Do machine learning predictors of microbial phenotype from genotype identify causal variants?

Morteza M. Saber[1], Maxwell Libbrecht [2], Leonid Chindelevitch[3] and B. Jesse Shapiro[1]

[1]Université de Montréal, [2]Simon Fraser University, [3]Imperial College London

[1]{morteza.mahmoudisaber,jesse.shapiro}@umontreal.ca, [2]maxwell_libbrecht@sfu.ca, [3]lchindel@ic.ac.uk

Introduction:

Mapping the relationship between genomic content of an organism and its phenotype is essential to precision medicine. In humans, mutations at specific sites of the genome have been proven to affect several clinical phenotypes. Similarly in prokaryotes, the recent expansion of genomic data repositories have paved the way for identifying the genomic elements underlying various clinically, environmentally and industrially important bacterial phenotypes [1–4]. Such discoveries has improved our knowledge of the molecular mechanisms underlying important microbial phenotypes such as antibiotic resistance and virulence; and thus has the potential to contributing to the development of novel drugs, vaccines and antibiotics.

Genome-Wide Association Studies (GWAS) have thus far been the most common type of analysis for genotype-phenotype mapping in bacteria which can be used to dissect the genetic components of any measurable and heritable phenotype in an unbiased hypothesis-free manner [5]. Traditional single-locus GWAS approach is defined as models testing the statistical significance of association between a phenotype and a single variant at a time, repeated for all variants across the genome while explicitly adjusting for population stratification and/or multiple-hypothesis testing. As an alternative approach, artificial intelligence (AI) in form of machine learning (ML) and deep learning have been used to build models to correlate genomic variations with phenotypes [6]. However, similar to polygenic risk score analysis [7], the main focus of AI methods is the accurate prediction of the microbial phenotype based on occurrence of genomic variation rather than thoroughly understanding the molecular mechanisms underlying it.

Nevertheless, the accurate prediction of phenotype by a specific model in a defined dataset do not guarantee the replicability of results in different dataset when the model is unable to identify true underlying genomic elements. This issue is particularly acute in bacterial genotype-phenotype mapping due to the unique characteristics of bacterial populations such as strong population stratification and bacterial genomes, and in particular, genome-wide linkage disequilibrium (LD) [5]. Models naïve to these confounding elements may achieve high prediction accuracies in the dataset used for model building; however, the achieved precision may not be generalizable to different dataset, an issue termed as 'overfitting' in AI context. An interpretable AI model capable of identifying the true causal markers not only guarantees higher generalizability of the model but also provides a means for experimental validation of the observed prediction accuracies and sheds light on the molecular mechanisms underlying the phenotype.

In this study, we evaluated the performance of various regression-based and decision tree-based ML approaches commonly used in bacterial genotype-phenotype mapping for their precision and accuracy in identifying true causal markers underlying simulated bacterial phenotypes. To evaluate the relative performance of AI and traditional genotype-phenotype mapping approaches, performance of interpretable ML models were compared with the linear mixed model-based GWAS approach implemented in GEMMA. GEMMA was selected because we previously found it to be the best performing non-ML method for bacterial GWAS [5]. Performances of all the models were benchmarked against bacterial genotype-phenotype simulations generated by BacGWASim [5]. We focused mainly on the effects of sample size, causal variant effect size and LD (recombination rates) as they are the most important variables determining the performance of GWAS tools [5, 8] while the other evolutionary parameters were kept constant.

With the rapid expansion of the application of AI in bacterial genotype-phenotype mapping, it is essential to evaluate the efficiency of these models in adjusting for confounding elements unique to bacteria. By systematic comparison of the performance of AI models with traditional bacterial GWAS tools across a range of realistic effect sizes, recombination rates and sample sizes, our study provides a basis for correct choice of genotype-phenotype mapping model across different evolutionary scenarios. Our work also inspects the heterogeneity of different ML models, in their power to identify true set of causal markers in high-dimensional bacterial genomic data.

Materials and methods:

Benchmark datasets

Total bacterial genomes including protein coding genes and noncoding regulatory elements were simulated using BacGWASim. Noncoding regulatory elements were included in the simulations as they have been suggested to play important roles in evolution of species [9, 10]. Eighteen causal markers with ORs (effect sizes) of 2, 3, 4, 7, 10, 11, 15 and 20 (natural logarithm in the range of 1 to 3) with minor allele frequency >0.1 were randomly chosen for phenotype simulation. The LD values between the selected markers were measured using bcftools [11] and markers with high correlation (r^2 >0.6) were discarded. This filtering step to remove strongly linked causal markers was included to ensure that these markers (including those of different effect sizes) were identifiable in the simulated datasets. We note that causal markers could still be linked to non-causal alleles, posing a challenge for ML methods to correctly identify the causal variant.

Ten sets of simulations were generated, each containing 100,000 markers with minor allele frequency >0.01 and for each simulation, a binary matrix with sample names as rows and marker ids as columns was produced.

Genotype-phenotype mapping tools

We trained the six interpretable ML models including l2-regularized logistic regression, l2-regularized support vector machine (SVM) [12], random forest [13], extreme gradient boosting (XGB) [14], light gradient boosting model (LGBM) [15] and kover [16] on the ten sets of simulations and calculated their mean power in correctly ranking the features based on the assigned effect sizes. Additionally, GEMMA v.98 [17] that is a linear mixed model-based GWAS tool was tested with pairwise variant-based genetic distances to correct for population stratification.

Results:

LightGBM ML model outperforms traditional GWAS in identification of causal markers

All the models were evaluated based on three performance metrics including 1) the power to identify causal variants in the top 50 ranked features, 2) the fraction of cumulative feature importance captured and 3) area under the curve (AUC) of precision-recall plot. Across the range of low (odds ratio ~ 1) to moderate (odds ratio ~ 2) and high odds ratio (~ 3), LGB model achieved the highest performance in identifying causal markers within the top 50 ranked features, respectively with the mean sensitivity of 0.57 , 0.82 and 0.87, followed by logistic regression and linear SVM. Using the same performance metrics, GEMMA respectively achieved the mean sensitivity of 0.17, 0.45 and 0.75 (Figure 1).

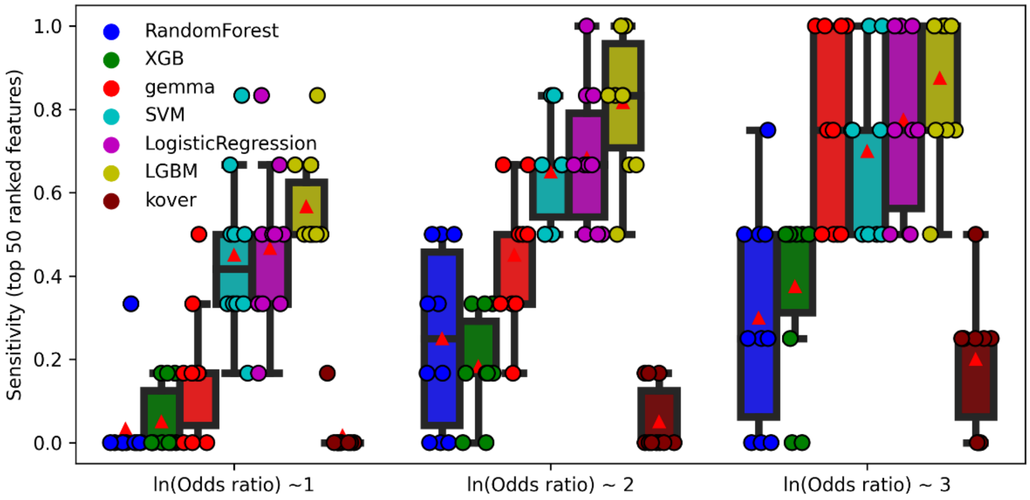

Figure 1 Power of genotype-phenotype mapping tools to identify causal variants. Sensitivity of ML models in comparison with GEMMA implementing a linear mixed model-based GWAS were compared across different categories of effect sizes. Mean values are shown by triangles.

Next, we checked whether the relative power of ML models depends on the number of top features considered in the analysis. To this end, the precision-recall scores for identifying causal markers with a mixture of effect sizes were calculated by including the range of 16 to 500 top ranked features and the corresponding AUC values were estimated. Consistently, the LGB model outperformed other models by AUC of 0.40 (standard deviation (SD) = 0.12), while GEMMA achieved average AUC of 0.17 (SD =0.08) (Figure 2).

Finally, models were evaluated for their efficiency to capture the quantitative importance of causal variants. After normalizing the estimated feature importance to sum to one, the fraction of cumulative feature importance captured by each model were estimated across categories of causal markers with different effect sizes. Our analysis showed that LGB model possessed the best performance by respectively capturing 0.36, 0.34 and 0.37 of cumulative initial feature importance assigned to causal markers across categories of effect sizes (Figure 3). Logistic regression and linear SVM models along with GEMMA all achieved poorer performances based on this metric.

In summary, our results indicates that in genotype-phenotype mapping of bacterial species with frequent genomic recombination such as Streptococcus pneumonia with moderate sample size (<500) and with causal markers distributed within a range of effect sizes (1<ln(odds ratio)<=3), ML can outperform traditional GWAS in adjusting for confounding elements and hence, identifying causal variants.

Discussion:

Accurate microbial genotype-phenotype prediction is a problem of high significance in precision medicine, but which poses great challenges for learning algorithms. Difficulties arise not only because of the unique characteristics of bacterial populations such as population stratification but also due to the relatively small size of available samples compared with the size of the genomic data available for each sample. Furthermore, interpretability of the prediction model is essential to fill the gap in our understanding on the molecular mechanisms underlying microbial phenotypes which is not possible with most of state-of-the-art deep learning-based AI models [16].

In this study we show that LGBM is the best performing ML model to adjust for confounding factors in bacterial genotype-phenotype mapping and achieves the best performance in identification of true causal markers (Figures 1-2). LGBM also captures the highest fraction of cumulative feature importance relative to other evaluated ML models (Figure 3). Notably, LGBM outperforms linear mixed model-approach implement in GEMMA which is the best performing traditional GWAS method based on all three evaluated metrics indicating that this model could potentially replace traditional GWAS tools for the purpose of accurate identification of genomic elements underlying bacterial phenotypes.

This work is still in progress, and the findings need to be replicated under different realistic evolutionary scenarios such as different recombination rates and sample sizes.

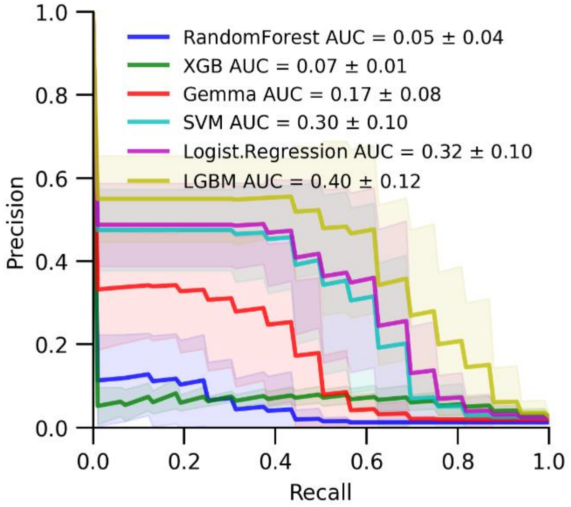

Figure 2. Precision-recall AUC. Area under the curve of precision-recall plots estimated across the range of top selected features ranging from 16 to 500.

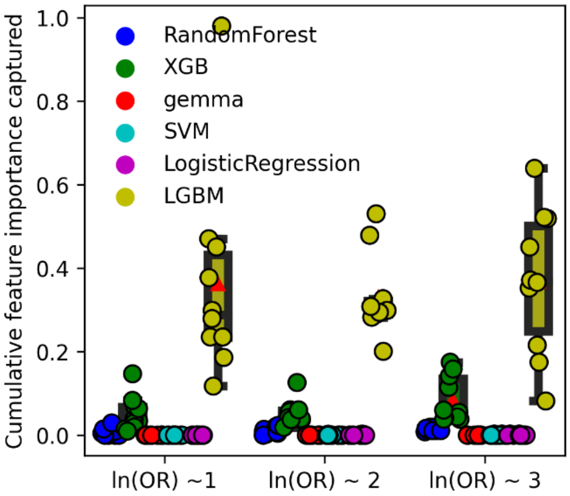

Figure 3. Power of genotype-phenotype mapping tools in capturing the quantitative importance of causal variants. Fraction of cumulative feature importance assigned to causal variants captured by different models were compared across different categories of effect sizes. (OR= odds ratio)

References:

1. **Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, *et al.*** Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *Elife*;6. Epub ahead of print 25 2017. DOI: 10.7554/eLife.26255.

2. **Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, *et al.*** Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet* 2013;45:1183–1189.

3. **Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, *et al.*** GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat Commun* 2019;10:2128.

4. **Berthenet E, Yahara K, Thorell K, Pascoe B, Meric G, *et al.*** A GWAS on Helicobacter pylori strains points to genetic variants associated with gastric cancer risk. *BMC Biol* 2018;16:84.

5. **Saber MM, Shapiro BJ**. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom*;6. Epub ahead of print 2020. DOI: 10.1099/mgen.0.000337.

6. **Levade I, Saber MM, Midani F, Chowdhury F, Khan AI, *et al.*** Predicting Vibrio cholerae infection and disease severity using metagenomics in a prospective cohort study. *J Infect Dis*. Epub ahead of print 1 July 2020. DOI: 10.1093/infdis/jiaa358.

7. **Choi SW, Mak TS-H, O'Reilly PF**. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* 2020;15:2759–2772.

8. **Chen PE, Shapiro BJ**. The advent of genome-wide association studies for bacteria. *Curr Opin Microbiol* 2015;25:17–24.

9. **Mahmoudi Saber M, Saitou N**. Silencing Effect of Hominoid Highly Conserved Noncoding Sequences on Embryonic Brain Development. *Genome Biol Evol* 2017;9:2037–2048.

10. **Saber MM, Adeyemi Babarinde I, Hettiarachchi N, Saitou N**. Emergence and Evolution of Hominidae-Specific Coding and Noncoding Genomic Sequences. *Genome Biol Evol* 2016;8:2076–2092.

11. **Li H**. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27:2987–2993.

12. **Hearst MA**. Support Vector Machines. *IEEE Intelligent Systems* 1998;13:18–28.

13. **Breiman L**. Random Forests. *Machine Learning* 2001;45:5–32.

14. **Chen T, Guestrin C**. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016;785–794.

15. **Ke G, Meng Q, Finley T, Wang T, Chen W, *et al.*** LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, et al. (editors). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. pp. 3146–3154.

16. **Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J, *et al.*** Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci Rep* 2019;9:4071.

17. **Zhou X, Stephens M**. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 2012;44:821–824.