Reproducibility of SNP-calling in multiple sequencing runs from single tumors

Dakota Z. Derryberry*[1], Matthew C. Cowperthwaite[3], and Claus O. Wilke[1,2]

[1]The University of Texas at Austin, Cell & Molecular Biology
[2]The University of Texas at Austin, Integrative Biology
[3]St. David's NeuroTexas Institute Research Foundation

May 14, 2015

Abstract

We examined 55 technical sequencing replicates of Glioblastoma multiforme (GBM) tumors from The Cancer Genome Atlas (TCGA) to ascertain the degree of repeatability in SNP-calling when using the same data analysis pipeline on two instances of sequencing the same tumor, and to determine what additional filtering might increase or decrease the total similarity. We analyzed 55 pairs of technical replicates using the same computational pipeline and measured the extent of the overlap between two replicates, that is how many specific point mutations were found in both replicates. We further tested whether additional filtering increased or decreased the size of the overlap. We found that about half of the putative mutations found in one sequencing run of a given sample are also found in the second, and that this percentage remains steady throughout orders of magnitude of variation in the total number of mutations called (23 to 10,966). We further found that using a SNP and subsequent filtering removes the overlap completely. We concluded that there is variation in the frequency of mutations in GBMs, and that while some parts of the SNP-caller preferentially removed putative mutations found in only one replicate, others removed primarily putative mutations found in both.

Introduction

The past six years have seen an explosion of data in cancer genomics, an effort led by TCGA, an archive of publicly-available data that includes sequencing of paired tumor-normal samples from a single patient for thousands of tumors (Network, 2008; Brennan

*dakotaz@utexas.edu

et al., 2013). TCGA's database includes hundreds of samples of each of several tumor types, including 528 and counting samples of Glioblastoma multiforme (GBM), the most common and deadly primary brain tumor. GBM has a median survival of 14 months and a 5-year survival of 5%, and prognosis for patients with this disease remains poor despite significant research investment, due to the difficulty of surgical resection and the limited number of effective chemotherapeutics (Wilson et al., 2014). To work towards the goal of improving patient outcomes using precision medicine, TCGA and other groups (Parsons et al., 2008) used large scale genomic data to discover genes and pathways mutated in GBM (Cerami et al., 2010), discover different GBM subtypes (Verhaak et al., 2010), and develop a variety of computational models to find GBM driver mutations (Gevaert and Plevritis, 2013).

But these and other efforts have produced very few results for patients. It may be that we simply need more new ways to analyze the data. Another factor, however, may be that the data are noisy. It is well known that sequencing and SNP-calling pipelines are not error-free. Given the heterogeneity in cancer genomes (Kumar et al., 2014; Friedmann-Morvinski, 2014), and the presence of functional low-frequency variants in GBM (Nishikawa et al., 2004), the signal to noise ratio in this data set may be even worse. Despite widespread use of the TCGA dataset, and significant monetary investment in collecting and analyzing the data, we found very little research concerning the best way to maximize the quality of the sequence data and SNP-calls. Because different SNP-calling pipelines have been known to return different results on the same data (Yu and Sun, 2013), we turned this question around and asked, will the same SNP-calling pipeline return the same results on two sequencing runs from the same tissue? Does SNP-calling software and filtering increase or decrease the degree similarity?

We answered these questions using data from TCGA, which includes 55 GBM tumors that were sequenced twice, once each with (i) the standard WGS protocol, and (ii) an additional amplification step before library prep, which we refer to as the WGA protocol. For each of these 55 technical replicate pairs, we compared the somatic variants found in the WGS and WGA replicates, before and after analyzing the variants using SomaticSniper (Larson et al., 2012). We found significant overlap (around 50%) between technical replicates, but also significant differences. As expected, the additional amplification step in the WGA protocol versus the WGS protocol added some putative mutations to the sample, so that on average these replicates had (i) more putative mutations, and (ii) a smaller percentage overlap between replicates. Still, the number of mutations in the WGS replicates varied by orders of magnitude, from 110 to 8,192. Contrary to expectations, the percentage overlap between technical replicates did not decrease with increasing numbers of mutations, suggesting real variation in mutation frequency between samples. This may be in part due to know mutational hotspots in some tumors (Wang et al., 2008). Our attempt to use SomaticSniper and some custom filters to increase the similarity between technical replicates was unsuccessful: filtering removed almost the entirety of the overlap between the replicates.

Results

How similar are two iterations of sequencing the same tumor?

DNA sequencing is not error-free. Error is introduced by mis-called bases in sequencing runs, and by mis-aligned bases during sequence analysis (Wall et al., 2014). Loss of heterozygosity also occurs in sequencing if only one allele of a polymorphic site is amplified. Cancer DNA is highly heterogenous, which makes for an additional source of error: a mutation present in only some tumor cells may or may not be present in a given sample at high enough frequency to be seen, so that a partially penetrant mutation may appear fully penetrant or not at all. In general, we would like to know how often these error occur. One way of investigating would be to use PCR to verify each individual mutation, but this method is expensive and time consuming. A cheaper alternative would be to sequence the tumor multiple times, and look at the similarity between replicates. Theoretically, any fully penetrant mutation will appear in all replicates, while errors due to (i) sequencing errors, (ii) amplification errors, or (iii) alignment errors will not. (Incompletely penetrant mutations would be present in some but not all samples, so this method does not address the difficulty of calling low-frequency variants in tumors.) This is the method used in most biological sequencing experiments–but not generally in cancer genomics, presumably due to the expense of sequencing multiple replicates for each of the hundreds of samples necessary for cancer genomics research. Still, even if researchers cannot verify every mutation or sequence multiple replicates for each tumor, it would be useful to know about what percentage of called mutations would or would not likely appear in additional sequencing replicates.

TCGA's GBM data set includes 2 technical replicates each of 55 tumors. In this case, the technical replicates are not identical. One protocol included additional amplification step (Network, 2008), and we refer to this replicate as the WGA replicate, and the other as the WGS replicate. Despite this difference, we still assumed that any fully penetrant mutation would appear in both replicates, while incompletely penetrant mutations and sequencing error might appear in only one. Thus overall, those putative mutations appearing in both replicates are more likely to be real somatic variants than those found in only one replicate. We further hypothesized that the WGA samples would have a greater number of amplification errors, and thus more putative SNPs per sample, than the corresponding WGS samples.

For each patient (n=55), we called mutations in both technical replicates and the patient's blood sample using the same computational pipeline (see Figure 1 and Methods): we downloaded TCGA bamfiles with CGHub (of California Santa Cruz, 2014), re-generated fastq files with picard (alecw et al., 2014), re-aligned the fastq files to hg19 with bwa (Li and Durbin, 2009), performed indel re-alignment and base recalibration with GATK (McKenna et al., 2010), and finally called somatic mutations with SomaticSniper (Larson et al., 2012). We then compared the VCFs produced by SomaticSniper for each of the two technical replicates, and calculated the number of somatic mutations called in

each replicate and the number of individual somatic mutations called in both replicates (hereafter, the length of the overlap, see Figure 2). We further calculated the percentage of of mutations in each replicate that occurred in the overlap between the two.

There were on average 844 putative mutations in the WGS replicates and 1,531 putative mutations in the WGA replicates (for more detail, see summary statistics, Table 1). Across all samples, the number of mutations in a given WGS replicate was correlated with the number of mutations in it's corresponding WGA replicate, cor=0.51, t=4.3, df=53, p-value=8.35e-05 (Figure 2). As expected, for each sample the WGA replicate (with the additional amplification step) had slightly more mutations overall, with a slightly smaller percentage appearing in the overlap (Figures 3 and 4). We further found that the percent overlap between the two samples, calculated by $WGA \cap WGS/WGS$ for WGS replicates and $WGA \cap WGS/WGA$ for WGA replicates, was fairly consistent, on average 31% in WGA replicates and 44% in WGS replicates (see summary statistics, Table 1). As expected, the distribution was narrower and taller in the WGS replicates, because on the whole the WGA samples had more amplification errors than the WGS samples (Figure 3).

Although the percent overlap in WGS and WGA samples remained fairly constant, and the number of possible mutations in WGS and WGA lists was correlated, the exact number of putative mutations varied across samples by orders of magnitude (less than 23 to over 10,966). Different cancers mutate at different rates: some pediatric cancers have very few mutations (Knudson, 1971; Chen et al., 2015), while some adult tumors show a mutator phenotype leading to vastly more mutations, usually resulting from errors in DNA repair pathways (Loeb, 2011). GBM specifically is thought to have a relatively low mutation rate (Parsons et al., 2008; Brennan et al., 2013), and while some of our samples had low mutation frequencies in line with this theory (29 out of 110 samples had a mutation frequency within a factor of 2 of the reported 3 mutations per Mbp genome), several samples also had mutation frequencies an order of magnitude greater (24 out of 110 samples had a mutation frequency greater than 30 mutations per Mbp genome). One possible explanation is a degraded DNA sample, or bad data. If this were the case, we would expect the percentage of the overlap between replicates (a measure of data quality) to decrease with the overall number of putative mutations. We found no correlation (cor=0.25, p=.06, not significant) between the number of putative mutations in the WGS replicate and the percentage of those mutations that were in the overlap between replicates (Figure 4), though it is possible that with more more samples this could become a weak correlation. Still, that the percentage overlap does not correlate strongly with the number of putative mutations in either replicate suggests that some samples may simply have a higher mutation frequency than others, or indeed than is generally supposed in GBM.

Does SNP-calling software increase or decrease the degree similarity between replicates?

The aim of sequencing tumors is generally to find somatic mutations, those mutations that arise in a tumor and are not present in the rest of the organism. To do this, researchers sequence a patient's tumor and blood and align the two to each other; differences between the two sequences are putative somatic mutations. Because sequencing is not error-free, and because a patient's blood tissue might also have somatic mutations different from the tumor, not all of the putative somatic mutations are mutations, and some method of distinguishing true somatic mutations from sequencing and other errors is needed. If possible, this is done by individually validating each putative mutation with PCR; however, due to restraints on time and money, as well as the amount of tumor tissue available and its purity, this is often not possible. In this case, we may use computational algorithms that attempt to distinguish somatic mutations from germ line mutations and sequencing errors (Larson et al., 2012; Alioto et al., 2014).

Software platforms to distinguish somatic mutations from germ line mutations and sequencing errors are plentiful, and each one typically employs multiple methods to identify true positive mutations. The two platforms used in this research, SomaticSniper (Larson et al., 2012) and GATK (McKenna et al., 2010), calculate one or more quality scores based on features of the dataset and the individual reads, and mutations with higher quality scores are considered more likely than those with low quality scores to be true somatic mutations. Additionally, after removing reads with low quality scores, there are additional filtering steps used to distinguish somatic mutations from errors of all sorts. The entries in Table 2 describe each of the ways we filtered the raw SNPs called in each sample. The first three entries are the quality scores generated by GATK and SomaticSniper (we simply remove from the dataset anything that fails to meet these thresholds). The last five are additional filters that we coded ourselves in Python. Each of these five filters represents a quality of the data or putative mutation that is generally thought to indicate that it is probably a false positive.

Our initial question was, does filtering the data increase or decrease the percentage of the sample that is overlapping between the two technical replicates? Put another way, does filtering out putative somatic mutations with these features increase or decrease the proportion of true somatic mutations in the remaining dataset? We found that after removing putative mutations tagged by any of the eight filters, the number of putative mutations per replicate decreased from 23–10,966 to 0–14. The size of the overlap between technical replicates decreased to 0–2 per sample, with 0 as the mode, so the overall percentage also decreases. We concluded that running all the filters on the data, in the absence of any other verification method, was counterproductive, because it removed all of the signal (as well as the noise).

We next looked at the individual effects of each of the five filters that we coded and one of the SomaticSniper filters (Variant Allele Quality, or VAQ). We first looked at the total number of putative SNPs removed by each filter (Figure 5). We found that different

filters removed different numbers of mutations, and that the lion's share of mutations were removed by the VAQ and LOH filters, which removed on average 309 and 539 putative SNPs, respectively (see Table 1 for summary statistics). Three other filters, those removing overlap with dbSNP and mutations within a 10 bp window of indels or other SNPs, removed 16, 28, and 46 putative mutations per sample (Figure 5, summary statistics in Table 1). The final filter, which removed putative SNPs with less than 10% coverage of the alternate allele, removed 1 putative SNP on average.

We next asked how many of the mutations removed by each filter were mutations present in the overlap between technical replicates, and how many were in just one sample? Put differently, what percentage of putative SNPs removed by a given filter were in the category more likely to be true positives (overlap), versus the category more likely to be false positives (only one replicate)? To answer this, we graphed the percent of the overlap (per sample) that was removed by each of the six filters (Figure 6). We found that the three filters removing overlap with dbSNP and putative mutations within 10bp of an indel or another SNP removed, on average, only 3%, 4%, or 3% of the overlap, or almost none, respectively (summary statistics in Table 1). In contrast, the VAQ filter (specific to SomaticSniper) and the LOH filter each removed 53% and 51% of the overlap, or about half each, respectively (Figure 6, summary statistics in Table 1). Thus, our evidence suggests that the filters removing overlap with dbSNP and putative mutations near other putative mutations are preferentially removing false positives, but the filters removing low VAQ and LOH are less discriminatory.

Finally, we asked whether the VAQ and LOH filters, responsible between them for removing most of the overlap, were removing the same or different putative SNPs in each sample. We plotted the percent of the overlap filtered out by the LOH filter against the number of putative mutations in the overlap (Figure 7), and the percent of the overlap filtered out by the VAQ filter against the number of putative mutations in the overlap (Figure 8). We found opposite trends: the percentage filtered out by VAQ decreased, and the percentage filtered out by LOH increased, with increasing length of overlap. The two graphs are almost, but not quite, perfect inverses. But given that (i) each of the LOH and VAQ filters removes about half of the overlap, and (ii) all or almost all of the overlap is removed every time, this is expected. We concluded that the LOH and VAQ filters were not removing the same putative mutations.

Discussion

GBM is an evolutionary disease that develops when mutations arise in glial cell lines and these mutated cells and their lineages co-opt the surrounding tissue and systems to the detriment of the organism as a whole. Treatment for GBM is difficult and has poor outcomes (Wilson et al., 2014), but may be improved by a more complete understanding of the somatic mutations present in GBM. Large-scale sequencing projects, like TCGA, make cancer sequencing data available to many researchers. These data are enormously

valuable to the research community, but their accuracy and reproducibility are unknown, and the knowing could only increase the utility of the data. We made a first pass at evaluating the repeatability of the data by comparing 55 technical replicates in the TCGA GBM dataset.

We looked at the similarity between technical sequencing replicates for 55 GBM samples in TCGA. We found that on average, about half of the putative mutations in the raw data for the WGS replicate (no amplification before library preparation) and about a third of those in the WGA replicate (with amplification before library preparation) were present in both replicates. The number of mutations present in both replicates was anywhere between 20 and 5,000 putative mutations. We found that the high number of putative somatic mutations in some, but not all, of the patient samples was repeatable across technical replicates. Further, samples with a higher frequency of putative mutations had equally similar technical replicates to those samples with a lower frequency of putative mutations. This suggests the possibility that a higher mutation frequency could be a feature of a subset of GBM tumors and not a data artifact.

Filtering the raw computational data using both quality scores from GATK and SomaticSniper as well as five additional custom filters eliminated more than half of the total number of putative mutations in all 110 samples, including most or all of those present in both replicates. On some level, this is unsurprising: these filters were designed to be used on wild type genomes, where it is generally assumed that more differences are due to error than to real changes (e.g. LOH in this case is more likely to be an error than a change). They are designed to reduce or eliminate differences between the reference and the sample. When we do cancer genomics, our goal is the opposite: to highlight changes. Therefor, it is possible that we need entirely different filtering protocols. There are also more theoretical reasons to consider altering the filtering protocols for cancer genomes. For example, multiple sources suggest that LOH mutations may be essential to cancer (Fujimoto et al., 1989). This, along with the data presented here, makes a strong argument for retaining these mutations in functional analyses rather than excluding them.

Of the six filters whose individual effects we examined, only two, those that removed Loss of Heterozygosity (LOH) mutations and putative mutations with low VAQ (calculated by SomaticSniper), removed primarily mutations that found in both technical replicates, and combined, they removed most or all of those mutations present in both replicates; the two filters removed almost disjoint sets of putative mutations. Of the remaining four filters, only three removed any appreciable number of putative mutations from the sample, and each of these preferentially removed mutations present in only one technical replicate. Two of these three filters, those removing putative mutations within a 10 bp window of putative indels or other putative SNPs, recognized a feature (clustered mutations) that suggests a local problem with the reads or alignment, and removed this data from consideration. The third removed overlap with dbGaP. Our analysis suggests that these three filters do clean up the data in a meaningful way. By contrast, it may be more useful to discard the two filters that removed principally data from the overlap of

the two replicates.

Several factors limit the conclusions we may draw from this analysis. First, in this analysis we used repeatability between technical overlaps (being in the WGS and WGA samples) as a measure of confidence in a putative SNP. This metric is potentially problematic for three reasons: (i) cancer is highly heterogeneous, and so a legitimate somatic SNP might show up in one replicate and not another; (ii) if the DNA sample is degraded to some extent, due to surgery conditions or some other factor out of the hands of the sequencing center, the same errors may appear in both replicates; and (iii) a putative mutation that is in both samples may be a somatic mutation that arose before the tumor (Tomasettia et al., 2013). Although repeatability does not and could not perfectly measure confidence that a putative mutation is a somatic mutation, it does make it more likely. Having an independent measure of confidence in SNP calls (repeatability), even an imperfect one, can help us gauge the accuracy of other measures (filtering).

Second, we looked at results from one SNP caller. There are many more equally widely used SNP callers available, which might do better or worse than the one we have chosen here. Also, we did not even look at every quality metric available in the SNP-caller we did choose (we did not look at the effects individually of the SomaticScore or the GATK quality score). In the future, we would like to evaluate other SNP-callers and other quality metrics.

We have shown that there is significant overlap between technical replicates of whole exome sequencing in the TCGA GBM dataset, comprising about 50% of putative SNPs in WGS samples and about 30% in WGA samples. This effect remains even at a high number of putative SNPs, suggesting that some GBMs may have significantly more somatic mutation than others. While PCR remains the gold standard for distinguishing true positives from false positives in sequencing, we looked at the effects of six data filters for the same purpose. We found that some filters remove principally those mutations found in one sample or the other, while other filers remove primarily those in the overlap. We suggest using only the first set in the future, when mutation validation by PCR is not an option.

Methods

Data and back-end processing

All sequence data came from The Cancer Genome Atlas (TCGA) Research Network's Glioblastoma multiforme (GBM) data set (Network, 2008). We downloaded three bamfiles for each of 55 patients using CGHub (of California Santa Cruz, 2014). For each patient, data consisted of one bamfile taken from blood DNA, and two bamfiles from tumor DNA, one for each technical replicate. In each case, the only difference in data collection for the two sets of tumor DNA was whether or not an amplification step was

performed prior to building a library (Network, 2008).

We created a pipeline for backend processing of all TCGA bamfiles, summarized in Figure 1, with commands given in Table 3. The custom python code to connect the pipeline is available in a public git repository (`https://github.com/clauswilke/GBM_genomics`). Our pipeline first regenerated fastq files (original reads) from the TCGA bamfiles, which were aligned to hg18, using picard (alecw et al., 2014). Next, we used BWA (Li and Durbin, 2009) to align the fastq files to hg19, and samtools (Li et al., 2009) to sort, index, and de-duplicate the new bamfile. We used GATK (McKenna et al., 2010) to do indel realignment and base recalibration, according to the standard best practices for genomics data (Team, 2015). Finally, we predicted somatic variants with SomaticSniper (Larson et al., 2012), with the output given in VCF format.

Filtering and data analysis

After generating a VCF file will all of the putative somatic variants for each replicate of each sample, we used custom python code (available in public git repository) to list the putative mutations in each VCF, and to calculate the overlap between technical replicates. We then filtered the lists of putative SNPs according to the eight filters described in Table 1, using a combination of command line options and custom python code (available in a public git repository). The eight filters were enacted as follows:

1. (GATK) The GATK quality score is automatically generated by GATK; GATK recommends discarding all putative mutations with a quality score below 40, which we did using the command line option `-q 40` when we called SomaticSniper (for the exact command, see Table 2). We did not at any point consider putative mutations removed by this filter, and did not consider its individual action.

2. (SS) The SomaticScore is a similar metric calculated by SomaticSniper. As recommended, we removed from consideration all putative mutations with a SomaticScore below 40 by using the command line option `-Q 40` when we called SomaticSniper (for the exact command, see Table 2). We did not at any point consider putative mutations removed by this filter, and did not consider its individual action.

3. (VAQ) The Variant Allele Quality (VAQ) score calculated by SomaticSniper is a third measure of this type. SomaticSniper recommends discarding putative mutations with a VAQ below 40, which we accomplished using a custom python script available in the public git repository. This recommendation is discussed in the Results and Discussion.

4. (LOH) It is general (but not universal) practice to disregard Loss of Heterozygosity (LOH) in large scale genomics, because LOH to easily results from sequencing errors. We excluded LOH variants using a custom python script available in the public git repository. This recommendation is discussed in the Results and Discussion.

5. (10bp-SNP) It is universal or near universal practice to exclude variants within 10 bp of another putative somatic variant, because clusters of putative mutations often indicate an error in reads or sequence alignment. We excluded 10bp-SNP variants using a custom python script available in the public git repository. This recommendation is discussed in the Results and Discussion.

6. (10bp-INDEL) It is universal or near universal practice to exclude variants within 10 bp of a putative indel, because clusters of putative mutations often indicate an error in reads or sequence alignment. We excluded 10bp-INDEL variants using a custom python script available in the public git repository. This recommendation is discussed in the Results and Discussion.

7. (dbSNP) It is universal or near universal practice to exclude variants that overlap with dbSNP, because the overlap often indicates an amplification error and not a true somatic variant. We excluded dbSNP variants using a custom python script available in the public git repository. This recommendation is discussed in the Results and Discussion.

8. ($< 10\%$) It is universal or near universal practice to exclude variants when the alternate allele coverage is less than 10%, because the low coverage of the alternate allele often indicates sequencing error. We excluded $< 10\%$ variants using a custom python script available in the public git repository. This recommendation is discussed in the Results and Discussion.

We used the same or substantially similar python scripts to calculate and compare the overlap between technical replicates before and after filtering. These scripts are also available in the public git repository. We plotted all data and did all statistics with standard plotting and statistics functions in R (Team, 2014). This code is also available in the public git repository.

References

alecw, brilliantred, mmccowan, nilshomer, tfenne. 2014. *picard*. URL `http://picard.sourceforge.net`.

Alioto TS, Derdak S, Beck TA, Boutros PC, Bower L, Buchhalter I, Eldridge MD, Harding NJ, Heisler LE, Hovig E, Jones DTW, Lynch AG, Nakken S, Ribeca P, Sertier AS, Simpson JT, Spellman P, Tarpey P, Tonon L, Vodák D, Yamaguchi TN, Agullo SB, Dabad M, Denroche RE, Ginsbach P, Heath SC, Raineri E, Anderson CL, Brors B, Drews R, Eils R, Fujimoto A, Giner FC, He M, Hennings-Yeomans P, Hutter B, Jäger N, Kabbe R, Kandoth C, Lee S, Létourneau L, Ma S, Nakagawa H, Paramasivam N, Patch AM, Peto M, Schlesner M, Seth S, Torrents D, Wheeler DA, Xi L, Zhang J, Gerhard DS, Quesada V, Valdés-Mas R, Gut M, Campbell

PJ, Hudson TJ, McPherson JD, Puente XS, Gut IG. 2014. A comprehensive assessment of somatic mutation calling in cancer genomes. *bioRxiv* **Epub available before publication, Dec 24, 2014**:–. doi:http://dx.doi.org/10.1101/012997.

Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanbom JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung W, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Binger DD, Meir EGV, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L, , Network TR. 2013. The somatic genomic landscape of glioblastoma. *Cell* **155**:462–477.

Cerami E, Demir E, Schultz N, Taylor BS, Sander C. 2010. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* **5**:e8918.

Chen X, Pappo A, Dyer MA. 2015. Pediatric solid tumor genomics and developmental pliancy. *Oncogene* **Feb 2, Epub ahead of print**:–. doi:0.1038/onc.2014.474.

Friedmann-Morvinski D. 2014. Glioblastoma heterogeneity and cancer cell plasticity. *Critical Reviews in Oncogenesis* **19**:327–336.

Fujimoto M, Fults DW, Thomas GA, Nakamura Y, Heilbrun M, White R, Story JL, Naylor SL, Kagan-Hallet KS, Sheridan PJ. 1989. Loss of heterozygosity on chromosome 10 in human glioblastoma multiforme. *Genomics* **4**:210–214.

Gevaert O, Plevritis S. 2013. Identifying master regulators of cancer and their downstream targets by integrating genomic and epigenomic features. *Pacific Symposium on Biocomputing* 123–134.

Knudson A. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences* **68**:820–823.

Kumar A, Boyle EA, Tokita M, Mikheev AM, Sanger MC, Girard E, Silber JR, Gonzalez-Cuyar LF, Hiatt JB, Adey A, Lee C, Kitzman JO, Born DE, Silbergeld DL, Olson JM, Rostomily RC, Shendure J. 2014. Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes. *Genome Biology* **15**:530–539.

Larson DE, Harris CC, Chan K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, , Dong L. 2012. Somaticsniper: Identification of somatic point mutaitons in whole genome sequencing data. *Bioinformatics* **28**:311–317.

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPP. **2009**. The sequence alignment/map format and samtools. *Bioinformatics* **25**:2078–2079.

Loeb LA. **2011**. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Reviews Cancer* **11**:450–457.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. **2010**. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research* **20**:1297–1303.

Network TCGAR. **2008**. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**:1061–1068.

Nishikawa R, Sugiyama T, Narita Y, Furnari F, Cavenee WK, Matsutani M. **2004**. Immunohistochemical analysis of the mutant epidermal growth factor, δEGFR, in glioblastoma. *Brain Tumor Pathology* **21**:53–56.

of California Santa Cruz U. **2014**. *CGHub User Guide, Release 4.2.1*. University of California, Santa Cruz, Santa Cruz, California. URL `https://cghub.ucsc.edu`.

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Jr LAD, Hartigan J, Smith DR, Strausberg RL, Marie SKN, Shinjo SMO, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. **2008**. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**:1807–1812.

Team GD. **2015**. *GATK Best Practices: Recommended workflows for variant analysis with GATK*. URL `https://www.broadinstitute.org/gatk/guide/best-practices`.

Team RC. **2014**. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Tomasettia C, Vogelstein B, Parmigiania G. **2013**. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences* **110**:1999–2004.

Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN, Network TCGAR. **2010**. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell* **17**:98——110.

Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PY, Schaefer C, Risch N. **2014**. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research* **24**:1734–1739.

Wang G, Carbajal S, Vijg J, DiGiovanni J, Vasquez KM. **2008**. Dna structure-induced genomic instability in vivo. *Journal of the National Cancer Institute* **100**:1815–1817.

Wilson TA, Karajannis MA, Harter DH. **2014**. Glioblastoma multiforme: State of the art and future therapeutics. *Surgical Neurology International* **5**. doi:10.4103/2152-7806.132138.

Yu X, Sun S. **2013**. Comparing a few snp calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* **14**. doi:10.1186/1471-2105-14-274.

Tables and Figures

Table 1: **TABLE CAPTION** Table explanation.

Quantity	average	median	min	max	stdev
No. mutations, WGS	844	328	110	8,192	1,306
No. mutations, WGA	1,531	694	23	10,966	2,230
% mutations in overlap, WGS	31%	31%	1%	74%	20%
% mutations in overlap, WGA	44%	45%	3%	71%	13%
No. putative mutations removed, VAQ	309	164	17	11,439	1,372
No. putative mutations removed, LOH	539	169	3	8,538	1,051
No. putative mutations removed, 10bp-SNP	46	19	0	1,515	115
No. putative mutations removed, 10bp-INDEL	28	16	0	535	45
No. putative mutations removed, dbSNP	16	8	0	348	33
No. putative mutations removed, <10%	1	1	1	36	3
% overlap removed, VAQ	53%	57%	0%	100%	28%
% overlap removed, LOH	51%	52%	0%	99%	29%
% overlap removed, 10bp-SNP	3%	2%	0%	20%	3%
% overlap removed, 10bp-INDEL	4%	4%	0%	22%	4%
% overlap removed, dbSNP	3%	2%	0%	47%	7%

Table 2: **SNP prediction filters.** This table shows the various methods (filters) used to predict which differences found in tumor alignments relative to blood alignments are real somatic variants, as opposed to sequencing errors or other variants.

filter	software	purpose
GATK	GATK	Removes putative SNPs with GATK quality scores less than 40 (as part of the GATK processing, with indel realignment and base recalibration)
SS	SomaticSniper	Removes putative SNPs with a SomaticScore less than 40
VAQ	SomaticSniper	Removes putative SNPs with SomaticSniper Varaint Allele Quality scores less than 20
LOH	SomaticSniper, python	Removes putative SNPs that are identified as loss of heterozygosity
10bp-SNP	python	Removes putative SNPs located within a 10 bp window of any other putative SNP
10bp-INDEL	python	Removes putative SNPs located within a 10 bp window of indels
dbSNP	python	Removes putative SNPs that overlap with dbSNP coverage
<10%	python	Removes putative SNPs if, in the tumor data, the percentage of reads covering the site with the alternate allele is less than 10%

Table 3: **Back-end Processing.** This table shows the software packages we used in data processing, what we used each piece of software for, and the command associated with it. The rows are in order of use.

software	purpose	command
picard	regenerate fastq files from bamfile aligned to hg18	`java -d64 -Xmx4g -jar SamToFastq.jar I=$pfx.bam F=$pfx.1.fastq F2=$pfx.2.fastq 2>&1`
bwa	align fastq files to hg19	`bwa aln -q 30 -t 8 $hgReference $fastq > $fastq.aln.sai`
bwa, samtools	convert aligned fastq files into new bamfile	`bwa sampe -a 600 -P -r "$RG" $hgReference $fastq1.aln.sai $fastq2.aln.sai $fastq1 $fastq2 | samtools view -bSh -o $outprefix.bam -`
samtools	sort and index new bamfile	`samtools sort -@ 16 $outprefix.bam $outprefix.sorted 2, samtools index $outprefix.sorted.bam 2`
samtools	remove duplicate reads from bamfiles	`samtools rmdup ../$tumorpfx/$tumorpfx.out.sorted.bam $tumorpfx.dedup.bam`
GATK	indel realignment	`java -d64 -jar $gatkJar -R $hgReference -T IndelRealigner -rf BadCigar -I $tumorpfx.dedup.bam -known $G1000.Mills -known $G1000.Phase1.Indels -targetIntervals $tumorpfx.intervals -o $tumorpfx.realn.bam`
GATK	base recalibration	`java -d64 -jar $gatkJar -nct 8 -T BaseRecalibrator -rf BadCigar -I $tumorpfx.realn.bam -R $hgReference -knownSites $dbSNP -o $tumorpfx.recal.grp`
samtools	index recalibrated bamfile	`samtools index $tumorpfx.realn.recal.bam`
SomaticSniper	call somatic mutations, generate VCF	`bam-somaticsniper -q 40 -Q 40 -J -s 0.001 -F vcf -f $hgReference $tumorbam $normalbam $tumorpfx.SS.vcf`

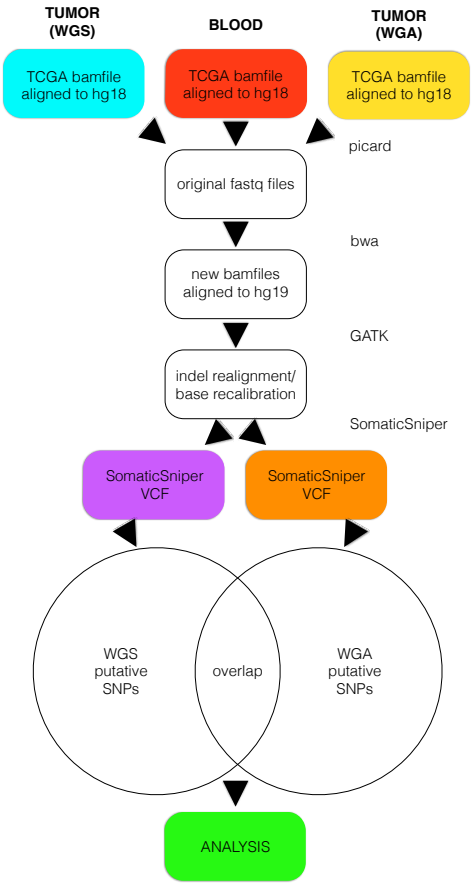

Figure 1: **Data processing pipeline.** For each of 55 patients, we began with a C484 tumor bamfile (WGS), a C282 tumor bamfile (WGA), and a normal bamfile, all aligned to hg18. For each bamfile, we used picard to regenerate fastq files, bwa to realign the fastq files to hg19, and GATK to recalibrate bases and indels. We used SomaticSniper to call somatic mutations (differences between the tumor and normal sequences) for each replicate. When we had a VCF for each replicate, we calculated the overlap between the two lists as the number of individual mutations which appeared in both replicates.

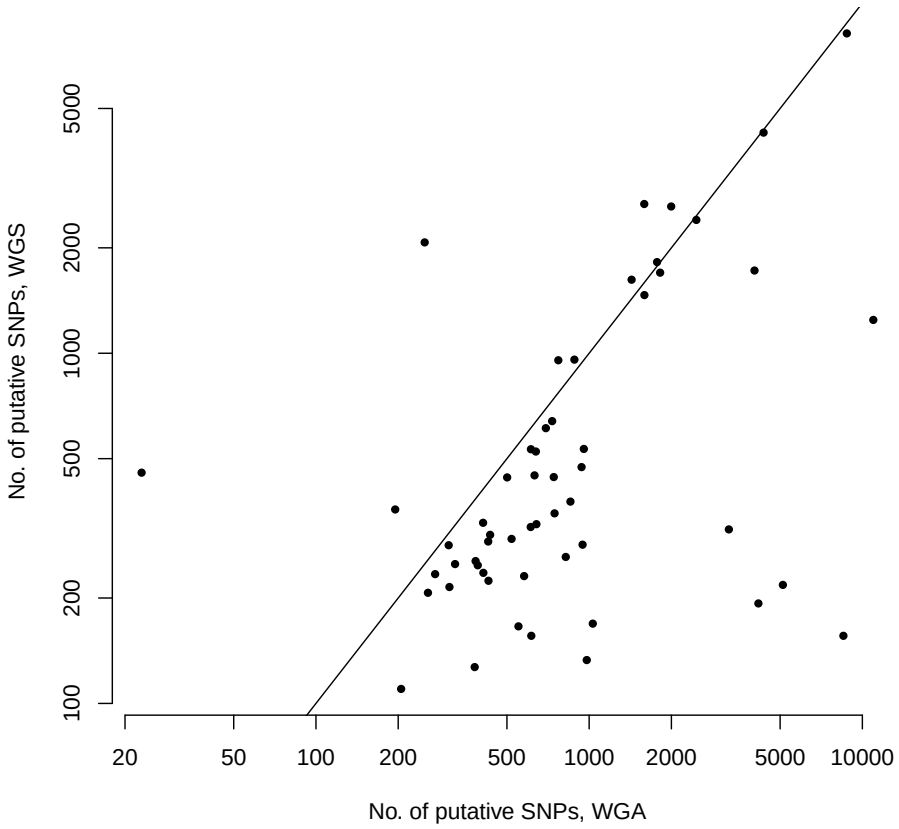

Figure 2: **Number of putative SNPs in WGS v. WGA, as called by SomaticSniper before filtering.** Each point represents data for a single patient. The line is $y = x$, so points falling below the line agree with the hypothesis that an additional amplification step produces more sequencing errors in a sample. The number of mutations found in one replicate correlates with the number of mutations found in the other replicate, cor=0.51, t = 4.3, df = 53, p-value = 8.35e-05.

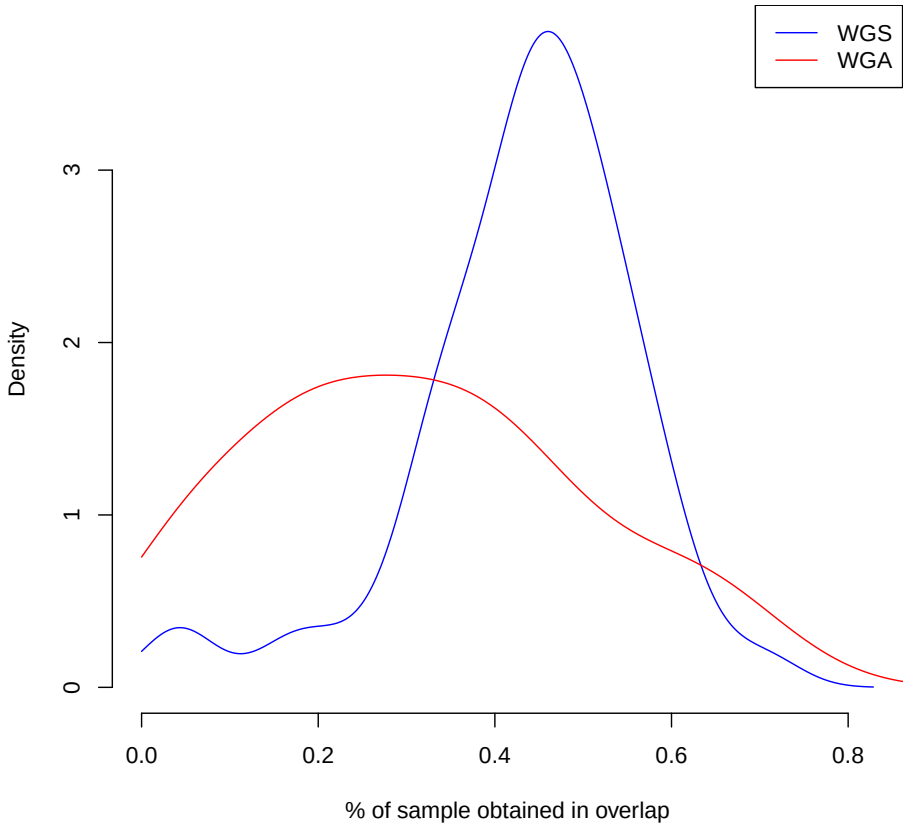

Figure 3: **Density of percentage overlap between WGS and WGA samples.** Density of the percentage of each WGS (blue) and WGA (red) sample that is present in the overlap between replicates for each patient. The WGS distribution is higher and narrower, showing that the WGS samples overall have a higher percentage overlap than the WGA samples, and less range in this parameter.

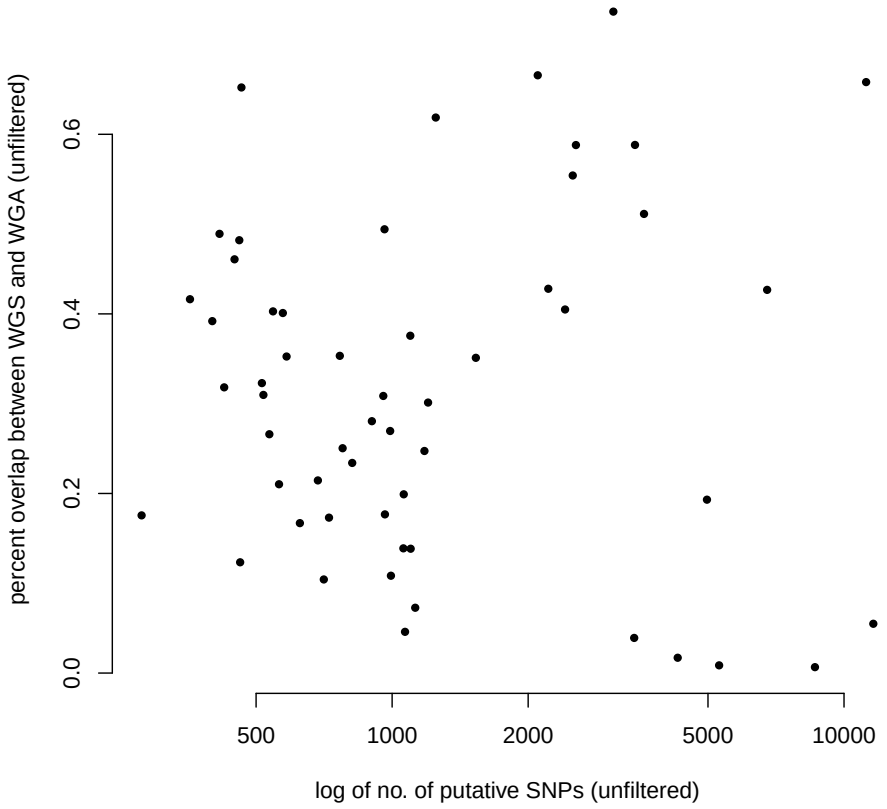

Figure 4: **Percent of overlap versus (log) number of putative SNPs per sample.** Plot of the percentage of a given WGS samples that is in the overlap between replicates against the number of mutations in that sample. (cor=0.25, p=.06, not significant)

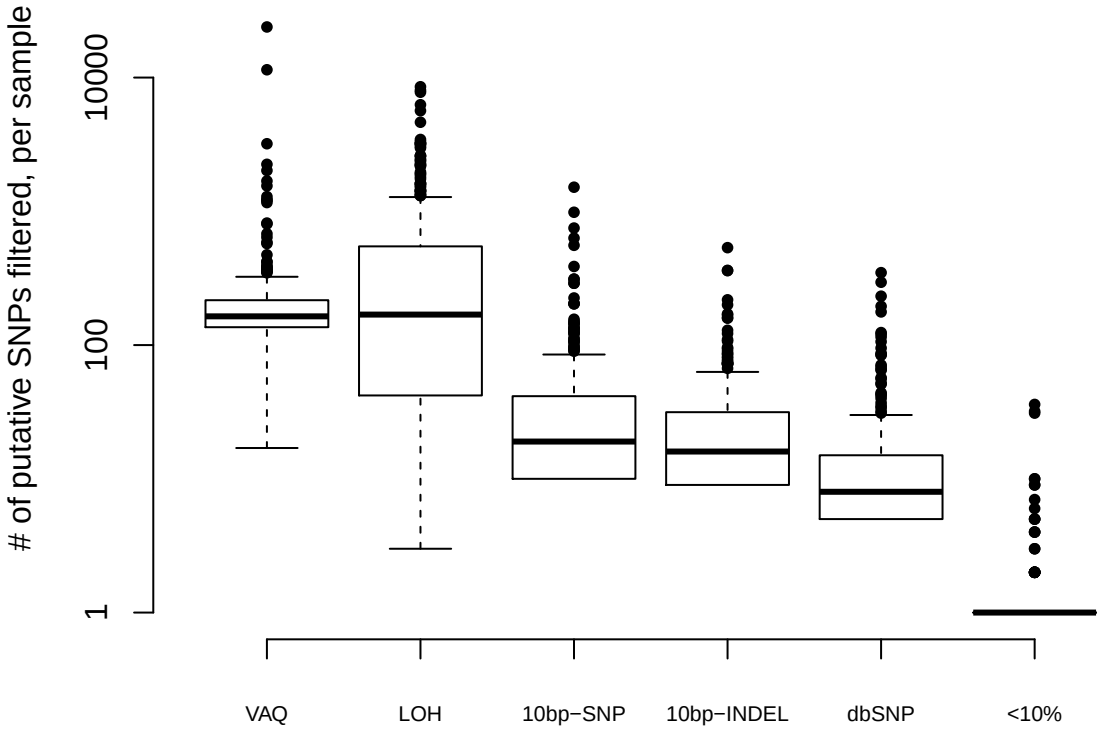

Figure 5: **Number of putative SNPs removed by each of six filters.** The $x-axis$ gives the name of each filter (detail on Table 1), and the $y-axis$ gives, on a log scale, the number of mutations removed by a given filter in a given sample. In all cases, the LOH and VAQ filters removed the majority of the putative SNPs for a given sample.

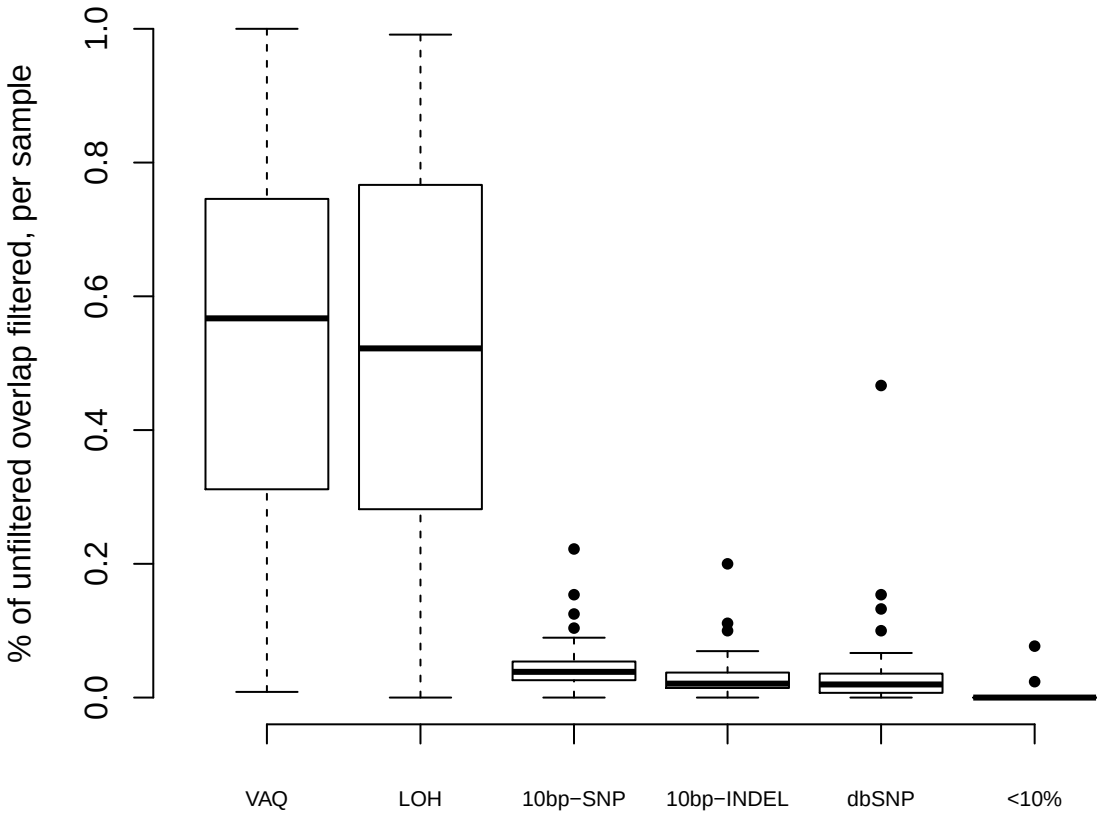

Figure 6: **Percentage of WGS/WGA overlap removed, by filter.** The $x - axis$ gives the names of the filters, and the $y - axis$ gives the percentage of the WGS/WGA overlap removed by the filter, per sample. Filters removing LOH and putative SNPs with low VAQ removed a significant percentage of the overlap; nothing else did.

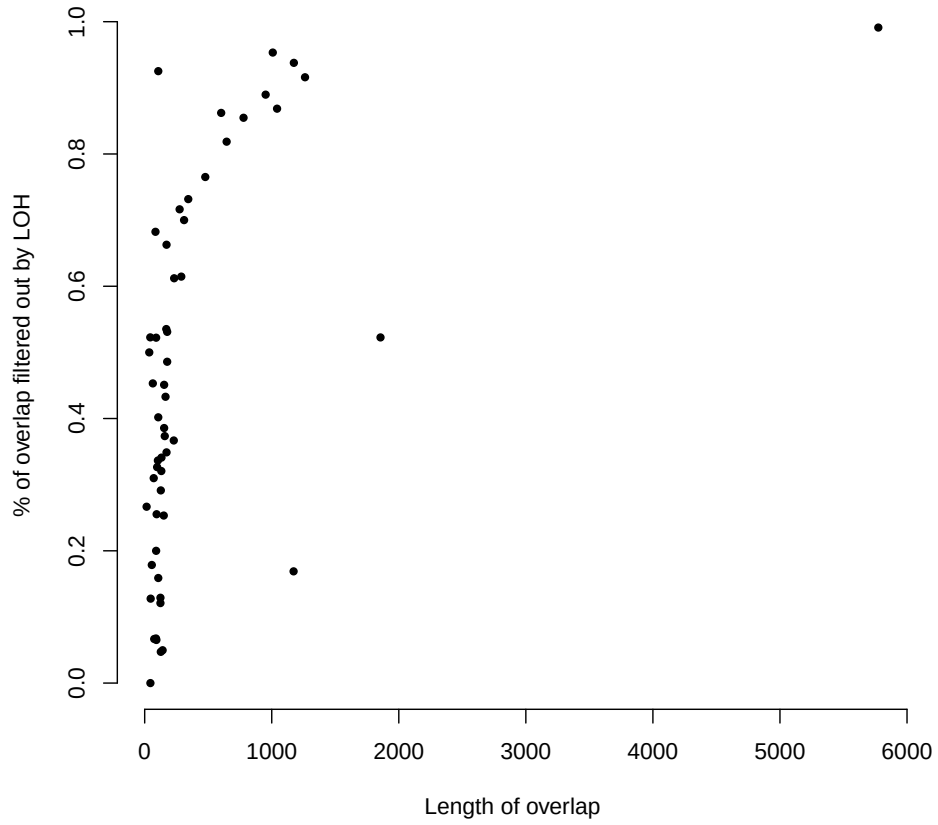

Figure 7: **Length of overlap of LOH segments.** As the length of the overlap between WGS and WGA increases ($x - axis$), the percentage of the overlap filtered out by LOH increases. This is the almost (but not quite) perfect inverse of the pattern observed in VAQ mutations (Figure 5).

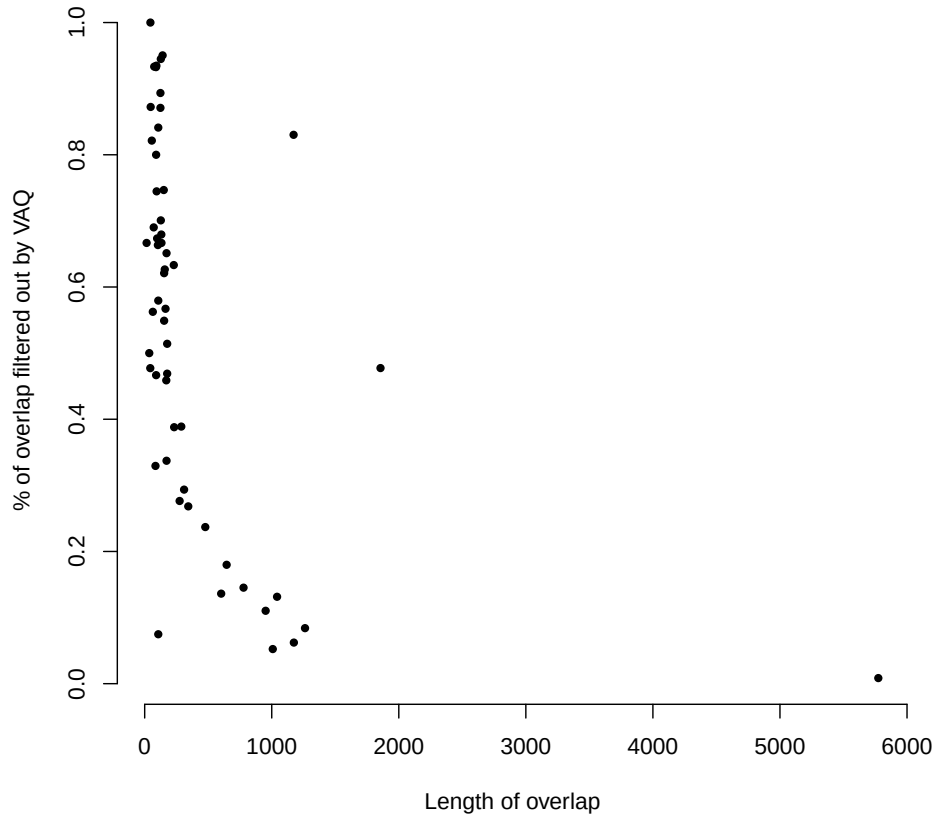

Figure 8: **Length of overlap of VAQ segments.** As the length of the overlap between WGS and WGA increases ($x - axis$), the percentage of the overlap filtered out by VAQ decreases. This is the almost (but not quite) perfect inverse of the pattern observed in LOH mutations (Figure 4).