Reproducibility of SNP calling in multiple sequencing runs from single tumors

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

June 6, 2014

1 Introduction

Intro goes here.

2 Results

2.1 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)

2.2 Unfiltered data: Time points

2.3 Filtering

3 Methods

3.1 Data and back-end processing

All sequence data came from The Cancer Genome Atlas (TCGA) Gliobastoma multiforme (GBM) data set. We downloaded raw reads in (XXX format) using (CGHub?) on (date) for 68 patients. For each patient, data consisted of one set of rads 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 (bowtie?) for alignment and (bedtools? whatever – quality control...) We used SomaticSniper to align each tumor sequence with it's corresponding blood sequence. We then filtered the SomaticSniper data using 8 custom filters. These removed (i) , (ii) , (iii) , (iv) , (v) , (vi) , (vii) , and (viii) . All custom code is available in a github repository, located at (webaddress). In addition to sequence data, filter 6 uses dbSNP, version 137, to find common SNPs among the putative mutaitons called by SomaticSniper.

3.2 Analysis

We used custom python scripts, available in the above github repository, to perform simple calculations and data operations. We used R [1] to do statistics and generate figures, and this code is also available in the git repository.

4 Discussion

What do I think of this?

5 Figures

References

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

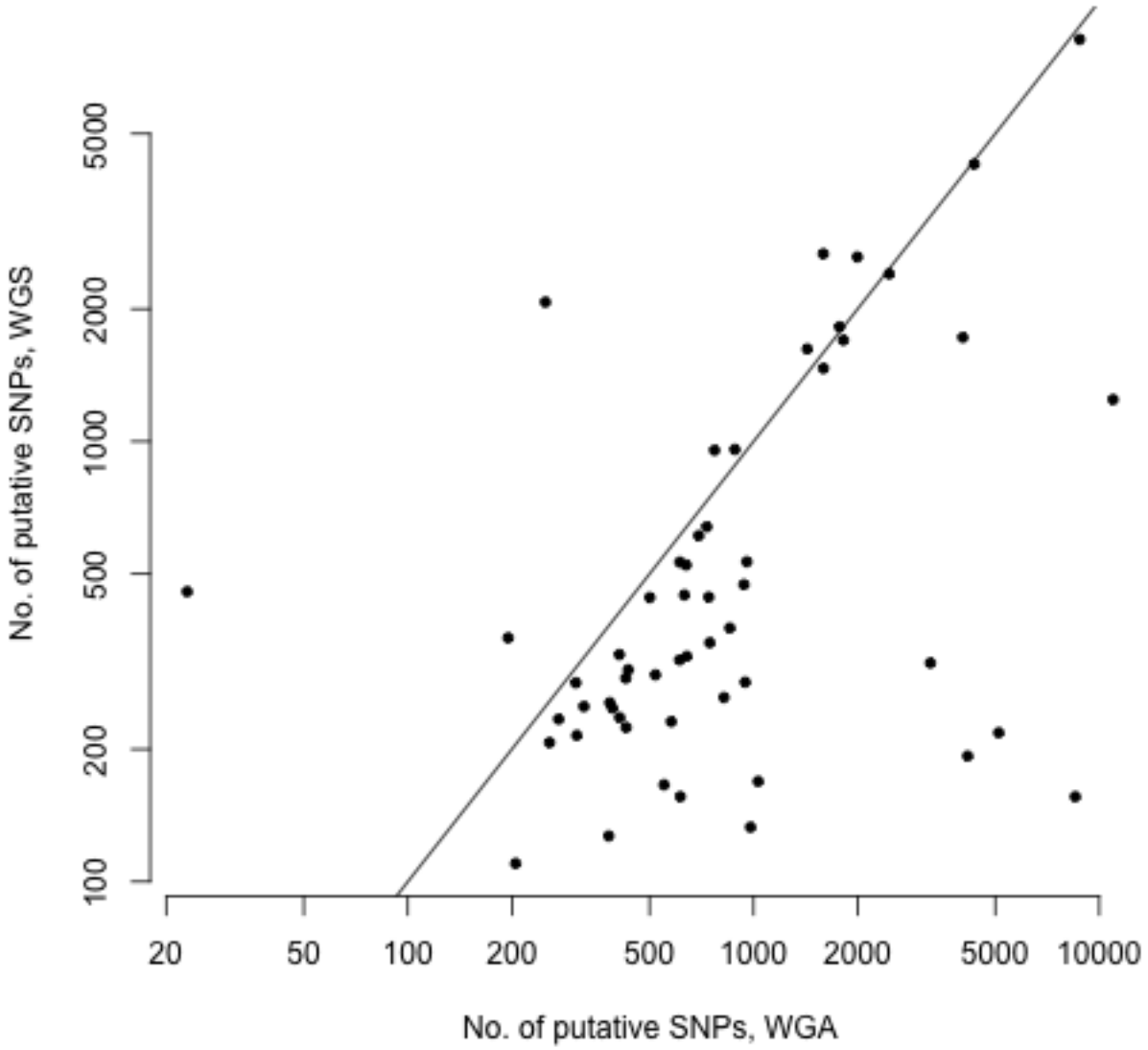

Figure 1: This figure plots the number of putative SNPs called by SomaticSniper (unfiltered) using the WGS v. WGA. Each point is a patient. The line is y=x, so points falling below the line agree with the hypothesis that whole genome amplification makes more mutations in a sample.

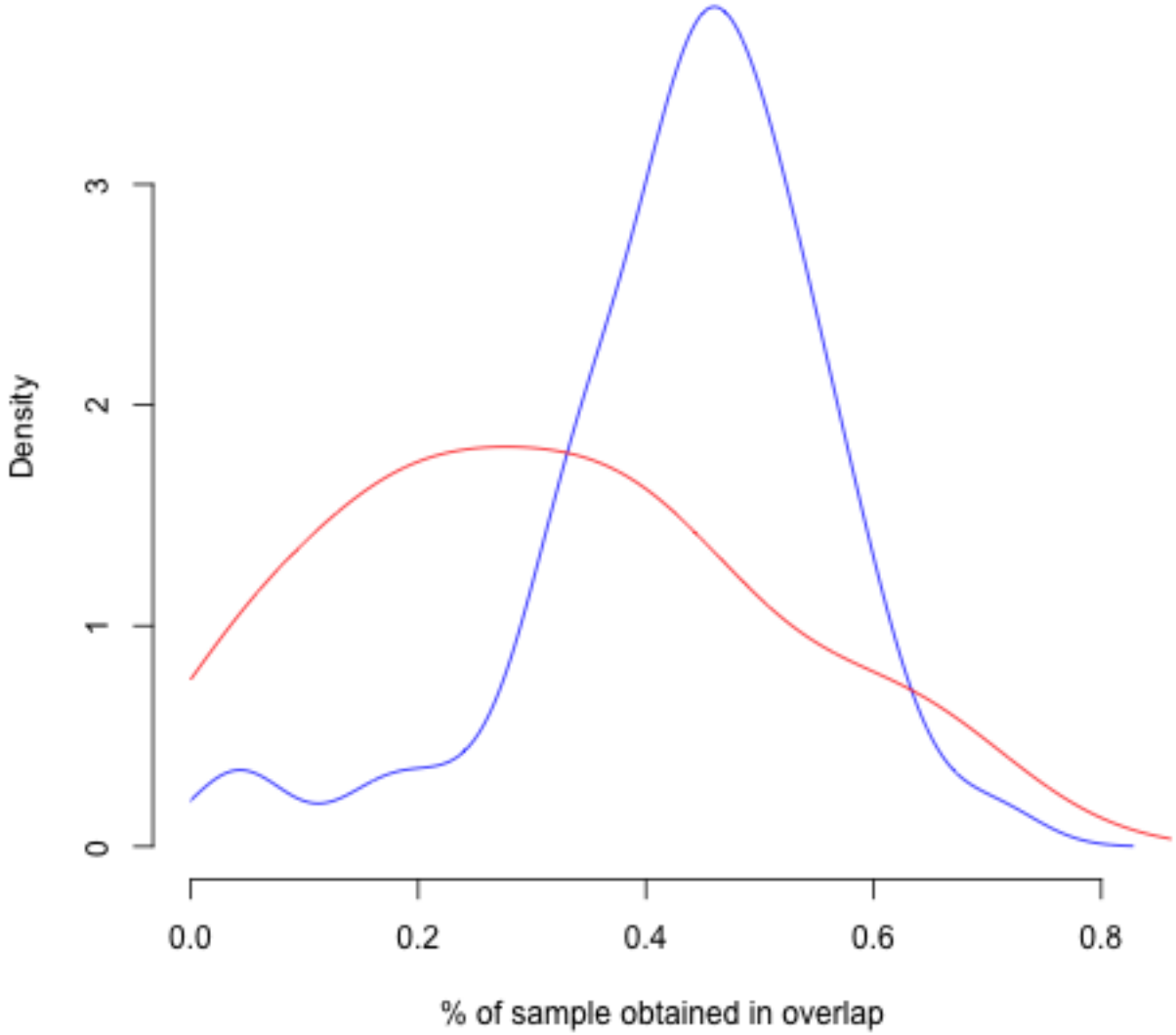

Figure 2: This figure shows the density of the percentage of each WGS (blue) and WGA (red) sample that overlaps with the other sample from the same patient. The WGS distribution is higher and narrower, showing that the WGS samples overall have a higher percentage overlap than the WGA samples, and less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

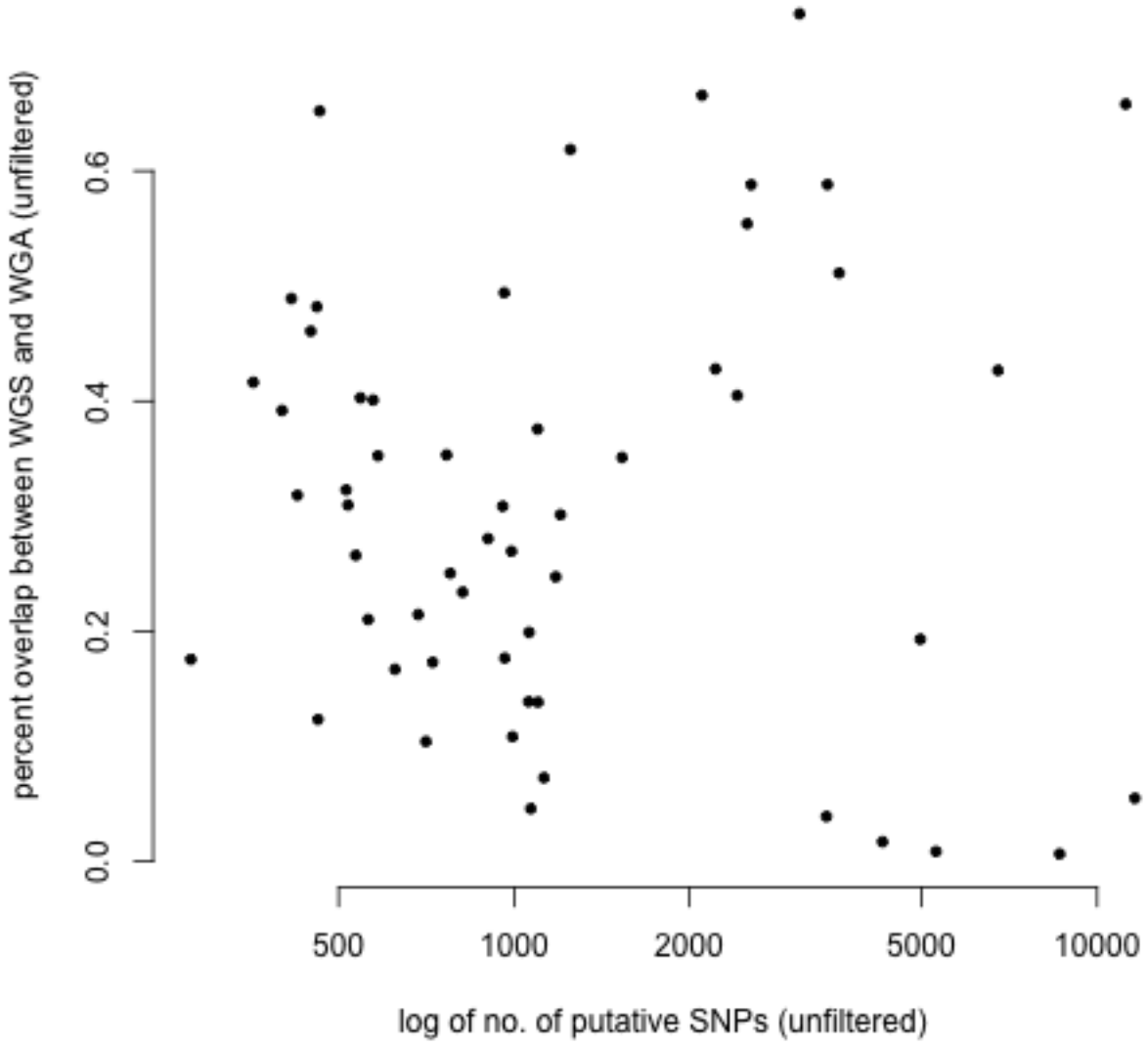

Figure 3: Plot of the percentage of the WGA samples that overlaped with the corresponding WGS samples (as a measure of sample quality) against the total number of putative SNPs in the WGS sample