Human Immunology 77 (2016) 307–312

Contents lists available at ScienceDirect

journal homepage: www.elsevier.com/locate/humimm

HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set

Kelly Nunes [a], Xiuwen Zheng [b], Margareth Torres [c], Maria Elisa Moraes [c], Bruno Z. Piovezan [c], Gerlandia N. Pontes [c], Lilian Kimura [a], Juliana E.P. Carnavalli [a], Regina C. Mingroni Netto [a], Diogo Meyer [a,*]

[a] University of São Paulo, Department of Genetics and Evolutionary Biology, São Paulo, Brazil
[b] University of Washington, Department of Biostatistics, Seattle, WA, USA
[c] JRM-Investigações Imunológicas, Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 26 June 2015
Revised 4 November 2015
Accepted 9 November 2015
Available online 12 November 2015

Keywords:
HLA
Imputation
1000 Genomes
Admixed populations
Relatedness

ABSTRACT

Methods to impute HLA alleles based on dense single nucleotide polymorphism (SNP) data provide a valuable resource to association studies and evolutionary investigation of the MHC region. The availability of appropriate training sets is critical to the accuracy of HLA imputation, and the inclusion of samples with various ancestries is an important pre-requisite in studies of admixed populations. We assess the accuracy of HLA imputation using 1000 Genomes Project data as a training set, applying it to a highly admixed Brazilian population, the *Quilombos* from the state of São Paulo. To assess accuracy, we compared imputed and experimentally determined genotypes for 146 samples at 4 HLA classical loci. We found imputation accuracies of 82.9%, 81.8%, 94.8% and 86.6% for *HLA-A*, *-B*, *-C* and *-DRB1* respectively (two-field resolution). Accuracies were improved when we included a subset of *Quilombo* individuals in the training set. We conclude that the 1000 Genomes data is a valuable resource for construction of training sets due to the diversity of ancestries and the potential for a large overlap of SNPs with the target population. We also show that tailoring training sets to features of the target population substantially enhances imputation accuracy.

1. Introduction

Technological advances and the availability of large-scale genomic data have boosted the development of tools for the imputation of genotypes at both the genomic scale and in specific genomic regions of interest. Imputation methods combine training sets containing subjects genotyped for a high density of SNPs (single nucleotide polymorphisms) with samples of interest genotyped for only a subset of these markers. Based on population genetic models and allelic correlation measures (e.g. linkage disequilibrium), imputation methods predict unobserved genotypes from those present in the training set.

While high resolution HLA typing is still the gold standard in the field, imputation of HLA alleles is becoming increasingly used. The main advantage of HLA imputation is that it provides information on HLA variants for studies involving large samples, and for which HLA typing was not performed (e.g. many GWAS studies). The imputed HLA allele calls allow the GWAS hits to be interpreted with additional biological context [1]. For example, by analyzing GWAS SNPs with genomewide significance in the light of an individual's HLA genotype, interactions can be tested for, and confounding effects can be controlled for (e.g. specific predisposing HLA alleles which are already known). Imputation can even provide, with a high reliability, the variant an individual carries at a specific amino-acid position, and this can be included in models testing for association between genotypes and disease phenotypes [2–4].

Given the complexity and costs associated with HLA genotyping and the increasing availability of genomewide SNP data, over the last years several methods have been developed with the goal of imputing the HLA alleles based on dense SNP data for the MHC region [3,5–7]. This is a challenging task, considering the large number of alleles of HLA genes, which makes methods more effective when: (a) the training set consists of a large number of samples [3,5]; (b) there is a suitable pairing among the population(s) that make up the training set and the sample of interest [8,9]; (c) the HLA alleles being queried in the target population are not rare.

* Corresponding author at: Departamento de Genética e Biologia Evolutiva, Rua do Matão, 277, São Paulo, SP 05508-090, Brazil.
E-mail address: diogo@ib.usp.br (D. Meyer).

http://dx.doi.org/10.1016/j.humimm.2015.11.004
0198-8859/

Choosing a suitable training set is critical to the success of the imputation methods. However, due to the high cost, it is not always possible to generate a training set tailored for the specific target population under study, so imputation is commonly made using public datasets as training sets such as the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) and the British 1958 birth cohort of Welcome Trust Case Control Consortium (http://www.ebi.ac.uk/ega/) [7]). The use of such resources can also be a challenge, since public datasets do not always have the populations related to those in the target sample, for which imputation is to be performed. This is especially critical for admixed populations, such as those from the Americas who carry Native American ancestry, which is underrepresented in public datasets. Because of the difficulties in obtaining Native American samples, an alternative is to use other admixed populations to make imputations for this ancestry component.

In recent years, one of the most widely used public resources for population genetic studies is the 1000 Genomes Project [10]. Phase I of the 1000 Genomes Project provided mainly low coverage sequencing data for two African, five European, three Asian, and four admixed populations from the Americas (African–Americans, Mexicans, Puerto Ricans and Colombians). These samples were recently genotyped at high resolution for the classical HLA genes [11], providing a valuable resource which integrates genomewide SNP data with HLA allele calls [12].

In the present study we examine the accuracy of HLA imputation in a highly admixed Brazilian population (with 40% African, 39% European and 21% Native American average ancestries [13]) using the 1000 Genomes HLA and SNP data as a training set. Our interest is motivated by the importance of admixed populations in studies with a focus on admixture mapping (e.g. [14]) and in understanding the role of introgression involving HLA genes (i.e. the observation that ancestry proportions in the HLA region deviate from genomewide averages for admixed populations, [15,16]).

In this study we do not intend to compare the performance of different HLA imputation methods, as others have done before (eg. [7,8]). Rather, we assess the performance of the 1000 Genomes data as a training set for imputation of highly admixed populations, and explore how the quantity of SNPs and ancestry of the individuals in the training set impacts imputation accuracy. We perform imputation using HIBAG [7], an ensemble classifier that has been shown to provide accurate imputation, and for which imputation models can be built using training sets of choice.

We find that the 1000 Genomes data provides HLA imputation of 83–94% accuracy at the two-field level. We compare imputation accuracy to that obtained when other training sets are used, or when individuals which are related to the target sample are included in the training set. Finally, we discuss how SNP density and geographic origin of populations making up the target sample contribute to imputation accuracy, in the context of an admixed population.

2. Materials and methods

2.1. The Brazilian admixed sample

We imputed HLA genotypes for highly admixed samples from Brazilian communities known as "*Quilombos*" from Vale do Ribeira region, São Paulo State. These were founded by runaway, abandoned and free slaves in the 18th century, and established in remote areas in the Atlantic Rainforest of Southeastern Brazil, where they subsequently admixed with Native Americans, adding a third ancestry component, in addition to African and European (Table S1). A total of 365 samples (referred to as the "QUI dataset") were genotyped using the Affymetrix Axiom Human Origins Array (600K SNPs), and a subset of 146 individuals were experimentally genotyped at HLA loci using PCR-SBT (Thermo Fisher) for *HLA-A*, *-B*. *-C* (exons 2, 3 and 4) and *-DRB1* (exon 2). The ethics committee of the *Instituto de Biociências da Universidade de São Paulo* approved this study and informed consent was obtained from all participants.

2.2. Data for training set using 1000 Genomes Project samples (1000g)

We selected 931 samples from the 1000 Genomes Project for which SNP [10] and HLA genotypes [11] were available: 126 African, 317 European, 265 East Asian, and 223 admixed samples from the Americas (53 African-American, 60 Colombian, 55 Mexican and 55 Puerto Rican; Table S2). The SNP data was mainly of low coverage genotype calls ([10]; ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/), and HLA typing was generated by sequence-based typing (PCR-SBT) for *HLA-A*, *-B*, *-C* and *-DRB1* genes ([11]; data available at (http://www.ncbi.nlm.nih.gov/gv/mhc/xslcgi.fcgi?cmd=cellsearch).

2.3. Training set of Zheng et al. [7] (UW)

To place our result in the context of previous studies, we also used the multi-ethnic training set specific to the Affymetrix Axiom Human Origins Array platform assembled by Zheng et al. [7], which consists of 2 different datasets (HAPMAP Phase 2 and HLARES) and includes more than 3000 samples (details in [7]) (Table S3).

2.4. Data cleaning and SNP selection for imputation analysis

We filtered the Quilombo (QUI) SNP dataset for genotype quality using R Package GWASTools [17]. We selected a total of 1238 SNPs that flanked the *HLA-A*, *-B*, *-C* and *-DRB1* genes in 500 kb windows. For the UW dataset, the 500 kb windows resulted in a set of 467 SNPs.

2.5. Building a multi-ethnic model for HLA allele imputation

HLA imputation was performed with Attribute Bagging, implemented in the HIBAG program, which averages over many classifiers (obtained by 100 bootstrap resamplings) to define HLA alleles with highest posterior probabilities [7]. This method has proven to be robust in a previous study with another admixed population [8]. We used HIBAG to build multi-ethnic models for *HLA-A*, *-B*, *-C* and *-DRB1*, with parameters to build the models used according to recommendations of the HIBAG authors [7]. We built three models (for both one and two field resolution), based on three different training sets: (a) 1000g; (b) 1000g with the inclusion of an additional set of 57 unrelated Quilombo samples (1000g+QUI); and (c) UW.

Details on how the unrelated individuals were selected to be added to the 1000g set are presented in Section 2.7. The models used in this study are available for download at www.ib.usp.br/genevol (1000g, 1000g+QUI) and www.biostat.washington.edu/~bsweir/HIBAG/ (UW model used in HIBAG).

2.6. Quantification of imputation accuracy

To assess the accuracy of imputation at each locus, we quantified the number of chromosomes with correctly called HLA alleles over the total number of imputations made (corresponding to 292 chromosomes for which experimentally generated HLA calls were available). We did not require a minimum posterior probability (implying a call threshold of 0%). For 1000g+QUI we adopted the