HIBAG: an R Package for HLA Genotype Imputation with Attribute Bagging

Xiuwen Zheng

March 25, 2013

Contents

1 Overview **1**

2 Features **3**

3 Examples **3**
3.1 The pre-fit HIBAG models for HLA imputation 3
3.2 Build your own HIBAG models for HLA imputation 4

4 Resources **5**

1 Overview

The human leukocyte antigen (HLA) system, located in the major histocompatibility complex (MHC) on chromosome 6p21.3, is highly polymorphic. This region has been shown to be important in human disease, adverse drug reactions and organ transplantation [1]. HLA genes play a role in the immune system and autoimmunity as they are central to the presentation of antigens for recognition by T cells. Since they have to provide defense against a great diversity of environmental microbes, HLA genes must be able to present a wide range of peptides. Evolutionary pressure at these loci have given rise to a great deal of functional diversity. For example, the *HLA–B* locus has 1,898 four-digit alleles listed in the April 2012 release of the IMGT-HLA Database [2] (`http://www.ebi.ac.uk/imgt/hla/`).

Classical HLA genotyping methodologies have been predominantly developed for tissue typing purposes, with sequence based typing (SBT) approaches currently considered the gold standard. While there is widespread availability of vendors offering HLA genotyping services,

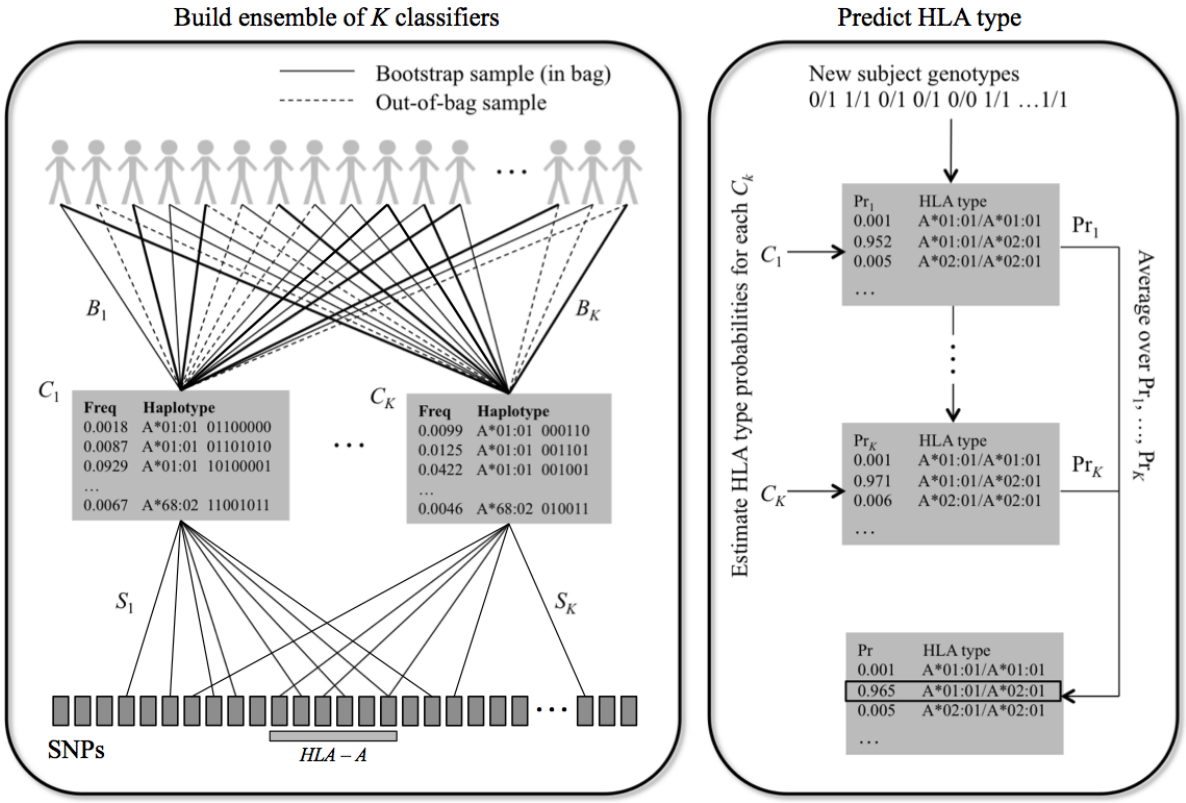

Figure 1: Overview of the HIBAG prediction algorithm. HIBAG is an ensemble classifier consisting of individual classifiers (C_k) with HLA and SNP haplotype probabilities estimated from bootstrapped samples (B_k) and SNP subsets ($\mathcal{S}_k$). The SNP subsets are determined by a variable selection algorithm with a random component. HLA type predictions are averaged over the posterior probabilities from all classifiers.

the complexities involved in performing this to the standard required for diagnostic purposes make using a SBT approach time-consuming and cost-prohibitive for most research studies wishing to look in detail at the involvement of classical HLA genes in disease. Previous studies have suggested that the existence of some HLA alleles can be predicted by a SNP-based tagging approach [3, 4]. However, SNP-based tagging does not offer a definitive solution to HLA genotyping by prediction since many HLA alleles are found on multiple haplotype backgrounds [5] that differ among populations.

Here we propose a new method for **H**LA **I**mputation using attribute **BAG**ging, HIBAG, that is highly accurate, computationally tractable, and can be used with published parameter estimates, eliminating the need to access large training samples [6]. It combines the concepts of attribute bagging with haplotype inference from unphased SNPs and HLA types. Attribute bagging is a technique for improving the accuracy and stability of classifier ensembles deduced

using bootstrap aggregating and random subsets of variables [7, 8, 9], as shown in Figure 1. In this case, individual classifiers are created which utilize a subset of SNPs to predict HLA types and haplotype frequencies estimated from a training data set of SNPs and HLA types. Each of the classifiers employs a variable selection algorithm with a random component to select a subset of the SNPs. HLA type predictions are determined by maximizing the average posterior probabilities from all classifiers. Compared to other methods like HLA*IMP and BEAGLE, HIBAG has only the minimal assumption of Hardy-Weinberg equilibrium (HWE).

2 Features

1. HIBAG can be used by researchers with published parameter estimates (http://www.biostat.washington.edu/~bsweir/HIBAG/) instead of requiring access to large training sample datasets.

2. A typical HIBAG parameter file contains only haplotype frequencies at different SNP subsets rather than individual training genotypes.

3. SNPs within the xMHC region (chromosome 6) are used for imputation.

4. HIBAG employs unphased genotypes of unrelated individuals as a training set.

5. HIBAG supports parallel computing with R.

3 Examples

3.1 The pre-fit HIBAG models for HLA imputation

```
library(HIBAG)

# Load the published parameter estimates from European ancestry
model.list <- get(load("European-HLA4.RData"))

#############################################################################
# Import your PLINK BED file
#
yourgeno <- hlaBED2Geno(bed.fn=".bed", fam.fn=".fam", bim.fn=".bim")
summary(yourgeno)

# HLA imputation at HLA-A
```

```
hla.id <- "A"
model <- hlaModelFromObj(model.list[[hla.id]])
summary(model)

# SNPs in the model
head(model$snp.id)
# "rs2523442" "rs9257863" "rs2107191" "rs4713226" "rs1362076" "rs7751705"
head(model$snp.position)
# 29525796 29533563 29542274 29542393 29549148 29549597

# plot SNP information
plot(model)

# best-guess genotypes
pred.guess <- predict(model, yourgeno, type="response")
# posterior probabilities
pred.prob <- predict(model, yourgeno, type="prob")
```

3.2 Build your own HIBAG models for HLA imputation

```
library(HIBAG)

# Import your PLINK BED file
geno <- hlaBED2Geno(bed.fn=".bed", fam.fn=".fam", bim.fn=".bim")
summary(geno)

# HLA genotypes, 01:02/02:01, 05:01/03:01, ...
train.HLA <- hlaAllele(geno$sample.id, H1=c("01:02", "05:01", ...),
    H2=c("02:01", "03:01", ...), locus="A")

# Selected SNPs, two options:
# 1) the flanking region of 500kb on each side,
#    or an appropriate flanking size without sacrificing predictive accuracy
  snpid <- hlaFlankingSNP(geno$snp.id, geno$snp.position, "A", 500*1000)
# 2) the SNPs in our pre-fit models
  model.list <- get(load("European-HLA4.RData"))
  snpid <- model.list[["A"]]$snp.id
# Subset training SNP genotypes
train.geno <- hlaGenoSubset(geno, snp.sel=match(snpid, geno$snp.id))

# Building ...
model <- hlaAttrBagging(train.HLA, train.geno, nclassifier=100, verbose.detail=TRUE)
summary(model)

# Save your model
```

```
model.obj <- hlaModelToObj(model)
save(model.obj, file="your_model.RData")

# Predict ...
# best-guess genotypes
pred.guess <- predict(model, newgeno, type="response")
# posterior probabilities
pred.prob <- predict(model, newgeno, type="prob")
```

4 Resources

1. Allele Frequency Net Database (AFND): http://www.allelefrequencies.net.
2. IMGT/HLA Database: http://www.ebi.ac.uk/imgt/hla.
3. GlaxoSmithKline (GSK): http://www.gsk.com.

References

[1] Takashi Shiina, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy Kulski. The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of human genetics*, 54(1):15–39, 2009.

[2] J Robinson, J Halliwell, H McWilliam, R Lopez, P Parham, and S Marsh. The IMGT/HLA database. *Nucleic Acids Res*, 41(Database issue):1222–1227, Jan 2013.

[3] PI de Bakker, G McVean, PC Sabeti, MM Miretti, T Green, J Marchini, X Ke, A J Monsuur, P Whittaker, M Delgado, J Morrison, A Richardson, EC Walsh, X Gao, L Galver, J Hart, D A Hafler, M Pericak-Vance, J A Todd, M J Daly, J Trowsdale, C Wijmenga, T J Vyse, S Beck, S S Murray, M Carrington, S Gregory, P Deloukas, and J D Rioux. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet*, 38(10):1166–1172, Oct 2006.

[4] I Evseeva, K Nicodemus, C Bonilla, S Tonks, and W Bodmer. Linkage disequilibrium and age of HLA region SNPs in relation to classic HLA gene alleles within Europe. *Eur J Hum Genet*, 18(8):924–932, Aug 2010.

[5] S Leslie, P Donnelly, and G McVean. A statistical method for predicting classical HLA alleles from SNP data. *Am J Hum Genet*, 82(1):48–56, Jan 2008.

[6] Xiuwen Zheng, Judong Shen, Charles Cox, Jonathan C. Wakefield, Margaret G. Ehm, Matthew R. Nelson, and Bruce S. Weir. HIBAG – HLA Genotype Imputation with Attribute Bagging. *the Pharmacogenomics Journal*, 10.1038/tpj.2013.18, 2013.

[7] Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

[8] Leo Breiman. Random forests. *Mach. Learn.*, 45:5–32, 2001.

[9] Robert Bryll, Ricardo Gutierrez-Osuna, and Francis Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36(6):1291 – 1302, 2003.