HLA Allele Imputation with Convolutional Neural Network

C. Chi1,2, L.F. Barcellos1

1) Genetic Epidemiology and Genomics Laboratory, UC Berkeley, Berkeley, CA; 2) Computational Biology Graduate Group, College of Engineering, UC Berkeley, Berkeley, CA

Introduction

- Human leukocyte antigen (HLA) genes in the major histocompatibility complex (MHC) encode antigen-presenting proteins within the host immune system.
- HLA alleles are highly polymorphic and many have large effect sizes in autoimmune and infectious diseases, but direct HLA typing is expensive.
- Existing HLA imputation methods have limitations due to accuracy or speed.
  - SNP2HLA: imputation not as accurate for less frequent alleles.
  - HIBAG: slow due to separate classifier for each HLA locus.
- Convolutional neural network (CNN) is suited to process data in the form of multiple arrays, including sequences such as genetic data.

Data Description

- Individuals of European ancestry from the Type 1 Diabetes Genetics Consortium (T1DGC), totaling 5,225 individuals.
- Individuals genotyped for 5,698 SNPs at the MHC with the Illumina 550K array.
- HLA alleles are highly polymorphic and many have large effect sizes in complex (MHC) encode antigen-presenting proteins within the host immune system.
- Individuals phaged at the 2-field resolution typed for HLA*A, HLA*B, HLA*C, HLA*DPA1, HLA*DPB1, HLA*DQA1, HLA*DQB1, and HLA*DRB1, totaling 296 unique HLA alleles.
- Removed 109 individuals with any missing HLA alleles.

Methods

**Data Processing:**
- Select SNPs flanking each HLA locus by ±250 kb as predictive features.
- For each SNP subsequence, transform consecutive SNPs into 5-grams (e.g. AGTCGATAGC → [AGTCG, ATAGC]).
- Construct 1-to-1 mapping between SNP 5-grams and natural numbers.

**ConvNet Architecture:**

- Embedding layer of dimension 8
- Batch normalization, 1D convolution of 64 filters of size 4, Max pool size 4
- Flatten, dropout rate 0.5
- Dense output with 32
- Softmax

- Overview of architecture
  - Embedding layer of dimension 8
  - Batch normalization, 1D convolution of 64 filters of size 4, ReLU, Max pool size 4
  - Batch normalization, 1D convolution of 64 filters of size 8, ReLU, Max pool size 4
  - Flatten, dropout rate 0.5
  - Concatenation: concatenate adjacent first-order layers for each locus
  - Dense output 32 with ReLU, dropout rate 0.5
  - Dense output softmax
  - Learning with Adam optimizer with learning rate 0.001; mini-batch of 512; early stopping with patience of 2 epochs.
  - Training (70%) and test (30%), randomly split by individuals

Results

**Comparison of imputation methods**

**Comparison of training times**

**Bootstrapped test accuracy error bars of ConvNet**

**Occlusion sensitivity analysis**

Conclusions and Future Directions

- HIBAG and ConvNet have comparable imputation accuracies, and appear more accurate than SNP2HLA.
- The ConvNet has the shortest training time of all imputation methods by as much as 0.5%.
- The imputation accuracy by HLA locus varies by at most 2%.
- Rare alleles are difficult to impute accurately.
- The ConvNet often uses SNPs around a tag SNP to learning the mapping between SNPs and HLA alleles.
- Future directions
  - Since HIBAG trains a model for each HLA locus, a fair comparison between ConvNet and HIBAG involves training a ConvNet independently for each HLA locus.
  - Robustness analysis of ConvNet performance against hyperparameters.

References