Reproducibility of SNP calling in multiple sequencing runs from single tumors

Dakota Z. Derryberry, Claus O. Wilke, Matthew C. Cowperthwaite

July 2, 2014

1 Abstract

A) What system are you studying? B) Why is it important? C) What is your research approach? D) (optional) How is your research different from previous work? E) What are the results of your research? F) What conclusions can we draw from these results?

2 Introduction

A) What is the general field of study? B) What is the specific question of the research? C) Why is (a) and/or (b) important? D) What has previously been done in this area? E) How does your research differ? F) (end with) a summary of the research described in this paper

Glioblastoma multiforme (GBM) is the most common and deadly primary brain tumor, with a median survival of 13.9 months and a 5-year survival of 5% [2]. 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 chemoteraputics. Like all cancers, GBM is an evolutionary disease caused by mutations and other alterations to normal glial cell lines. However, following substantial research efforts dedicated to identifying all of the mutations in an individual patient's tumor to try and identify causal variants, few precise causal elationships have been identified. There are two theories behind this descrepancy: First is the view that "GBM growth is driven by a signaling network with functional redundancy that permits adaptation" [2]. Second is the notion that batch effects and sequencing protocols, and thus the data on which genomic analyses are based, is imperfect. In reality, both of these views are probably correct. In this paper, we use data from doubly-sequnced GBM patients to investigate key attributes of the data and some of the most common computational processing pipelines.

The largest effort behind this research is that of The Cancer Genome Atlas' Research Network, who first sequenced 206 GBMs in 2008 [7]. In the intervening years, this data set has grown to contain over 500 tumors with multiple types of molecular data, all made publicly available [2]. Each sample in the database contains data from a single patient's tumor and blood samples, allowing researchers to directly compare the two, and then compare sets of differences between tumor and "normal" tissue between patients. While most of the 600 sequencing samples in teh TCGA are from different patients, some 63 patients appear in the database twice, 55 having been seuqenced once each with two different DNA preparation methods, and 13 having been sequenced at two different times. In this paper,

Since it's publication in 2008, TCGA data has been used by researchers to investigate a variety of phenomena in GBM. (subtyping, drivres, pathways, targets) These successes aside, other results have caused some researchers to question what exactly is measured by the TCGA sequencing data. (the apper that showed all four subtypes in one tumor) (the more recent over time paper) (maybe another about heterogeneity) Sample preparation and sequencing methods, as well as computational analytical methods, indubitably influence sequencing results. In this paper, we begin to ask how.

WGS and WGA are different... We have 55 samples of each... Expectation is X, and we find as follows...

Same thing for the 13 samples...

Everyone must filter, (describe why)... We begin with the one SNP-caller SomaticSniper, and compare the filtered to the unfiltered date... We find as follows...

3 Results

3.1 Unfiltered data: WGA v. WGS

For 55 samples, TCGA has two identical data sets except for the sample preparation: one is whole genome sequencing (WGS) and the other is whole genome amplified sequencing (WGA). We calculated the number of mutations in each of these 110 samples (2 each from 55 patients). The number of mutations per sample ranged considerably, from XX to XX in the WGS samples and from XX to XX in the WGA samples. That the range is higher among the WGA samples is expected, because the amplification process may introduces mutations thereby artificially inflating the number of mutations in the dataset. To test this, we plotted the number of putative SNPs in the WGS and WGA sample preparation for each individual patient. As expected, for most patients, there were more mutations in the WGA sample than in the WGS sample (Figure 1). Surprisingly, not all exceptions had night numbers of mutations. One WGA sample with fewer than 50 mutaions had nearly 500 in the corresponding WGS sample; and one WGS sample with more than 200 mutations had only 200 in the corresponding WGA sample.

We next looked at the overlap between the called putative SNPs in the WGS and WGA samples. Although amplification may introduce new mutations, the mutations found in the WGS sample for each patient should be (mostly) a subset of those found in the WGA sample for that same patient. Although WGS may itself introduce some new mutations, the percent of WGS mutations that overlap with those found in the WGA sample should be significantly higher than the percent of WGA mutations found in the overlap with WGS. To test this, we calculated the overlap between the WGS and WGA samples for each patient. As expected, WGS samples have fewer non-overlapping putative SNPs than WGA samples (Figure 2).

As seen in figure one, most WGS samples have around 500 putative mutaitons, and most WGA samples have around 1000, but there are a number of exceptions. We hypothesized that these samples, some containing as many as 10,000 mutations, were mostly corrupted data. To test this, we plotted the percentage of the WGA samples that overlaped with the corresponding WGS samples (as a measure of sample quality) against the total number of putative SNPs in the WGS sample (Figure 3). We expected a negative correlation, which would indicate that as the quality of the sample worsened, the number of mutations increased. Instead, we got a completely random distribution. This could indicate... (still trying to decide what I think this indicates)

3.2 Unfiltered data: Time points

3.3 Filtering

4 Methods

4.1 Data and back-end processing

All sequence data came from The Cancer Genome Atlas (TCGA) Research Network's Gliobastoma multiforme (GBM) data set [7]. We downloaded raw reads in .bam files from 68 patients using CGHub [10]. For each patient, data consisted of one set of reads taken from blood DNA, and two sets of reads taken from tumor DNA. In 55 cases, the two sets of reads from tumor DNA were one set of reads from whole genome sequencing (WGS) and one set of reads from whole genome sequencing with amplification (WGA). In 13 cases, the two sets of reads from tumor DNA were one set of reads pre-readiation treatment and one set post-radiation treatment. We developed a pipeline to align all reads to HG19. This pipeline used picard [1], BWA [4] for alignment, SamTools [5], GATK [6], and custom python code. We used SomaticSniper [3] to align each tumor sequence with it's corresponding blood sequence, then filtered the SomaticSniper data using 6 custom filters. These removed putative SNPs (i) with a SomaticScore less than 40, (ii) within 10 bp from a high-quality indel, (iii) within 10 bp of another putative SNP, (iv) with a variant allele quality less than 20, (v) with a dbSNP [9] identifier and

fewer than 10x coverage, and (vi) that were loss of heterozygosity. All custom code is available in a github repository, and is available upon request.

4.2 Analysis

We used custom python scripts, available in the above github repository, to perform simple calculations and data operations. We used R [8] to do statistics and generate figures, and this code is also stored in a github repository and available upon request.

5 Discussion

A) Start with a brief (one paragraph or less) summary fo your research, what you did and what you found B) Put your results into perspective, discussing (i) what they imply and (ii) how they compare to existing literature C) Address potential shortcomings. What could have gone wrong, which parts of the research might be misleading, are there any other caveats? D) (optional) Describe implications for future research; are there obvious extensions? E) (optional) End with a paragraph that again summarizes your research and highlights the most important findings

6 Figures

References

[1] mmccowan nilshomer tfenne alecw, brilliantred. picard, May 2014.

[2] Aaron McKenna Benito Campos Houtan Noushmehr Sofie R. Salama Siyuan Zheng Debyani Chakravarty J. Zachary Sanbom Samuel H. Berman Rameen Beroukhim Brady Bernard Chang-Jiun Wu Giannicola Genovese Ilya Shmulevich Jill Barnholtz-Sloan Lihua Zou Rahulsimhan Vegesna Sachet A. Shukla Giovanni Ciriello W.K. Yung Wei Zhang Carrie Sougnez Tom mikkelsen Kenneth Aldape Darell D. Binger Erwin G. Van Meir Michael Prados Andrew Sloan Kieth L. Black Jennifer Eschbacher Gaetano Finocchiaro William Friedman David W. Andrews Abhijit Guha Mary Iacocca Brian P. O'Neill Greg Foltz Jerome Myers Daniel J. Weisenberger Robert Penny Raju Kucherlapati Charles M. Perou D. Niel Hayes Richard Gibbs Marco Marra Gordon B. Mills Eric Lander Paul Spellman Richard Wilson Chris Sander John Weinstein Matthew Meyerson Stacey Gabriel Peter W. Laird David Haussler Gad Getz Lynda Chin Cameron W. Brennan, Roel G.W. Verhaak and TCGA Research Network. The somatic genomic landscape of glioblastoma. *Cell*, 155:462–477, 2013.

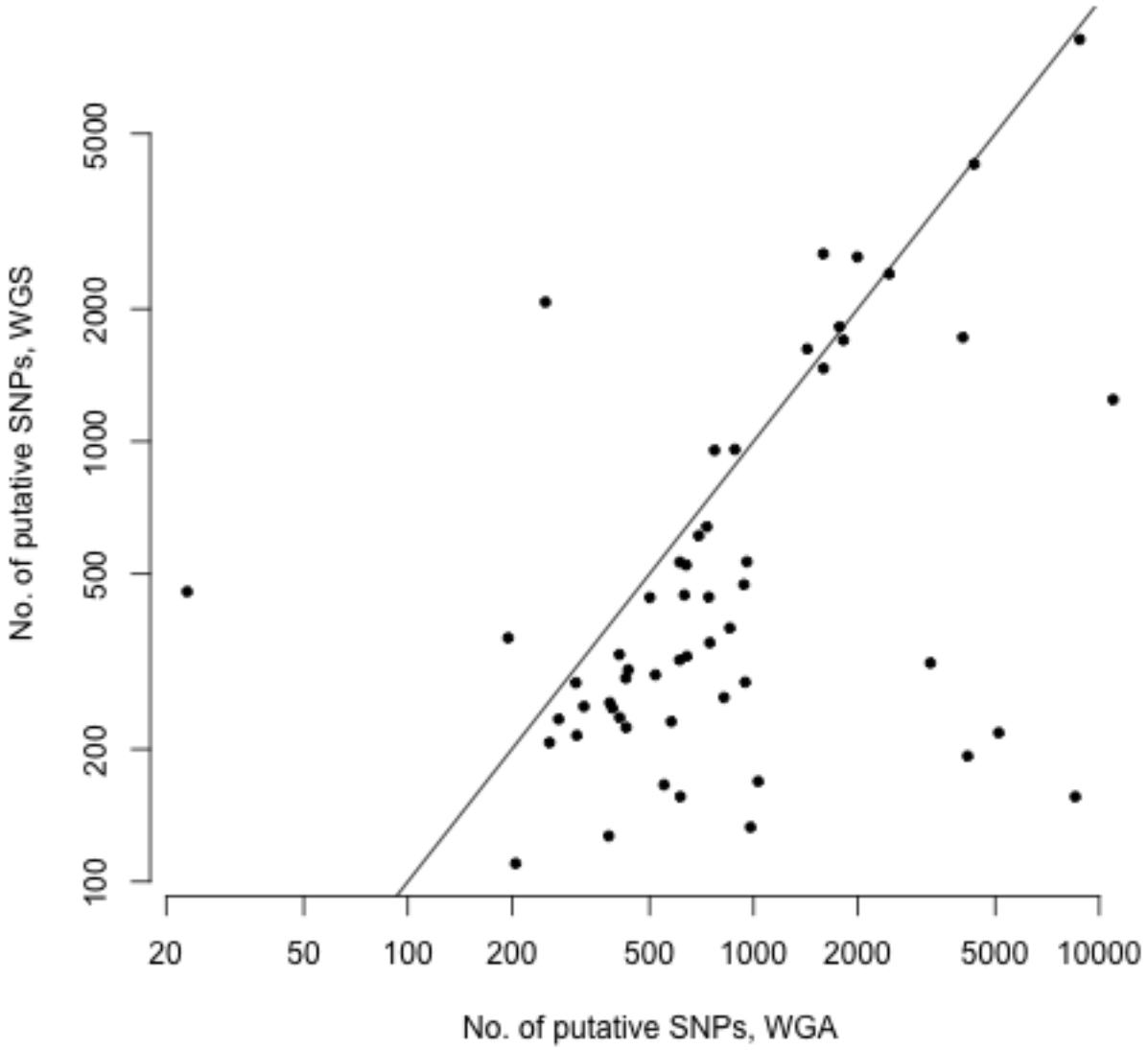

Figure 1: This figure plots the number of putative SNPs called by SomaticSniper (unfiltered) using the WGS v. WGA. Each point is a patient. The line is y=x, so points falling below the line agree with the hypothesis that whole genome amplification makes more mutations in a sample.

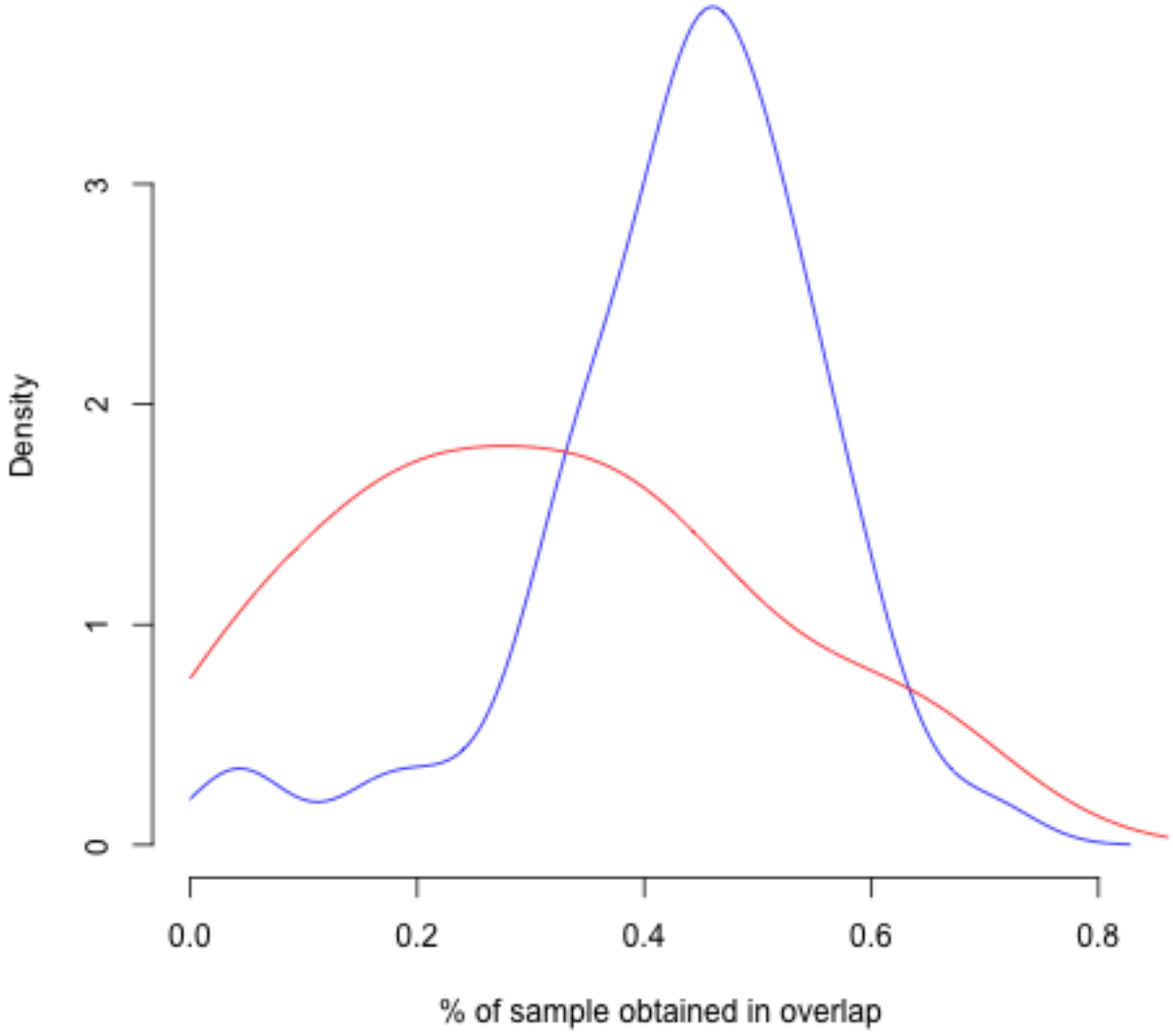

Figure 2: This figure shows the density of the percentage of each WGS (blue) and WGA (red) sample that overlaps with the other sample from the same patient. The WGS distribution is higher and narrower, showing that the WGS samples overall have a higher percentage overlap than the WGA samples, and less range in this parameter.

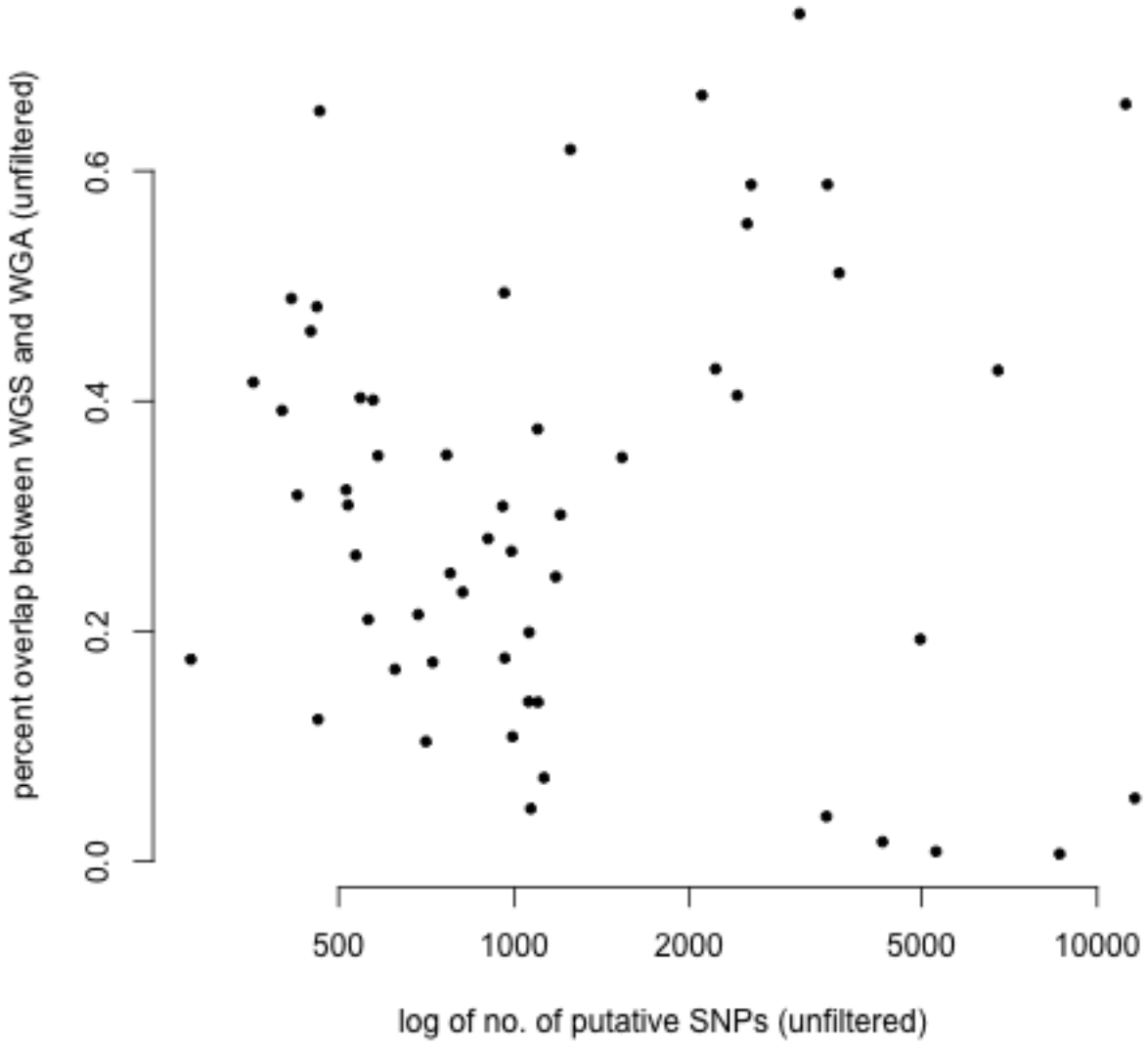

Figure 3: Plot of the percentage of the WGA samples that overlaped with the corresponding WGS samples (as a measure of sample quality) against the total number of putative SNPs in the WGS sample

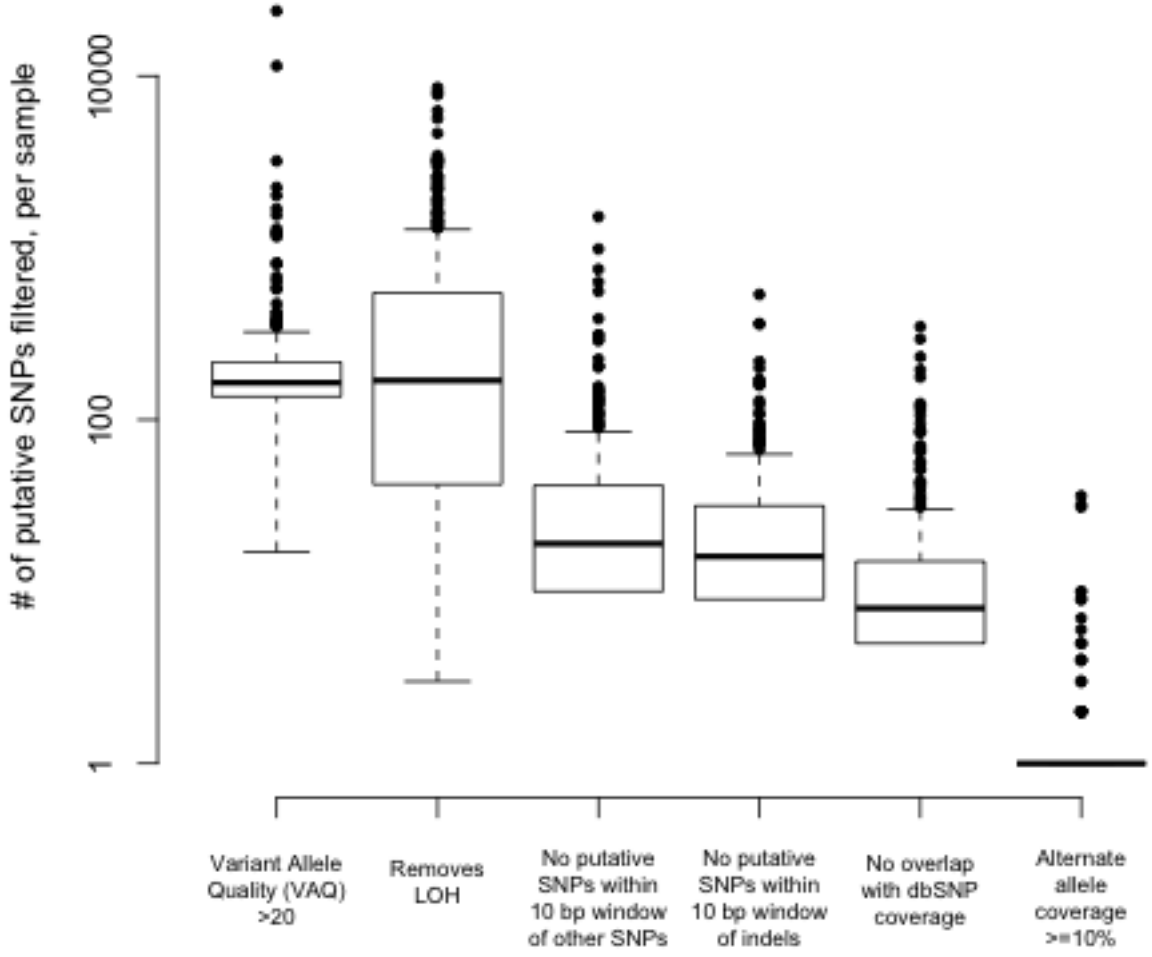

Figure 4: Boxplot of the total number of putative SNPs removed by each of six filters, in each of 379 sequencing samples, from 311 patients. The x-axis gives the name of those filters that removed putative SNPs, and the y-axis gives, on a log scale, the number of mutations removed by a given sample.

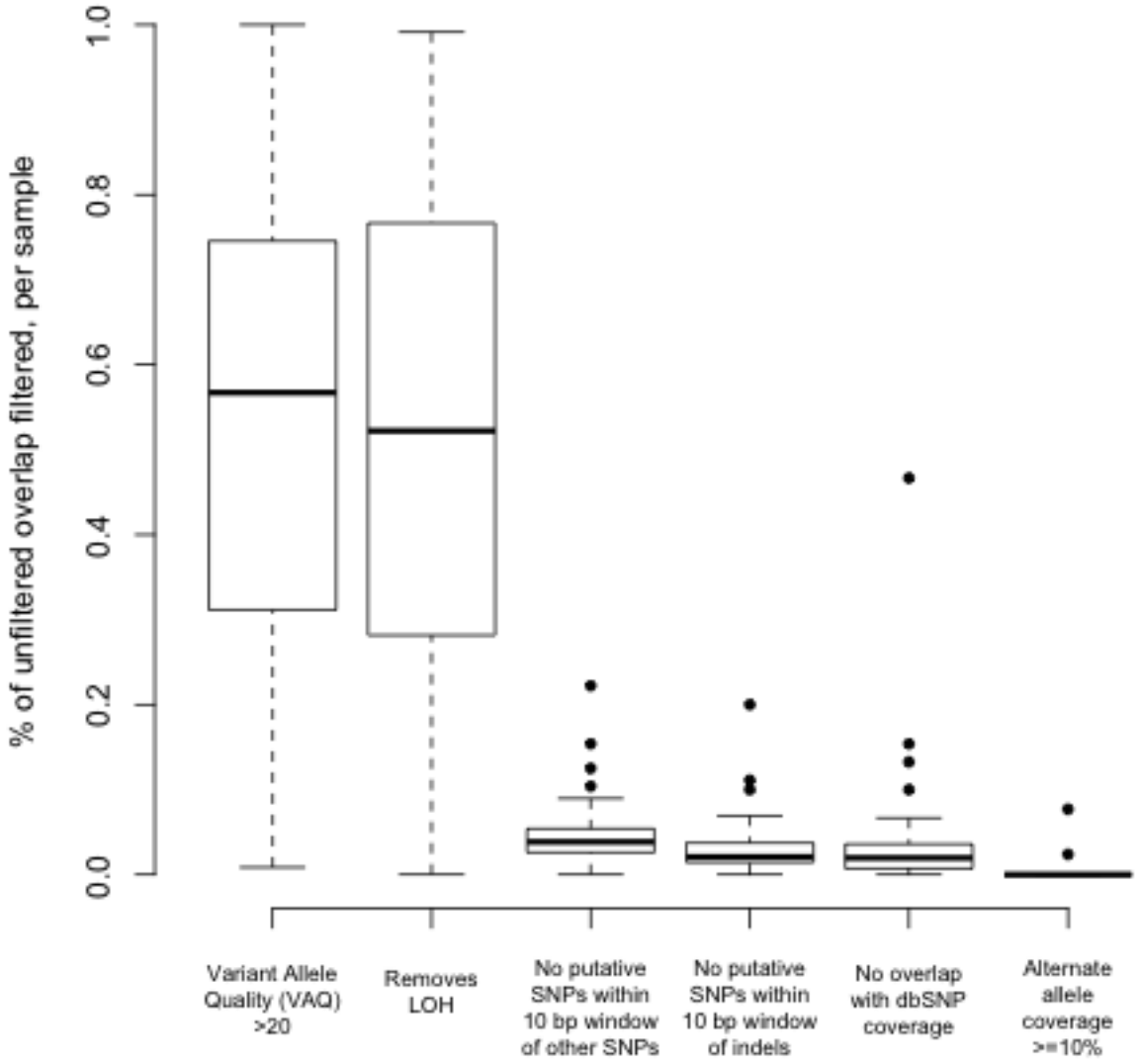

Figure 5: Boxplot of the percentage of putative SNPs making up the overlap between the WGS and WGA sequences of a single sample that was removed by a given filter. The x-axis describes the filters, and the y-axis gives the percentage of the overlap removed.

[3] Ken Chan Daniel C. Koboldt Travis E. Abbott David J. Dooling Timothy J. Ley Elaine R. Mardis Richard K. Wilson David E. Larson, Christopher C. Harris and Li Dong. Somaticsniper: Identification of somatic point mutaitons in whole genome sequencing data. *Bioinformatics*, 28:311–317, 2012.

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

[5] Alec Wysoker Tim Fennell Jue Ruan Nils Homer Gabor Marth Goncalo Abecasis Richard Durbin Heng Li, Bob Handsaker and 1000 Genome Project Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25:2078–2079, 2009.

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

[7] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[9] Kholodov M Baker J Phan L Smigielski EM Sirotkin K Sherry ST, Ward MH. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29:308–311, 2001.

[10] University of California, Santa Cruz, Santa Cruz, California. *CGHub User Guide, Release 4.2.1*, 2014.