Walking the cost-accuracy tightrope: balancing trade-offs in data-intensive genomics

Kathryn Leung
Princeton University

Meghan Kimball
DePaul University

Jason Pitt (Advisor)
National University of Singapore

Anna Woodard (Advisor)
University of Chicago

Kyle Chard (Advisor)
University of Chicago

Introduction 1

Scientific applications (e.g., those based on machine learning or ensemble models) allow researchers to continually improve accuracy at the expense of increased computational cost. Thus, it is possible to dynamically trade off analysis cost and accuracy; however, measuring and predicting cost and accuracy in a way that can be used to quantify these trade-offs is challenging. We present a predictive cost and accuracy model and use these models to create visualizations which can be used to quantify and selectively balance the cost-accuracy trade-off.

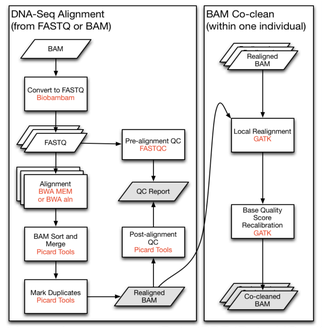
Data-intensive genomics 2

The Genomic Data Commons (GDC) applies computational pipelines to identify somatic variants within whole exome sequencing and whole genome sequencing data. Somatic variants are generated by comparing allele frequencies in normal and tumor sample alignments, annotating each mutation, and aggregating mutations from multiple cases into one project file.

Using 10,000 samples of somatic mutation files from the GDC we explore the trade-offs between cost and accuracy when applying an ensemble [4] of four variant callers: muse, mutect, varscan, somaticsniper.

GDC analysis pipelines include:

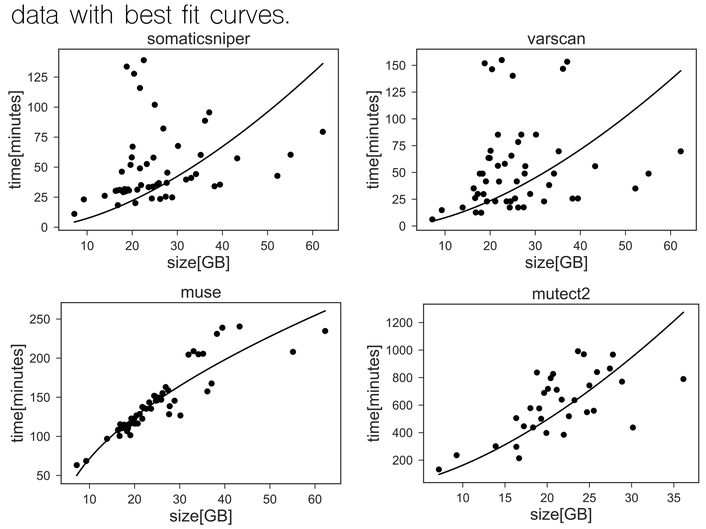
- Genome Alignment
- Alignment Co-Cleaning
- Somatic Variant Calling
- Variant Annotation
- Mutation Aggregation
- Aggregated Mutation Masking



https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf

Cost Models 3

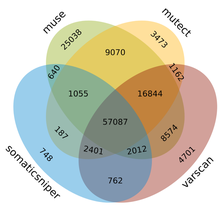
We implemented the GDC pipeline with Parsl [1] and measured the execution time of each variant caller for different input data sizes on a campus cluster (1288 nodes each with 12 cores and 128 GB of RAM using shared flash storage). Using a least squares fit, we fit the data with best fit curves.



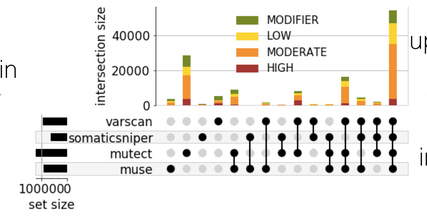
Data Visualization and Filtering 4

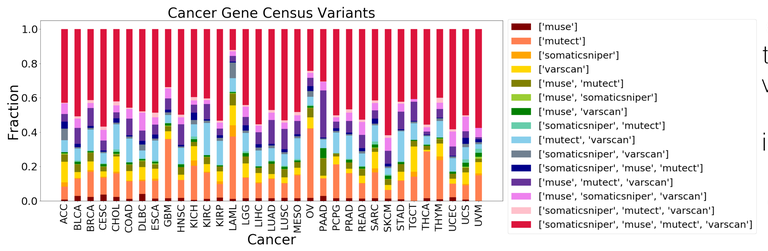
a) 3-way Venn diagrams showing number of variants in each variant caller intersection for breast cancer (BRCA)

b) 4-way Venn diagrams showing number of variants in each variant caller intersection for breast cancer (BRCA)



c) 4-way upset plot [3] with color filtering by protein impact level (BRCA)



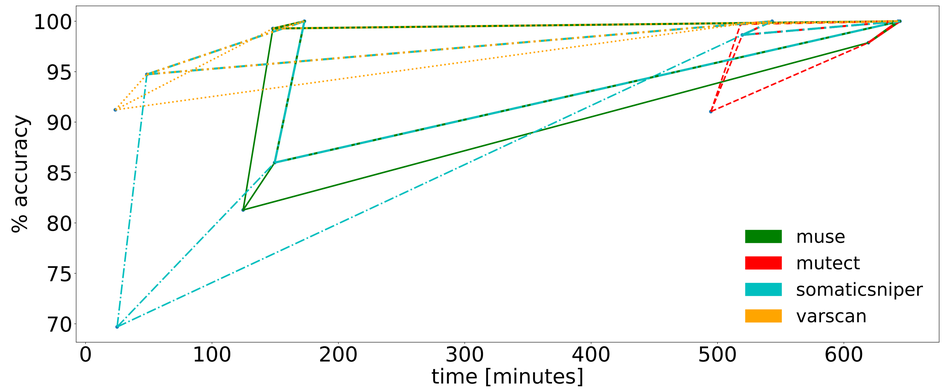


d) Stack plot showing the proportions of total variants found by each variant caller intersection for variants that are related to cancer (Cancer Gene Census)

We can see from our preliminary visualization/analysis that from our set of variant callers, the majority of variants from our dataset can be found by mutect and varscan.

Accuracy/Cost Analysis 7

We can now visualize ensemble strategies of variant callers. The following figure shows the average accuracy (percentage of all real variants found by each ensemble) vs. cost (execution time) for all possible ensembles of variant callers on a fixed input size. The points of intersection for the different colored lines (where each color corresponds to a variant caller) represent the accuracy/time for that combination of variant callers. Our results show that, on average, the GDC could achieve 99% accuracy in approximately half the time by optimally selecting variant callers.



We can reconstruct this plot as an edge weighted directed graph, and assign weights to each edge:

$$w_{(i,j)} = \frac{accuracy_j - accuracy_i}{time_j - time_i} | i \subset j; i, j \in \mathcal{V}$$

where i represents the baseline variant caller(s),j represents the union of i and the variant caller being added, *V* is the set of all subsets of callers

Cost/Accuracy tradeoff algorithm:

Add source node at (0,0), *i* ,that connects starting points of all variant callers → Choose next variant caller *j* , $\max_j w_{(i,j)}$ → $j = i$ → Repeat until desired accuracy is achieved or budget is reached

Quorum Rule 5

Without ground truth for the GDC data, we implemented the standard practice quorum of two rule [2] to generate a functional truth data set to evaluate our methods: If any two or more variant callers called a variant we considered it to be 'real'.

n is the number of total variant callers, *m* is the number of callers, *q* is the quorum.

(1) The number of intersections in an n-way Venn diagram $\sum_{i=1}^{n} \binom{n}{i} = 2^n - 1$

(2) The number of real variant intersections $\sum_{i=q}^{n} \binom{n}{i}$

(3) The number of intersections found by *m* callers $\sum_{i=1}^{n} \binom{n}{i} - \sum_{i=1}^{\infty} \binom{n-m}{i}$

(4) The number of real variant intersections found by *m* callers $\sum_{i=q}^{n} \binom{n}{i} - \sum_{i=q}^{\infty} \binom{n-m}{i}$

For our analysis, *n* = 4, *m* = 3, and *q* = 2; consequently, the second term in rule (4) is always 0. This means that we achieve 100 percent accuracy with any three callers.

Accuracy Modeling 6

We created random forest models using scikit-learn to predict whether a specific additional variant caller will yield an increase in accuracy based on features of the data (e.g., cancer type, substitution type) and from the number of variants identified by previously applied variant callers.

The average performance of our models for all possible baseline caller(s) is shown in the table below. Because some baseline callers, like mutect and varscan, perform significantly better than the other callers our data is often skewed towards there being a zero increase in accuracy when adding additional variant callers. To reconcile this issue, we applied a random undersampling algorithm to improve the performance of our models.

baseline variant caller(s)	accuracy	precision	recall
muse	0.67	0.69	0.81
mutect	0.68	0.55	0.68
somaticsniper	0.75	0.78	0.90
varscan	0.65	0.47	0.68
muse, mutect	0.65	0.21	0.66
muse, somaticsniper	0.63	0.63	0.66
muse, varscan	0.66	0.11	0.70
mutect, somaticsniper	0.71	0.24	0.67
mutect, varscan	0.71	0.05	0.76
somaticsniper, varscan	0.70	0.49	0.68

Researchers can use these models in conjunction with our cost models to guide their decision on whether they should run an additional variant caller.

Results 8

We have shown that the canonical approach in genomics research of applying as many variant callers as possible to achieve the highest accuracy is often cost-inefficient through our visualization and analysis of the GDC data. We have also shown that we can predict cost and accuracy for four variant callers. Our models allow researchers to optimize the cost-accuracy trade-off using an edge weighted digraph, and our workflow can be generalized to any number of other variant callers. Our methods are applied here to genomics, but they are applicable to other domains that use ensemble methods.

Acknowledgements: We would like to thank Matthew Baughman for his invaluable guidance on this project. This work is supported by the NSF CCF-1757964/1757970 REU award (BigDataX). This work is partially supported by NSF 1816611.The computational work for this article was (fully/partially) performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References:
[1] Yadu Babuji Et al. 2019. Parsl: Pervasive Parallel Programming in Python. In 28th ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC). https://doi.org/10.1145/3307681.3325400 babuji19parsl.pdf. [2] K Ellrott Et al. 2018. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst 6, 3 (03 2018), 271–281. https://doi.org/10.1016/j.cels.2018.03.002 [3] Alexander Lex Et al. 2014. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics (InfoVis '14) 20, 12 (2014), 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248 [4] Vassily Trubetskoy Et al. 2014. Consensus Genotyper for Exome Sequencing (CGES): improving the quality of exome variant genotypes. Bioinformatics 31, 2 (09 2014), 187–193. https://doi.org/10.1093/bioinformatics/btu591