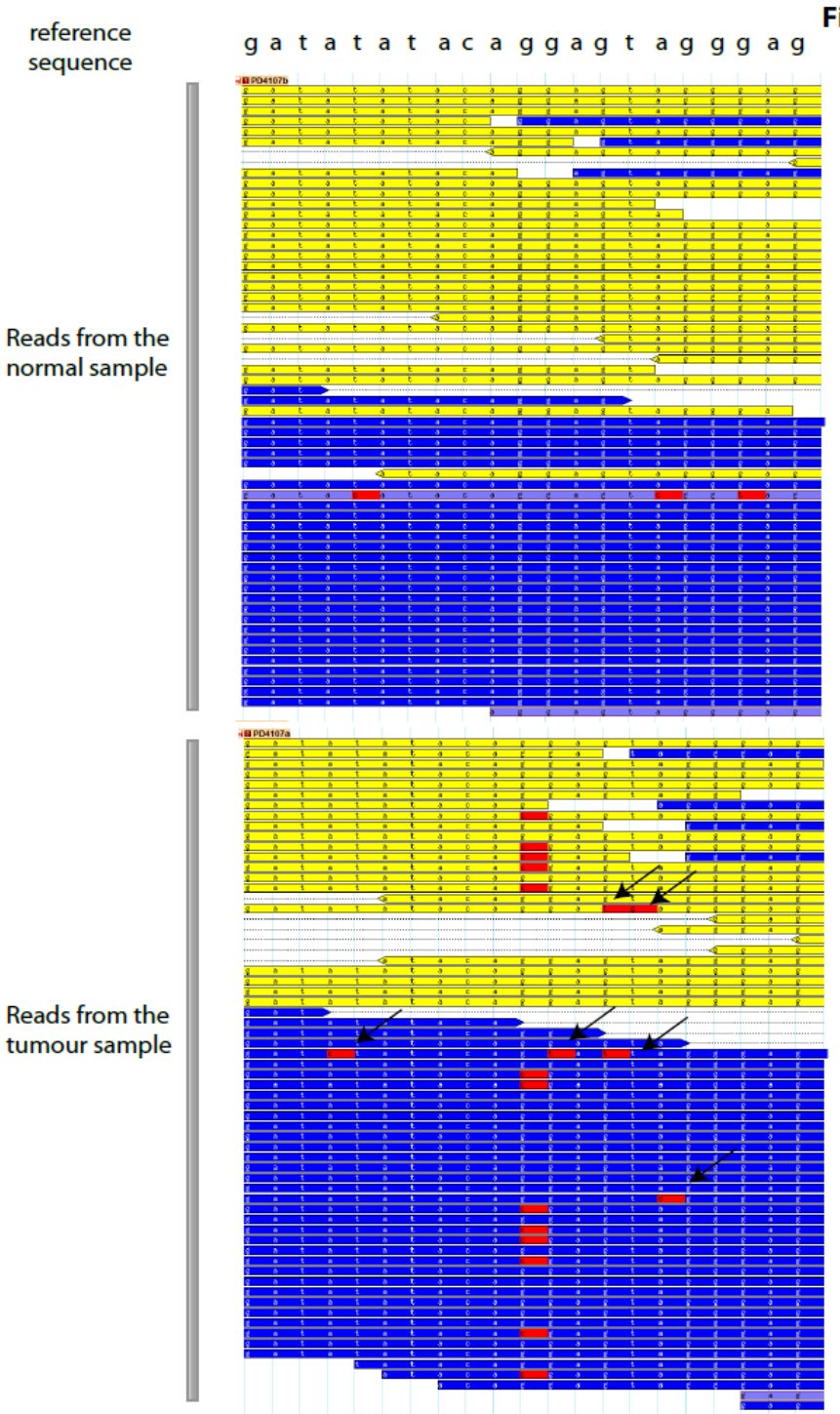


Figure 3.4: Reads in G-browse, the genome browser used to view short read sequences. Blue and yellow reads represent next-generation sequencing reads in the forward and reverse directions respectively. Each base in the reference genome is re-sequenced many-fold. The intensity of the colour reflects the mapping quality of each read. The dotted line joins each read to its read-mate. The reads on top represent reads from the matched normal and the reads below represent the tumour sample. Each read represents sequencing information from a single DNA molecule. A sequenced base which correctly matches the reference genome is not highlighted. In contrast, a base which is different to the reference genome appears (a mismatch) appears red. Here, 13 of the 57 reads in the tumour carry a G>C mismatch at the same genomic coordinate whilst no reads in the normal carry the same mismatch, corresponding to a somatic heterozygous change at this location. Note that 6 other mismatches can be seen within the same screenshot (arrows) in the tumour which represent mismatches arising as random sequencing errors or arising from mismapped reads. However, the mutation probability estimates of these randomly distributed errors are not sufficient to being called as a somatic variant.

250 variants were selected for validation by Sanger sequencing at this stage in order to identify the true PPV of CaVEMan and to identify the nature of the false positive variants that remained. Of these, 58% were confirmed as somatic (Figure 3.5a). 42% showed no evidence of the somatic variant by Sanger sequencing and were declared false positives. Of these false positive calls, 3% fell within the vicinity of germline indels, 7% were within or immediately adjacent to repeat tracts, 7% were germline single-nucleotide polymorphisms (SNPs) and 12% showed a systematic sequencing artifact characterised by unidirectionality of reads on which variants were called (Figure 3.5b). The identification of false positives and subsequent determination of causes of mis-calls were critical for development of more post-processing filters and will be described in more detail in the following sections. A further 13% showed no immediately discernible pattern initially, but as the dataset improved in its specificity, more subtle patterns emerged and became amenable to post-processing.

Figure 3.5

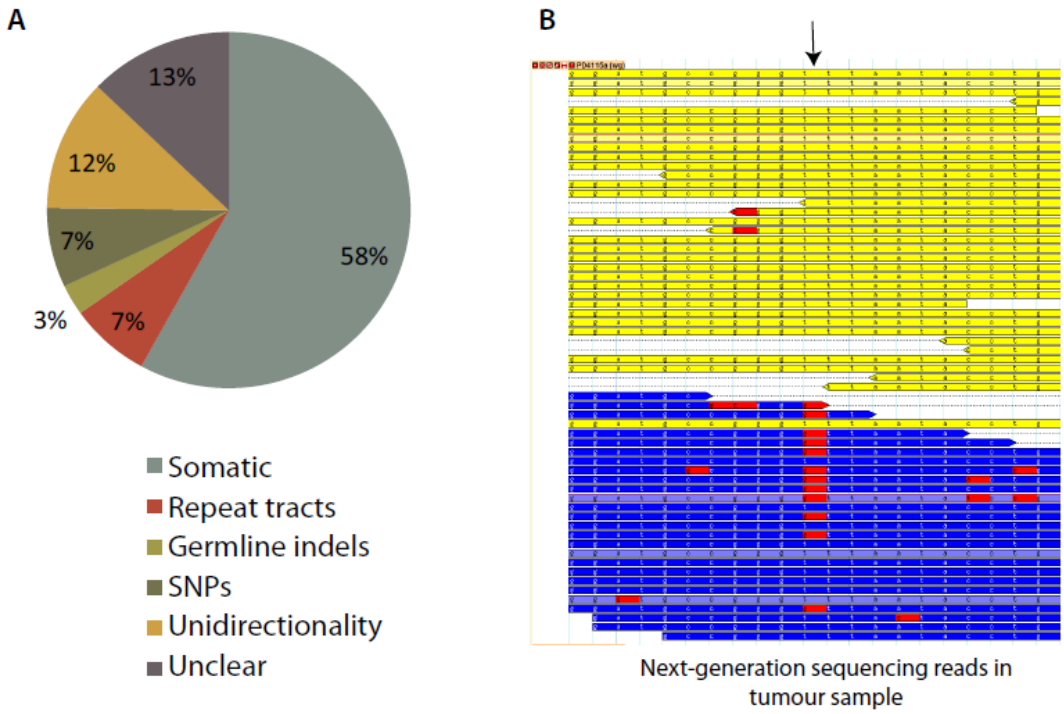


Figure 3.5: False positive calls revealed. (A) A breakdown of the false positive variants for the first iteration of validation of CaVEMan variants. (B) An example of a false positive call appearing in a unidirectional manner (only systematically on blue or forward reads) and present in tumour as well as normal reads (not shown). The variant was always the same as the preceding base in the reference genome in the direction of sequencing of the read. Here, a T>G variant following a string of G's.

3.6 THE PRINCIPLE OF DEVELOPING FURTHER POST-PROCESSING FILTERS

From the false positive variants identified in the above experiment, it was possible to classify variants that showed recurrent patterns. Some false positive variants, for example, occurred near homo-polymer or microsatellite repeat tracts, in regions of excessively low or high sequence coverage, at particular sequence motifs, at particular positions in sequencing reads (at the very ends) or near germline indels.

A post-processing filter was developed for each of these reasons and tested individually. For each filter, the reason for the filter was decided, a boolean relationship outlined and the code tested. For each test, it was necessary to ensure:

- That the expected false positives were removed
- That the known true somatic variants remained
- That there were no other unexpected changes due to errors in writing the code.

If a filter was deemed to be appropriate, it was implemented and the next filter was introduced. This procedure of "training" of filters was also performed on several other genomes in order to not over-fit filters to one sample (Figure 3.6).

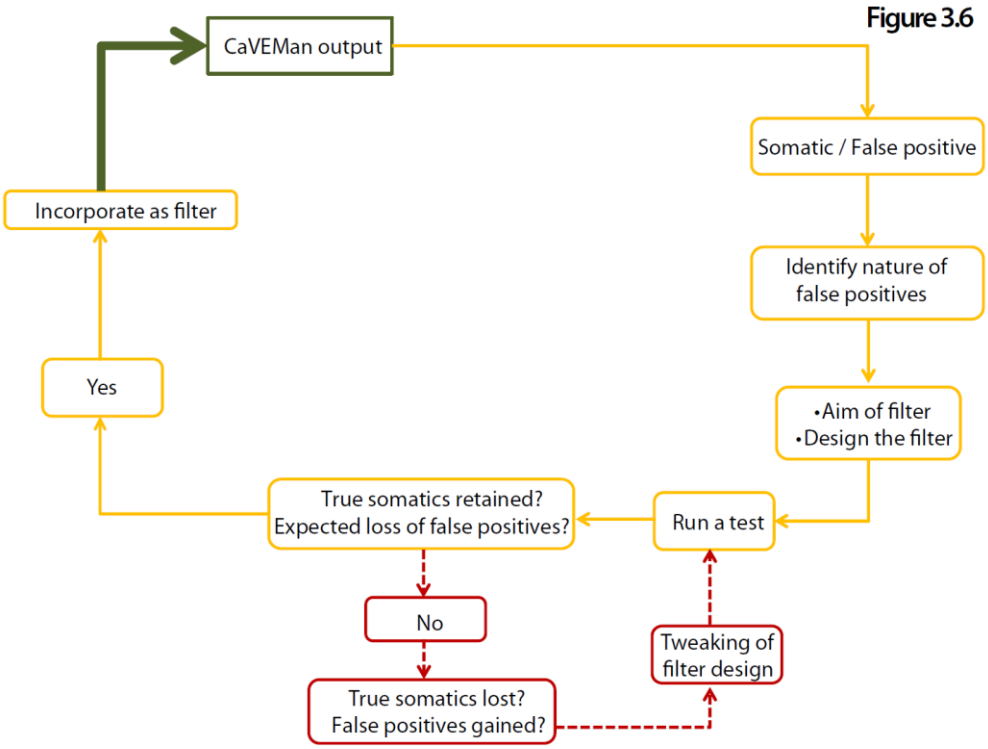


Figure 3.6: The principle of developing post-processing filters.

3.7 THE DEFINITIONS OF INDIVIDUAL POST-PROCESSING FILTERS FOR SUBSTITUTIONS

The final list of post-processing filters comprised twelve filters altogether. These could broadly be classified into three main categories.

- Filters dependent on intrinsic thresholds of sequencing/mutation-calling
- Filters for removal of systematic sequencing artifacts caused by the next-generation sequencing reaction
- Filters for genomic features that result in errors of mis-mapping

The table below provides a more detailed description of all of the filters (Table 3.1). Many of the filters take base qualities or mapping qualities into account which were described in the introduction (Figure 3.1).

Table 3.1: The reasons for and the definitions of each post-processing filter used in substitution-calling of the twenty-one breast cancer genomes

CATEGORY	NAME OF FILTER	DEFINITION	RATIONALE
Intrinsic threshold	***Mutation probability threshold***	The mutant allele probability score based on the core algorithm was equal to or above 0.95	Variants with a lower probabilistic score were simply less likely to be true somatic variants
Intrinsic threshold	***Read depth***	At least a third of bases in the tumour sample reporting the mutant allele had to exceed or equal a base quality of 25	Randomly erroneous variant bases due to the occasional fall in sequencing efficiency produced lower base qualities. True somatic variant bases had the same base qualities as other bases representing the reference allele. For a variant allele to be considered a true somatic variant, it had to be well-represented in the tumour sample, with good base qualities on several reads.
Intrinsic threshold	***Average mapping quality***	The mean mapping quality of reads reporting the mutant allele had to exceed 20	Some reads, particularly those where a germline SNP was present somewhere in the read mate, could map erroneously in highly homologous regions. If a read could map with equal or almost equivalent likelihood in more than one locus in the genome, then the mapping quality of the read reflected this lack of uniqueness. In essence, these were likely to be a cluster of mismapped reads.
Systematic sequencing artifacts	***Read Position***	The mutant allele failed this flag if it was present in less than 8 reads AND only represented on the last third of a read or only last third and first 8% of any read	Sequencing qualities and the reliability of base calls were known to fall towards the ends of reads. As a result, mismatches appeared to be more common towards the end of reads. This flag was designed to detect recurrent mismatches at the very

			ends of reads.
Systematic sequencing artifacts	***Matched normal***	The mutant allele failed this flag if it was present at base qualities exceeding 15 in more than 5% of reads in the matched normal sample	This flag was intended for removing remaining germline SNPs which had escaped initial exclusion.
Systematic sequencing artifacts	***Panel of other normals***	The mutant allele failed this flag if it was present in at least 5% of reads in at least 2 samples from the panel of randomly selected normal samples	Systematic sequencing artifacts should not discriminate between tumour and normal samples. However, they may only happen in a small fraction of reads. This flag was designed to identify those recurrent sequencing artifacts that arose intermittently in Illumina next-generation sequencing. In order to avoid the possibility of removing recurrent somatic events occurring in a subclonal population in a cancer, a randomly selected panel of normals was used to screen out recurrent sequencing artifacts.
Systematic sequencing artefacts	***Pentameric motif***	The mutant allele failed this flag if all reads carrying the variant but one were unidirectional (on forward or reverse strands only) AND the variants were only present in the last half of the read AND The reads carrying the mutant allele contained the motif GGC[A/T]G in the same sequencing direction as the variant AND the mean base quality for every base after the variant was calculated for each read and was less than 20	A systematic sequencing artefact was occurring following a specific sequencing motif characterised by GGC[A/T]G. Furthermore, the base qualities for all the bases following the putative variant usually fell well below expected. This pattern was exploited for the purposes of removing this sequencing artifact which was inexplicably worse for some tumours than others.
Systematic	***Phasing***	The mutant allele itself was required	Systematic sequencing artefacts that

sequencing artefacts		to have a mean variant base quality of more than or equal to 21 and was not unidirectionally represented.	resulted in next-generation sequencing polymerases going out of phase at some sequencing cycles. This was particularly predisposed at certain sequence motifs (-GGGG). The result was usually mutant alleles represented unidirectionally and of the same base as the immediately preceding allele in the direction of sequencing, in the reference genome. These variant alleles were usually of low base quality.
Genomic features	***Simple repeat***	The mutant variant call was failed if it fell within a simple repeat or within the immediate 5bp flanking the boundaries of a simple repeat as defined by UCSC	Mismapping of reads frequently occurred in and around simple repeats generating miscalls within or immediately flanking simple repeats.
Genomic features	***Centromeric microsatellite***	The mutant variant call was failed if it fell within the boundaries of a centromeric repeat as defined by UCSC.	Mismapping of reads frequently occurred in centromeric microsatellites generating miscalls.
Genomic features	***HiSeq coverage***	The mutant variant call was failed if it fell within a genomic window where the coverage in 2 or more genomes in a panel of normal genomes, exceeded 8 SD of the average of the coverage for those genomes *or* if it fell within parts of the genome which were consistently in the top 5% of coverage of HiSeq sequenced genomes as defined by UCSC (Pickrell et al., 2011).	Some repetitive sequences which are polymorphic in number of copies have been collapsed into a single copy in the human reference genome. When individual genomes are sequenced and mapped back to the collapsed reference genome, this results in excessively high coverage, increasing the likelihood for the accumulation of sequencing artifacts.
Genomic features	***Germline indels***	The mutant allele must not fall within the boundaries or be within ± 4bp of a germline indel as detected	Reads which ended in indels were more likely to map the very tip of the read within the indel and erroneously call it a

		by the indel-detecting algorithm.	mismatch than to map it correctly with a gap. If this occurred in multiple reads, this was effectively called as a substitution variant.

Figure 3.7

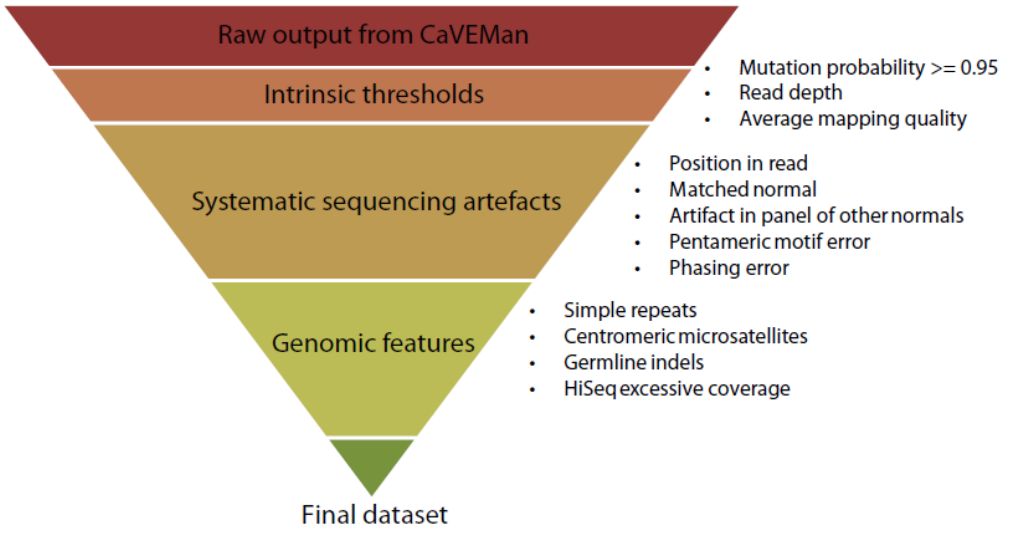


Figure 3.7: A schematic of the number of substitution variants following post-processing. The final curated dataset was always a small fraction of the total number of substitutions called.

Because each filter was applied independently for each variant, some variants could fail on multiple filters. In fact, the majority of raw substitution variants failed on multiple filters, attesting to the low likelihood of these variants being true somatic variants (Table 3.2). The final tally of substitution variants was always substantially fewer than the original raw output of the core CaVEMan substitution-calling algorithm for each genome (Figure 3.7, Table 3.2).

A revealing analysis of the effectiveness of each filter was seen in the number of variants that were removed exclusively by each filter (Figure 3.8). This demonstrated that the *Panel of other normals* was one of the most effective filters, removing the largest number of variants uniquely. This was followed by the *Matched normal* filter and the *Read Position* filter.

Sample	Raw calls	Failed more than one filter	Failed one filter	Final number of variants	Fraction of somatic variants from raw output
PD3851a	67917	58909	7226	1782	0.03
PD3890a	76235	58649	11462	6124	0.08
PD3904a	61665	49753	6304	5608	0.09
PD3905a	100027	82520	12920	4587	0.05
PD3945a	61668	44899	6461	10308	0.17
PD4005a	76186	61313	8769	6104	0.08
PD4006a	89525	69808	10523	9194	0.10
PD4085a	94504	84875	6956	2673	0.03
PD4086a	86594	77697	6698	2199	0.03
PD4088a	46420	41964	2751	1705	0.04
PD4103a	81750	70576	5814	5360	0.07
PD4107a	103870	86902	6677	10291	0.10
PD4109a	81007	65815	5304	9888	0.12
PD4115a	81136	63866	7316	9954	0.12
PD4116a	76191	59506	8659	8026	0.11
PD4192a	100127	85638	10570	3919	0.04
PD4194a	46466	40507	4475	1484	0.03
PD4198a	106246	89756	11938	4552	0.04
PD4199a	85204	68122	10150	6932	0.08
PD4248a	138435	120443	15456	2536	0.02

Table 3.2: Summary of substitution variants: From raw output to final datasets. Note that PD4120a, the deep-sequenced cancer, has not been included in this analysis of samples sequenced to 30-40-fold coverage.

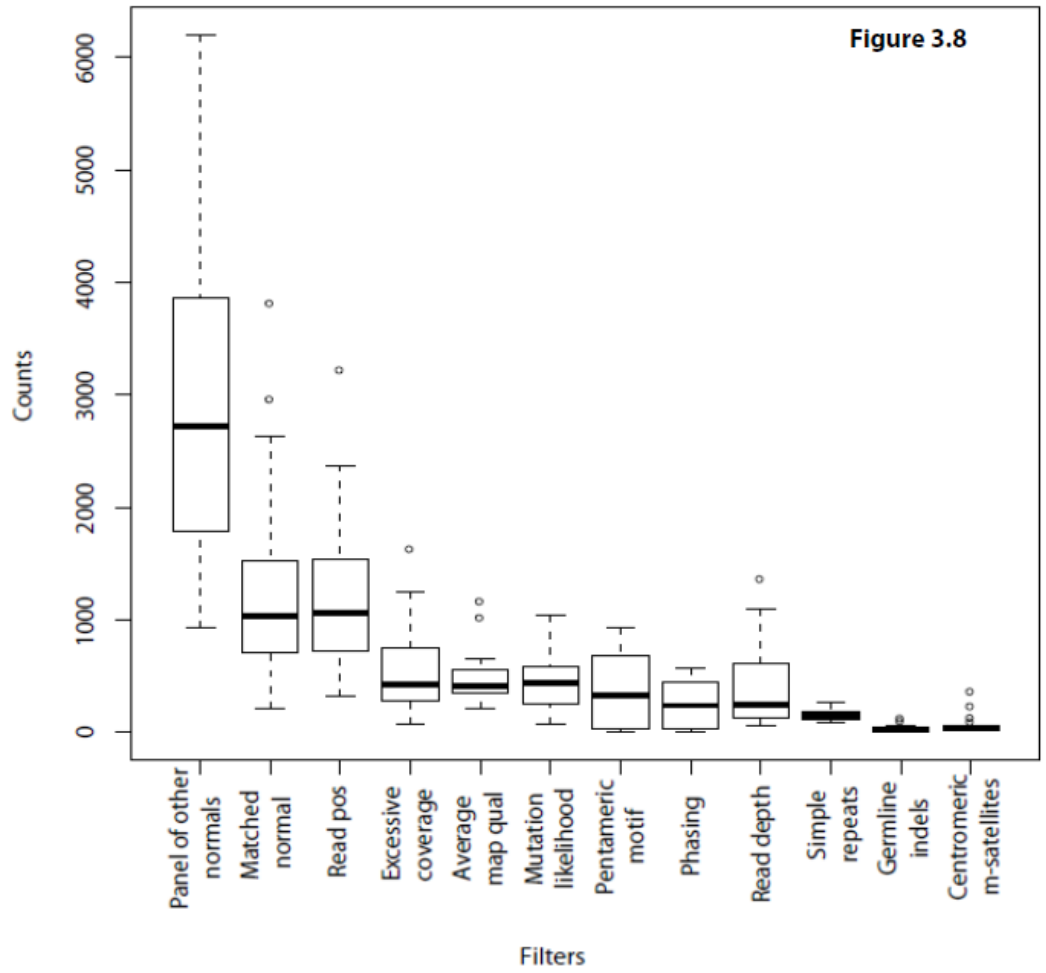


Figure 3.8: Variants removed exclusively by each filter. Total number of substitution variants removed by each filter exclusively on the vertical axis. Bottom and top of boxes in boxplots represent 25th and 75th percentiles with middle thick band at 50th percentile. Whiskers represent lowest and highest datapoints within 1.5 of the interquartile range. Small circles are outliers.

3.8 THE FRACTION OF THE GENOME WHERE MUTATIONS CAN NEVER BE CALLED

There were regions in the genome which were filtered out by virtue of being in zones of automatic exclusion. The fraction of the genome that was potentially filtered out did not simply represent the number of variants removed but was informative for the non-variant sites in the reference genome where mutations could never be called. The fraction of the genome affected by the relevant filters is documented in Table 3.3. The *germline indel* flag also contributed a proportion of genome in which no variants could be called. However, because germline indels vary between individuals, the coordinates involved in this filter was variable between cancer genomes. In general, ~1% of the genome was excluded by this filter.

Filter	Number of bases removed in the genome (bp)	Proportion of genome
Simple repeats	82,688,560	2.52
Centromeric repeats	1,660,347	0.06
HiSeq coverage	3,073,270	0.11

Table 3.3: Fraction of the genome effectively excluded by relevant filters

3.9 FINAL POSITIVE PREDICTIVE VALUE (PPV) OF CAVEMAN FOR THE DATASETS

To evaluate the improvement of the PPV of the substitution-calling process, ~400 substitution variants were re-sequenced using an orthogonal sequencing technology, in particular, Roche 454 pyrsequencing.

The PPV for each cancer genome at the point of having the first three filters and later when all twelve filters were in place is shown in Figure 3.9. The average positive predictive value for twenty cancer genomes was in the region of 92.1%.

Figure 3.9

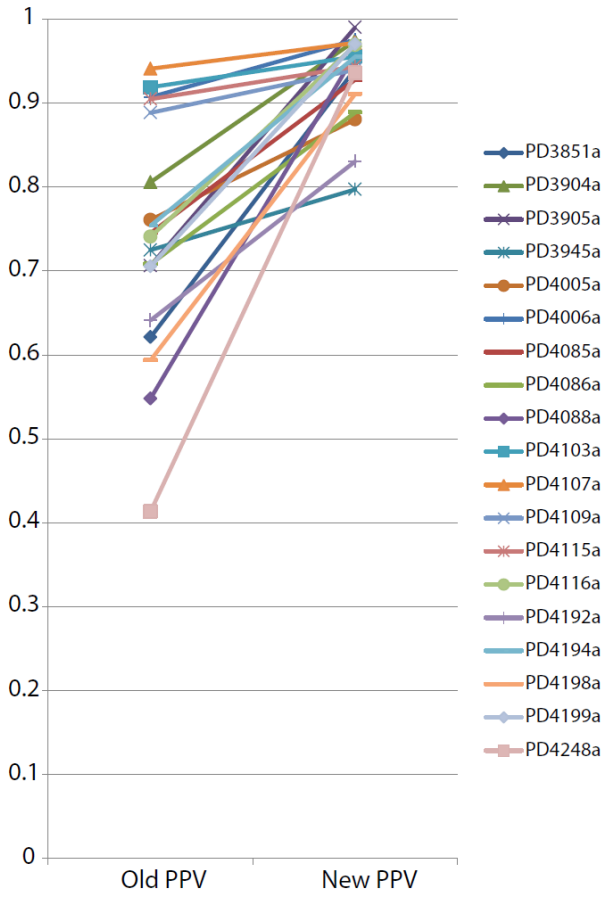


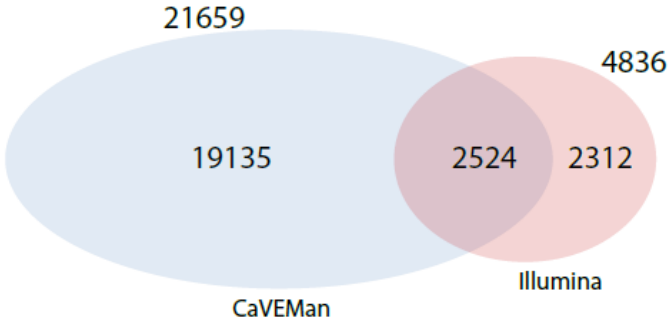
Figure 3.9: Improvement in positive predictive value for each cancer genome at the start of the experiment with three filters in place (*Mutation Probability Threshold*, *Read Depth* and *Matched Normal* filters) and later in the experiment with twelve filters in place (PPV is positive predictive value). Only 19 of 20 samples are shown here as PD3890a, was used as the sample for training many of the filters and so was excluded. The 21st sample, PD4120a, was sequenced to ultra-high depth and was therefore also excluded.

The fine-tuning of this large-scale process is expected to result in a trade-off between the gain in specificity and the loss in sensitivity. A comparison of these two parameters can be seen in PD3890a, which was sequenced at Illumina© and in which substitutions were called by an alternative caller. For the marked enhancement of the positive predictive value (56% to 90%), there was a loss of sensitivity (97% to 94.7%), at least for PD3890a.

Figure 3.10

Minimal filters

Sensitivity	Positive predictive value
0.970	0.58



All filters

Sensitivity	Positive predictive value
0.947	0.90

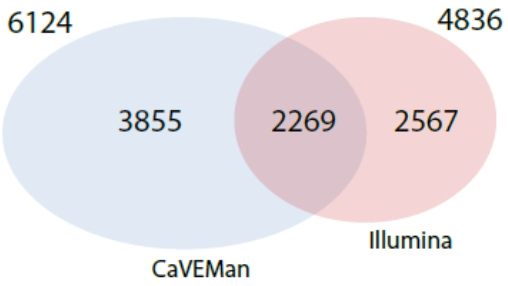


Figure 3.10: A comparison of the sensitivity and positive predictive value of PD3890a before and after development of all post-processing filters.

3.9.1 Positive predictive value does not correlate with sequencing coverage but correlates with degree of normal tissue contamination as predicted by the ASCAT (copy number algorithm)

The breast cancer genomes were assessed for whether the final PPV correlated with sequence coverage in tumour or normal. Neither of these appeared to show a correlation with the specificity of variant calling (Figure 3.11). Instead, the PPV of CaVEMan did appear to correlate with the degree of normal tissue contamination as predicted by ASCAT (the copy number algorithm used for this study). The general trend was that as aberrant cell fraction increased (and the normal contamination decreased), the PPV also increased.

Figure 3.11

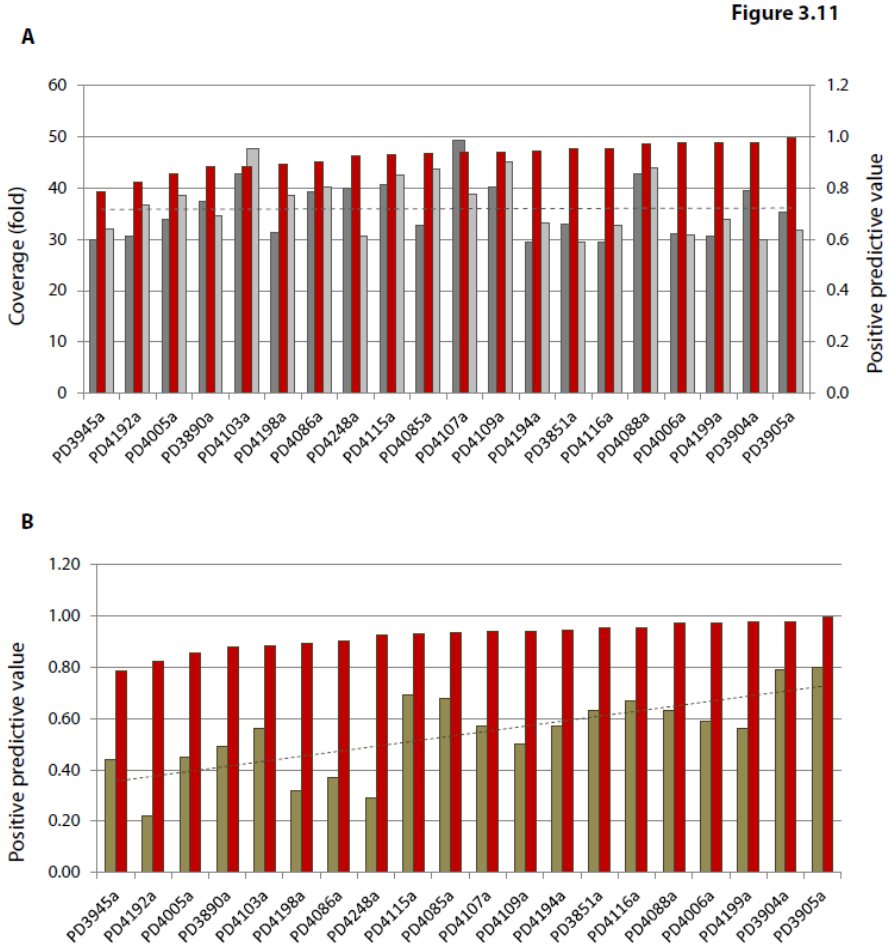


Figure 3.11: (A) No correlation was seen between positive predictive value (PPV) and tumour/normal sequencing coverage. Dotted line represents linear trend for tumour coverage (R^2=0.0002). (B) A correlation was appreciable from the comparison between PPV and aberrant cell fraction (R^2=0.5328). Dark grey = tumour coverage, light grey = normal coverage, red = PPV, tan = aberrant cell fraction. Only 20 of the 21 breast cancers were included in this analysis as PD4120a was sequenced to ultra-high coverage, and not all the filters designed were applied to this cancer.

3.10 SENSITIVITY OF DETECTION OF VARIANTS RELATIVE TO EXOMES

Four of the 21 breast cancer genomes were involved in a high-coverage (~100-fold) screen of coding sequences (exome screen, see section 2.3.1.2 for description) of 100 breast cancers (PD4103a, PD4107a, PD4109a and PD4120a). In order to gauge the sensitivity of mutation-detection in coding regions, the intersection between genome variants and exome variants was sought in three of the four cancers (PD4120a was an outlier having been whole genome sequenced to ~188-fold coverage and thus was not included in this analysis). For the three genomes, on average, 76.6% of variants detected through exome sequencing were detected in the whole genome sequences of the same cancers (range 68-82%).

The converse comparison was also performed. In each breast cancer, a proportion of variants in the coding sequence were called in the genome and missed in the exome screen. On average 22.3% of variants were missed by the exome screen ranging from 11.6-36.2%. Those variants that were missed in the exome screen were almost always due to a lack of coverage by the pull-down experiment in that region of the exome-sequenced cancer.

3.11 COMPARING CAVEMAN TO OTHER AVAILABLE MUTATION CALLERS

Although several mutation callers are available, none provides the level of (publicly available) post-processing that has been developed for these 21 breast cancer genomes. Comparing the dataset here with the raw output from other mutation callers does not therefore constitute a fair comparison. A version of Somatic Sniper was used to call variants in PD4107a but generated an enormous number of mutations (~450,000) as no post-processing was available at the time (http://gmt.genome.wustl.edu/somatic-sniper/current/). An alternative somatic single nucleotide variant caller which did have some post-processing options, MuTect (https://confluence.broadinstitute.org/display/CGATools/MuTect) generated an excess of 2.5 to 8 fold more variants for 3 breast cancers tested (Table 3.4). This was despite adopting the most stringent of post-processing filters available.

Table 3.4: Comparison between MuTect and CaVEMan, using three genomes as examples.

Sample	MuTect variants	CaVEMan variants	Overlap-ping variants	Variants missed by MuTect	Proportion of variants missed by MuTect	Average variant allele fraction of overlapping variants	Average variant allele fraction of variants missed by MuTect	Aberrant cell fraction	Tumour ploidy	Variants missed by CaVEMan which appear real	Variants missed by MuTect which appear real
PD4192a	20307	3919	3078	841	0.21	0.17	0.16	0.22	4.68	0	0.34
PD4198a	14618	4552	4142	410	0.09	0.21	0.18	0.32	3.05	0	0.48
PD4199a	17499	6932	6542	390	0.06	0.28	0.21	0.56	1.69	0	0.54

In order to evaluate the performance of MuTect and CaVEMan relative to each the other, a cohort of variants were sampled and visually assessed. In an ideal situation, these cohorts would have been validated. Of the variants missed by CaVEMan but were present in MuTect, none were real. Interestingly, between 17-28% of these were previously seen in CaVEMan but filtered out on the *Panel of Normals* filter alone. It is therefore likely that the vast majority of the excess of variant calls made by MuTect are false positive calls.

Assessing the variants present in CaVEMan and missed by MuTect, between 34-54% of variants looked real on visual inspection with many of the true variants being present at a lower variant allele fraction both in regions which were diploid as well as regions that were polyploid. This suggests that the sensitivity of variant detection by CaVEMan was higher for subclonal variants as well as variants which occurred on a single allele in of a multi-allele region in the clonal population.

3.12 INSERTIONS/DELETIONS AND REARRANGEMENTS

A similar methodical process of elimination of potential false positives was performed on the insertions/deletions. However, the indel-calling algorithm, Pindel, worked in a relatively simple way in its method of detecting variants. Pindel does not work on a probabilistic model and does not perform a comparison between tumour and normal. Therefore, a set of crude filters were designed in order to reduce the total number of variants.

Validation experiments on this filtered dataset revealed that the positive predictive value was still relatively low (40%-60%). As a result, only validated indels have been presented for downstream investigation, leaving a smaller but purer cohort of variants for analysis. The same principle applied to the detection of structural variants.

3.13 SUMMARY OF THE ANALYSIS PROCESS USED TO GENERATE THE FINAL DATASET

Following multiple iterations of validation and post-processing, the final analysis process was one which showed a high degree of interdependency (Figure 3.12). The final datasets used and described in the subsequent chapters therefore comprised:

- all the filtered substitutions with a subset of variants which were validated (Appendix 1)
- validated insertions/deletions (Appendix 2)
- validated rearrangements (Appendix 3)

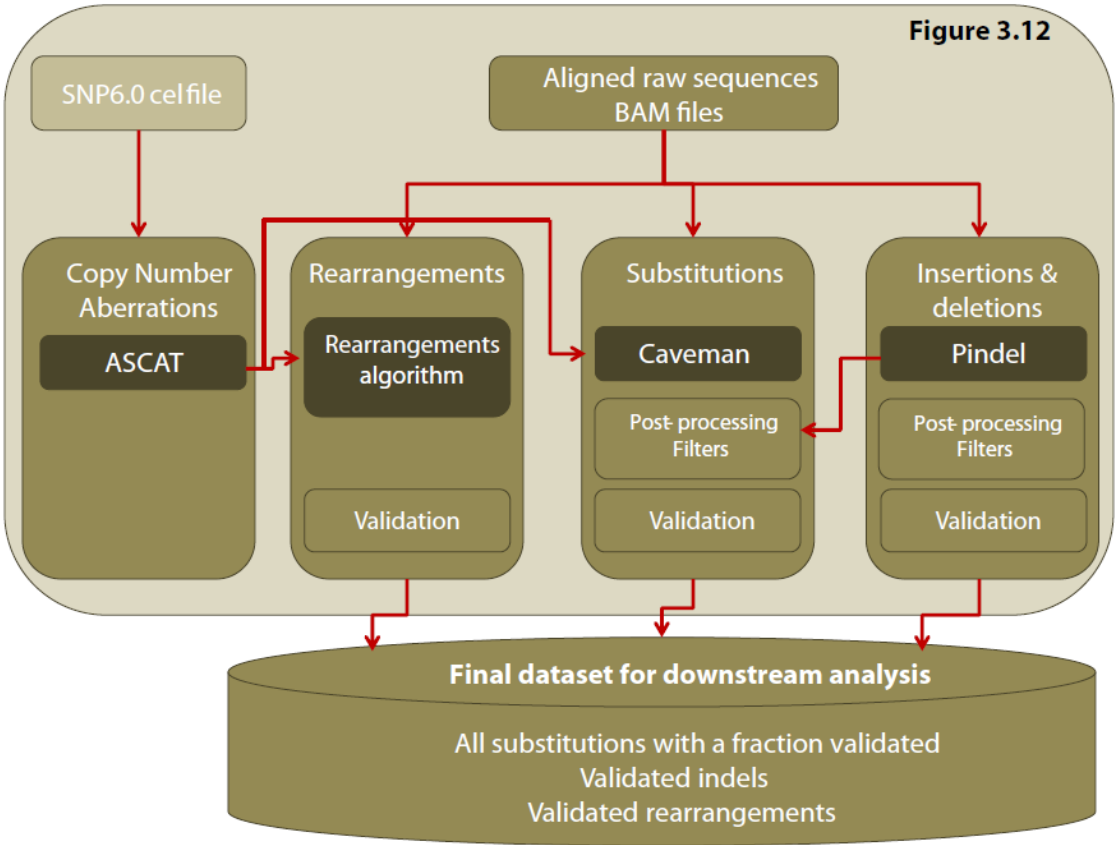


Figure 3.12: A schematic of the final analysis pipeline used to obtain a list of variants for downstream analysis in this thesis

3.14 DISCUSSION

This chapter was dedicated towards the development of post-processing filters required to obtain a final curated dataset that was essential for the detailed analysis performed later in this thesis, particularly for substitutions. Here, a systematic approach of identification of false positives, the reasons why they occur and the development of a collection of post-processing filters, were described. The positive predictive value (PPV) was used as a measure of the effectiveness of each of the post-processing filters.

In all, twelve post-processing filters were designed, reducing the dataset substantially and increasing the positive predictive value remarkably, with a minor cost to sensitivity. These filters could be classified into three main categories: those which involved intrinsic thresholds of the algorithm, those which were designed to remove systematic sequencing artifacts and those which were necessary to remove erroneous calls due to genomic features which caused mapping errors of short-read data. The final average positive predictive value for this cohort of breast cancers was ~92%.

3.14.1 A fair comparison between different mutation callers would involve comparing datasets *after* post-processing

Although other substitution-callers exist, none are known to consider complicating factors associated with the complexity of cancer: tumour heterogeneity, degree of contaminating normal cells and abnormalities of ploidy. Inclusion of these additional parameters in probability estimates in CaVEMan allowed base-by-base adjustments and in theory, increased the likelihood of calling true somatic substitution variants, particularly those which were present in a minor subclone in a cancer or those occurring on only one allele in a polyploid region of a cancer. This increased sensitivity is reflected in the variants missed by other callers but present by CaVEMan substitution-calling, which were all at a low variant allele fraction. Furthermore, depending on the nature of the biological specimen studied (e.g. cell lines), these parameters could be tuned in order to maximize the likelihood of detection of somatic variants.

Presently, despite application of the highest stringency filters (of which there are very few if any for some callers), the total number of mutations called by alternative callers are markedly more than by CaVEMan. Given that the curated dataset obtained here had twelve post-processing filters applied, a fairer comparison to other substitution-callers would require application of equivalent post-processing filters. Furthermore, a more comprehensive comparison between the performance of

CaVEMan relative to other substitution-callers would possibly require a degree of validation of those variants missed by CaVEMan. This has not been performed as part of this thesis due to time constraints.

3.14.2 Balance between sensitivity and specificity: two sorts of datasets

The set of mutations obtained from any large-scale genomic experiment will always comprise a set of true somatic variants and a larger set of false positive calls. The degree to which a dataset is filtered will depend entirely on the question sought. In exome-sequencing experiments of cancers, targeted enrichment of the coding sequence and higher sequencing coverage in these protein-coding exons is primarily aimed at identification of driver events and demands as high a measure of sensitivity as possible. Although this may result in a high number of false positive calls, the total burden of mutations is still relatively low and amenable to validation in order to isolate true somatic events. The same approach would overwhelm a genome-sequencing experiment. Because the focus in this genome-sequencing project was on seeking genome-wide signatures and related less to detection of cancer genes, it was imperative for specificity to be set as high as possible in order to reduce the likelihood of detection of false positive signatures.

In the near future and for large-scale genome sequencing projects in cancer, it may be necessary for some combination of both approaches to be used. Perhaps, the core algorithm could be run with a set of “high sensitivity” filters concentrating on the coding sequence in order to detect all important coding mutations, as well as “high specificity” filters for the whole genome, in order to obtain a complete catalogue of variants from a single sequencing experiment.

3.14.3 Scope for improvement of individual filters

There is likely to be scope for improvement of some filters. First, there was considerable overlap in the variants removed by some filters, particularly between the “Read Position” and “Germline indel” filters. However, each also removed a definite and mutually exclusive subset. Hence, it was difficult to justify removing either as a filter. Second, more time could have been spent on improving the sensitivity lost with each filter. This would have required several more iterations of each filter and for this thesis, had to be balanced with the timeline of getting an adequately curated dataset. Nevertheless, enhancements to the current set of filters are expected in the near future.

Furthermore, post-processing filters developed here had been trained to accommodate cancer genomes sequenced to 30X-50X coverage of 100bp reads, with equivalent depth in the matched

normal. The efficacy of these filters is likely to be affected by genomes with significantly different levels of coverage between tumour and normal. The use of proportions was favoured over the use of absolute values particularly when defining read depth in the post-processing filters, but this was not always possible (e.g. *Read Position* filter). Therefore, filters which are sensitive to variation in coverage may become less effective if the coverage in the tumour is not at 30-50X. Distinctions based on proportional distance along each read were also made in some filters and this could be adversely affected by shorter read lengths of 50 or 75bp reads. Therefore, subtle differences in experimental approach may affect the application of these filters and could possibly be factored into the design of each filter, in the future.

3.14.4 The moving target: future optimization will be necessary

Any improvements to the core algorithm will necessitate further optimization of the substitution-calling process. In addition, changes in sequencing technology and chemistry resulting in vastly increased yields per lane of sequencing is likely to give rise to other novel sequencing artifacts and will require thoughtful application of new filters, or adaptations to old ones, in order to manage new problems.

3.14.5 The performance of callers on indels and rearrangements

This chapter has focused on developing post-processing filters for calling substitutions. The performance of the core algorithms and current filters for indels and rearrangements was much less desirable, with poorer specificity for both of these mutation classes. As a result, confidence can only be placed on validated variants and only these validated indels and rearrangements were used for downstream analysis.

Other approaches could be considered for the near future. Local reassembly is a feature used by the Broad Institute (GATK) to improve the mapping of reads overlying indels. This is an approach that has not been explored in this thesis. Because suitably stringent post-processing filters are not available for GATK, one possibility would be to perform primary indel-calling using Pindel and then perform local reassembly across these indels to improve mapping characteristics of the informative reads before post-processing. Another approach that is described is to use multiple callers on the same dataset and to simply use the variants which are overlapping.