The Malaria DREAM challenge: team TPOT's sub-challenge 2 write-up

This manuscript (permalink) was automatically generated from trang1618/plasmodium-falciparum@4c484ef on August 13, 2019.

Authors

- **Alena Orlenko**
  0000-0003-1757-293X · desmidium · desmidium
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Trang T. Le**
  0000-0003-3737-6565 · trang1618 · trang1618
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Weixuan Fu**
  0000-0002-6434-5468 · weixuanfu · weixuanfu
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Jason H. Moore**
  0000-0002-5015-1099 · EpistasisLab · moorejh
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

Abstract

Challenge link: https://www.synapse.org/#!Synapse:syn16924919/wiki/583955

Objective: Predict the parasite clearance rate of malaria parasite isolates based on in vitro transcriptional profiles.

Data source: [1]

Methods

Data harmonization

For this subchallenge, the training dataset's characteristics were very different compared to those of the test set. Importantly, training set contains *in vivo* transcription as described in [1] while test set contains *in vitro* transcription. Other differences in data collection, synchronization, microarray platforms, preprocessing steps, etc. have been detailed on the challenge website. We first impute the missing data points using the KNN imputation method with k = 10 via the fancyimpute package. Next, to adjust for batch effects between the two datasets, we apply the *ComBat* algorithm [2] from the *sva* R package [3] on the transcription data. We assess the effect of the adjustment by examining the principal component analysis (PCA) plots on the raw and processed data.

Gene feature selection

Using the STatistical Inference Relief (STIR) algorithm [4], we selected the genes with adjusted STIR P values < 0.05. Specifically, with the MultiSURF neighborhood, STIR nearest-neighbors to select features whose association with an outcome may be due to epistasis

Automated machine learning for model training

[TPOT details here…] We used balanced accuracy as our scoring function.

Test sample selection

While there is one single sample per isolate in the training set, there are eight samples for each of the 32 isolates in the test set (most have 2 biological replicates, 2 time points: 6 hours and 24 hours post invasion, perturbed with 5nM DHA (DHA) or perturbed with DMSO (UT)). Therefore, to obtain one prediction per isolate, we need to take caution in selecting which test samples to predict on. First, because the training samples are unperturbed, we discard the perturbed samples in the testing set, keeping only the controls (UT). Second, we analyze the developmental stage of each biological replicate at seperate timepoints. […] Transcriptional profiles in the test set were compared against the 3D7 sample from [5].

Non-genetic features consideration

Because all samples from the test data were collected from the Thailand-Myanmar border, for now, we ignore the `Country` feature. In the future, however, perhaps we can place more weights on the training samples that are geographically closer to this region. We also ignore the `Kmeans.Grp` feature that is cluster groups corresponding to three types of transcription profile based on parasite developmental stage.

Availability

Detailed preprocessing, modeling and analysis code for this study is available at https://github.com/EpistasisLab/malaria-challenge.

References

1. **Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance**
S. Mok, E. A. Ashley, P. E. Ferreira, L. Zhu, Z. Lin, T. Yeo, K. Chotivanich, M. Imwong, S. Pukrittayakamee, M. Dhorda, … Z. Bozdech
Science (2014-12-11) https://doi.org/f3ph2f
DOI: 10.1126/science.1260403 · PMID: 25502316 · PMCID: PMC5642863

2. **Adjusting batch effects in microarray expression data using empirical Bayes methods**
W. Evan Johnson, Cheng Li, Ariel Rabinovic
Biostatistics (2006-04-21) https://doi.org/dsf386
DOI: 10.1093/biostatistics/kxj037 · PMID: 16632515

3. **The sva package for removing batch effects and other unwanted variation in high-throughput experiments**
Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, John D. Storey
Bioinformatics (2012-01-17) https://doi.org/fx9kw2
DOI: 10.1093/bioinformatics/bts034 · PMID: 22257669 · PMCID: PMC3307112

4. **STatistical Inference Relief (STIR) feature selection**
Trang T Le, Ryan J Urbanowicz, Jason H Moore, Brett A McKinney
Bioinformatics (2018-09-18) https://doi.org/gfbdwx
DOI: 10.1093/bioinformatics/bty788 · PMID: 30239600 · PMCID: PMC6477983

5. **The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum**
Zbynek Bozdech, Manuel Llinás, Brian Lee Pulliam, Edith D Wong, Jingchun Zhu, Joseph L DeRisi
PLoS Biology (2003-08-18) https://doi.org/cd68hh
DOI: 10.1371/journal.pbio.0000005 · PMID: 12929205 · PMCID: PMC176545