Inferring Piecewise Ancestral History from Haploid Sequences

Russell Schwartz[1], Andrew G. Clark[2], and Sorin Istrail[3]

[1] Celera Corp.
Rockville MD 20850, USA
present address: Department of Biological Sciences
Carnegie Mellon University
Pittsburgh, PA 15213, USA
russells@andrew.cmu.edu
[2] Department of Molecular Biology and Genetics
Cornell University
Ithaca, New York 14853, USA
ac347@cornell.edu
[3] Applied Biosystems
Rockville MD 20850, USA
Sorin.Istrail@celera.com

Abstract. There has been considerable recent interest in the use of haplotype structure to aid in the design and analysis of case-control association studies searching for genetic predictors of human disease. The use of haplotype structure is based on the premise that genetic variations that are physically close on the genome will often be predictive of one another due to their frequent descent intact through recent evolution. Understanding these correlations between sites should make it possible to minimize the amount of redundant information gathered through assays or examined in association tests, improving the power and reducing the cost of the studies. In this work, we evaluate the potential value of haplotype structure in this context by applying it to two key subproblems: inferring hidden polymorphic sites in partial haploid sequences and choosing subsets of variants that optimally capture the information content of the full set of sequences. We develop methods for these approaches based on a prior method we developed for predicting piece-wise shared ancestry of haploid sequences. We apply these methods to a case study of two genetic regions with very different levels of sequence diversity. We conclude that haplotype correlations do have considerable potential for these problems, but that the degree to which they are useful will be strongly dependent on the population sizes available and the specifics of the genetic regions examined.

1 Introduction

Since the release of draft consensus human genome sequences [8, 24], much attention has turned to studying the variations in the genome that distinguish one

S. Istrail et al. (Eds.): SNPs and Haplotype Inference, LNBI 2983, pp. 62–73, 2004.

person from another. These variation occur predominantly in the form of single nucleotide polymorphisms (SNPs) at which a single DNA base pair has two common variants in the population. It is hoped that characterizing these variations and correlating them with phenotype through case-control association studies can assist in locating genes or specific genetic variants that influence human disease risk. Association-based studies are likely to allow much finer-scale mapping than was possible with the traditional pedigree-based linkage studies (See, for example, Cardon and Bell [2]). Furthermore, they are potentially much better suited to tracking down genetic influences on common, complex diseases, which are believed to have substantial genetic components but for which these components are obscured by strong environmental influences and the likely interaction of many distinct genetic factors [19]. Despite some successes [13], however, these benefits so far remain largely hypothetical.

One leading approach to improving the power of these studies is to rely on correlations between physically close SNPs by grouping SNPs into conserved haplotypes. By performing association studies on these haplotypes instead of individual SNPs, we can greatly reduce the number of distinct hypotheses being considered in an association study, allowing us to lower our standards of proof for each hypothesis and thereby increase the power of the study. Some statistical approaches have attempted to incorporate haplotype structure directly into association tests for this purpose [14, 21, 15, 12]. An alternative strategy is to characterize the haplotype structure in a recombining population independently of its application to a particular association test, a strategy that has been pursued from many directions [11, 7, 26, 27, 25, 28, 3, 20, 23].

One widely adopted approach to the direct characterization of haplotype structure stems from a seminal study by Daly et al. [4], which suggested that the human genome could be decomposed into segments of low haplotype diversity separated by regions inferred to be frequent sites of recombination. Locating the "haplotype blocks" of low diversity would allow one to reduce the complexity of the inference problem by working with haplotype block alleles instead of individual polymorphic sites. In addition, the use of "haplotype tagging" SNPs, which are subsets of SNPs in a block adequate to characterize a large fraction of its population diversity, could potentially significantly reduce the cost of conducting association studies without substantially hurting their power [10, 29]. A variety of methods have been proposed for defining optimal block decompositions [4, 10, 18, 9, 6]. Optimal block decompositions can be efficiently computed for a broad class of objective functions, a task that can also simultaneously yield minimal informative SNP subsets for the given block decomposition [30]. Block decompositions also yield straightforward algorithms for inferring missing sites based on best-matching alleles within each block.

While these discrete block decompositions are algorithmically convenient, though, they do not capture all of the available information that might be useful for downstream analyses. There are often correlations between consecutive blocks [6], suggesting the presence of longer-range information than is captured by the block decompositions. Similarly, there may be finer structure the decom-

positions obscure. These additional sources of information can be expected to be much more pronounced when rarer alleles are included than in the Daly et al. and Gabriel et al. studies, as may be required for fine-scale mapping of genetic influences on common diseases. These factors argue for examining the feasibility of approaching the downstream problems using haplotype inferences that do not rely on discrete block decompositions.

In this work, we assess the value of block-free predictions of shared haplotype ancestry in characterizing the information content of haploid sequences. We accomplish this by applying a prior method of ours for block-free piecewise inference of ancestry [20] to two key sub-problems in the design and analysis of inference studies: inferring missing data and choosing informative SNP subsets. Inferring missing data based on sequence context provides a good test of our general ability to detect and apply statistical correlations found through haplotype context information. Locating reduced subsets of SNPs that allow us to characterize the remaining missing sites with high accuracy gives us an approximate idea of how much we can hope to reduce assay sizes while still adequately capturing the available sequence diversity. We develop simple methods for both problems and evaluate their performance using cross-validation studies of two real genetic datasets. We conclude that haplotype data provides a considerable amount of information usable for such problems but that there may be significant limits to what can be accomplished given reasonable sizes of population samples, depending on the specific genetic region examined.

2 Methods

2.1 Predicting Sequence Ancestry

We predict sequence ancestry using a method introduced in Schwartz et al. [20]. That work presented the problem of ancestry inference in terms of what we call "haplotype coloring," coloring each site of a sequence so as to indicate from which of a set of ancestral sequences it is most likely to have descended at each polymorphic site. We presented two methods for the problem. The first, which was simultaneously introduced by Ukkonen [23], predicts ancestry by building on block decompositions. That method first finds a block decomposition for a set of sequences, then joins alleles in adjacent blocks using a maximum matching algorithm. The second method, which forms the basis of the present work, uses a restricted hidden Markov model (HMM) to represent the possible ways of combining ancestral sequences to yield a modern population.

The basis of the HMM coloring method is optimization of a function expressing the probability of generating an observed sequence and coloring given site-specific frequencies for ancestral haplotypes. To generate a modern sequence under this model, we first sample among all possible ancestral sequences, each of which has a characteristic starting probability. At each subsequent site, we either continue with the current sequence or undergo a recombination event, with a global uniform probability of recombination. If a recombination event

occurs between two sites, then we resample among all potential ancestral sequences according to the characteristic site-specific frequency for each sequence to chose the ancestor at the next site. We also allow, at each polymorphic site, a uniform mutation probability of the generated sequence differing from its predicted ancestor at that site. This model is similar to that used by Stephens et al. [22] for the related problem of haplotype phase inference, although it differs in the use of a global recombination probability and site- and sequence-specific ancestral haplotype frequencies. These changes are intended to reduce the data dependence of the learning methods by reducing the number of parameters to be inferred from quadratic to linear in the number of sequences, a change that comes at a cost in generality of the model. The log of the probability implied by this model, normalized by a factor independent of the coloring and frequencies, yields the following objective function:

$$G = f_{h_1} + \sum_{j=1}^{n} mD(s_{h_j}, \sigma_j) + \sum_{j=2}^{n} (f_{h_j} + r)D(h_j, h_{j-1}) + log((1 - e^r) + e^{r+f_{h_j}})(1 - D(h_j, h_{j-1}))$$

where

- n is the number of polymorphic sites per sequence
- f_{ij} is the frequency with which haplotype i is chosen following a recombination between sets $j-1$ and j.
- h_j is the color assigned to site j of the target sequence
- s_{ij} is allele value of site j of reference sequence i
- σ_j is the allele value of site j of the target sequence
- m is the log prior probability of mutation at any site
- r is the log prior probability of recombination between any two sites
- $D(a, b)$ is 0 if $a = b$ and 1 if $a \neq b$

In the above equation, the first term is the contribution of the probability of starting with particular color. The second term is the sum of mutation contributions accounting for errors in matching the predicted ancestors. The third is a sum of the log probabilities of two possible events at each site: choosing a new haplotype following a recombination event or sticking with the prior haplotype either because there was no recombination or because there was a recombination with a sequence sharing the same common ancestor at that site. The set of values $h_1 \ldots h_n$ maximizing G is the maximum probability coloring for a single target sequence. Because sequences are colored independently of one another, finding the optimal coloring for each target sequence will yield the globally optimal coloring for all sequences for a given set of frequencies. We solve for this objective by a Viterbi-like dynamic programming algorithm in which, for each site j and ancestral sequence k, we find the optimal coloring of a target sequence on sites 1 through j terminating with color k.

We determine likely ancestral sequences from a modern population and establish their site-specific frequencies by an expectation-maximization algorithm [1]. The algorithm takes as input a single reference population and finds a locally optimal set of frequencies for that population and coloring in terms of those frequencies. Optimal coloring is achieved given a set of frequencies by applying the above dynamic programming algorithm for each sequence independently. Locally optimal frequencies are derived at each site by a steepest descent search for frequency values maximizing the contribution to G at that site. We initialize the method by assuming that each sequence in the data set is ancestral and that its frequency at all sites is equal to the measured frequency of the full sequence in the modern population. We then repeatedly apply the discrete coloring algorithm followed by the continuous frequency optimization until the method converges on a solution. This method is similar to the standard Baum-Welch method, although it is complicated by the interdependencies between transition probabilities introduced by our attempts to reduce the data dependence of the method. For more details on the haplotype coloring models and algorithms, see Schwartz et al. [20].

2.2 Missing Data Inference

We can adapt the block-free HMM method straightforwardly to the problem of missing data inference with a minor modification of the probability model allowing for the scoring of unknown sites. We treat an unknown site value as a match to any allele at that site, thus eliminating the potential for mismatch penalties at the unknown site. This model is equivalent to assuming that all known alleles are equally likely at an unknown site, and thus each possible match incurs the same penalty. This assumption allows us to color sequences with unknown values using the dynamic programming algorithm described above, with only the change that $D(s_{h_j}, \sigma_j) = 0$ if s_{h_j} or σ_j is unknown. This change has no effect on the asymptotic run time of the coloring algorithm.

Given the coloring of an unknown target sequence in terms of a reference population, we can then fill in likely missing values. For a polymorphic site of unknown allele value that has been assigned a color, we predict that the true allele value is the most frequent one among all known alleles assigned the same color in the reference population used in the EM frequency estimation. It might be more sound in terms of the model to chose the allele corresponding to the predicted ancestral sequence at the polymorphic site, even if it is not the most common allele for its predicted descendants. We chose the most common value, however, to allow for the possibility that the reference ancestral sequences may themselves have unknown values which can nonetheless be inferred based on their predicted descendants.

2.3 Informative SNP Selection

While much of the field has focused on SNP selection by blocks, creating an algorithmically convenient framework at the cost of some loss in usable infor-

mation content, we choose in this work to deal with the more general block-free framework and accept a cost in our ability to solve optimally and efficiently for the informative SNP selection problem. This decision gives us the freedom to work with a difficult objective function that closely matches a reasonable measure of the value of the chosen SNP set. We define the optimal SNP set of a given size to be the SNP set that maximizes the number of sites outside that set that would be predicted correctly in our training data based only on sites in the chosen SNP set, using our coloring-based missing data inference algorithm of section 2.2.

As we do not have an exact algorithm to find an optimal SNP set for our objective function, we worked with a simple heuristic method using a genetic algorithm. The hallmark of a genetic algorithm is a method for "mating" established solutions to produce new solutions that may yield better answers. Our mating method proceeds as follows:

1. Choose two solutions at random, S_1 and S_2, presumed to each contain k SNPs.
2. Chose a random crossing SNP site j.
3. Construct a new solution $S_3 = \{s | (s < j \wedge s \in S_1) \vee (s \geq j \wedge s \in S_2)\}$
4. If $|S_3| < k$ then add $k - |S_3|$ SNPs chosen uniformly at random from among all $s \notin S_3$ to S_3.
5. If $|S_3| \geq k$ then remove $k - |S_3|$ SNPs chosen uniformly at random from among all $s \in S_3$ from S_3.
6. Return S_3 as a new candidate solution.

The mating method forms the core of an algorithm for heuristically generating and testing possible solutions to attempt to find an optimal or near-optimal choice. We begin the overall algorithm by constructing an initial set of 20 candidate solutions by choosing 20 SNP sets uniformly at random among all possible subsets of k our n SNPs. We then assign a score to each candidate solution by hiding all sites not in the candidate in our training set, predicting the hidden sites using the algorithm of section 2.2, and counting the fraction of hidden sites correctly predicted from the training data. Among the 20 ranked SNP sets, we keep the top half and produce an equal number of new sets by mating the 10 old sets we keep, yielding 10 old and 10 new sets. We repeat this process for 100 rounds for each value of k, finally returning the highest-scoring SNP set in the final round as our approximate optimum.

While the use of a genetic algorithm for a computational genetics problem may seem an odd choice, we would in fact expect the model of meiotic mutation and recombination used by a genetic algorithm to be a good match to the nature of information locality in actual recombining sequences. As a result, the genetic algorithm appears likely to be a reasonable choice as heuristic methods go.

3 Results

We evaluated the methods using two real datasets: a set haploid sequences from 22 biallelic variations (21 SNPs and one two-base deletion polymorphism) in

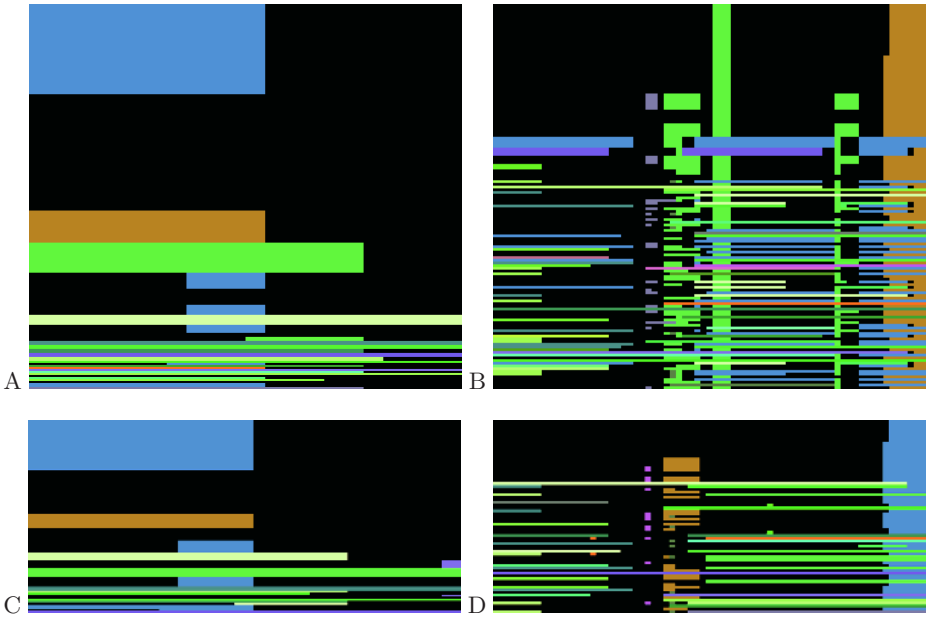

Fig. 1. Colorings of APOE and LPL datasets. In each image, rows of color represent distinct sequences and columns represent distinct polymorphic sites. Like colors within an image represent an inference of descent from a common ancestral sequence. Colors do not in general have any correspondence between images. A: coloring of the full APOE dataset. B: coloring of the full LPL dataset. C: coloring of the APOE training set. D: coloring of the LPL training set.

192 chromosomes for the apolipoprotein E (APOE) gene [16, 5] and a set of computationally inferred haplotypes from 71 SNPs in 142 chromosomes for the lipoprotein lipase (LPL) gene [17].

We first established colorings for the two datasets. For both data sets, we used a mutation parameter of -4. For APOE we used a recombination parameter of -1 and for LPL a recombination parameter of -0.5, selected empirically based on a visual analysis of the colorings they yielded. Figures 1A and 1B show the colorings yielded for these parameters for each data set. To evaluate the method, we further randomly split each data set into two equal-sized subsets of chromosomes: a training set and a testing set. Figures 1C and 1D show the coloring derived for the same parameter sets using just the training data. For APOE, the training data alone yields the major features of population-wide haplotype structure found in the full data set. For LPL, some noticeable structural elements from the full population are detected from the training data alone, while others are missed.

We next used the data sets to evaluate the ability of the coloring methods to perform missing site inference. For each fraction of hidden sites, in increments of

5%, we created an artificial data set from the testing data by creating ten copies of each sequence in the testing data and independently at random hiding the chosen fraction of hidden sites in each sequence. Finally, we applied the block-free coloring method to infer an HMM on the training set and used this HMM to color the artificial testing set sequences and fill in hidden sites based on the coloring. Finally, we evaluated the accuracy of the predictions compared to the true values of the hidden sites in the testing data.

Figure 2 shows the effectiveness of the methods on the two data sets data. We used two measures of quality: accuracy per base in predicting individual hidden sites and accuracy per sequence in predicting all hidden sites in a given sequence correctly, each as a function of the fraction of sites hidden. For APOE, per-base accuracy remains consistently high for all numbers of hidden sites, largely reflecting the fact that many polymorphisms have relatively rare minor variants in this data set. Per-sequence accuracy, however, shows a gradual decline as a greater fraction of sites are hidden. LPL shows a more nuanced profile, consistent with its greater diversity in the population. Per-base accuracy shows a slight decline with the fraction of sites hidden until reaching a plateau of about 80% accuracy at about 65% of sites hidden. Full-sequence accuracy falls off far more steeply, dropping rapidly up to about 20% hidden sites, then more gradually decreasing to zero.

We evaluated the effectiveness of the informative SNP selection method with a similar protocol to that used for missing data inference, judging the ability of the methods to infer hidden sites when those sites are computationally chosen. We used the same partition of APOE and LPL data sets into testing and training sets and the same program parameters as with missing site prediction. Instead of randomly hiding sites, however, we used the SNP selection method described in section 2.3 to choose the sites to be hidden based on the training data. We then evaluated our ability to predict the hidden sites using the testing data. We repeated this analysis for each possible number of hidden SNPs sites (0–22 for APOE and 0–71 for LPL).

Figure 3 shows the results of the missing site prediction using informative SNP selection for the two data sets. The graphs show a substantial improvement in accuracy compared to the results for randomly hidden sites. In both cases, the gradual decay approaching a plateau seen with per-sequence accuracy for random data is replaced by a slow decay at lower numbers of hidden sites followed by an increasingly steep decline for larger numbers of hidden sites. Per-site accuracy is also substantially improved for both until almost all sites are hidden. The benefits of selected versus random hidden sites is more pronounced for APOE than for LPL. In APOE, the accuracy drops minimally until 80% of sites are hidden, then falls sharply. For LPL, there is a continual decay, but it is much slower than when random hidden sites are chosen. The jerkiness of the LPL graph compared to the APOE graph appears primarily to reflect the fact that LPL's training set is less predictive of the haplotype patterns in its testing set than is the case with APOE.

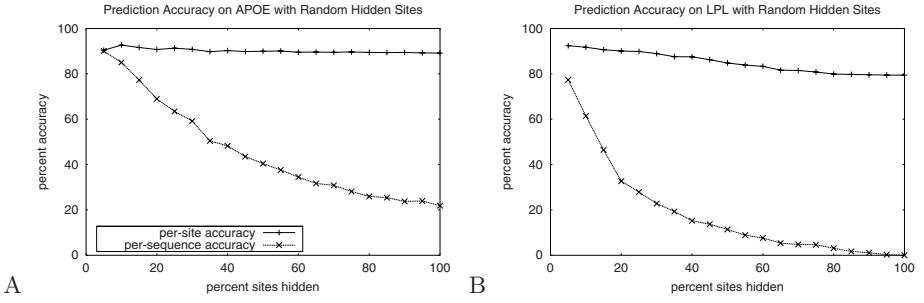

Fig. 2. Results of missing data inference. Graphs show accuracy in predicting individual sites and complete sequences as a function of the fraction of hidden sites for each data set when sites are hidden independently at random. A: results on APOE. B: results on LPL.

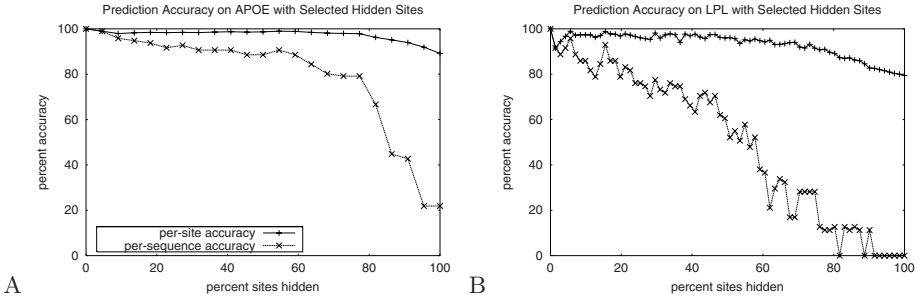

Fig. 3. Results of SNP selection. Graphs show accuracy in predicting individual sites and complete sequences as a function of the fraction of hidden sites for each data set when hidden sites are computationally selected to optimize per-site accuracy in the training data. A: results on APOE. B: results on LPL.

4 Discussion

We have presented computational methods for approaching some key problems in understanding haploid genetic data and applying it to the design and analysis of association studies. We showed how a previously developed method for piece-wise ancestry prediction could be adapted to the problems of inferring missing data and choosing informative SNP subsets. We further showed results of the ancestry prediction itself and its application to the two problems for two gene regions. These results indicate that there is substantial redundant information contained in sequences of polymorphic sites that we can exploit by using haplotype sequence context. They also suggest, though, that the value of such methods is likely to vary significantly between different genetic regions and may be severely limited for some regions if population samples are of inadequate size.

We can draw several general conclusions from our results about the prospects for haplotype-based studies. The comparison of randomly hidden to selectively hidden SNPs indicates that the right choice of SNPs can substantially improve our ability to infer the hidden sites, lending support to the idea of choosing haplotype tagging SNPs as a way to reduce assay costs [10]. On the other hand, the results, for LPL in particular, suggest that there may be limits on our ability to predict missing sites for reasonable sizes of data sets. Although the ability to predict hidden sites is a more stringent test than might be necessary for our true goal of association study design, it does provide an estimate of how well any given subset of sites captures the information content of the full set. For APOE, we can characterize about 80% of the population with only 25% of the SNPs and about 90% of the population with about 50% of the SNPs, although we require almost all SNPs to get close to 100% of the population. For LPL, the picture is more pessimistic, with hardly any resolving power until about half of the SNPs are used and with almost all SNPs required to reliably distinguish over 90% of sequences. The inability of the method to capture the last 10% of population diversity when only a few SNPs are hidden largely reflects the fact that the training data is not adequate for characterizing a significant fraction of sequence variants that are found only in the testing data. For any initial set of polymorphic sites sampled and any population sample size, we can expect similar absolute limits on the value of informative SNP selection methods determined by how well the sampled data characterizes the variability in the full population. As the differences between our results on the two data sets illustrate, the magnitude of the problem of inadequate sampling can vary considerably depending on local properties of particular genes of interest. These conclusions seem unlikely to be dependent on the specifics of our methods, but rather are likely to apply to all methods for these problems. They should therefore be considered in future assessments of methods for informative SNP selection, by performing cross-validation in evaluating methods and by explicitly building models of statistical significance of chosen SNP sets into the methods themselves.

There are many avenues by which this work can be continued. The basic models for ancestry detection are in some ways highly simplified. For example, they do not adjust for variations in recombination or mutation propensity across the genome or explicitly account for many specific genetic processes, such as hyper-variable sites, recombination hotspots, or gene conversion. More detailed models that incorporate a greater range of existing knowledge about the origins of genetic variation may perform better in practice. On the other hand, the models are too dependent on unknown parameters — particularly the user-supplied mutation and recombination rates — and are likely to be more practically useful if these parameter can be eliminated or automatically inferred. The actual methods for the problems could also likely be significantly improved, particularly the crude heuristic method used in this work for informative SNP selection. We might also consider extensions of these methods for unphased data or pooled data sets from multiple individuals. The measures used to test accuracy — per-site and per-sequence — are both imperfect and could use refinement. More

generally, the practical importance of these problems suggests a need for established benchmarks of performance under various conditions, such as levels of genetic diversity and population sample size.

References

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probability functions of Markov chains. *Annals Math Stat*, 41:164–171, 1970.

[2] L. R. Cardon and J. I Bell. Association study designs for complex diseases. *Nature Reviews Genetics*, 2:91–99, 2001.

[3] N. H. Chapman and E. A. Thompson. The effect of population history on the lengths of ancestral chromosome segments. *Genetics*, 162:449–458, 2002.

[4] M.J. Daly, J.D. Rioux, S.F. Schaffner, and T.J. Hudson. High-resolution haplotype structure in the human genome. *Nat Genet*, 29:229–232, 2001.

[5] S. M. Fullerton, A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor, J. H. Stengaard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C. F. Sing. Apolipoprotein E variation at the sequence haplotype level: implications for the origins and maintenance of a major human polymorphism. *Am J Hum Gen*, 67:881–900, 2000.

[6] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altschuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

[7] R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequence with recombination. *Journal of Computational Biology*, 3/4:479–502, 1996.

[8] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[9] A.J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctute meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*, 29:217–222, 2001.

[10] G. C. Johnson, L. Esposito, B. J. Barret, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbrigde, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29:233–237, 2001.

[11] J. Kececioglu and D. Gusfield. Reconstructing a history of recombinations from a set of sequences. In *Proceedings of the Fifth ACM-SIAM Symposium on Discrete Algorithms*, pages 471–480, 1994.

[12] J. S. Liu, C. Sabatii, J. Teng, B. J. B. Keats, and N. Risch. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res*, 11:1716–1724, 2001.

[13] K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, 33:177–182, 2003.

[14] M. S. McPeek and A. Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Gen*, 65:858–875, 1999.

[15] A. P. Morris, J. C. Whittaker, and D. J. Balding. Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am J Hum Gen*, 67:155–169, 2000.

[16] D. A. Nickerson, S. L. Taylor, S. M. Fullerton, K. M. Weiss, A. G. Clark, J. H. Stengaard, V. Salomaa, E. Boerwinkle, and C. F. Sing. Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res*, 10:1532–1545, 2000.

[17] D. A. Nickerson, S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson, J. H. Stengaard, V. Salomaa, E. Vartiainen, E. Boerwinkle, and C. F. Sing. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet*, 19:233–240, 1998.

[18] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 294:1719–1722, 2001.

[19] N. J. Risch and K. R. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, 1996.

[20] R. Schwartz, A.G. Clark, and S. Istrail. Methods for inferring block-wise ancestral history from haploid sequences: The haplotype coloring problem. In *Lecture Notes in Computer Science 2452 (Proceedings of the Second International Workshop on Algorithms in Bioinformatics)*, pages 44–59, 2002.

[21] S. K. Service, D. W. Temple Lang, N. B. Freimer, and L. A. Sandkuijl. Linkage disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am J Hum Gen*, 64:1728–1738, 1999.

[22] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001.

[23] E. Ukkonen. Finding founder sequences from a set of recombinants. In *Lecture Notes in Computer Science 2452 (Proceedings of the Second International Workshop on Algorithms in Bioinformatics)*, pages 277–286, 2002.

[24] G. Venter, M. A. Adams, E. W. Myers, et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[25] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, 8:69–78, 2002.

[26] C. Wiuf and J. Hein. On the number of ancestors to a DNA sequence. *Genetics*, 147:1459–1468, 1997.

[27] C. Wiuf and J. Hein. The ancestry of a sample of sequences subject to recombination. *Genetics*, 151:1217–1228, 1999.

[28] S. Wu and X. Gu. A greedy algorithm for optimal recombination. *Lecture Notes in Computer Science*, 2108:86–90, 2002.

[29] K. Zhang, P. Calabrese, M. Nordborg, and Sun F. Haplotype block structure and its applications to association studies: Power and study designs. *Am J Hum Gen*, 71:1386–1394, 2002.

[30] K. Zhang, M. Deng, T. Chen, M.S. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA*, 99:7335–7339, 2002.