BOSTON UNIVERSITY

Genomics-enabled blood antigen phenotype prediction

Tim Farrell
Bioinformatics Program, Boston University

Introduction

Despite the great advances in genomic sequencing technologies and their accompanying techniques, automated clinical-grade interpretations of this data remain challenging.[1,2] This is primarily due to high levels of variation in such data across both patients and technologies.[3,4,5]

As of June 2015, there have been 80 million variants identified in the human genome, the majority of which have unclear clinical significance.[6] Last year, the National Institute of Standards and Technology reported that variant and genotype calling cannot be done confidently in ~23% of the human genome, due to the high discordance across technologies.[7] Ultimately, these challenges prevent the development of clinical sequencing analysis pipelines that do not require final manual curation by experts.

Genomic classification is a framework that has recently demonstrated promise for overcoming challenges of variation in cancer genomics research.[8,9] Here we apply a supervised genomic classification approach in the context of the highly-polymorphic structurally-variant Rh blood antigen system to demonstrate the generalizability of this framework for clinical phenotype prediction.

Methods

General

As our input data, we use whole genome sequencing alignment data from 93 patients along with their associated antigen serological test results as phenotype labels. Over 16 feature type-sets, we extract feature data, train a multi-label decision tree classifier and test its performance across 10 iterations of 10-fold cross-validation.

Feature Selection

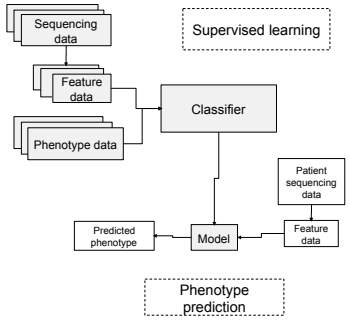
The 16 feature type-sets were developed based on three criteria.

1. Genomic positions selected.
   a. Whole exome.
   b. Only positions identified in variant databases as associated with different genotypes.
2. Quantifier used at each genomic position:
   a. Max base coverage.
   b. Mean base coverage.
   c. Variance base coverage.
   d. Categorically coding for the base called at that position.
3. Whether the feature data was encoded or not.

System Concept

The ultimate goal of this work is to implement the methods described in a clinical sequencing analysis pipeline. Herein a model, previously trained on labeled genomic data, would be applied to predict patient phenotype based on features of their genome.

Assuming that in the coming years genomic data will be routinely included in a patient's electronic health record, such a system would eliminate the need for potentially costly laboratory tests.



Discussion

With further feature set development and optimization, we believe this model can achieve clinical-grade performance and be in the position to replace the laboratory test upon whose data it was trained. Assuming that in time genomic information will be stored routinely in patient electronic health records, this could potentially reduce the aggregate cost of diagnostic laboratory procedures.

Additionally, this study demonstrates generally how machine learning frameworks may be used to build effective and reliable bioinformatics-informed clinical decision-making aids.

Future Work

There are two more advanced techniques that could greatly improve these preliminary results:

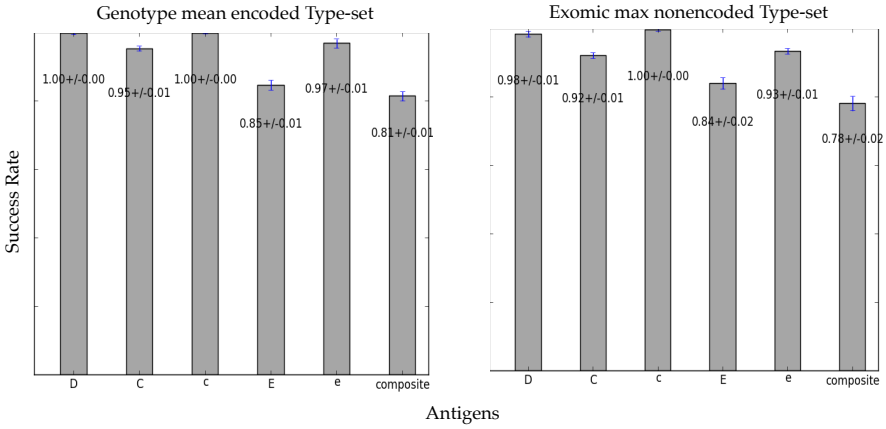
(1) Use well-established bioinformatics tools to characterize better characterize genomic structure and combine into a representative feature set.

(2) Use feature learning techniques to engineer more informative feature sets.

(3) Use more advanced classifiers.

Results

Two Best Performing Feature Typesets



The best feature typeset 'genotype mean encoded' achieved an 0.814 +/- 0.014 success rate for combined phenotype calls (i.e. accurately predicting all phenotypes in one prediction) and much greater success rate for individual antigen phenotype calls, within the composite prediction. The second best feature typeset 'exomic max nonencoded' achieved a highly similar success rate. Of note, 6 of the 16 feature typesets achieved a success rate over 0.75.

Relevant Literature

[1] Jameson JL and Longo DL. 2015. *Precision medicine – personalized, promising and problematic.* N Engl J Med. 372(23): 2229-2234.

[2] Biesecker LG and Green RC. 2014. *Diagnostic clinical genome and exome sequencing.* N Engl J Med. 370:2418-25.

[3] Roberts NJ et al. 2012. *The predictive capacity of personal genome sequencing.* Sci Trans Med. 4, 133ra58.

[4] McLauglin HM, et al. 2014. *A systematic approach to the reporting of medically relevant findings form whole genome sequencing.* BMC Medical Genetics 15:134. doi:10.1186/s12881-014-0134-1.

[5] Baker M. 2012. Structural variation: the genome's hidden architecture. Nat Methods. 9(2): 133-139.

[6] Rehm HL, et al. 2015. *ClinGen – the clinical genome resource.* N Engl J Med 372;23.

[7] Zook JM, et al. 2014. *Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.* Nat Biotechnol 30(2): 246-251.

Acknowledgments

Thanks to Bill Lane (BWH) for providing the datasets and for advising on Rh blood antigen system. And to Peter Tonellato (HMS) for general advising.