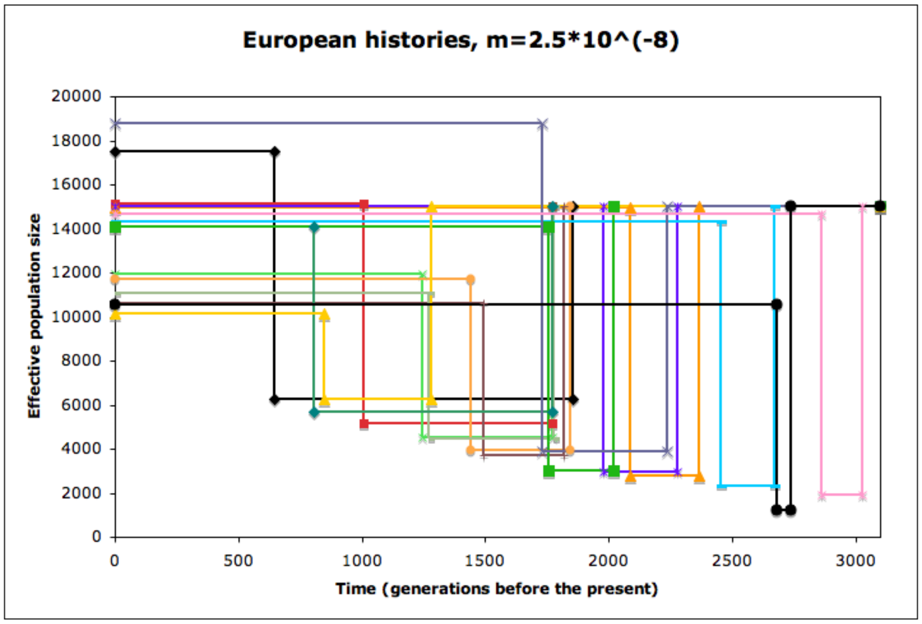

Figure 21: A population bottleneck is required to fit the CEU ROH data–either a shallow, recent bottleneck or a deeper, more ancient bottleneck.

Asian HapMap allele frequency spectrum shows more evidence of genetic drift than the European HapMap allele frequency spectrum [30].

It remains an open problem to mathematically describe the set of histories that fit the distribution of ROHs in each ethnic group. However, in showing that such histories exist, we achieve our aim of predicting the length distribution of ROHs in real genome data and validating the theory that we use to compute IBD probabilities.

8 Discussion

In this paper, we attempt to very precisely model patterns of linkage disequilibrium in genetic data, capturing its decay over long regions of the genome instead of assuming that certain blocks or pairs of loci assort independently. Rather than adding a new LD model to the myriad that exist already, we work

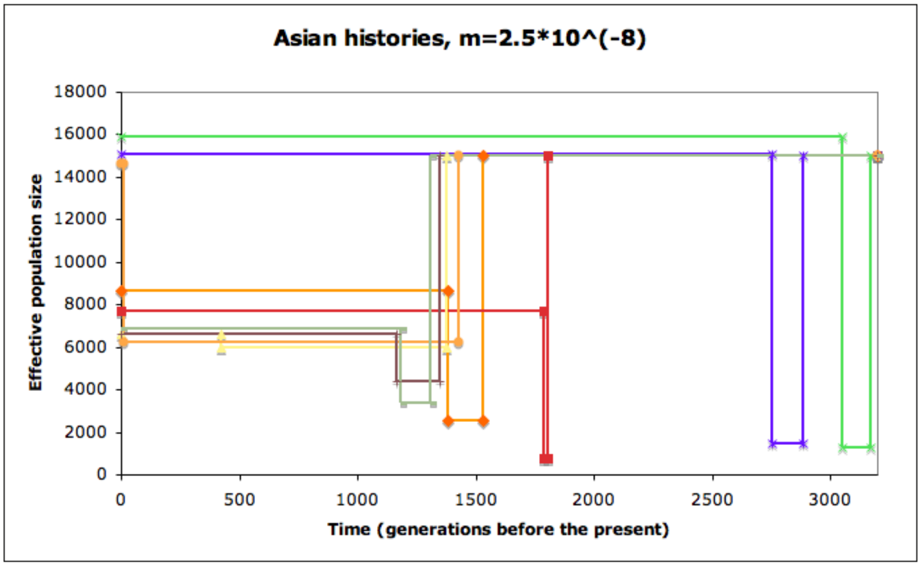

Figure 22: A bottleneck is required to fit Asian ROH data. These good fit histories have smaller recent effective population sizes than the histories that fit the CEU data (see Figure 21).

as closely as possible to the neutral coalescent with recombination, deriving results within a theoretical framework whose advantages and limits are well understood already. Our claims about patterns of homozygosity in the genome make no assumptions that weren't published as part of the sequentially Markovian coalescent (SMC) model [43]; while we do not account for recombination hotspots or other point irregularities, the results should be quite accurate across the genome as a whole, as seen empirically in Section 7.

While it was necessary to relax the SMC somewhat to model LD in diploid alignments with uncertain phasing, our model approaches the coalescent in the limit of high LD where phasing and imputation attain high accuracy. While it is possible to check our computations with a coalescent simulator such as MS, our method makes it much quicker to compute a length spread of IBS probabilities; without it, each data point on each of our plots would have to be obtained from a separate coalescent simulation with memory and storage requirements that get quite large for long IBS alignments. Efficiency was key when we had to compare many population histories to find a matches for the African, European, and Asian homozygosity data, and also when we considered the best alignment between a test sequence and some member of a reference panel.

The accuracy values that we compute for various imputation panels assume an idealized population of effective size 10,000, closer to the truth for Africans than for non-Africans. However, it is possible to condition instead on any piecewise-constant population size history, as is briefly explored in the last section of the paper, if e.g. the goal is to impute only Europeans. We concentrate less on evaluating the capabilities of an existing panel than on building a framework for making informed decisions about the design of future panels.

Our results do posit lower bounds on the amount of variation that should not be imputable using HapMap, as we estimate that a 120-reference panel has only a 70% chance of imputing a one kilobase mini-haplotype with 99% accuracy and an 80% chance of imputing it with 90% accuracy. Accuracy is likely higher when the aim is to impute only common SNPs, but it can only be higher insofar as those common SNPs fail to tag the rarer SNPs around them. Our rough estimate of HapMap variation coverage is close to that obtained by Bhangale, et al., who resequenced 1.6 megabases in each of the HapMap individuals and reported that only 60-80% of the SNPs they found were within $r^2 > 0.8$ of a tag SNP [8].

Besides showing that it is feasible to study imputation with coalescent theory, we hope that the tricks and shortcuts we've developed might help make coalescent theory more applicable to other hard problems. Important as imputation is, it is not the only setting where understanding IBS could be useful. As seen in Section 7, real hets are distributed differently from false positive hets, such that hets scattered in regions of high homozygosity are very likely to be false positives. It is easy to compute the false positive probability of a het that is L bases away from the nearest het, and it might be useful to incorporate this result into base calling software. A long stretch of near-homozygosity provides a lot of evidence that the region is IBD (see Figure 8 for a plot of p(IBS) versus length), and there is less than a 1.5% chance that an IBD region will contain seven or more mismatches (the number of mismatches being Poisson-distributed with a mean of $\mu/\rho = 2.5$). When the sequencing error rate is 10^{-5} and a 10^7-base region contains about 100 scattered hets, the likely truth is that the region is IBD and most of those hets are errors.

A challenge for the future will be to address the effect of recent exponential population growth. Only the frequency of the longest IBS regions should be affected, but this will be enough to make a 1000-haplotype reference panel less perfect than it appears in figures 10 and 11; Ionita-Laza, et al. predict that more than 3,000 individuals will be needed to find all variants with frequency 0.1% based on allele frequency data from the outbred CEU and YRI sequence data [26]. One way around this issue would be to conduct GWASs in moderately isolated populations where exponential growth has been very recent and 1000 references do constitute a perfect imputation panel. Family-based linkage studies have been successful for some time at discovering functional variants that affect few people but shed valuable light on disease mechanisms, and our results suggest that IBD mapping need not be limited to groups as small as families. A project the size of 1000 Genomes [1] should make it possible to essentially know the sequence of every individual in a small population like Iceland, which would, in principle, make it possible to test for functionality all across the genome, including at a large pool of rare and untaggable SNPs that are invisible even in GWASs conducted with the help of imputation from HapMap.

A Base-calling methods

The sequences used to validate our method were re-called with the hope of reducing the frequency of errors. The call sets are available for download at ftp://ftp.sanger.ac.uk/pub/rd/humanSequences.

Raw Illumina read data were obtained from NCBI's sequence read archive (see Table 13) and EMBL's European read archive. They were aligned by BWA (0.5.5), using human reference genome build 36, which masks pseudoautosomal regions on Y by including unassembled contigs and the Epstein-Barr virus genome (AC:NC_007605). Low quality bases were trimmed from the 3'-ends of Illumina short reads by applying BWA option '-q 15.' Because SJK base qualities were overestimated, they were recalibrated using the Genome Analysis Toolkit after discarding SNPs known from dbSNP-129. Default BWA-SW algorithms were used for capillary reads.

The 'pileup' command of the *samtools* software package was used to create each autosome's diploid consensus. The consensus was then filtered, the following kinds of bases being marked low-confidence calls:

1. Read depth is more than twice or less than half of the depth estimated at loci genotyped in HapMap3

2. Reads covering the locus have root mean square mapping quality below 25

3. There is a predicted short indel less than 10 base pairs away

4. Inferred consensus quality is below 20 (Illumina data) or 10 (capillary

data)

5. Out of the 35 reference sequence 35-mers that overlap with the site, fewer than 18 can be mapped elsewhere with at most one mismatch

References

[1] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." *Nature* (2010) 1061–1073.

[2] C.A. Anderson, D. Brocklebank, and A.P. Morris. "A comparison of reference panels for imputation in genome-wide association studies." *The European Journal of Human Genetics*, to appear.

[3] C.A. Anderson, F.H. Pettersson, J.C. Barrett, J.J. Zhuang, J. Ragoussis, L.R. Cardon, and A.P. Morris. "Evaluation the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms." *The American Journal of Human Genetics* (2008) 112–119.

[4] S.-M. Ahn, T.-H. Kim, S. Lee, et al. "The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group." *Genome Research* 19(9) (2009) 1622–1629.

[5] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dougan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, and R. Wooster. "THe COSMIC (Catalogue of Somatic Mutations in Cancer) database and website." *British Journal of Cancer* 91 (2004) 355–358.

[6] J.C. Barrett and L.R. Cardon. "Evaluating the coverage of genome-wide association studies." *Nature Genetics* 38(6) (2006) 659–662.

[7] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, et al. "Accurate whole genome sequencing using reversible terminator chemistry." *Nature* 456 (2008) 53–59.