# CSE 280A Project 9 Report

Sharad Venkateswaran Yein Kim Ross DeVito

March 17, 2021

## 1 Project Overview

The haplotype assembly problem is the task of determining phased haplotypes given a reference genome and several overlapping reads from an individual. Single-molecule sequencing technologies, such as Pacifica Biosciences and Oxford Nanopore Technologies (ONT), produce longer but more error prone reads. While these longer reads cover more SNV sites which can aid haplotyping, the higher read error rate can introduce noise and false variantion that may hinder haplotype assembly.

Longshot [3], a variant-calling tool implemented by Professor Bansal, is designed for haplotype assembly on these long error-prone reads. Given a reference genome and reads for an individual, Longshot identifies potential variant sites, allelotypes each read, then uses HapCUT2 [2] to predict the haplotypes of the individual. These steps are shows as **a**, **b**, and **c** in figure 1. Sites that are homozygous are not useful for haplotype phasing, so only sites determined to be heterozygous by step **b** will be phased by HapCUT2 in step **c**. Due to read errors, some sites that are actually homozygous may be deemed heterozygous by **b** and passed to HapCUT2, where they will likely increase the error. Our goal for this project was to identify these false variants (sites that appear heterozygous in ONT reads after step **b** but are actually homozygous) between steps **b** and **c** so that they could be removed to improve the accuracy of the final haplotype phasing.

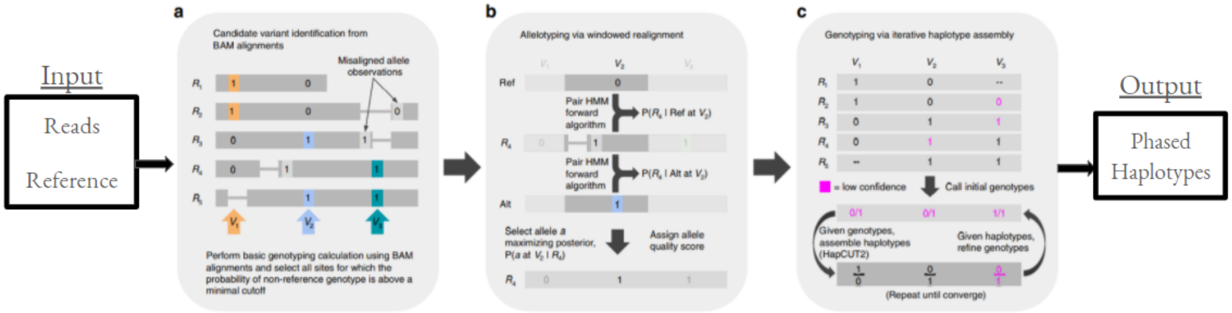

Input
Reads
Reference
a
Candidate variant identification from BAM alignments
Misaligned allele observations
$R_1$
$R_2$
$R_3$
$R_4$
$R_5$
$V_1$
$V_2$
$V_3$
Perform basic genotyping calculation using BAM alignments and select all sites for which the probability of non-reference genotype is above a minimal cutoff
b
Allelotyping via windowed realignment
Ref
Pair HMM forward algorithm
P($R_4$ | Ref at $V_2$)
P($R_4$ | Alt at $V_2$)
Alt
Select allele *a* maximizing posterior, P(*a* at $V_2$ | $R_4$)
Assign allele quality score
c
Genotyping via iterative haplotype assembly
■ = low confidence
Call initial genotypes
0/1
0/1
1/1
Given genotypes, assemble haplotypes (HapCUT2)
Given haplotypes, refine genotypes
(Repeat until converge)
Output
Phased Haplotypes

Figure 1: Longshot's data pipeline flow. The goal of the project is to identify and remove false variants between the steps **b** and **c**, before HapCUT2 is run to phase the haplotypes.

## 2 Data

### 2.1 Data overview

To evaluate our intervention, we used data for chromosome 20 of Genome in a Bottle sample GM24385 [6], a 45 year old male who is labeled "Ashkenazi Son." Oxford Nanopore Technologies provides a Nanopore sequencing dataset of prealigned reads for individual GM24385 [4], which we used as the input to Longshot.

The GRCh38 human reference sequence [5] from NCBI was used as the reference genome when running Longshot.

To determine which sites predicted to be heterozygous by step **b** were true and false variant sites, we needed the true haplotypes for our sample. This ground truth was available from the Genome in a Bottle (GIAB) project for sample GM24385 in the form of a corresponding BAM and VCF file.

## 2.2 Data preprocessing

For our task we consider just the sites predicted to be heterozygous by step **b** of Longshot and the reads of fragments at these sites. To get this information, we run Longshot on the ONT reads with the GRCh38 reference for the intermediate data files it saves between steps. Of interest to us are *2.0.realigned_genotypes.vcf* and *fragments.txt*, both produced by step **b**.

The VCF file provides the positions and unphased genotypes of sites as predicted by Longshot step **b**. The genotypes are used to filter the sites down to just sites predicted to be heterozygous by step **b** (e.i. sites Longshot predicts are variant sites). We then reduce the sites we used to just sites were there is ground truth phased genotype information from the GIAB ground truth sequence using the GIAB BED file. Predicted sites that were heterozygous in GIAB ground truth VCF were labeled as true variants and the other predicted sites as false variants.

The *fragments.txt* file provides a fragment matrix where each row corresponds to a read fragment and each column corresponds to a site. The value of a matrix entry is 0 if the corresponding allele matches the reference, 1 if it is the alternate allele, and nan if the site is not covered by the fragment. For all non-nan entries in the matrix, there are associated Phred read quality scores that can form a corresponding quality matrix. We used the subset of columns from the fragments and quality matrices that corresponded to the predicted variant sites we had labels for.

For our tests, we used the first 5 million base pairs of chromosome 20. This resulted in 5,393 labeled sites, of which 1,003 (18.6%) were false variants and 4,390 (81.4%) were true variants. The fragments matrix consisted of 11,332 rows corresponding to fragments and the 5,393 columns corresponding to the labeled sites. 1,101 fragments (9.7%) did not have any reads (e.i. just nan values) for this subset of columns, and overall the fragments matrix is 99.5% nan values. The distributions of counts of reads by fragment and site with the all nan fragments removed is given by figure 2.

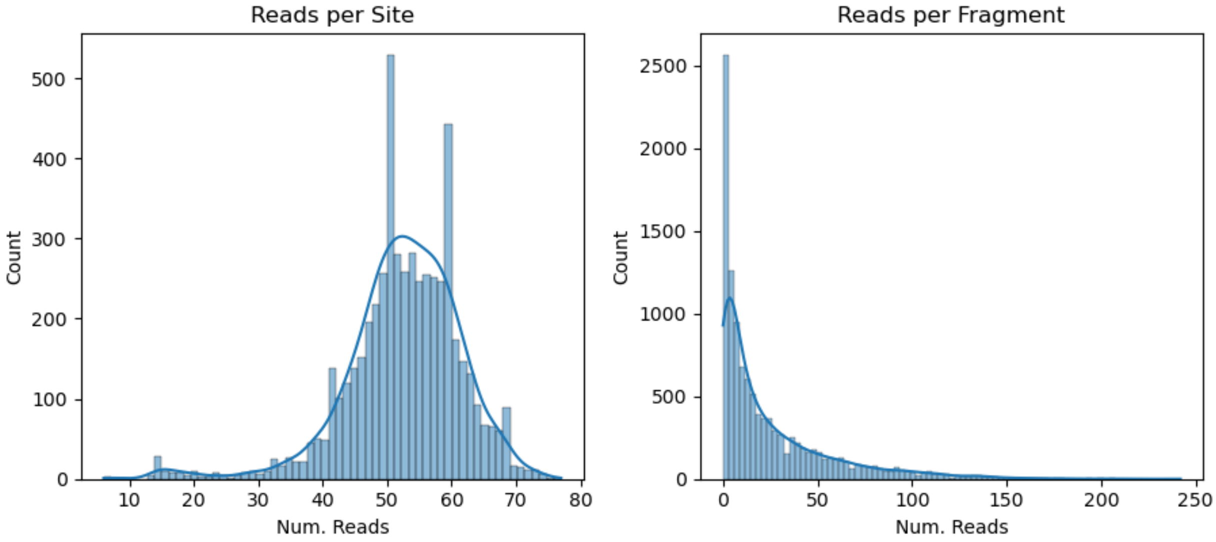
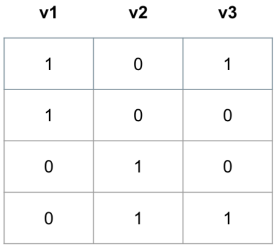


Figure 2: Distribution of counts of reads (read if 0 or 1 in fragment matrix, not read if nan) by fragment and site.

# 3 Methods

## 3.1 Motivation

We first make an observation that motivated our methodology. Consider the following small fragment matrix with three variants and four reads. We can observe that all the variants have the same allele frequency of two zero's and two one's. A per-variant approach is very limited when identifying false variants in this case. Instead, we consider each pair of variants, keeping in mind that a pair of bi-allelic variants can have values of either $(0,0),(1,1)$ or $(0,1),(1,0)$. We see that two pairs including $v_3$ has all the four possible values. This implies that $v_3$ has low correlations with the other variants and is likely a false variant.



| v1 | v2 | v3 |
|---|---|---|
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |

Figure 3: Motivating example to develop pairwise correlation scores. Despite the same allele frequency, $v_3$ is likely a false variant due to its low correlation with other variants.

## 3.2 Flow

We emulate the HapCUT algorithm's graph-based approach [1]. We first construct a graph where a vertex corresponds to a variant site. A pair of vertices is connected by an edge if they are covered by a fragment. The weight of an edge is the correlation score between the two variants.

Then we employ two approaches to identify a set of false variants. The first approach is average edge comparison. For each variant, we compute the average edge weight and predict those with low average values as false variants. This is based on our assumption that false variants tend to be less correlated with other variants. The second approach is max-cut based as in the HapCUT paper. We use a heuristic to pre-filter variants whose average scores are above the median as true variants. This pre-filtering method is mindful of the problem setup where the ratio of true variants is significantly higher than that of false variants. Also, halving the number of vertices in the graph will reduce the complexity of max-cut algorithm.

## 3.3 Correlation scores

We experimented with six different correlation scores as the edge weights. Normalized and adjusted versions of these scores were tried where available. For asymmetric measures, the in, out, and combined edge weights were all tested.

- **Estimated misread ratio**: Given that a pair of bi-allelic variants has values of either $(0,0),(1,1)$ or $(0,1),(1,0)$, compute the number of minority paired allele reads over the total number of reads. Given the two columns (variants) in the matrix $V_1$ and $V_2$, this can be computed as the minimum of the hamming distance between $V_1$ and $V_2$ and that between $V_1$ and $\sim V_2$ where $\sim V_2$ is $V_2$ with flipped 0's and 1's.
- **Absolute phi coefficient**: Measure of association between binary variables. Undefined when one variable (site) has only one value.
- **Weighted entropy**: Asymmetric measure of uncertainty; Given two variants $V_1$ and $V_2$, computed as $P(V_1=0)H(V_2|V_1=0)+P(V_1=1)H(V_2|V_1=1)$
- **Homogeneity**: Uniformity of clusters where one site is cluster labels and the other values.

- **Rand index**: Measure of similarity between clustering where both sites are cluster labels.
- **Mutual information**: Mutual dependency between two variables

We visualize the distribution of the correlation scores for true and false variants as below. Overall, we observe that true and false variants are separable based on the scores despite some outliers.

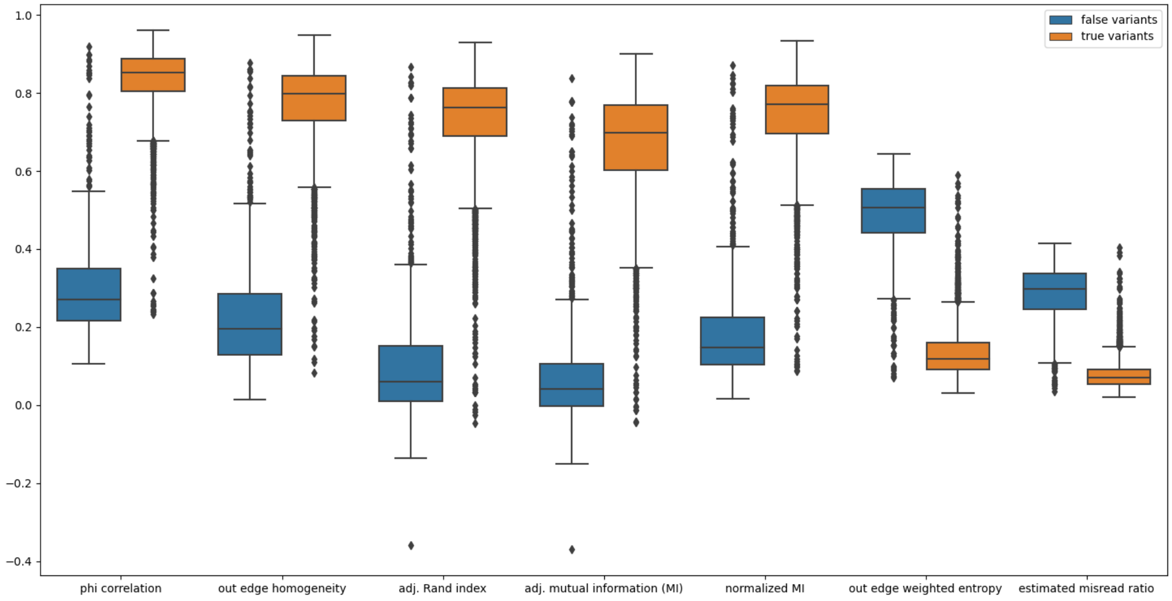


Figure 4: Distribution of correlation score values. True variants are colored orange and false variants blue. Note that true variants have lower values for out edge weighted entropy and estimated misread ratio, consistent with our notion that true variants have lower entropy and estimated misread ratio. We negate the two scores when assigning them as edge weights in the methodology.

# 4 Results

## 4.1 Performance comparison

From the table 1, phi correlation average edge weight comparison has the best performance across all the evaluation metrics. Overall, scores adjusted to account for chance, such as adjusted rand index, adjusted mutual information and normalized mutual information, outperformed their adjusted versions. Estimated misread ratio has lower AUC values than other correlation scores and performs better with the average edge weight comparison than with the max-cut algorithm.

## 4.2 Deep dive into false positives and false negatives

In this section, we investigate false positives and false negatives to understand the weaknesses of our method. We focus on the estimated misread ratio based max-cut approach as it is the most intuitive correlation score. We first need to clarify the terms, “false positives” and “false negatives.” The primary goal of the project is to identify false variants. Hence, false positives are true variants in the ground truth label that are predicted as false variants. False negatives, on the other hand, are false variants in the ground truth label that are predicted to be true variants. To better understand the average estimated misread ratio distribution of false positives and false negatives, we visualize them along with the correctly predicted variants on a sample of 650 variants in figure 6. True variants from the ground truth are colored as blue while false variants from the

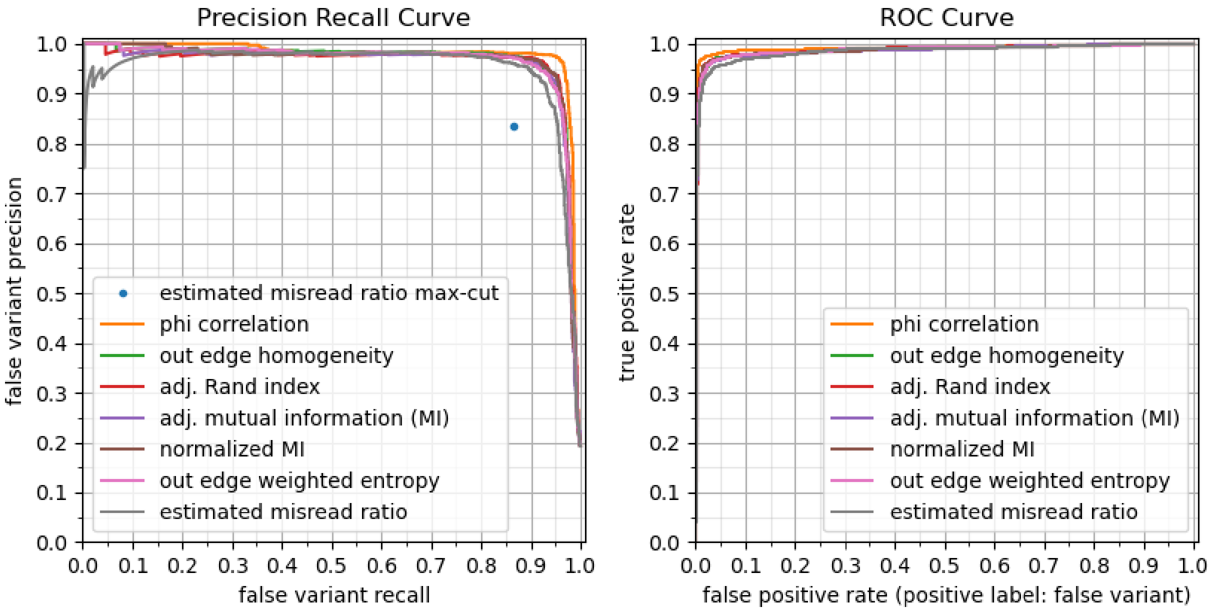


Figure 5: PR and ROC curves. All results shown except *estimates misread ratio max-cut* use the mean edge weight method.

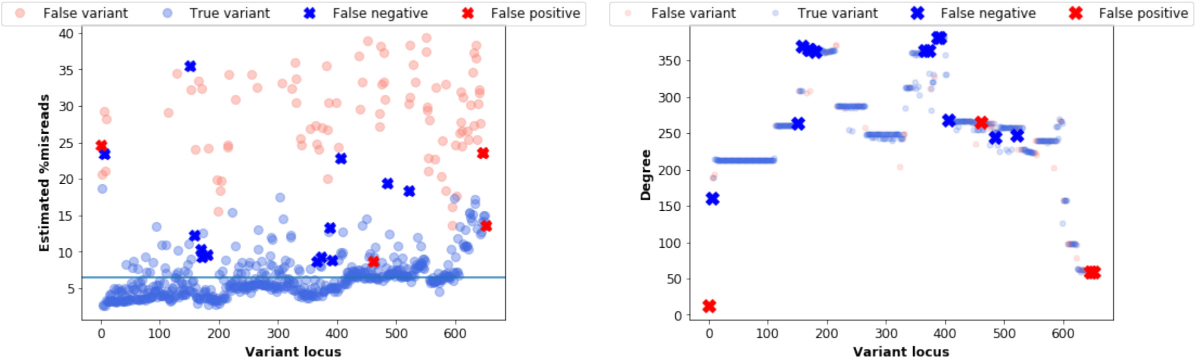
| **Method** | **AUC** | **false variant prec.** | **false variant recall** | **false variant f1** |
|---|---|---|---|---|
| phi correlation | 0.990 | 0.972 | 0.958 | 0.965 |
| out edge weighted entropy | 0.986 | 0.945 | 0.929 | 0.937 |
| out edge homogeneity | 0.987 | 0.951 | 0.927 | 0.939 |
| adjusted Rand index | 0.987 | 0.936 | 0.955 | 0.946 |
| adjusted mutual information | 0.986 | 0.948 | 0.937 | 0.942 |
| normalized mutual information | 0.986 | 0.941 | 0.949 | 0.945 |
| estimated misread ratio | 0.984 | 0.936 | 0.916 | 0.926 |
| estimated misread ratio max-cut | - | 0.834 | 0.866 | 0.850 |

Table 1: All methods except *estimates misread ratio max-cut* use the mean edge weight method. Phi correlation scores highest by all four measures.

ground truth are colored as red. Variants marked as ‘x’ are incorrectly predicted. The light blue horizontal line represents the median estimated misread ratio used in the pre-filtering step.

We first note the effectiveness of the pre-filtering method: predicting variants whose average edge weight is below the median as true variants. As the majority of true variants have low estimated misread ratio, they are correctly labeled as true variants even before running the max-cut algorithm. Whether this pre-filtering method will be applicable to other correlation scores is to be studied. We also observe that most of the false predictions are made on variants near the “boundary” between true and false variants. Therefore, even with the max-cut approach, it is challenging to identify false variants that have relatively low estimated misread ratio.

We also visualize the degree or the number of neighboring vertices of each variant in figure 6. Interestingly, we observe that the degree of false positives is low while that of false negatives is high in general. We hypothesize that the scores of small-degree-true-variants connected to false variants can be easily polluted and falsely predicted as false variants. On the other hand, false variants adjacent to a lot of true variants can have relatively high correlation scores and be classified as true variants. These outliers are not easily captured in the current method.



(a) Average misread ratio distribution

(b) Degree or the number of adjacent vertices

Figure 6: Deep dive into false positives and false negatives. False positives are true variants in the ground truth label that are predicted as false variants. False negatives are false variants in the ground truth label that are predicted to be true variants

# 5 Conclusion

We were surprised that the max-cut approach was less effective than the average edge weight approach. This may be because we computed a one-off max-cut as opposed to identifying and removing variants iteratively. Another potential degrading factor is the presence of multiple connected components if there are disjoint variant sets not covered by a read. Applying max-cut on a per-connected-component basis may improve the performance. Finally, we applied the max-cut approach solely on the estimated misread ratio, whose performance is worse than other correlation scores in the average weight comparison approach. We can extend the max-cut approach to other correlation scores for a more thorough evaluation.

Furthermore, the current correlation scores only consider the allele calls between the two sites. However, Longshot provides allele quality scores, an indication of how likely a given allele call is to be correct. We can leverage this quality information to improve our correlation scores with a likelihood based approach. This may improve our performance on blind spots where false variants are adjacent to a lot of true variants and true variants have low degrees.

Lastly, we would like to hook up our intervention implementation to Longshot to assess how the Longshot performances changes after identifying and removing false variants. This will help us understand how compatible our intervention method is with the HapCUT2 algorithm.

## 5.1 Contributions

RD wrote code to process Longshot output and GIAB ground truth into a fragment matrix with variant true/false labels, then implemented and evaluated different correlation scores. YK worked on implementing the estimated misread ratio score and the max-cut method. SV contributed to making the slides and drafting the report.

# References

[1] V. Bansal and V. Bafna. “HapCUT: an efficient and accurate algorithm for the haplotype assembly problem”. In: *Bioinformatics* 24.16 (2008), pp. i153–i159. DOI: 10.1093/bioinformatics/btn298.

[2] Peter Edge, Vineet Bafna, and Vikas Bansal. “HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies”. In: *Genome Research* 27.5 (2016), pp. 801–812. DOI: 10.1101/gr.213462.116.

[3] Peter Edge and Vikas Bansal. “Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing”. In: *Nature Communications* 10.1 (2019). DOI: 10.1038/s41467-019-12493-y.

[4] *GM24385 Dataset Release*. Sept. 2020. URL: https://nanoporetech.github.io/ont-open-datasets/gm24385_2020.09/.

[5] Valerie A. Schneider et al. “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly”. In: (2016). DOI: 10.1101/072116.

[6] Justin M. Zook et al. “Extensive sequencing of seven human genomes to characterize benchmark reference materials”. In: *Scientific Data* 3.1 (2016). DOI: 10.1038/sdata.2016.25.