CALCULATING HAPLOTYPE FREQUENCIES, HLA DIVERSITY AND DONOR UNIQUENESS IN A LARGE MIXED-RESOLUTION DATASET

Jeremy E Stein [1], Hazael Maldonado Torres [1], James Robinson [1, 2], Steven GE Marsh [1,2]

1 Anthony Nolan Research Institute, London, United Kingdom, **2** UCL Cancer Institute, London, United Kingdom

INTRODUCTION

The population genetics of HLA is challenging due to the high level of variation in human populations. In addition, most registry datasets contain HLA typings generated by different techniques, from serology to the latest sequencing technologies. Further to this, a number of these techniques may produce ambiguous typing "strings". For some samples there may be no data recorded for some loci: HLA-C, -DRB1 or -DQB1. Analysing datasets of mixed-resolution and with missing data is therefore a complex problem.

These datasets can be used to generate haplotype frequencies. Accurate haplotype frequencies form the basis of many search and prediction algorithms used by donor registries. Calculating haplotype frequencies is important because they can be used to both predict the missing data and resolve the ambiguous HLA types. The haplotype frequencies can also be used to resolve phase ambiguity, where the haplotypes a person possesses are unknown. This occurs even in the highest resolution data, derived using the latest sequencing techniques. The challenge is that the size of dataset required to provide accurate haplotype frequencies often contains mixed-resolution data and missing values.

The established method for calculating haplotype frequencies is Expectation-Maximisation (EM) (1), but when used on a large, ambiguous and incomplete dataset it requires careful implementation to avoid the exponential growth of memory and CPU usage that would result from using a "brute force" approach. This is because a brute force approach would generate every possible pair of haplotypes given the ambiguous input. However, in large datasets with even a small amount of low-resolution or missing typing, this would quickly exceed memory and computational time limits. An alternative method is required that permits the analysis to be completed within reasonable time and without exceeding available memory.

METHODS AND MATERIALS

Haplotype frequency estimation

Various techniques have been used to avoid the problems that would be caused by a brute force approach (2, 3). The most important is to add loci one at a time (or hierarchically), with trimming of low probability diplotypes after every step. In the latest version of the Anthony Nolan software, Cactus, the analysis can be run on very large datasets without splitting, reducing the scope for sampling error and improving the prediction of missing types. The latest method was developed using both Python 3 and C++, and the individual steps are controlled using the Snakemake workflow engine (4). Figure 1 shows an overview of the process.

Validation of haplotype frequency estimation

Synthetic datasets of 5,000,000 individuals, unambiguously typed over 5 loci, were generated using known haplotype frequencies for several different populations. These were analysed using Cactus, producing the original haplotype frequencies in return, with only small differences due to sampling error. Using 16 cores, these datasets took approximately 30 minutes to complete.

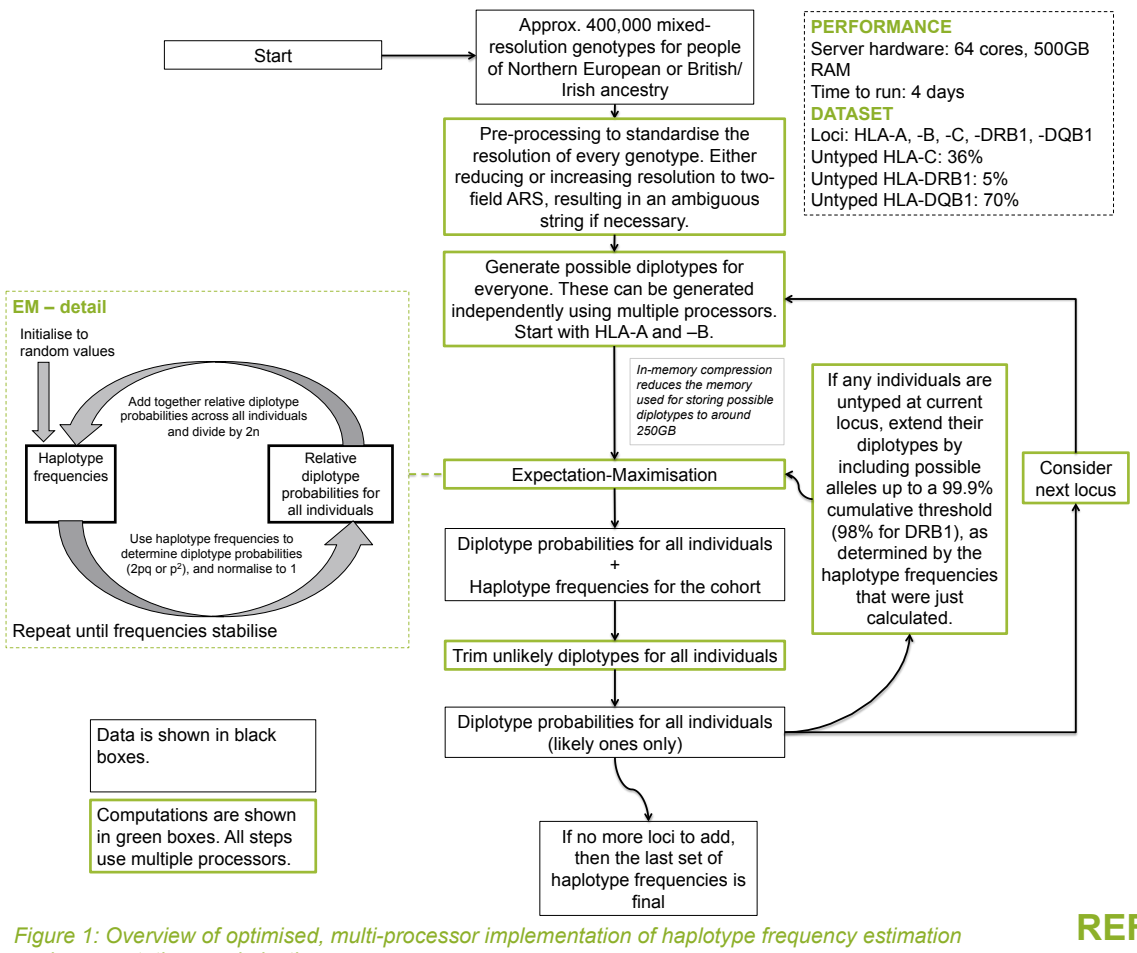

Figure 1: Overview of optimised, multi-processor implementation of haplotype frequency estimation by expectation-maximisation

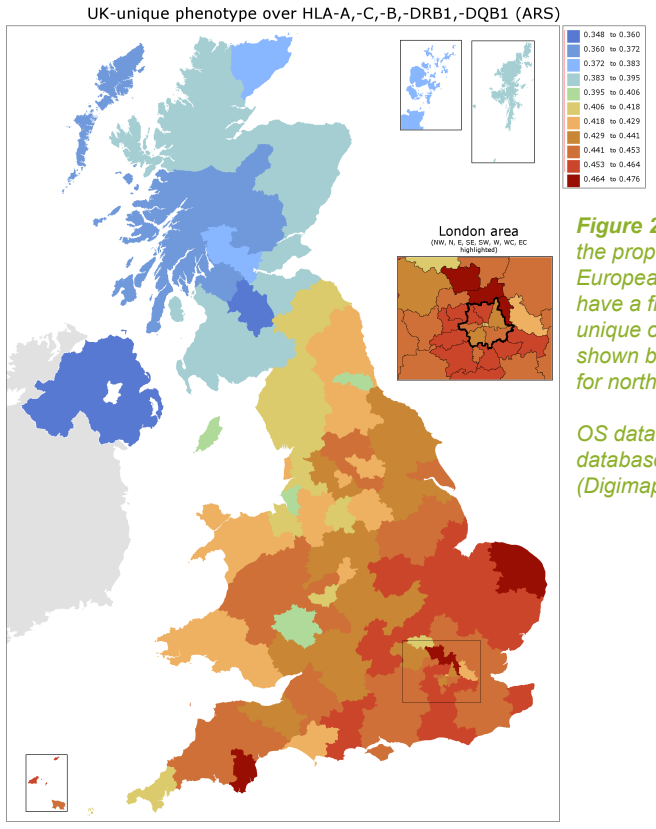

Figure 2: *Area of Differentiation Ratio; the proportion of donors of Northern European or British/Irish ancestry who have a five locus HLA phenotype that is unique on the Anthony Nolan register, shown by UK postcode area (e.g. "NW" for north-west London)*

OS data: © Crown Copyright and database rights 2007. Ordnance Survey (Digimap Licence)

INTERPRETATION AND FURTHER USES OF RESULTS

The latest optimisations of the EM method were used to analyse a dataset of approximately 400,000 individuals on the Anthony Nolan register who describe themselves as of Northern European or British/Irish ancestry. This dataset contains both mixed-resolution data and missing values. From this data, the software predicted the Antigen Recognition Site (ARS) resolution phenotypes (for HLA-A, -B, -C, -DRB1, -DQB1) for each individual. As shown in figure 1, the EM method produces diplotype predictions for all individuals, and these can be reduced down to phenotype predictions by ignoring the phasing information and summing the frequency for diplotypes that map to identical phenotypes. This can be used to determine which individuals are unique on the register.

Since there can be several candidate phenotypes for each person, these are first resolved using the method described by Gragert *et al* (5) to produce simulated actual phenotypes for all individuals. For example, if a person could have two phenotypes with an 80:20 probability split, then when generating a set of simulated phenotypes, there is an 80% chance that the first phenotype will be selected, and a 20% that the second will be. To account for the random nature of this process, the simulation is run multiple times and the analysis is run on each simulated dataset. These must then be combined, in this case by taking an average of the number of unique people found in each dataset. This information, when combined with postcode information, allows a map to be drawn showing what proportion of people on the Anthony Nolan register are unique in each UK postcode area, as shown in figure 2. This showed that in donors of Northern European or British/Irish descent, there is a gradual trend of increasing diversity moving from the north-west to the south-east of the British Isles: in Northern Ireland and Scotland, a lower proportion of donors have five-locus phenotypes that are unique in the UK, when compared to southern and eastern England.

CONCLUSIONS

The latest optimisations of the EM method implemented in Cactus have produced haplotype frequencies for a dataset of approximately 400,000 people on the Anthony Nolan register who describe themselves as of Northern European or British/Irish ancestry. Using the established Expectation-Maximisation technique as a basis for the algorithm, the latest implementations make extensive use of multiple processors without exceeding the memory available to a large server. These haplotype frequencies can be used for predicting the HLA phenotype of people with low-resolution or missing types, and those predictions are more accurate if the frequencies come from the same population as that person. Including as many people as possible in this analysis reduces the effect of sampling error when calculating haplotype frequencies and is especially important when dividing the dataset into small regions, such as in the analyses presented here.

REFERENCES

1. Excoffier, L, & Slatkin, M. (1995). Molecular Biology and Evolution, 12, 921-927.
2. Gragert, L, Madbouly, A, Freeman, J, & Maiers, M. (2013). Human Immunology, 74, 1313-1320.
3. Stein, J. E, Maldonado Torres, H, Robinson, J, Marsh, S. G. E. (2014). International Journal of Immunogenetics, 41, 431.
4. Köster, J, & Rahmann, S. (2012). Bioinformatics, 28, 2520-2522.
5. Gragert, L, Albrecht, M, DiPrima, S, Maiers, M. (2014). Tissue Antigens, 84, 29.

Reg charity no 803716/SC038827 978PA/0416

None of the authors have any conflict of interest to declare