HLA Allele Imputation with Convolutional Neural Network

C. Chi[1,2], L.F. Barcellos[1]

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[1].
  - HIBAG: slow due to separate classifier for each HLA locus[2].
- Convolutional neural network (CNN) is suited to process data in the form of multiple arrays, including sequences such as genetic data[3].

Data Description

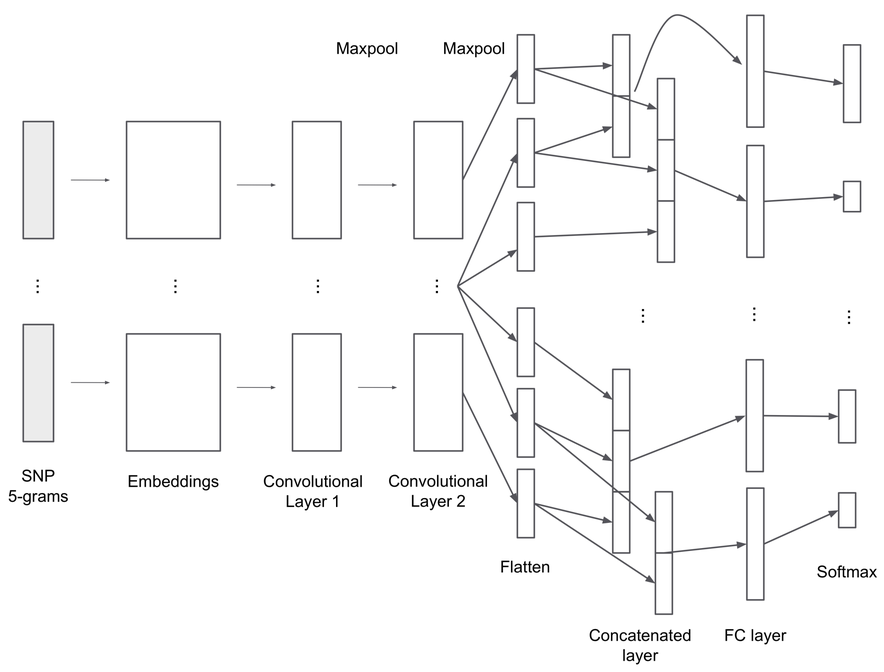
- Individuals of European ancestry from the Type 1 Diabetes Genetics Consortium (T1DGC), totaling 5,225 individuals[4].
- Individuals genotyped for 5,698 SNPs at the MHC with the Illumina 550K array.
- HLA alleles at the 2-field resolution typed for *HLA*A, HLA*B, HLA*C, HLA*DPA1, HLA*DPB1, HLA*DQA1, HLA*DQB1*, and *HLA*DRB1*, totaling 296 unique HLA alleles.
- Removed 109 individuals with any missing HLA alleles.

Methods

Data Processing:

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

Convnet Architecture:



Architecture of ConvNet, with input SNP 5-grams and softmax prediction layer for each HLA locus. The embedding layer and convolutional layers are shared between the loci, then branches off for each locus for prediction. FC layer = fully-connected layer.

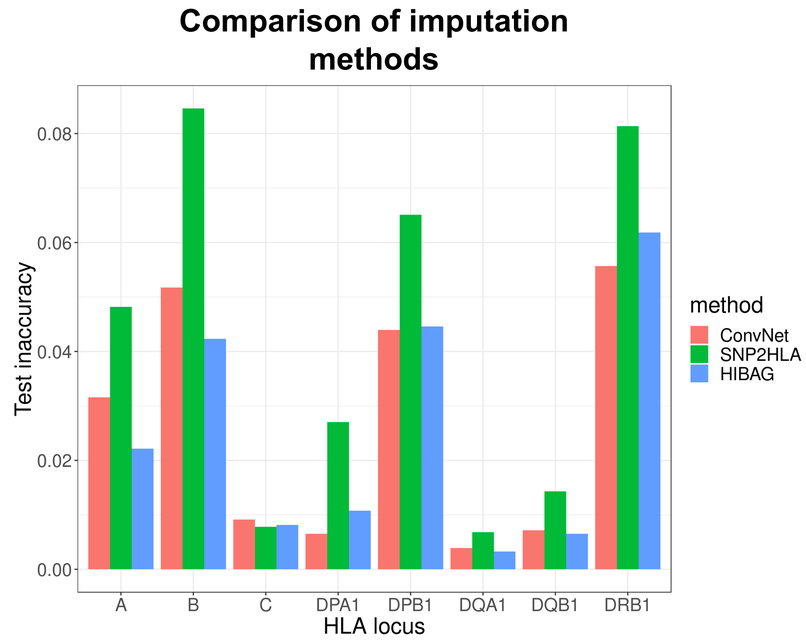
- Overview of architecture
  - Embedding layer of dimension 8
  - Batch normalization, 1D convolution of 64 filters of size 4, ReLU, Max pool size 4
  - Batch normalization, 1D convolution of 64 filters of size 8, ReLU, Max pool size 4
  - Flatten, dropout rate 0.5
  - Concatenation: concatenate adjacent 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

References

1. Jia X, Han B, Onengut-Gumuscu S, Chen W-M, Concannon PJ, Rich SS, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. Tang J, editor. PLoS One. 2013 Jun 6;8(6):e64683.
2. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG - HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2014;14(2):192–200.
3. Lecun Y, Bengio Y, Hinton G. Deep learning. Vol. 521, Nature. Nature Publishing Group; 2015. p. 436–44.
4. Rich SS, Akolkar B, Concannon P, Erlich H, Hilner JE, Julier C, et al. Overview of the Type I Diabetes Genetics Consortium. Genes Immun. 2009 Dec;10 Suppl 1:S1-4.

Results

A



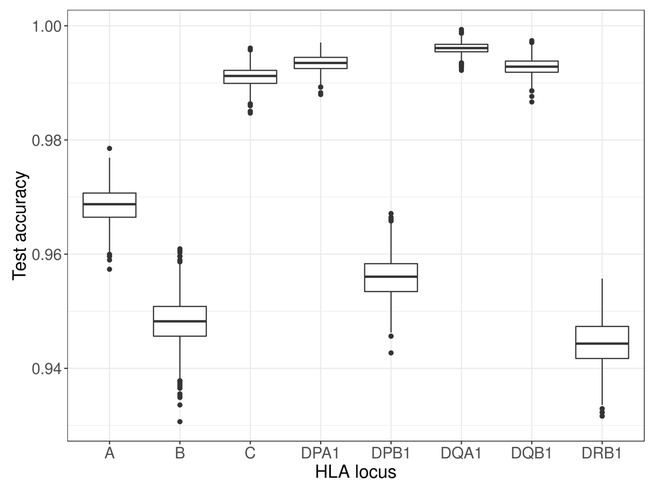
B

Comparison of training times

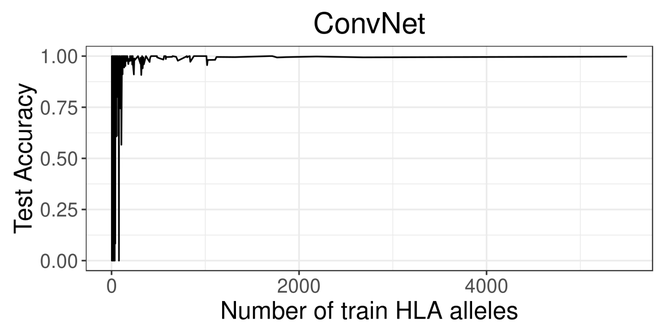
Method	Training time (minutes)
ConvNet	12 [563]
SNP2HLA	628
HIBAG	1940

*For ConvNet the minutes in brackets ”[ ]” is the phasing time for the test dataset using BEAGLE.

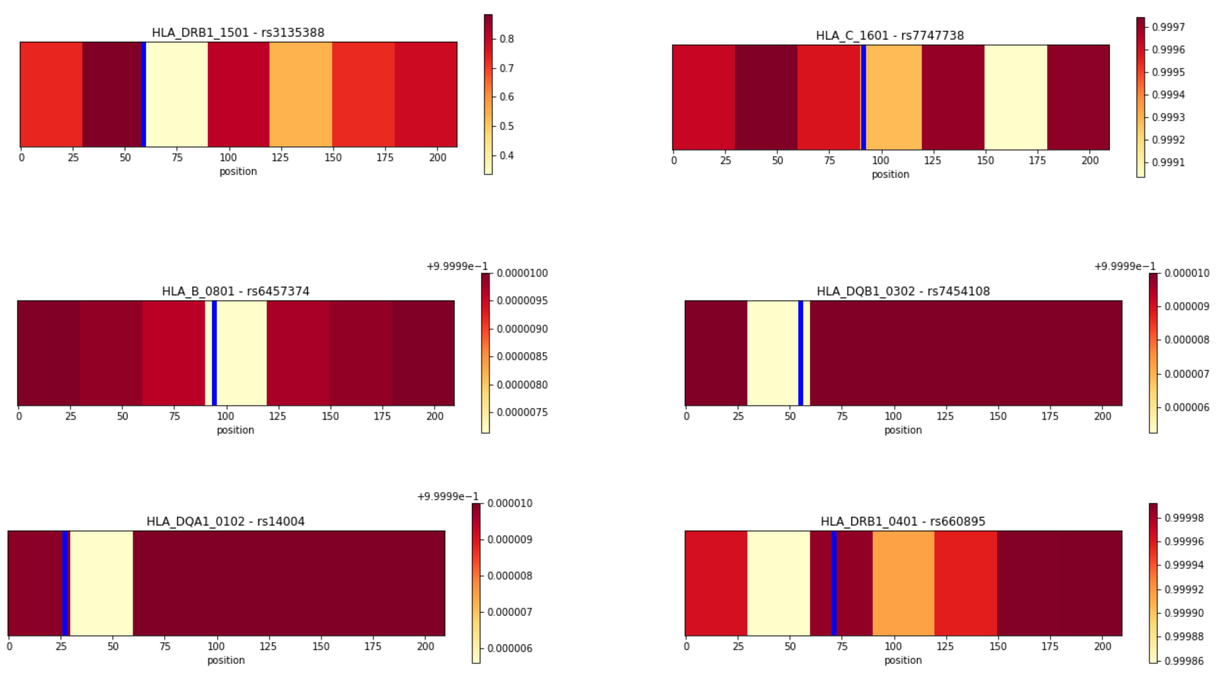
C



D



E



Sections of SNP 5-grams were independently blocked out to assess how the probability of the true class changes, the more the probability decreases, the more important the blocked out SNP 5-grams are to predicting the right allele. For each HLA allele selected, its tag SNP is marked with a blue line.

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.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship DGE 1106400.