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

ABSTRACT

Scientific applications often exhibit a trade-off between cost and accuracy. However, measuring and predicting cost and accuracy in a way that users can understand these trade-offs is challenging. To address these needs, we present predictive cost and accuracy models for data-intensive genomics applications. We use these models to create a trade-off graph, which researchers can use to selectively trade-off cost and accuracy.

ACM Reference Format:
Kathryn Leung, Meghan Kimball, Jason Pitt (Advisor), Anna Woodard (Advisor), and Kyle Chard (Advisor). 2019. Walking the cost-accuracy tightrope: balancing trade-offs in data-intensive genomics. In *Supercomputing '19: The International Conference for High Performance Computing, Networking, Storage, and Analysis. Nov 17–22, 2019, Denver, CO.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Exploding data volumes combined with adoption of data- and compute-intensive methodologies are transforming scientific computing. While iterative machine learning-based techniques and ensemble methods enable new discoveries, they also create new challenges in determining when models are sufficiently accurate and result in a trade-off between accuracy and computation cost.

To explore the cost-accuracy trade-off, we investigate the computationally expensive variant calling analysis of genomes. Variant calling identifies single nucleotide variants within an individual genome relative to the population at large. There are dozens of interchangeable variant callers, with no scientific consensus on which is the "best". Researchers and repositories, such as the Genomic Data Commons (GDC), often combine results from several variant callers and use ensemble-based approaches to improve accuracy [3].

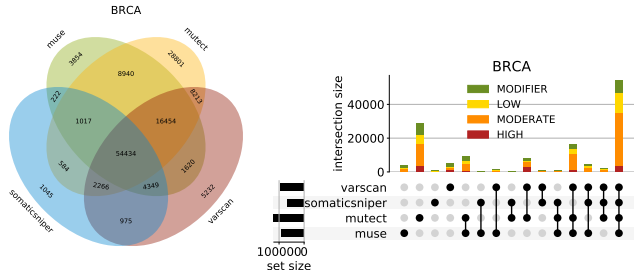
2 METHODOLOGY

In this study, we use GDC variant calling data, which consists of approximately 10,000 samples from 33 different cancers, analyzed using four different variant callers (muse, mutect, somaticsniper, varscan). We implement GDC pipelines in Parsl [1] and measure the computation time for different input sample sizes. We then develop models that predict the accuracy of variant calling ensemble models and represent the resulting cost-accuracy trade-off as an edge weighted digraph.

2.1 Ensemble-based analyses

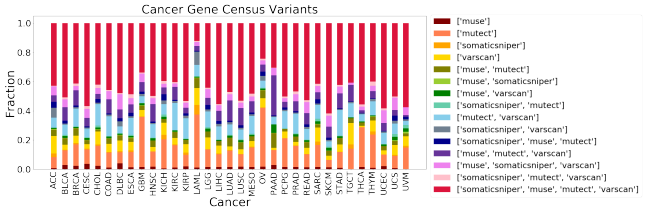
To better understand the benefits of ensemble-based approaches we first explored the GDC dataset. Figure 1 shows three visualizations of the performance of four variant callers on the aggregate of breast cancer (BRCA) samples. We observe that the variant callers identify a disjoint set of variants. Increasing the number of variant callers increases the number of identified variants. However, approximately half of all variants can be identified using only two callers.

In the absence of manually curated truth data for the GDC dataset, we apply a consensus approach as our metric for accuracy: if two or more variant callers identified a variant, we considered it to be "real" [2]. Using this approach, we found that that a significant fraction of the samples had a zero or near zero percent increase in accuracy when adding variant callers. Thus, when considering additional cost, it is important to consider which samples would benefit from additional processing.



(a) Venn diagram for each variant caller for BRCA

(b) Upset plot with color filtering by protein impact levels for BRCA



(c) All cancer types filtered by genes from the Cancer Gene Census

Figure 1: Performance of the different intersections of variant callers on the aggregate of samples.

SC '19, Nov 17–22, 2019, Denver, CO

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2.2 Cost model

To quantify the cost of each variant caller we measured the computational time on different input data sizes. We ran our experiments on ASPIRE1, a cluster with 1288 nodes (dual socket with 12 cores and 128 GB RAM per node). Initial results indicated file system contention, as a result we modified our experiments to make use of shared flash storage. Figure 2 shows a least-squares fit, implemented using scipy curve_fit, demonstrating the relationship between data size and execution duration for each caller. We use these fitted lines to predict run time as a function of input data size.

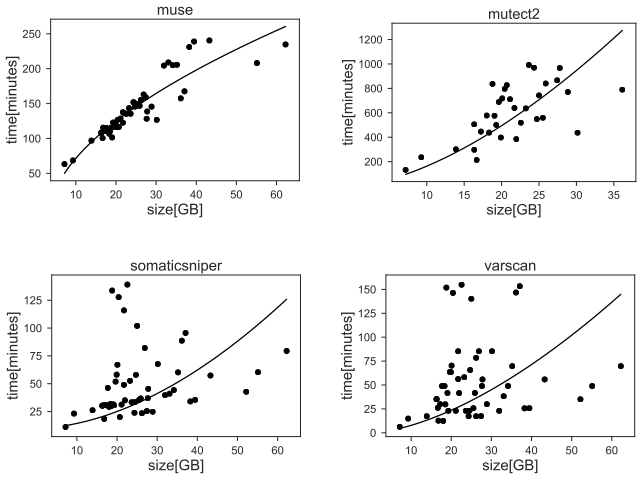


Figure 2: Cost (execution time) as a function of input data size plotted with a least-squares fit for four variant callers.

2.3 Accuracy model

We created random forest models using scikit-learn to predict whether an additional variant caller will yield an increase in accuracy based on features of the data (e.g., cancer type, allele substitution type) and from the number of variants identified by previously applied variant callers in the ensemble. Table 1 shows the accuracy, precision, and recall of our models.

Table 1: Average accuracy, precision, and recall when predicting whether a specific caller will increase accuracy.

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

Overall the results show reasonable predictive performance; however, we note that some combinations resulted in trivial predictive problems. For example, since mutect and varscan perform significantly better than muse and somaticsniper, as seen in Figure 1, our data is skewed towards there being a zero increase when adding muse or somaticsniper after mutect or varscan. To remedy this issue, we applied a random undersampling algorithm to our training set. This improved the results of our predictions, particularly by increasing the recall of our models. We focus on maximizing the recall of our models, because it is likely that researchers would be more willing to incur additional costs (from occasional execution of variant callers that yield no new variants) rather than to miss potentially important variants.

2.4 Cost-accuracy model

We combined the cost and accuracy models above to visualize the average accuracy across samples and cost for a given input size for all ensembles of variant callers. Figure 3 shows the increase in accuracy and cost for different orderings of variant callers. This graph can be used to quantify the cost-accuracy trade-off and optimize the ensemble. Our results show that, on average, we can achieve 99% accuracy at approximately half the cost by optimally selecting variant callers.

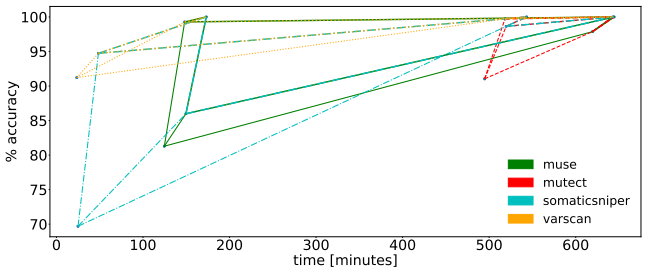


Figure 3: Visualization of accuracy vs. runtime

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

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

where i represents the baseline variant caller(s), j represents the union of i and the variant caller being added, $\mathcal{V}$ is the set of all subsets of callers. In order to traverse the graph, we add a root node at cost and accuracy of 0 with edges to each individual variant caller. We then iteratively select edges with the highest weight. The algorithm continues until cost or accuracy exceed a given threshold.

3 SUMMARY

We have shown that the common approach in genomics of applying as many variant callers as possible to achieve the highest accuracy is often cost-inefficient. Our predictive cost and accuracy models allow researchers to optimize the cost-accuracy trade-off using an edge weighted digraph, and our algorithm can be generalized to any number of other variant callers. Our methods are applied here to genomics, but they are also applicable to other ensemble models.

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

[2] Kyle Ellrott Et al. 2018. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* 6, 3 (2018), 271–281. https://doi.org/10.1016/j.cels.2018.03.002

[3] Vassily Trubetskoy Et al. 2014. Consensus Genotyper for Exome Sequencing (CGES): improving the quality of exome variant genotypes. *Bioinformatics* 31, 2 (2014), 187–193. https://doi.org/10.1093/bioinformatics/btu591