Exploring Human Exome Copy Number Variation

Yuchen Gao, Blaine Rothrock, Ryan Konz
School of Informatics and Computing, Indiana University

Abstract

Copy number variation (CNV) is one form of DNA structural variation that can account for differences among humans. CNVs have been implicated in physical appearance, disease, and drug susceptibility. Our study generated a list of regions of DNA that have high copy number variation within the Great Britain population. We used genomic data from the Thousand Genomes Project and generated virtual machines (VMs) through the FutureGrid project to process the data in parallel.

Fig. 1: FutureGrid iDataPlex

Purpose

We are exploring new approaches of human genome analysis through the large scale use of data, the restriction of CNV analysis to exonic regions, and the utilization of cloud computing to reduce time costs. Using these methods, we are discovering genes of high copy number variation within the Great Britain population.

Methods

Thousand Genomes Project

- Large-scale genome sequencing collaboration
- Collection of genetic data of 27 ethnic populations
- Publicly available data for 1000s of individuals

ExomeCNV

- Calculates copy number using raw sequencing data from TGP
- Limits analysis to coding regions of DNA
- Exons account for 1% of genome, but 85% of disease-causing mutations

FutureGrid

- India test-bed hosted by IU
- Creation of numerous virtual machines running custom Ubuntu image
- Parallel processing of data via cloud computing

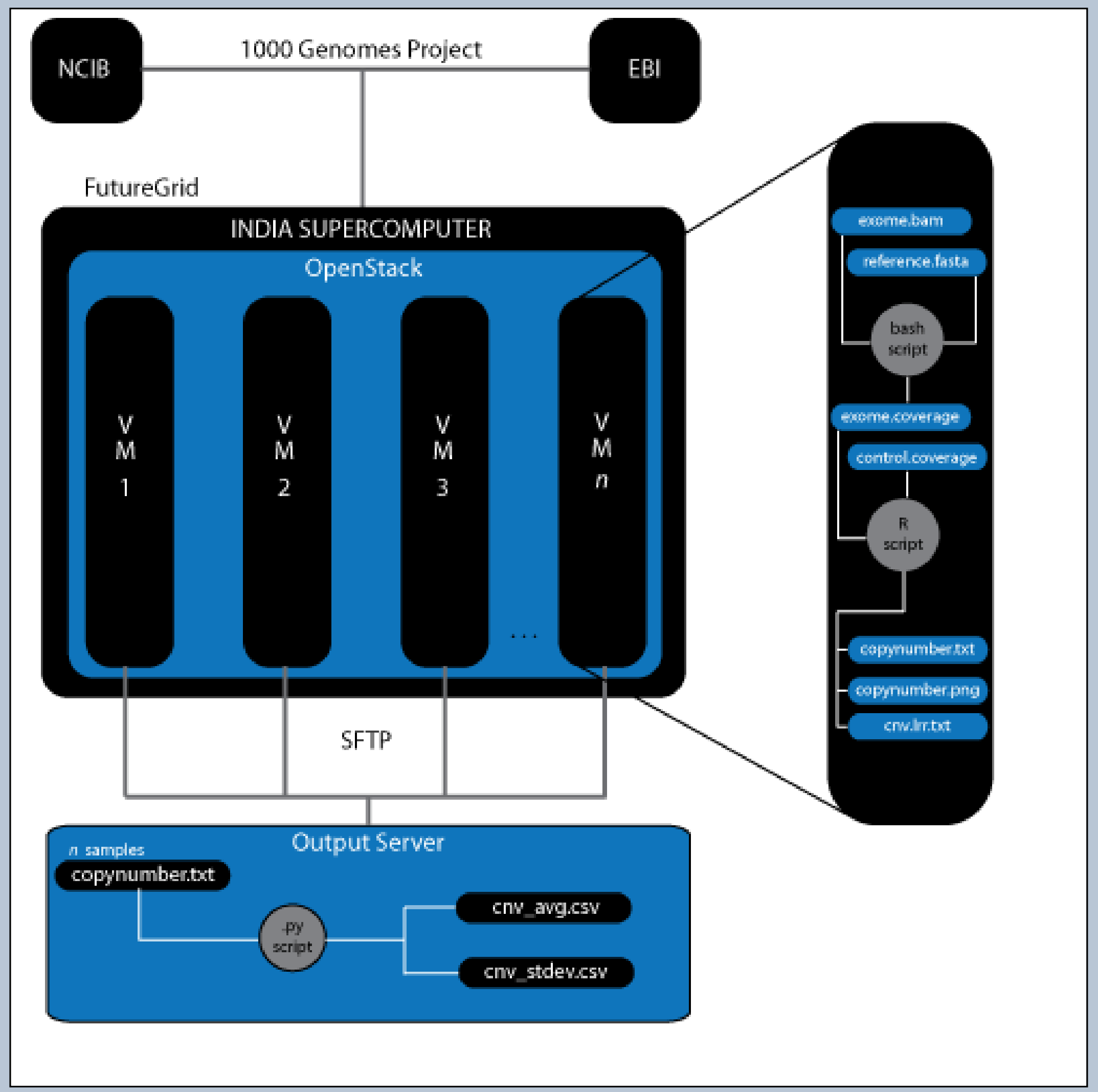

Fig. 2: Workflow of exome CNV analysis

Results

Visual representation of the copy number variation in chromosome 20 of one individual

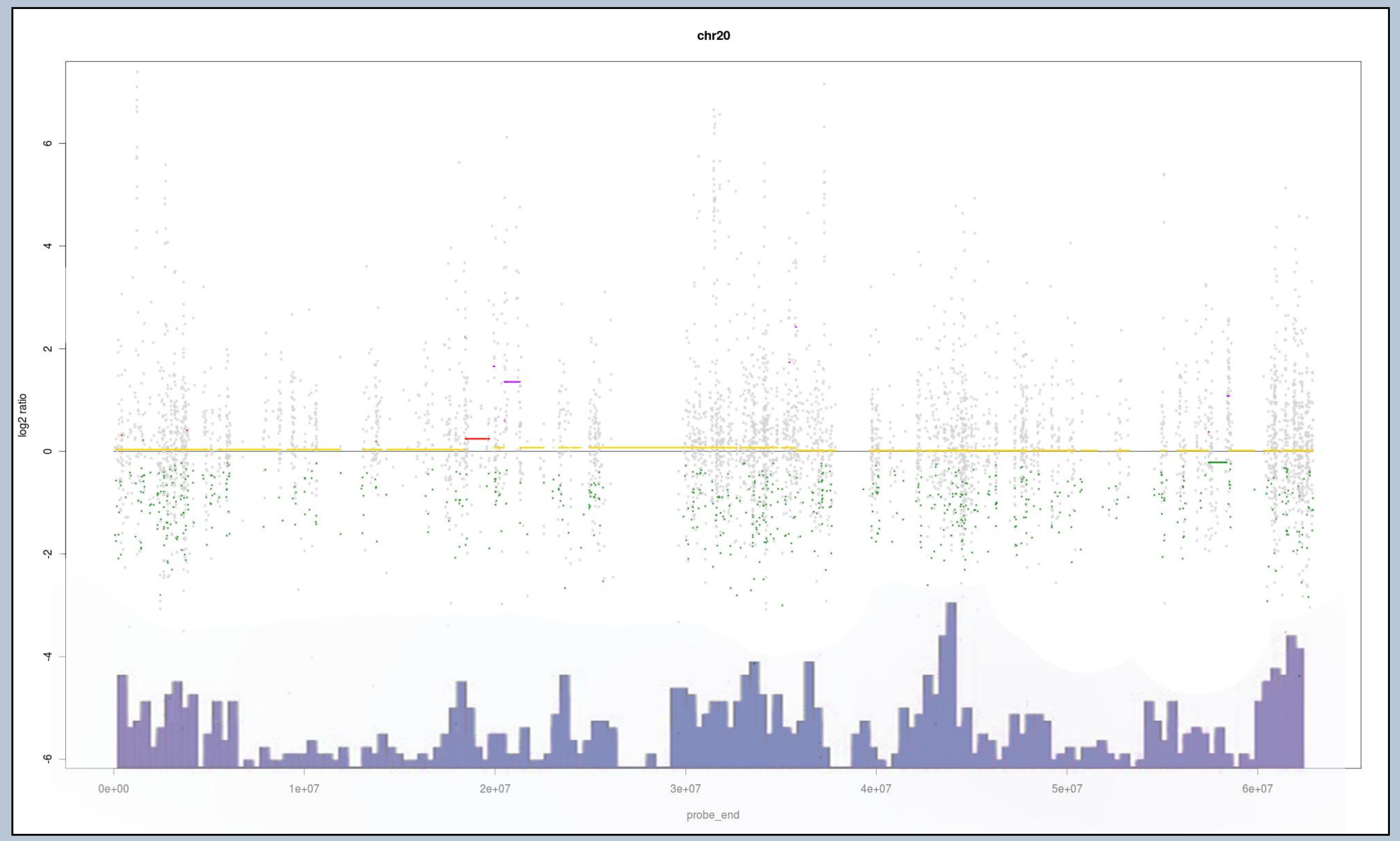

Fig. 3: Copy number in chromosome 20 of one sample

Genes found to have highest copy number and copy number variation using chromosome 20 of 50 samples of the Great Britain population.

Chromosome region	Mean CN	Gene
58444852 - 58497514	3.8	SYCP2
13694969 - 13695810	3.74	ESF1
10618331 - 10620603	3.68	JAG1
31573541 - 31592239	3.68	SUN5
54944444 - 54945396	3.66	AURKA

Table 1: Chromosome 20 regions with highest average copy number

Chromosome region	CN StDev	Gene
7961715 - 7963268	0.1829	TMX4
56072223 - 56073757	0.1815	CTCFL
57251242 - 57254580	0.1814	STX16
37128089 - 37128276	0.1801	RALGAPB
42355055 - 42355642	0.1778	GTSF1L

Table 2: Chromosome 20 regions with highest copy number standard deviation

Results

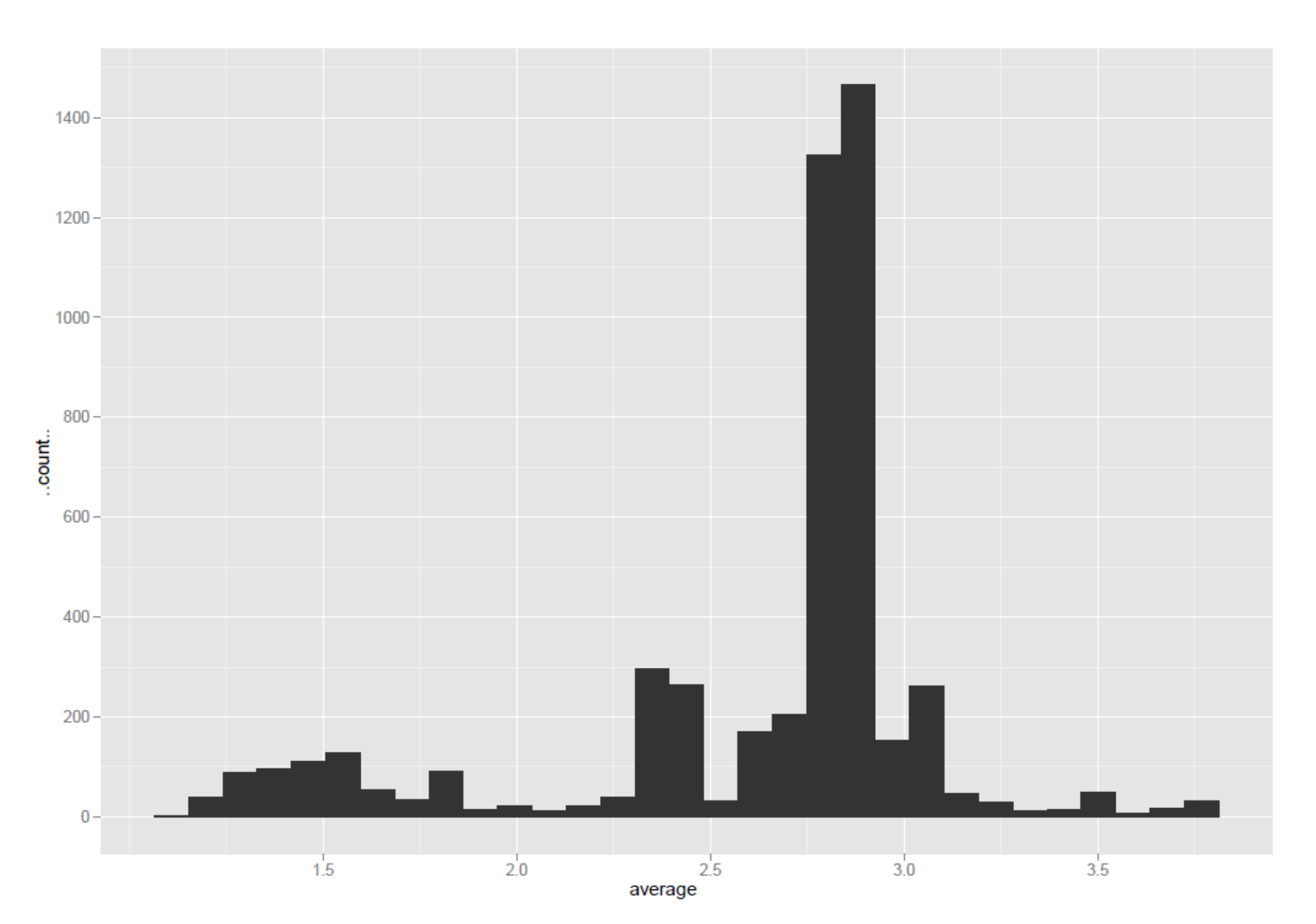

Fig. 4: Frequency of mean CNs on chromosome 20

Discussion/Conclusion

- We were able to successfully implement our workflow for exome analysis
- Instability of the FutureGrid system limited our analysis to chromosome 20
- ExomeCNV is highly-dependent on coverage and may not be suited for large-scale analysis
- Our experimental results provide interesting information but require further refinement of data

References

[1] The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467(7319), 1061-1073.*.

[2] Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., et al. (2011). Exome Sequencing-Based Copy-Number Variation and Loss of Heterozygosity Detection: ExomeCNV. *Bioinformatics.*

Acknowledgements

We'd like to thank our mentors Andrew Younge and Yicheng Feng for their advice on our project.