Variant Phasing with 10X Genomics

http://software.10xgenomics.com/genome-exome/downloads/latest

Patrick Marks, William H Heaton, Michail Schnall-Levin

July 11, 2016

1 Introduction

We take as input a pre-determined set of balletic variants. We label alleles $A_{i,p}$ where $i \in 1, ..., N$ indexes the variant, and $p \in 0, 1$ is an arbitrary label for the two alleles of the variant.

The set of alleles that come from the same parent chromosome is referred to as a haplotype, and are arbitrarily labeled $H_0 and H_1$. The goal of the phasing algorithm is to determine which allele from each variant came from each parent chromosome. The phasing result can be described by a ternary variable for each variant $X_i \in 0, 1, 2$ where $X_i = 0$ indicates the $A_{i,0} \in H_0$ and $A_{i,1} \in H_1$ and $X_i = 1$ indicates that $A_{i,0} \in H_1$ and $A_{i,1} \in H_0$ and $X_i = 2$ indicates that either $A_{i,0} \in H_0, H_1$ or $A_{i,1} \in H_0, H_1$ thus the variant is homozygous and the other allele does not exist. It is by this third label that we filter variants that are false positive and correct the genotype of homozygous variants that have falsely been called heterozygous.

Neighboring variants on the genome are often separated by distances longer than the read-pair length, causing very short phase blocks. Long input fragments covering a small fraction (0.001-0.0001) of the genome are exposed to each barcode, so the probability that a barcode contains a molecule from both haplotypes is very small[5].

We case the solution to the phasing problem as a search for the maximum likelihood phasing vector:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{X}}{\operatorname{argmax}} P(\boldsymbol{O}|\boldsymbol{X})$$

where $\boldsymbol{O}$ denotes the sets of barcoded reads observed, and $\boldsymbol{X}$ is the phasing result we wish to infer.

Read pairs are aligned to the genome as usual. Reads are grouped by the attached barcode sequences. Reads with a common barcode are partitioned into groups that are likely to have originated from a single genomic input fragment, and thus provide evidence that the alleles covered by the reads came from the same haplotype.

We compute the probability of the observed reads covering variant i from fragment f as:

$$\log P(O_{i,f}|A_{i,p}) = \sum_{r \in O_{i,f}} 1(X_i \neq 2)(1(S_r = A_{i,p})(1{-}10^{-Q_r/10}){+}1(S_r \neq A_{i,p})(10^{-Q_r/10})){+}1(X_i = 2)(0.5)$$

where r sums over reads, $1(S_r = A_{i,p})$ is the indicator function testing if the rth sequence S_r match allele $A_{i,p}$ and $1(X_i = 2)$ is the indicator function testing if the assignment of the ith variant is 2, also known as “not heterozygous”. The probability assigned is derived from the inverse-Phred transformed quality value of relevant read base Q_r.

The data from a fragment f come from one of three cases. First two cases are that variants not assigned $X_i = 2$ have the property that alleles present are only from H_0 or only from H_1. These cases are the typical case and have a high prior probability, governed by the fraction of the genome present in each partition. The third case is that multiple input DNA molecules covering the same locus from both haplotypes were present, so either allele is equally likely to be observed:

$$P(O_{i,f}...O_{N,f}|\boldsymbol{X}, H_f = 0) = \prod_i 1(X_i = 2)(0.5) + 1(X_i \neq 2)P(O_{i,f}|A_{i,X_i})$$

$$P(O_{i,f}...O_{N,f}|\boldsymbol{X}, H_f = 1) = \prod_i 1(X_i = 2)(0.5) + 1(X_i \neq 2)P(O_{i,f}|A_{i,1-X_i})$$

$$P(O_{i,f}...O_{N,f}|\boldsymbol{X}, H_f = M) = \prod_i 0.5$$

These equations give the probability of the observed reads from fragment f at variant location i, X_i, and fragment haplotype H_f. Observations are independent given the variant parity and fragment haplotype. The prior probability of the third case is α - the probability that a partition contains input DNA molecules from both haplotypes at a locus. We can compute the overall likelihood by summing over the three cases:

$$P(O_{i,f}|X_i) = (1-\alpha)(P(O_{i,f}...O_{N,f}|\boldsymbol{X}, H_f = 0)+P(O_{i,f}...O_{N,f}|\boldsymbol{X}, H_f = 1))+\alpha P(O_{i,f}...O_{N,f}|\boldsymbol{X}, H_f = M)$$

Fragments are independent given the variant parity X_i letting us form the overall objective function as:

$$P(\boldsymbol{O}|\boldsymbol{X}) = \prod_f P(O_{i,f}, ..., O_{N,f}|\boldsymbol{X})$$

2 Optimization

We optimize the overall objective function using a hierarchical search of the phasing vector $\boldsymbol{X}$.

Initially we break up $\boldsymbol{X}$ into local chunks of $n = 40$ variants and determine the relative phasing of the block using beam search of the assignments of $X_k, X_{k+1}, ..., X_{k+n}$. Where k is the first variant in the local block. Beam search is a standard method that has existed for a long time (see http://en.wikipedia.org/wiki/Beam_search).

The relative phasing of neighboring blocks is found greedily, yielding a candidate phasing vector $\boldsymbol{X}$. Finally $\boldsymbol{X}$ is iteratively refined by swapping the phase of individual variants. When refinement converges, we are left with our estimate of the optimal phasing configuration $\hat{\boldsymbol{X}}$.

3 QV Testing

W can compute estimates of the accuracy of the phasing configuration by computing the likelihood ratio between the optimal configuration $\hat{\boldsymbol{X}}$ and some alternative configuration $\boldsymbol{X_{alt}}$. The confidence is then reported as a Phred-scaled quality value:

$$Q(\boldsymbol{X_{alt}}) = -10\log_{10}(\frac{P(\boldsymbol{O}|\boldsymbol{X_{alt}})}{P(\boldsymbol{O}|\hat{\boldsymbol{X}})})$$

where X_{alt} here is the sum of the alternative phasing and the alternative option of the variant being not heterozygous (label 2).

There are two classes of errors we consider: short switch errors and long switch errors. Short switch errors are single variants that are assigned the wrong phasing in an otherwise correctly phased region - to measure the short switch confidence of variant i we flip X_i and add $X_i = 2$ to form X_{alt}. If X_i is already label 2, we consider it versus the summation of $X_i = 0 + X_i = 1$ as X_{alt}. If the short switch confidence is low, the variant is marked as not phased in the output. If the short switch confidence is high for label 0 or 1, it is marked as phased. And if the short switch confidence is high for label 2, we attempt do decide whether the variant is more likely homozygous reference (false positive) in which case it is filtered, or homozygous alt allele in which case the genotype is corrected.

Long switch errors occur when two neighboring blocks of variants $..., X_{i-2}, X_{i-1}$ and $X_i, X_{i+1}, ...$ are correctly phased internally, but have the wrong relative phasing between the two blocks. In this case we say a long switch error occurred at position i. We test this long switch confidence at position i by inverting the phase of X_j for all $i \leq j$ if $X_j \neq 2$. When the long switch confidence falls below a threshold we start a new phase block – variants in different phase blocks are not called as phased with respect to one another[1].

4 Related Work

HAPCUT [cite], and HASH [cite] are two relevant pieces of prior art. In HASH, the authors formulate a probabilistic objective similar to ours. Their algorithm uses single long Sanger reads as the input fragments so they don't need to consider the case that a fragment carries data from both haplotypes.

HASH uses Markov-Chain Monte Carlo (MCMC) to explore the posterior distribution of the phasing configuration, as opposed to the direct combinatorial optimization scheme we use. HASH uses a graph partitioning scheme to select sets of variants that are well-connected by by fragment data, and add MCMC moves that invert the phasing of the groups, allowing the Markov chain to converge more quickly. Either optimization scheme should find a near-optimal solution given enough run time. The MCMC approach is probably less efficient because it randomly explores[3].

HAPCUT is a follow-up method by the same group which uses combinatorial algorithms to optimize the minimum error correction (MEC) objective. It has very little in common with our method. It appears that this method was developed as a much faster alternative to the HASH method[2].

As for variant filtering via phasing inconsistencies, complete genomics long fragment read technology utilized a system by which they filtered all variants that had information by which they could phase but were not confidently phased. The paper is somewhat unclear on the method, but this is the best interpretation of these authors. Their method of phasing was largely undocumented in the paper. This strategy is likely to filter many true variants as well as false variants. And it is not addressing the case of homozygous alt variants that were falsely called heterozygous. It will filter these variants instead of updating their genotypes[4].

References

[1] Sasan Amini, Dmitry Pushkarev, Lena Christiansen, Emrah Kostem, Tom Royce, Casey Turk, Natasha Pignatelli, Andrew Adey, Jacob O Kitzman, Kandaswamy Vijayan, et al., *Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing*, Nature genetics **46** (2014), no. 12, 1343–1349.

[2] Vikas Bansal and Vineet Bafna, *Hapcut: an efficient and accurate algorithm for the haplotype assembly problem*, Bioinformatics **24** (2008), no. 16, i153–i159.

[3] Vikas Bansal, Aaron L Halpern, Nelson Axelrod, and Vineet Bafna, *An mcmc algorithm for haplotype assembly from whole-genome sequence data*, Genome research **18** (2008), no. 8, 1336–1346.

[4] Brock A Peters, Bahram G Kermani, Andrew B Sparks, Oleg Alferov, Peter Hong, Andrei Alexeev, Yuan Jiang, Fredrik Dahl, Y Tom Tang, Juergen Haas, et al., *Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells*, Nature **487** (2012), no. 7406, 190–195.

[5] Grace XY Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, et al., *Haplotyping germline and cancer genomes with high-throughput linked-read sequencing*, Nature biotechnology (2016).