Characterizing Mapping Quality Recalibration Approaches in a Variant Graph Genomics Tool

Jeffrey Chan[1], Adam Novak[2], Benedict Paten[2]
[1]Department of Computer Science, University of Puerto Rico Rio Piedras Campus
[2]UC Santa Cruz Genomics Institute, University of California, Santa Cruz

Motivation

Identifying DNA patterns can tell us useful information about any living being. Closely related organisms have similar DNA, while distantly related organisms have few similarities. Humans have extremely similar genomes; studying differences can help to identify particular variants that can cause illness. Vg is a variant graph-based alignment tool for DNA mapping using graph genome references; these graphs capture variation information from populations, which allows more accurate genome studies.

Background

Vg is a set of tools for working with genome variation graphs. These graphs consist of a set of nodes and edges, where each node represents a DNA sequence; edges, connections between two nodes, can be seen as concatenations of two sequences. We built VG graphs with genome references and their sequence variations. Because of the variation we have multiple paths through the graph, this means from a particular sequence you could have multiple edges to take. An essential part of vg is mapping DNA reads into the graph; that means searching for the position where the sequence is most similar to the reference graph. Mapping is challenging because genomes can be very repetitive, and in each repetition, sequences can vary. In addition to mapping, vg calculates a mapping quality score; this score is the probability of the mapping being wrong.

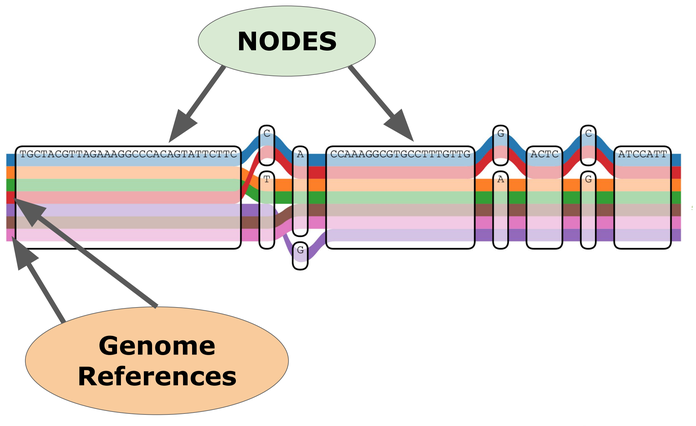

Figure 1: *Sequence TubeMaps visualization of a vg graph. (Garrison, et al. 2017) This visualization was created using the Sequence TubeMaps Tool. https://github.com/vgteam/sequenceTubeMap*

Approach

In this work, we create and benchmark models to predict the probabilities of mappings being wrong and compare our recalibration models against each other and against the original mapping quality scores. To build our dataset, we simulate sequences with errors from the reference graph and map these new sequences back into the graph, then label those mappings as correct or incorrect. We train our models to calculate when a mapping is wrong, then extract the probabilities from those predictions. Using these probabilities, we calculate mapping quality scores and compare them against the original scores calculated by vg using the Brier score.

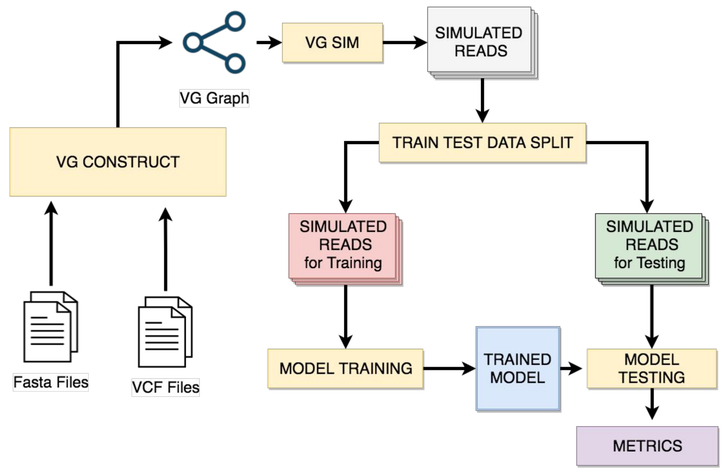

Figure 2: *Evaluation workflow of our models*

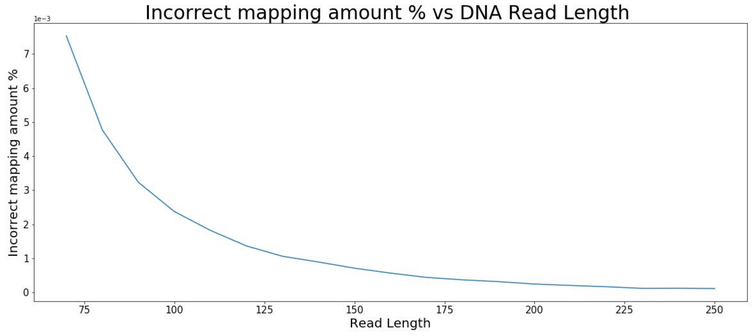

Figure 3: *Incorrect mapping percentage by DNA sequence length.*

To view code: https://github.com/binarySequoia/VG_Recal

Results

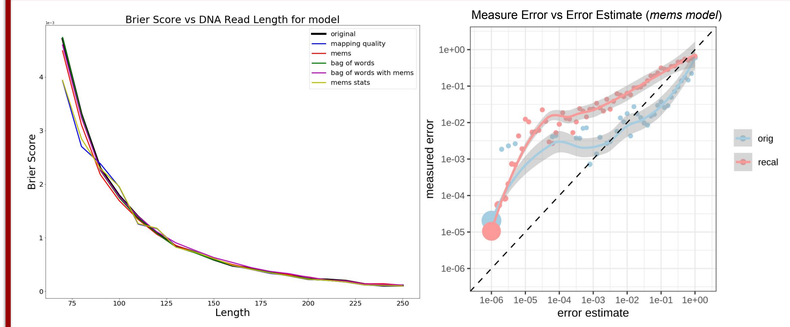

Figure 4: *Left plot shows Brier Score vs Length for different models.*
Figure 5: *Right plot shows a Q-Q plot of original and bag of words of mems data with a logistic regression*.

Discussion

We test 5 different models with logistic regression using mapping quality information, mems, sequences, mems stats and a combination between mems and sequences. Our experiments show that logistic regression with mems improves by 5.23% the original mapping score given by vg in reads of length 100 base pairs but is not able to generalize well across lengths. But the Q-Q plot shows that the mems model has over confidence about its predictions.

Future Work

- Multiple error ranges
- Generalize across graph reference
- Variant calling experiment with real reads

References

Garrison, Erik, et al. "Sequence variation aware genome references and read mapping with the variation graph toolkit." bioRxiv (2017): 234856.

Acknowledgment

This material is based upon work supported by the Wellcome Trust (grants 206194 and 207492), National Institutes of Health (5U41HG007234, R25MD010399), the W.M. Keck Foundation (DT06172015) and the Simons Foundation (SFLIFE# 35190). I would also want to thank Humberto Ortiz-Zuazaga, Patricia Ordóñez. Zia Isola.