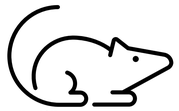
Comparison of Genotype Phasing Algorithms

Christa Lehr, Vinaydeep Singh, Sahar Al Seesi
Computer Science Department
Southern Connecticut State University

Introduction

The human genome is a diploid genome, which means there are two copies of each chromosome, except for the X and Y chromosomes. There are two copies (alleles) of each of the genes laying on these chromosomes. Allele Specific Expression Estimation (ASE) is the problem of estimating the expression level of each gene at the allele level, which means finding whether one of the two alleles or both alleles of a gene are actively being transcribed into RNA to generate proteins and whether they are expressed at the same or different levels. Transcriptome sequencing is used to address this problem. The analysis starts by comparing the sequencing data to a reference genome that represents all individuals in a species but is does not completely match given the variations between individuals. The reference genome is haploid, including the sequence of one copy of each chromosome. This poses challenges in addressing ASE, where we are interested in identifying difference of expression between the gene alleles coming from the diploid genome. One way to address this issue is through creating a diploid reference from the individual being studied, through first finding where and how this individual genome vary from the reference genome. Then phasing these variations arranges the alleles at different positions into two groups to allow us to create the diploid reference of the individual. In this research project, we are comparing phasing algorithms. We will report on execution requirements and accuracy of results of the multiple phasing algorithms.

Methodology



How to create and phase your own synthetic hybrid mouse in one semester or Less!

What is Phasing?

```
   1   2
AATCGTCCTG
   T   C
   A   G

AATCGTCCTG
AAACGTGCTG   Solution 1: T/A, C/G are pairs, but TC and AG are on the same allele.

AATCGTGCTG
AAACGTCCTG   Solution 2: T/A, G/C are pairs, but TG and AC are on the same allele.
```

CompSci
Southern Connecticut State University

Results

Chromosome	WhatsHap		HapCut2	
	Switch Error Rate	Total Heterozygous Percentage Phased (All Chromosomes)	Switch Error Rate	Total Heterozygous Percentage Phased (All Chromosomes)
Chr1	0.45%		0.37%	
Chr2	0.48%		0.43%	
Chr3	0.53%	91.27%	0.42%	93.51%
Chr4	0.50%	**Time for Program to Run**	0.41%	**Time for Program to Run**
Chr5	0.43%		0.36%	
Chr6	0.56%	2 hour 6 minutes	0.47%	40 minutes
Chr7	0.52%		0.47%	
Chr8	0.43%		0.37%	
Chr9	0.57%		0.50%	
Chr10	0.40%		0.34%	
Chr11	0.45%		0.39%	
Chr12	1.11%		0.97%	
Chr13	0.56%		0.46%	
Chr14	0.52%		0.39%	
Chr15	0.41%		0.33%	
Chr16	0.51%		0.43%	
Chr17	0.62%		0.54%	
Chr18	0.36%		0.30%	
Chr19	0.42%		0.36%	
ChrX	0.55%		0.36%	
ChrY	7.90%		8.66%	

Conclusion

- Two algorithms compared: WhatsHap and HapCut2
- HapCut2 outperforms WhatsHap
  - Lower error rates
  - Lower run times
  - Higher percentage of phased genotypes

Future Work

- Add HaploMaker to comparison
- Analyze WhataHap, HapCut2, and HaploMaker for their indels phasing

References

[1]T. M. Keane, L. Goodstadt, N. A. Furlotte, E. Eskin, C. Nellaker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edward, T. Grant Belgard, P. L. Oliver, P. Danecekl, R. E. Mcintyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. V. Der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, M. A. White, R. Durbin, I. J. Jackson, A. Czechanski, J. Afonso Guerra-Assuncao, L. Rae Don-Ahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Bir-Ney, J. Flint, K. Wong, D. J. Adams, B. Yalcin, A. Heger,A. Agam, G. Slater, and M. Goodson, "Mouse genomic variation and its effect on phenotypes and gene regulation," Nature (London), vol. 477, no. 7364, pp. 289–294, 2011
[2]P. Edge, V. Bafna, and V. Bansal, "Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies," Genome research, vol. 27, no. 5, pp. 801–812, 2017.
[3]M. Martin, M. Patterson, S. Garg, S. O. Fischer, N. Pisanti, G. W. Klau, A. Schöenhuth, and T. Marschall, "Whatshap: fast and accurate read-based phasing," bioRxiv, 2016. [Online]. Available: https://www.biorxiv.org/content/early/2016/11/14/085050

Southern Connecticut State University ERSP
Early Research Scholars Program

This research was done under the Early Research Scholars Program