The relationship of identity by state to identity by descent and imputation accuracy in population sequencing data

Kelley Harris
Emmanuel College
University of Cambridge

Dissertation submitted for the degree of Master of Philosophy

March 29, 2011

PREFACE

This dissertation is my own work. The research was not done in collaboration except where marked explicitly in the text.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Richard Durbin, for suggesting these research questions and providing so much guidance through the process of answering them and writing up this report. He and Heng Li, who created the genome call sets used in Section 7, will be my coauthors on a manuscript based on the work described in this dissertation. Besides thanking Heng for providing the data I needed to complete the project, I would like to thank everyone in the Durbin group for welcoming me and teaching me so much over the past year. In particular, Kimmo Palin and Jared Simpson helped me learn to use the Sanger computer system, and Kimmo, as well as Aylwyn Scally, helped me edit previous versions of this writeup. In writing and researching, I also benefitted from the helpful input of my thesis committee: Carl Anderson, Vincent Plagnol, Chris Tyler-Smith, and Eleftheria Zeggini, and would like to thank Gilean McVean and Jeffrey Barrett for taking on the finished product to read. I am grateful for the financial support I received from the Harvard College Herchel Smith Postgraduate Fellowship program, and to the Wellcome Trust Sanger Institute for hosting me. Finally, I'd like to thank my friends for always supporting me and my parents, Glenn Harris and Anne Katten, for doing that and so much more.

Abstract

When two DNA sequences look *identical by state* (IBS) along a string of genotyped markers, the DNA between the markers is often *identical by descent* (IBD), meaning inherited from a recent common ancestor without recombination. This fact makes it possible to scan the whole genome for functional variants without typing every base directly, making use of information about unobserved bases that is provided by the states of observed bases. We attempt to quantify that information here, using coalescent theory to predict how strongly various degrees of IBS imply IBD, taking into account the density of genotyped markers and past effective population size.

In addition to calculating the probability of IBD between IBS haploid sequences, we consider the problem of matching an unphased diploid sequence to a reference haplotype panel. The results have bearing on the practices of haplotype phasing and genotype imputation, both of which become more reliable when the ends of an IBS alignment are not assumed to be IBD. To compute p(IBD|IBS) when phasing ambiguity is an issue, it was necessary to develop a new approximation to the neutral coalescent with recombination: a further simplification of the sequentially Markovian coalescent [43].

Computing p(IBD|IBS) by our method is closely related to predicting the length distribution of homozygous stretches in the genome. After accounting for sequencing errors, we predict this distribution correctly, as judged by data from eleven complete human genomes. We also predict the length distribution of segments that appear homozygous based on thinned marker data, noting that the probability of sequence IBS given "thinned IBS" is a natural measure of imputation accuracy.

The probability of IBS given thinned IBS varies with sequence length in a way that is very ethnically distinctive, as judged by data from five Africans, four Europeans, and two Asians. We are able to account for these differences in terms of past changes in effective population size: an out-of Africa bottleneck followed by a shallower, more recent Asian bottleneck. We predict that IBS implies IBD most strongly in historically

outbred populations, and that extra care should be taken when inferring IBD in bottlenecked populations.

Returning to the problem of imputation, we estimate the accuracy spread of the imputation calls that can be made from a panel of n reference haplotypes. For an idealized population of effective size $N = 10,000$, we find that the 120-reference HapMap should omit a significant amount of genetic variation; that given a 1-kilobase stretch of a genotyped individual's DNA, a 120-reference panel gives us only a 70% chance of imputing the sequence of that stretch with 99% accuracy. However, we find that a 1000-haplotype panel should enable near-perfect imputation in a population that has been isolated from recent exponential population growth, and such perfect imputation would allow for precise genetic mapping in groups much larger than extended families.

Contents

1 Introduction 7

2 Computing the probability of identity by state 15

2.1 Constant effective population size 15

2.2 Non-uniform mutation and recombination 19

2.3 Correcting for changes in effective population size 19

3 The age distribution of maximal IBD segments 20

4 The probability of IBD given diploid IBS with uncertain haplotype phasing 23

5 Using identity by state to phase and impute haplotypes 34

5.1 The probability of IBD in the central subset of an IBS alignment 36

6 The effect of underestimating the linkage between markers when computing IBS probabilities 39

7 Empirical validation using genome sequence data 44

8 Discussion 55

A Base-calling methods 59