Consensus Genotyper for Exome Sequencing: Improving the Quality of Exome Variant Genotypes

Vassily Trubetskoy[1], Ravi Madduri[2], Alex Rodriguez[2], Jeremiah Scharf[3], Paul Dave[2], Ian Foster[2] , Nancy Cox[1] , Lea Davis[1]

1) Section Genetic Medicine, University of Chicago, Chicago, IL; 2) Computation Institute, University of Chicago, Chicago, IL; 3) Department of Neurology, Massachusetts General Hospital, Boston, MA

THE UNIVERSITY OF CHICAGO

Abstract

With the rise of next generation sequencing methods, there are now a variety of tools available to researchers for the detection and genotyping of sequence variants. However, the concordance among variant sets between these disparate approaches has been shown to be poor. Recently, researchers in the machine learning community have shown that combining the output of multiple models can dramatically improve performance of a classifier. Collectively, these techniques are referred to as ensemble methods. Here we describe a variant calling approach based on an ensemble of variant calling algorithms which we call Consensus Genotyper for Exome Sequencing (CGES). Our method employs a two-stage voting scheme among a set of three algorithm implementations, GATK2.0, Freebayes, and Atlas2.0, which were used to identify variant sites and determine genotypes in the study. While the ensemble method is can accept variants from any variant calling algorithm, these were chosen for their widespread adoption and diverse strategies. We apply CGES to a dataset consisting of 123 samples sequenced at the Center for Inherited Disease Research (CIDR) using the Agilent SureSelect3 Exome Capture and Illumina sequencing technology. Samples were drawn from extended pedigrees, allowing us to compare individual and family based quality metrics across all algorithms. The CGES approach is shown to outperform its constituent parts in many key quality metrics without a significant loss in the number of variant sites called. In particular we are able to achieve a threefold reduction in Mendelian inconsistencies between the best performing variant caller and our consensus approach (CGES = 240.14/trio and Atlas2.0 = 699.59/trio). For callers with comparable QUAL scores, our CGES set of variants has an average QUAL score 11% (GATK) and 70% (Freebayes) higher than the unfiltered output set of each respective variant caller. Additionally, the consensus set outperforms all individual callers in the study with regard to expected exome-wide transition-transversion ratio(CGES = 3.07 and Atlas2.0 = 2.98). For the purpose of accessible, efficient, and reproducible analysis, we provide implementation of CGES as a stand alone command line tool, as well as a set of parallel Galaxy tools and workflows for accessible and efficient use by the research community (see Implementing a High Performance, Reusable Consensus Calling Pipeline for Next Generation Sequencing using Globus Genomics, Madduri et al., ASHG 2013).

Introduction

Our goal is to identify high confidence variants in Exome sequence data by finding variation for which a collection of calling approaches agree. The Consensus Genotyper for Exome Sequencing (CGES) method employs a two stage voting scheme among variant call sets: site level consensus, and genotype level consensus. A high level schematic of the work flow is shown in Figure 1.

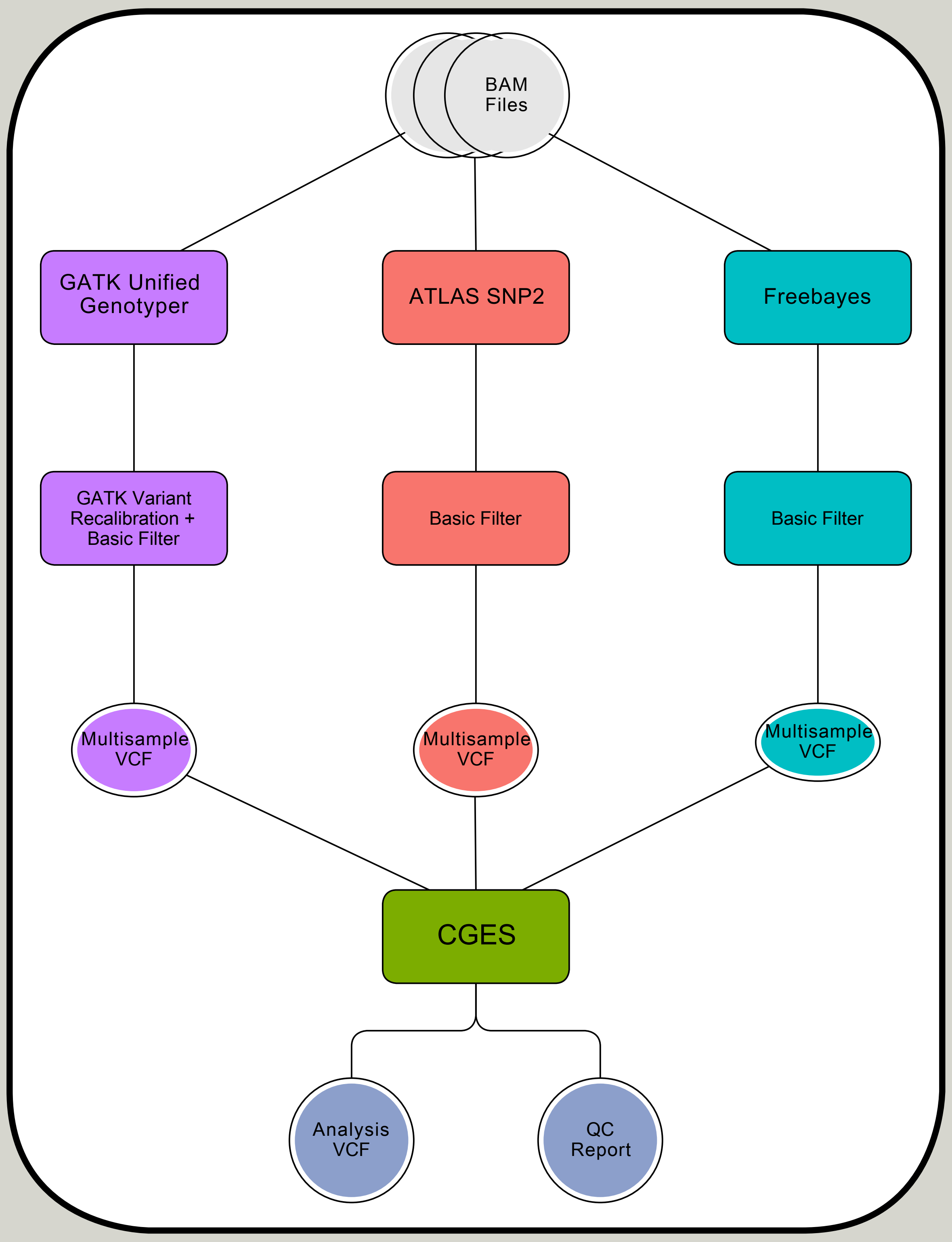

Figure 1 : *Schematic showing the high level work flow for identifying consensus genotypes. Our implementation leverages three callers and some basic quality filtering. The final CGES stage produces a high quality variant call set and a quality report.*

Method Details

Given an initial set of alignments (BAM files), variants are called using multiple algorithms. Here, three calling algorithms used:

- ATLAS-SNP2[1]
- GATKv2.0 Unified Genotyper[2]
- Freebayes[3]

After initial variant calling, individual branch sets are sent through basic quality filters. These filters restrict variant sites to on-target regions defined by the capture product, and a minimum QUAL score of 10. For the GATK's Unified Genotyper, we apply the GATK's recommended best practices by recalibrating variants using dbSNP135, Hapmap3, and 1000 Genomes data sets. The final set of GATK variants are filtered to the %99.0 quality tranche. The resulting variant sets (VCF files) are passed to CGES, which proceeds in nested stages. First, we identify all variation common to the three call sets. Next, for each site in this consensus set, we identify individual genotypes which match between all callers. Missing genotypes are ignored unless called missing in all branches. If a genotype is discordant among all branches, it is flagged and reported as missing.

Application to Exome Sequencing: Tourette/OCD Study

Data description:

- 123 samples sequenced using the Agilent SureSelect3 Exome Capture and Illumina sequencing technology at the Center for Inherited Disease Research (CIDR).
- Study design included extended pedigrees, with 22 trios in the final set.
- Sequenced to 80x average coverage.
- Alignment and alignment calibration was performed by CIDR using BWA, Picard and Samtools.

Variant Site Concordance

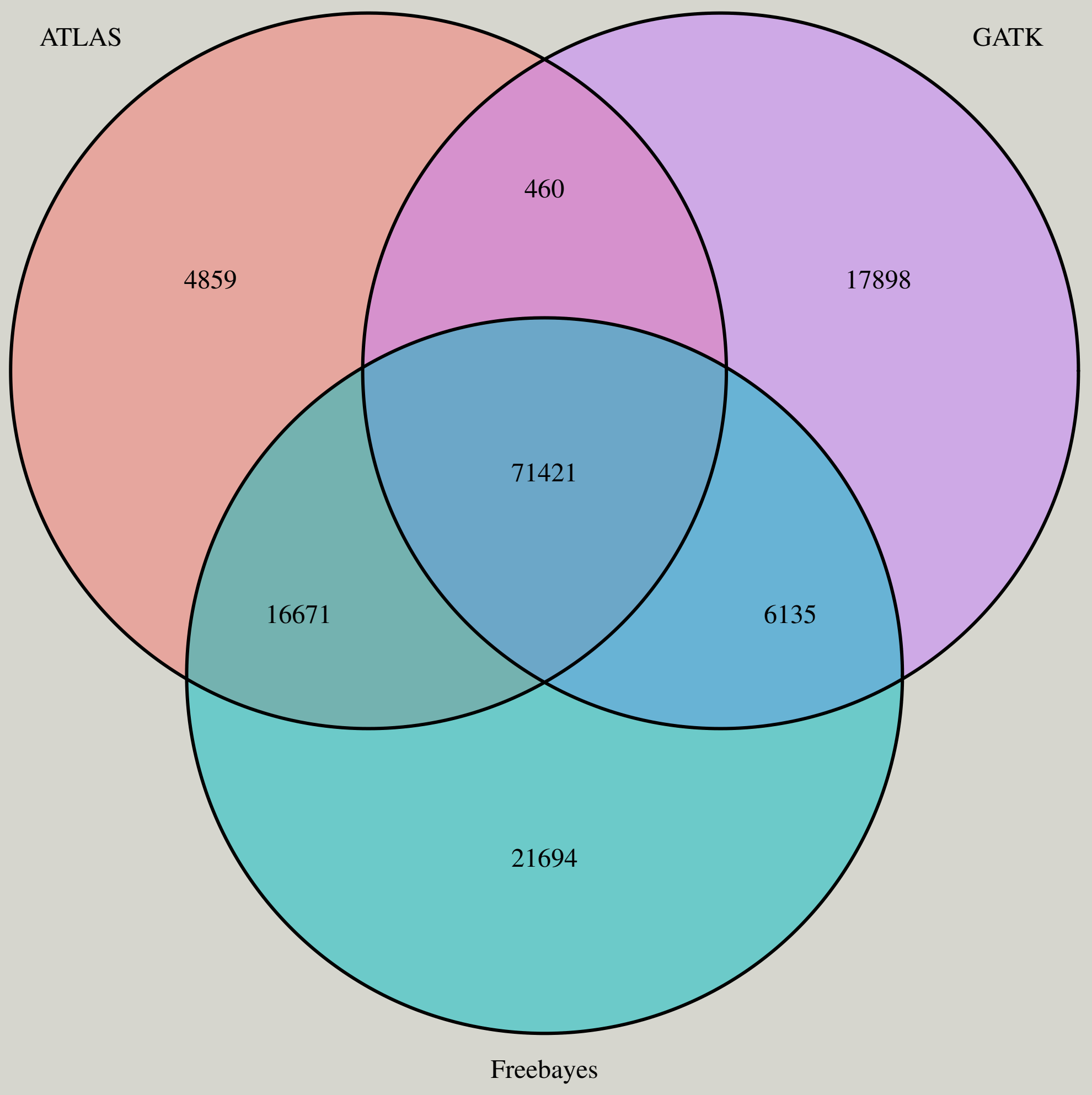

Figure 2 : *Venn diagram showing overlapping variant sites among callers. CGES results fall at the intersection of all sets. Constituent callers produced largely concordant variant sites.*

Mendelian Inconsistencies

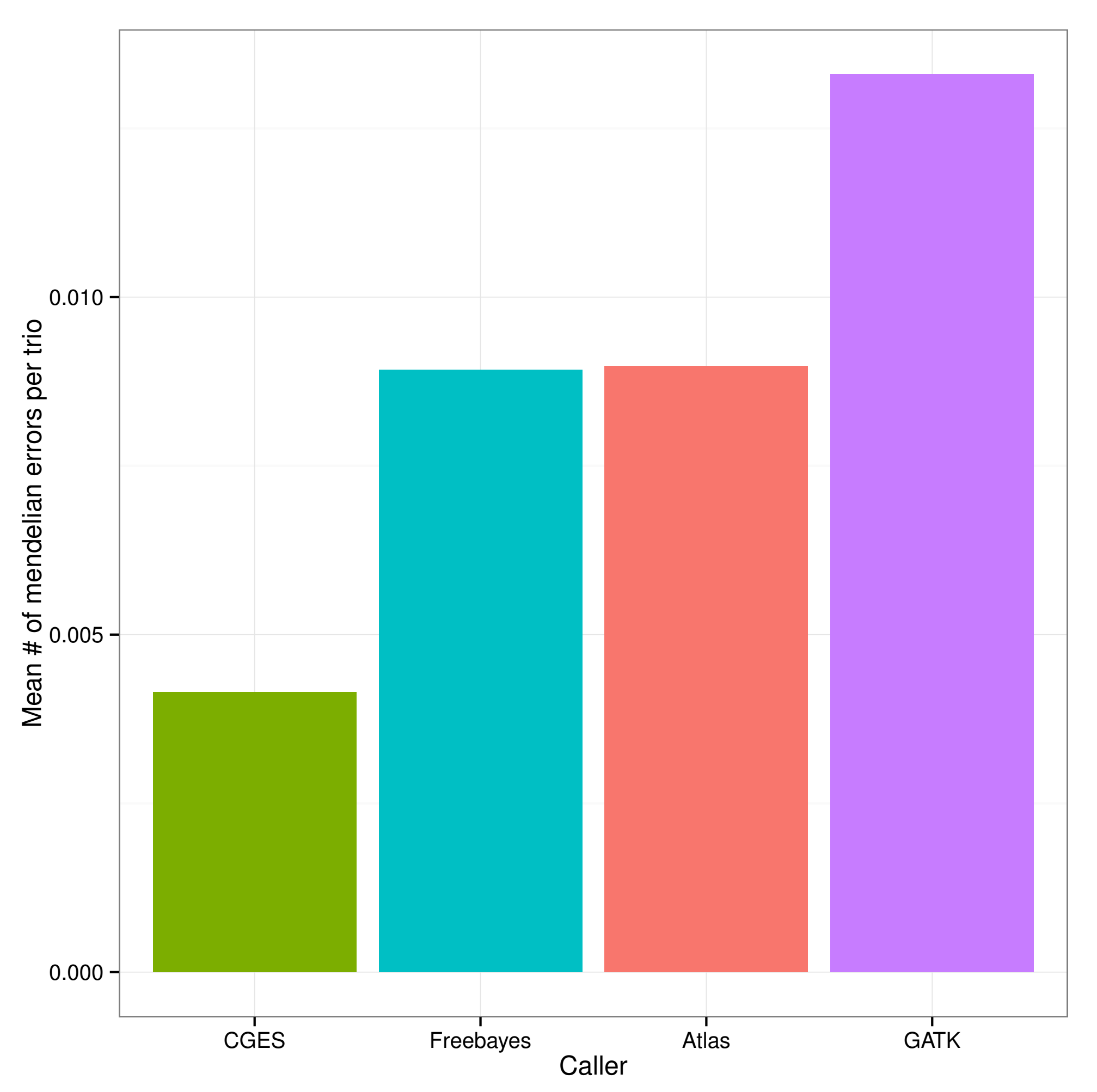

Figure 3 : *Rates of Mendelian inconsistencies for the variant sets. Inconsistency counts were generated per trio using PLINK[4], and the result is shown as a proportion of total genotypes in each set.*

Transition-Transversion Ratio

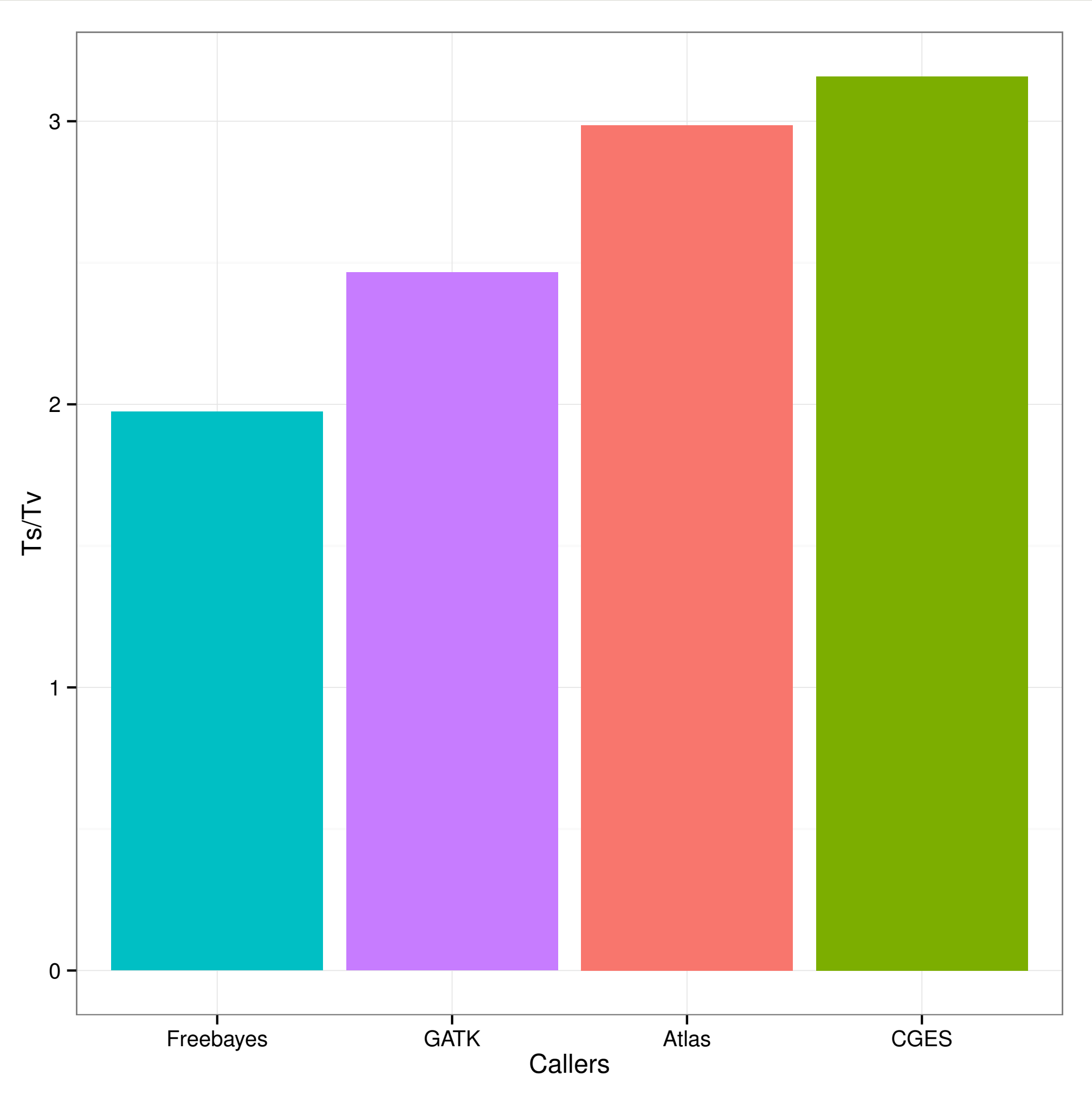

Figure 4 : *Transition/transversion mutation ratio for each variant call set. Protein coding regions have observed ratios in the range 3.0-3.5[5]. Variant sets with lower ratios are interpreted as having a higher rate of false positives.*

Variant Rediscovery

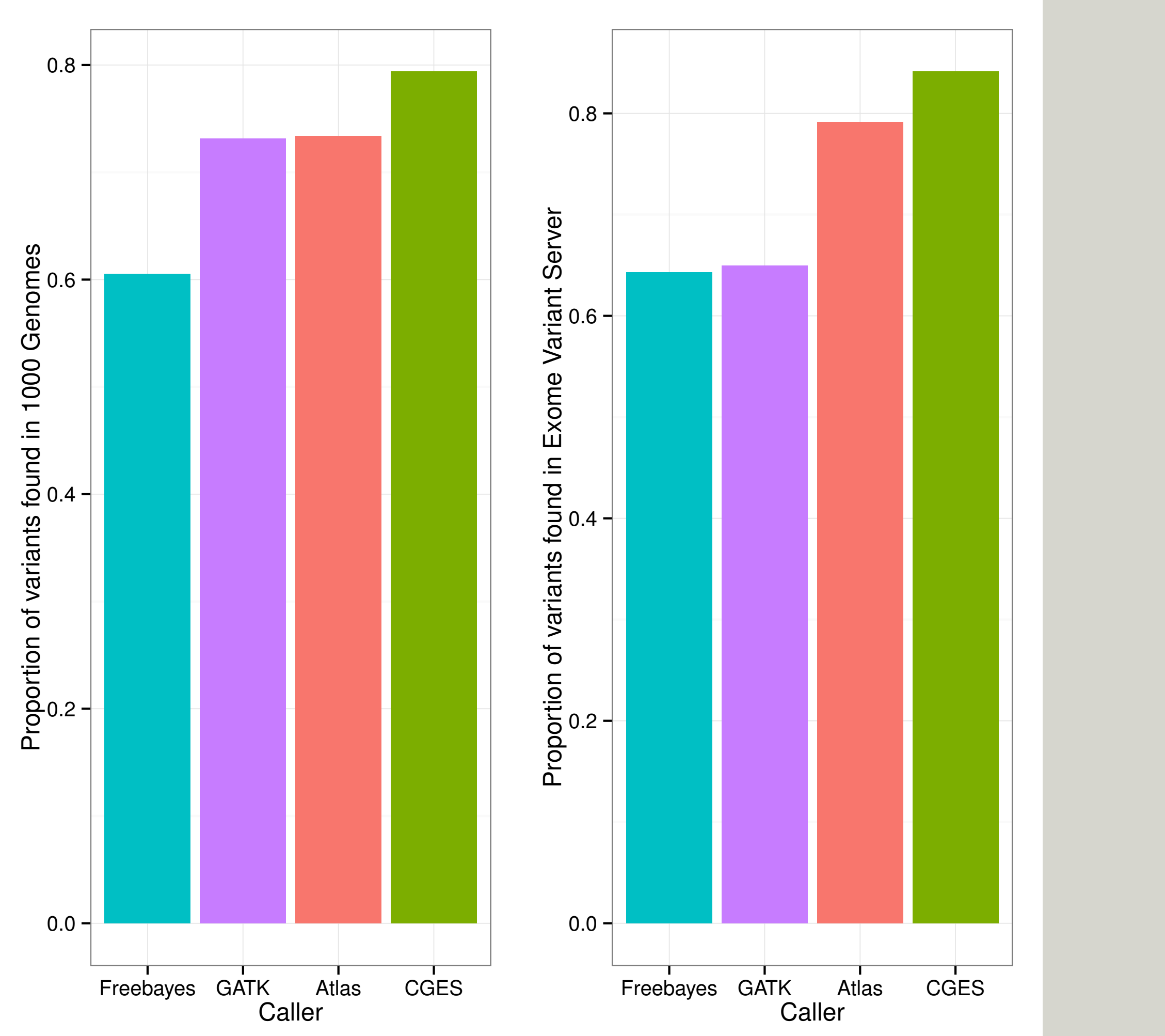

Figure 5 : Rates of variant rediscovery using 1000 Genomes[6] and Exome Variant Server[7] data sets as references. 1000 Genomes established best practices for sequence data, and the Exomes Variant Server provides a high quality publicly available Exome sequencing data set. CGES produces the highest ratio of previously-observed-to-novel variant sites.

Results & Discussion

Each caller produced largely concordant variant sites, as show in the Venn diagram in Figure 2. Importantly, ATLAS produced the smallest amount of exclusive variation, and performed best among constituent callers in all quality metrics presented here.

Improvements in quality:

- Reduced rate of Mendelian inconsistencies, as shown in Figure 3.
- Higher Transition-Transversion ratio, as shown in Figure 4.
- Higher rates of previously observed variants, as shown in Figure 5.

Trade-offs include:

- High computational cost, with many passes over initial alignments.
- Reduced rate of false positives at the expense of higher rate of false negatives.

In practice these trade offs have real impact: the algorithm is intractable given a naive serial implementation for all interesting data sets, and CGES cannot improve the rate of false positives. The latter is important, as this algorithm is an informed reduction of existing call sets and cannot identify variation outside of its input. Given this limitation, CGES makes most sense for study designs which are highly sensitive to false positives. More generally, CGES is constrained by the quality of input data. This means that each branch should be tuned for a specific data set in order to obtain the highest possible quality consensus results. Finally, access to distributed computing resources is important for studies using sequencing technology or large sample sizes.

Conclusion

The CGES method performs an informed reduction of variant call sets which produces a high quality subset of variants.This set of variants improves key quality metrics such as: rate of Mendelian inconsistencies, Exonic Ts/Tv ratio, and rate of previously observed variants. This improvement in quality comes from a reduction of false positives while not improving the false negative rate.

While this improvement in quality comes with additional computational burden, our implementation using Globus Genomics (citation) allows us to efficiently scale our analysis by leveraging Amazon's Elastic Compute Cloud. Currently, Galaxy tools and work flows are tied to our galaxy instance at `http://cox.galaxycloud.org/`. Access can be arranged, please contact me at: trubetskoy@uchicago.edu. Additionally, CGES is freely available as a command line tool at: `https://github.com/vtrubets/galaxy.consensus`.

References

[1] Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, *et al.*. A snp discovery method to assess variant allele probability from next-generation resequencing data. *Genome research* 2010; 20(2):273–280.

[2] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.*. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research* 2010; 20(9):1297–1303.

[3] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* 2012; .

[4] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, *et al.*. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 2007; 81(3):559–575.

[5] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, *et al.*. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics* 2011; 43(5):491–498.

[6] Siva N. 1000 genomes project. *Nature biotechnology* 2008; 26(3):256–256.

[7] Server EV. Nhlbi go exome sequencing project (esp). *Seattle, WA. http://evs. gs. washington. edu/EVS/. Accessed July* 2013; .