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

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

1. UT Austin, Cell and Molecular Biology, 2. St. David's Hospital NeuroTexas Research Institute, 3. UT Austin, Integrative Biology

Introduction

TCGA's public 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. [1,2] Large-scale sequencing projects like TCGA make cancer sequencing data available to many researchers. These data are an enormously valuable resource, but their accuracy and reproducibility are unknown, and the knowing could only increase the utility of the data. We make a first pass at evaluating the repeatability of the data by comparing 55 technical replicates in the TCGA GBM dataset.

Methods

For 55 tumors, we examine two technical replicates each: one replicate of normal whole exome sequencing (WGS), and another replicate with an additional amplification step (WGA).

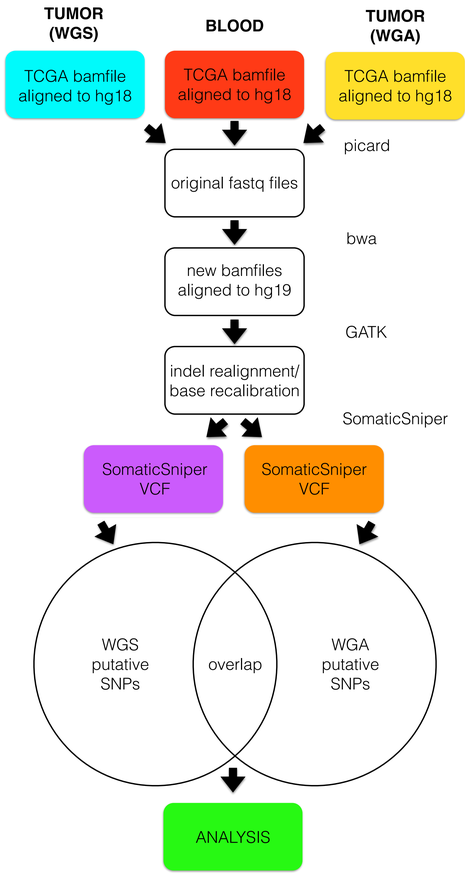


Figure 1: *Processing pipeline.* For each replicate, we downloaded TCGA bamfiles aligned to hg18 using the CGHub utility, regenerated fastq files using picard,[3] aligned fastq files to hg19 using bwa,[4] did base recalibration and indel realignment with GATK,[5] and finally variant calling with SomaticSniper.[6] We filtered variant calls and examined the overlaps between technical replicates using custom python scripts (in a git repository, available upon request). We generated all figures and statistics in R.[7]

~1/2 of SNPs are found in both samples

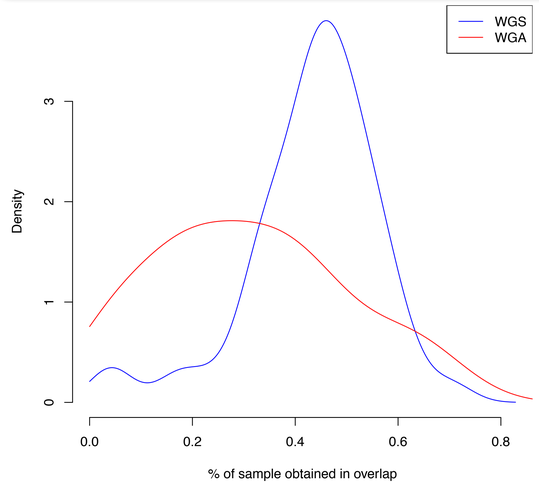


Figure 2: *Density of percentage overlap between WGS and WGA samples.* The density of the percentage of each WGS (blue) and WGA (red) sample that is present in the overlap between replicates for each patient. The WGS distribution is higher and narrower, showing that the WGS samples overall have a higher percentage overlap than the WGA samples, and less range in this parameter.

Mutation frequency varies

We tested the hypothesis that if the majority of putative SNPs in samples with a large number of putative SNPs are errors, then the percent overlap between replicates (a measure of data quality) should decrease while the number of putative SNPs increases.

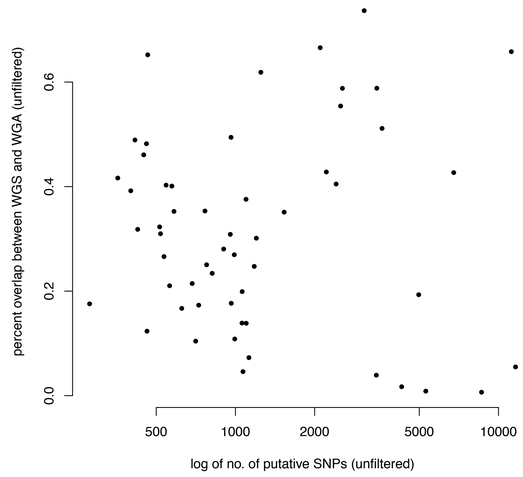


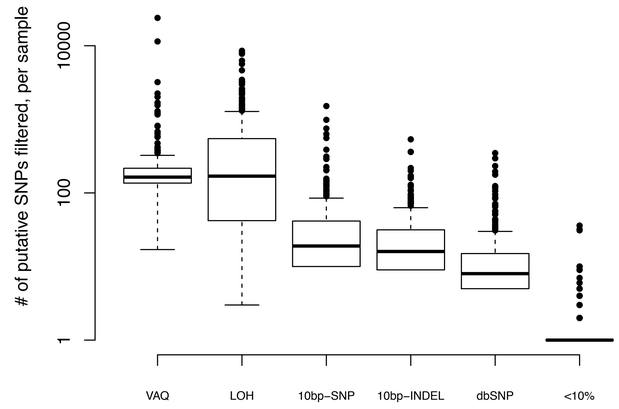
Figure 3: *Percent of overlap versus (log) number of putative SNPs per sample.* Plot of the percentage of a given WGS samples that is in the overlap between replicates against the number of mutations in that sample. (cor=0.25, p=.06, not significant)

Conclusions

- We found that about half of putative SNPs in one technical replicate were also present in the other technical replicate.
- A higher-than-expected number of putative somatic mutations found in some, but not all, of the patient samples was repeatable across technical replicates, suggesting the possibility that a higher mutation frequency could be a feature of a subset of GBM tumors.
- We note that these filters were designed to reduce or eliminate differences between the reference and the sample, but in cancer genomics our goal is to highlight changes. Therefore, it is possible that we need to modify our filtering protocols.

Not all filters are created equally.

We counted the number of putative SNPs removed by each filter, per sample:



And the percent of the overlap removed by each filter, in each sample:

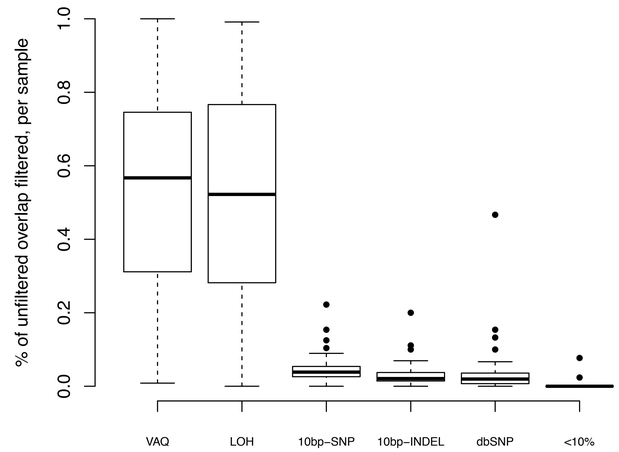


Figure 4: *Putative SNPs removed, by filter.* We found that only three filters removed any appreciable number of putative mutations from one and not both technical replicates.

References and Acknowledgements

1. The Cancer Genome Atlast Research Network. 2008. **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** Nature 455:1061–1068.
2. Brennan CW, et al. 2013. **The somatic genomic landscape of glioblastoma**. Cell 155:462–477.
3. alecw, brilliantred, mmccowan, nilshomer, tfenne. 2014. **picard.** URL http: // picard.sourceforge.net.
4. Li H, Durbin R. 2009. **Fast and accurate short read alignment with burrows-wheeler transform.** Bioinformatics 25:1754–1760.
5. McKenna A, et al. 2010. **The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** Genome Research 20:1297–1303.
6. Larson DE et al. 2012. **Somaticsniper: Identification of somatic point mutaitons in whole genome sequencing data.** Bioinformatics 28:311–317.
7. R Development Core Team. 2014. **R: A Language and Environment for Statistical Computing.** R Foun- dation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.