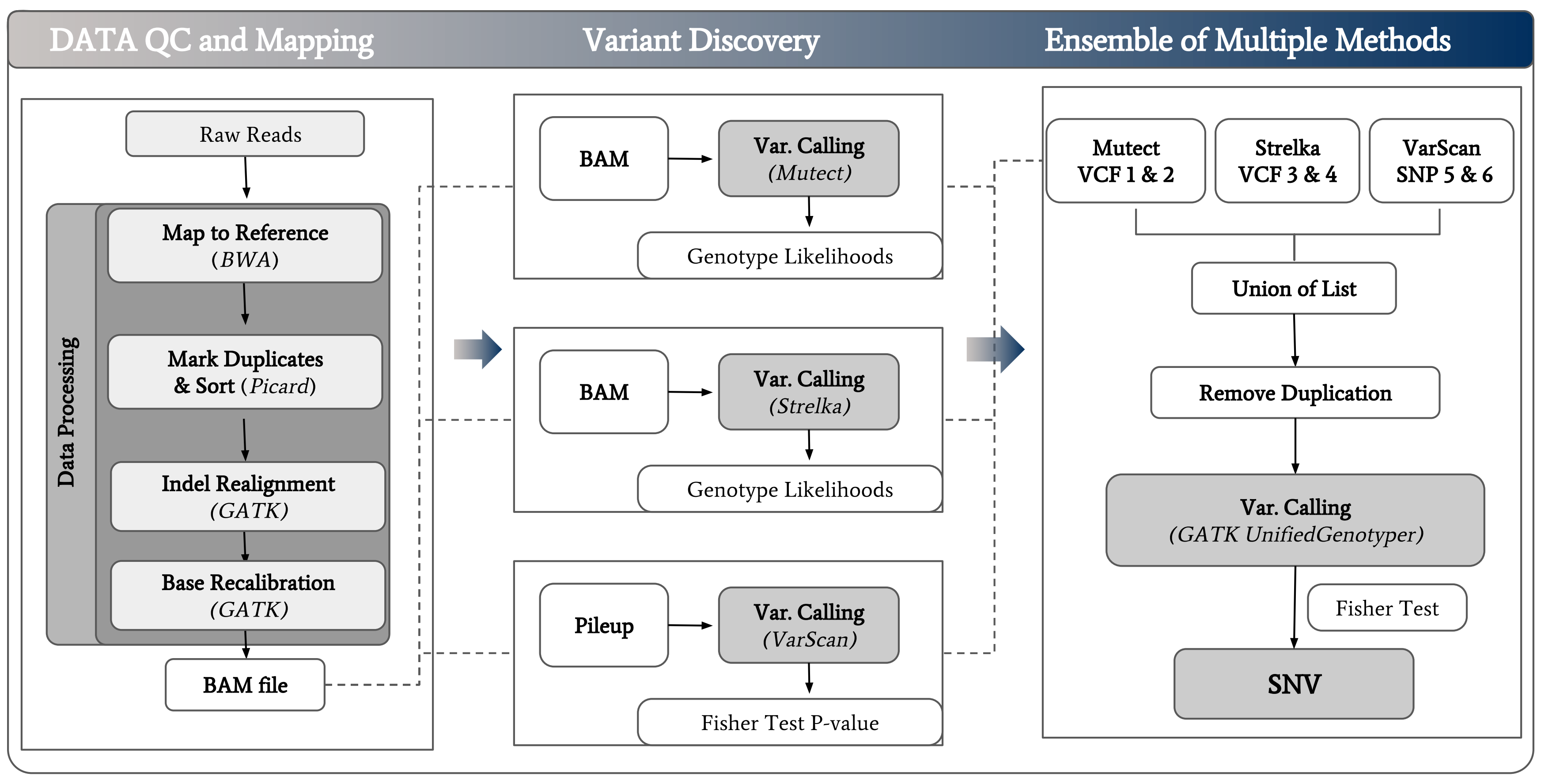
Finding the missing variants: Integrated somatic mutation detection from tumor-normal sequencing data using multiple callers

Yu Wang, Jun Z Li*, Department of Human Genetics, University of Michigan, Ann Arbor, MI.

Introduction

Multiple computational methods have appeared for calling somatic mutations based on next-generation sequencing data of matched tumor-normal pairs. The idea underlying most of these methods is to identify new alleles in the tumor sample or altered allele frequencies in the tumor that are statistically robust given the sequencing depth. However, several commonly adopted methods, such as ***Mutect***, ***Varscan*** and ***Strelka***, often detect very different sets of somatic variants. We sought to understand the source of such differences and found that some of these methods failed to consider 1) **the case of loss-of-heterozygosity (LOH)**, thus could not report Ref/Ref (or Alt/Alt) genotype in the tumor when the normal is Ref/Alt. Others relied on the Ref/Ref as the baseline genotype and 2) **failed to call the Ref/Alt in the tumor when the normal is Alt/Alt**. Further, many of these methods 3) **assumed no tumor-normal mixing** and would not leverage read count data to detect quantitative differences between the normal and tumor allele frequencies. We are developing a workflow to combine different callers result and overcome all mentioned problems.

Methods and Data

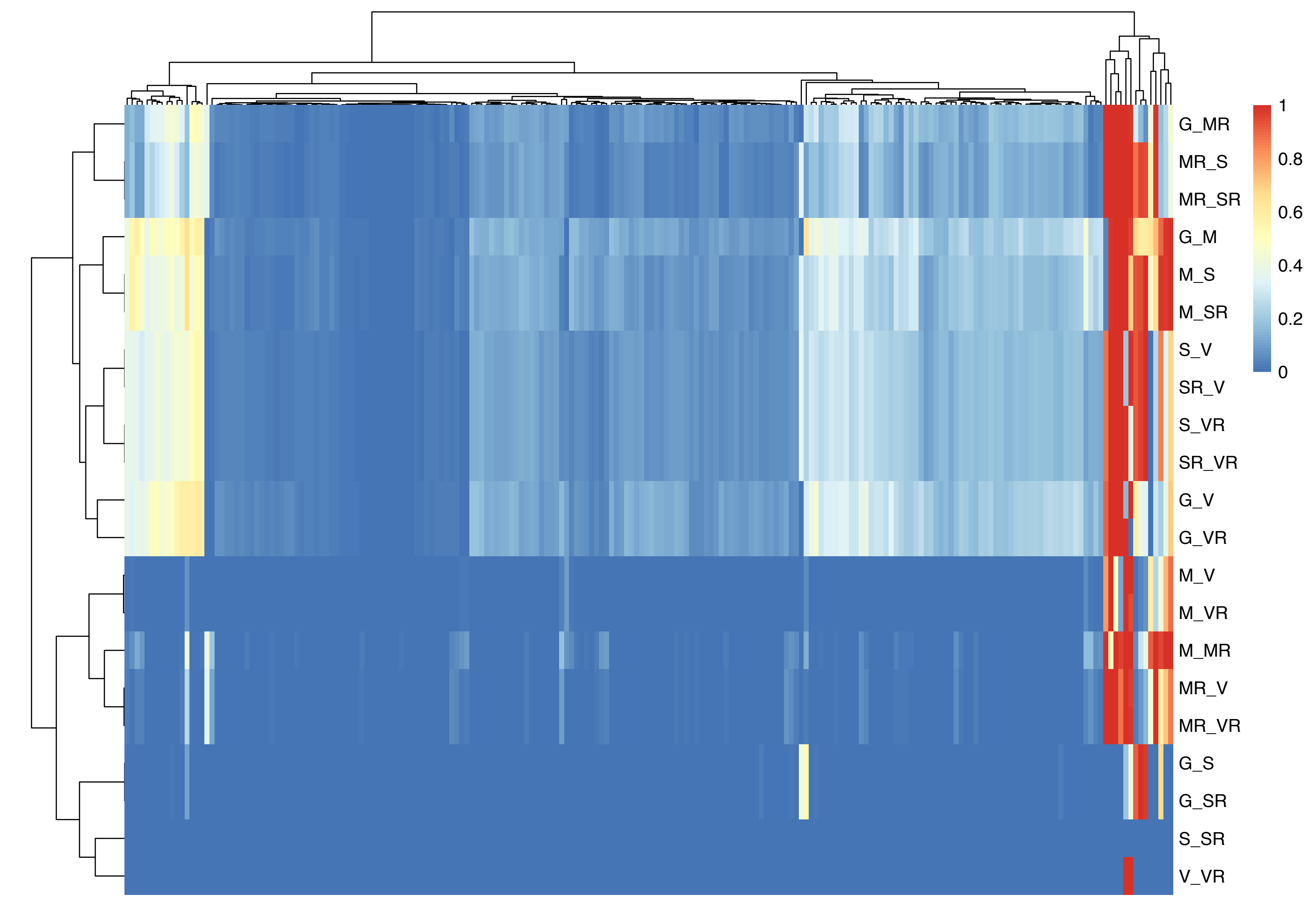


Workflow for combinations of SNV callers

We used an in-house exome sequencing data for a T-N pair at an average depth of 300X to compare Mutect, Varscan, and Strelka, and characterized their striking differences in the category of variants detected and their different cutoff values. We have since established a unified approach that combines the respective strengths of these methods. Our approach is to create a union list of called variants from these methods, combined with a parallel series of calls based on T-N swapping, and re-test the T-N allele frequencies using the Fisher's Exact test. The results thus obtained covered a much more complete catalog of somatic variants in all possible T-N allelic combinations and detected changes that occurred only in a fraction of the cells.

Parameters for ***Mutect***(1.1.4): max normal ref allele <= 4 (default 1, 1% of total) , downstream read depth limitation to 10000;
Parameters for ***Varscan***(2.2): default parameters; mpileup (q =1, remove duplicated reads first);
Parameters for ***Strelka*** (1.0.15) : isSkipDepthFilter = 1, set to 1 to skip depth filtration for whole exome data;
Parameters for ***GATK*** (3.4.46): downstream read depth limitation to 10000 (default 1000), min base quality 20.

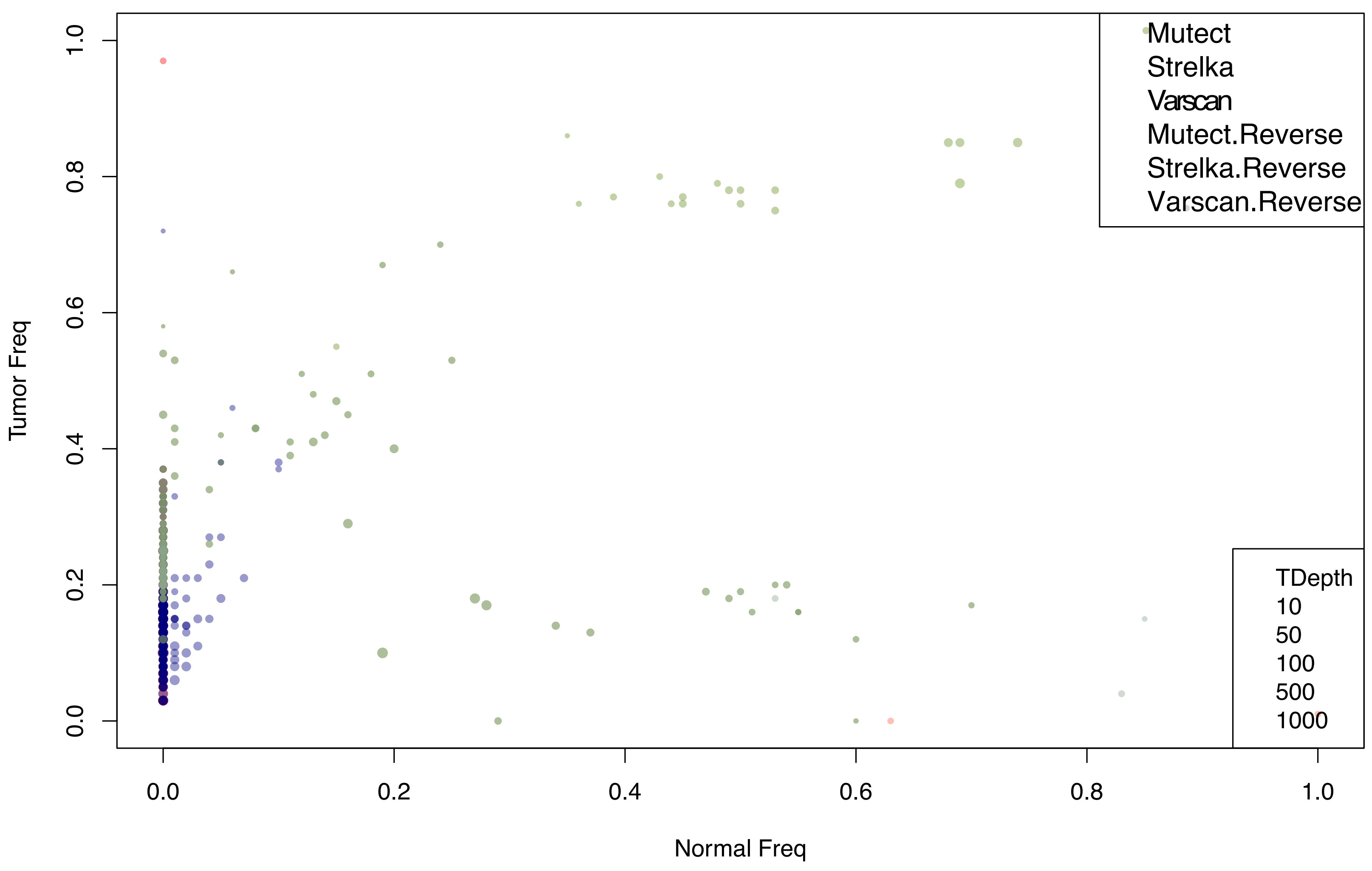
Comparison of different callers



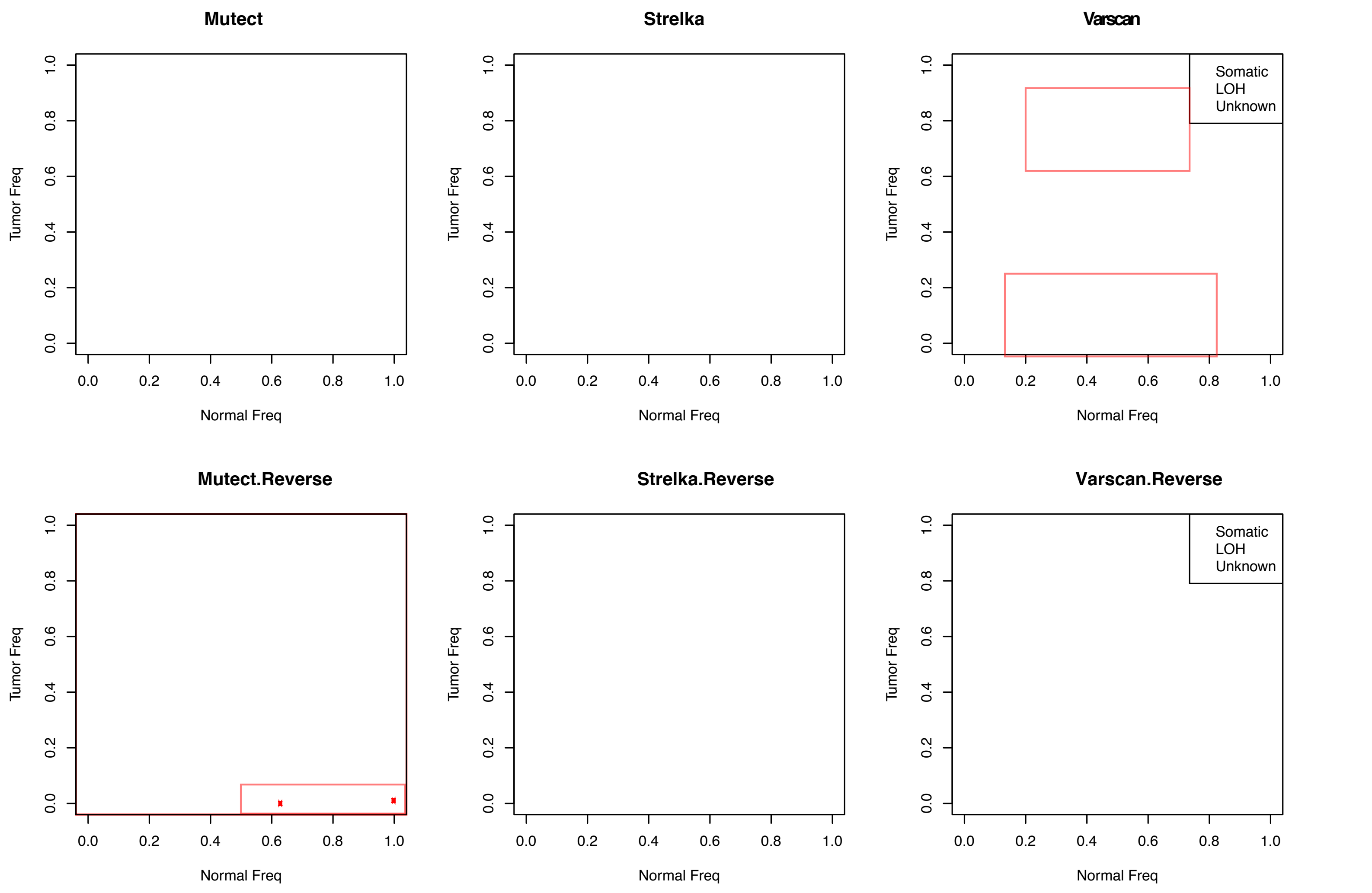
Correlation of read counts on four callers

X-axis: 211 variants;
Y-axis: Fisher Exact Test p-value between each two methods,
use read counts of normal reference, normal variant, tumor reference, and tumor variant.
G: GATK,
M: Mutect, MR: Mutect Reverse Normal-Tumor,
S: Strelka, SR: Strelka Reverse Normal-Tumor,
V: Varscan, VR: Varscan Reverse Normal-Tumor.

Result



Results combinding six methods, (p-value <= 1e-5)



Results of each method, (p-value <= 1e-5)

Mutation Counts of each method, (p-value <= 1e-5)

Mutect	Mutect Reverse	Strelka	Strelka Reverse	Varscan	Varscan Reverse	Input for GATK	P-value (<=1e-5)
469	397	315	297	4062	3512	6636	1177

Conclusion

By considering existing problems from SNP calling, we developed a strategy of combinding three methods with normal-tumor as input, three addtional methods with tumor-normal as input. After combining all results together, we used GATK unifiedGenotyper to call four counts on these lists: reference normal, variant normal, reference tumor, and variant normal. Then we applied fisher exact test on these number and filter in sites with p-value <=1e-5.

Our approach is superior to those that simply combines multiple variants lists by majority vote, as it is built on the actual underlying allele count differences. Its sensitivity is directly linked to allele frequency alterations and the sequencing depth and averted the problem that different methods have distinct sensitivities and sometimes incomplete consideration of possible mutation types.

Reference

McKenna, Aaron, et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research 20.9 (2010): 1297-1303.
Saunders, Christopher T., et al. "Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs." Bioinformatics 28.14 (2012): 1811-1817.
Koboldt, Daniel C., et al. "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing." Genome research 22.3 (2012): 568-576.
Cibulskis, Kristian, et al. "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples." Nature biotechnology 31.3 (2013): 213-219.

WeChat

Website

Contact

Email: Yu Wang (yulywang@umich.edu)
Jun Z Li (junzli@med.umich.edu)

Website: www.cancergenome.us
Twitter@Wang_yu_