Stat 550 Final Project

Hugh Chen

June 2017

1 Introduction

Linkage disequilibrium is a measure of allelic association between loci (figure 1) which can arise from mutation, population structure, drift, and more. These causal factors are important considerations when modeling linkage disequilibrium, however they may prove challenging to successfully incorporate into a new model. On the other side of things, linkage disequilibrium is also tightly related to recombination and random mating, both of which break down linkage disequilibrium. In theory, if we could use all of these biological factors in modelling linkage disequilibrium, our model would be very robust. In practice however, keeping track of so many complex, ostensibly interrelated, variables is no easy feat. In my project, I go through the Li and Stephens model, which focuses on modeling one variable (recombination) in order to capture linkage disequilibrium in a biologically motivated model [9].

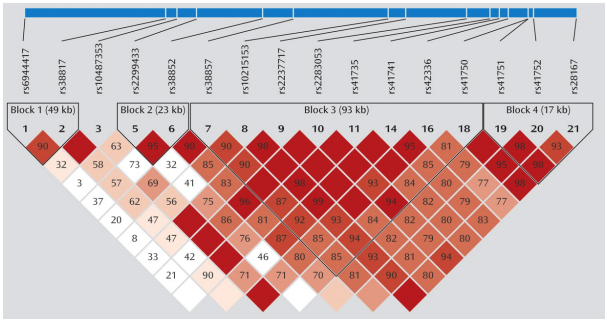

Figure 1: Figure representing linkage disequilibrium.

https://estrip.org/articles/read/tinypliny/44920/Linkage_Disequilibrium_Blocks_Triangles.html.

In terms of the broader context, it's important to motivate the problem of linkage disequilibrium. LD is used in "order to understand past evolutionary and demographic events, to map genes that are associated with quantitative characters and inherited diseases, and to understand the joint evolution of linked sets of genes" [11]. Linkage disequilibrium is important in studying genetics because it contains so much information. Information about population history, breeding patterns, geographic subdivisions, natural selection, mutation, and more is all reflected in these correlated loci [11]. One specific example of LD's usefulness is LD mapping, which uses the fact that rare marker alleles with strong linkage disequilibrium with a monogenic disease locus have to be closely linked to the causal locus. Knowing the causal locus for a disease could potentially allow people to detect diseases early and potentially help anyone who might suffer from this disease.

Li and Stephens proposed a new statistical model for linkage disequilibrium among multiple SNPs in a population [9]. At the time, many methods for analyzing linkage disequilibrium suffered from one or more of three limitations:

1. They may consider LD for only pairs of sites rather than for many sites simultaneously. Modelling linkage disequilibrium for only pairs of sites is problematic because these methods fail to approximate the underlying model well. One could imagine ignoring effects beyond pairwise ones would be inaccurate and potentially yield less informative measures.

2. They may assume "block-like" structures which are not always appropriate. "Block-like" structures, where LD tends to be high among contiguous sets of markers, may be expected to a certain degree. These "blocks" can arise either from chance or they can be the consequence of variations in local recombination rate. It would be useful from both a basic scientific and a modelling point of view to not have a priori assumptions of these blocks in order to be able to distinguish between them without needing experimental confirmation.

3. They may fail to incorporate relevant biological mechanisms (e.g. recombination rate). This is problematic since the problem of linkage disequilibrium is inherently biological. Leveraging prior biological knowledge is appealing from not only a biological perspective but also a computational one. Modelling according to the recombination rates allows us to approximate more complex joint distributions according to conditional ones, which Li and Stephens shows we can approximate quite nicely.

2 Paper Description

The most successful approaches at the time of Li and Stephen's (2003) paper were coalescent theory, first introduced in Kingman (1982) which was generalized to include recombination in Hudson (1983) [8, 5]. The coalescent model is a continuous-time Markov chain with a finite set of states that serves to describe family relationships among a sample of members from a larger haploid population [8]. According to Li and Stephens, the coalescent model based approaches were based on simplistic assumptions about demographics and were not practical for the purpose of inference. Another popular model around the time of Li and Stephen's (2003) paper were "composite-likelihood" methods, such as Hudson (2001) and Fearnhead and Donnelly (2002), for estimating recombination rates by multiplying pairwise likelihoods [6, 4].

2.1 Li and Stephens Model - π

The Li and Stephens model focuses on recombination rates. They model the joint distribution over samples conditioned on the recombination rate which is a hyperparameter in their model:

$$\begin{aligned} P(h_1, \cdots, h_n|\rho) &= P(h_1|\rho)P(h_2|h_1;\rho)\cdots P(h_n|h_1, \cdots, h_{n-1};\rho) \\ &= \hat{\pi}(h_1|\rho)\hat{\pi}(h_2|h_1;\rho)\cdots \hat{\pi}(h_n|h_1, \cdots, h_{n-1};\rho) \end{aligned}$$

Where Li and Stephens approximate the conditional distributions with $\hat{\pi}$ for the sake of tractability. They call this model a "product of approximate conditionals" (PAC) model with a likelihood of $L_{PAC} = \hat{\pi}(h_1|\rho)\hat{\pi}(h_2|h_1;\rho)\cdots\hat{\pi}(h_n|h_1, \cdots, h_{n-1};\rho)$. In this model, one of the arguably most important features is their choice of approximation. In the paper, they discuss several candidates, but ultimately settle on a version equivalent to one discussed by Fearnhead and Donnelly (2001) which satisfies five properties [3]:

1. Of the haplotypes that have been seen so far, the next haplotype should match ones that occurred frequently with higher probability. This is a natural probabilistic assumption.

2. As the number of samples considered increases, the probability of seeing an entirely new haplotype should decrease. As the pool of considered samples grows, the haplotypes previously encountered will only stay the same or increase. Under independence, when considering a fresh haplotype, it should have lower probability of being novel.

3. As the product of sample size and mutation rate ($\theta = N\mu$) increases, the probability of encountering novel haplotypes increases, since this is the expected number of novel haplotypes. If there are more novel haplotypes, the probability of encountering novel haplotypes increases.

4. If the next haplotype is not exactly the same as a previously seen haplotype, it should be similar to an existing haplotype (i.e. a previously seen haplotype).

5. Based on recombination, new haplotypes should not only resemble existing haplotypes, but are likely to match over "contiguous genomic regions" where the length of the regions may depend on local recombination rates.

The first three properties are based on the Ewens sampling formula which considers a neutral locus in a randomly mating population in an 'infinite-alleles' mutation model", evolving with constant size N and mutation rate μ per generation. Each mutation creates a novel haplotype. In Stephens and Donnelly (2000), the proposed $\hat{\pi}$ has next haplotypes that differ by M mutations from a random previously seen haplotype, where M has a geometric distribution that reproduces the Ewens sampling formula in the case of the infinite-alleles mutation model. Then, the final $\hat{\pi}$ is based on Fearnhead and Donnelly (2001) which extended the Stephens and Donnelly model to incorporate property 5 [3]. In the FD $\hat{\pi}$, the next haplotype is an imperfect mosaic of the first k haplotypes, with fragment size dependent on the recombination rate. Li and Stephens present a version of Fearnhead and Donnelly's approximation that is slightly quicker to compute and easier to understand (π_A).

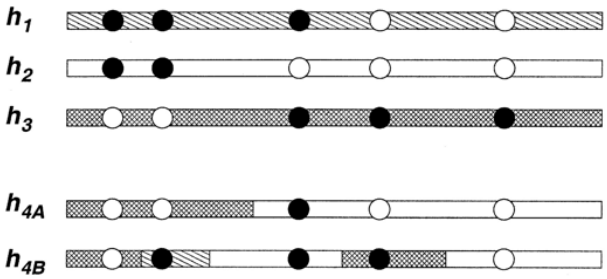

Figure 2: Figure representing the approximate conditional distribution π_A.

Taken from Li and Stephens (2003) [9].

In their paper, they use figure 2 to explain their approximation to the conditional distributions $\pi_A(h_{k+1}|h_1, \cdots, h_k)$. Above, h_{4A} and h_{4B} are built from parts of the set of previously seen haplotypes (i.e. h_1, h_2, h_3 in this example), which embodies the first and second properties. Importantly, the segments in h_{4A} and h_{4B} from the previously seen haplotypes are not perfect copies. For example, the third allele on h_{4A} and h_{4B} is from a part of the haplotype taken from h_2, but the allele is different. The copying process is modeled as a Markov process along the chromosomes with jumps occurring at a rate of ρ/k. This embodies the second, fourth, and fifth properties. As k increases, we jump less frequently and use more of the existing haplotypes. At different loci, the rates ρ may vary, but we can easily account for that by modifying the value of ρ. The third property is captured simply because we use a pool of encountered haplotypes to construct our next haplotype. Using this π_A is efficient because it is naturally modeled as a hidden Markov model. Li and Stephens not that π_A clearly depends on the order in which the haplotypes are encountered. Depending on the ordering of the haplotypes, our h_1, h_2, h_3 may be entirely different. And thus, the model may be entirely different as well. They solve this problem empirically by randomizing the ordering and averaging their likelihood function over several random orderings of the haplotypes.

2.2 Li and Stephens Model - ρ

One straightforward application of the way Li and Stephens modeled linkage disequilibrium is to use their model to estimate recombination rate. They first discuss estimating a constant recombination rate by numerically maximizing L_{PAC} for different generated data sets and comparing $\hat{\rho}_{PAC}$ with the true underlying ρ. In their simulated data sets, they varied the number of haplotypes, the number of markers typed, and the value of ρ. For these data sets they had fairly accurate estimate of the true ρ, but found biases under different simulation settings. In particular, the bias was especially dependent on the average spacing of the markers, with a tendency to overestimate ρ when markers are closely spaced and underestimate ρ when markers are far apart. Additionally, they compare their point estimate $\hat{\rho}_{PAC}$ with a composite-likelihood method for estimating the recombination rate which multiplies together likelihoods for every pair of SNPs (Hudson 2001). Perhaps unsurprisingly, they found that the composite-likelihood estimate for ρ, $\hat{\rho}_{CL}$, was better than $\hat{\rho}_{PAC}$ for small numbers of SNPs. For small datasets, the composite-likelihood estimate is exactly the maximum likelihood estimate. Conversely, the Li and Stephens estimate $\hat{\rho}_{PAC}$ had less variability than $\hat{\rho}_{CL}$ in datasets with large number of SNPs.

Then, Li and Stephens move on to discuss a variable recombination rate. They analyzing a simple model where crossovers in a single meiosis can occur as a Poisson process with variable rate at position x and specify it in two ways:

1. A simple single-hotspot model, where

$$c(x) = \begin{cases} \lambda\bar{c} & a \leq x \leq b \\ \bar{c} & o.w. \end{cases}$$

   Above, $\bar{c}$ represents the background rate of crossover and a and b represent the hotspot's boundaries. The PAC likelihood for this model goes from the constant case of L_{PAC} to $L_{PAC}(\bar{\rho} \equiv 4N\bar{c}, a, b, \lambda)$. For this model, it is possible to numerically obtain the maximum likelihood estimates using their "product of approximate conditionals" likelihood. This model of recombination does make unrealistic assumptions that the background recombination rate and the hotspot recombination rate is constant. Additionally, modeling only one recombination hotspot is not a very robust assumption either.

2. A more general model, where $c(x) = \lambda_j\bar{c}$ if x is between markers j and $j+1$. Once again $\bar{c}$ is a background rate of crossover and λ_j is a multiplier for the crossover rate. Here we have a likelihood with more parameters: $L_{PAC}(\bar{\rho}, \lambda_1, \cdots, \lambda_{S-1})$, where S is the number of SNPs. For this model, estimating the maximum likelihood estimates is more difficult because the parameters are unidentifiable (the MLEs are not unique). In order to address this problem, Li and Stephens assume a "prior" distribution that the λ_j are independent and identically distributed such that there is an occasional deviation from the background recombination rate with a factor of 10 or more. This reflects their prior beliefs about the λ_j. As a whole, this model makes few assumptions and allows more of a flexible parameterization of linkage disequilibrium. Unfortunately, the extra parameters can result in a reduction in the precision of the estimates.

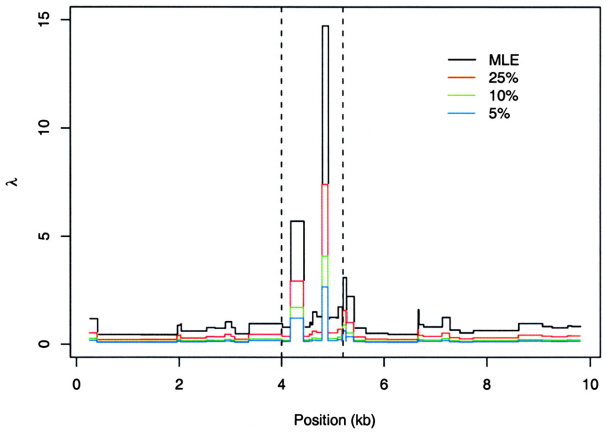

Figure 3: Figure representing the estimated recombination rates for TAP2.

Taken from Li and Stephens (2003) [9].

In several simulated experiments, Li and Stephens find that they have fairly good type I error rates (≈ 0.05) and power (≈ 0.90) for data with a hotspot. They see reduced power or inflated type I error rates under scenarios involving population expansion. As another simulated experiment, they fit the general model to data and end up capturing recombination hotspots fairly well, with a fairly informative depiction of hotspots over many SNP positions. As another evaluation, they use a population sample of 30 individuals from the United Kingdom to look at the location of a putative recombination hotspot in the human TAP2 gene, which we can see in figure 3. Their estimate for the average magnitude of the hotspot using this simple single-hotspot model is 12 times the background rate, which agrees well with an analysis by Jeffreys et al (2000) that likewise identified a region of increased crossover intensity [7]. They repeat one more experiment for lipoprotein lipase data (from Clark et al 1998, Nickerson et al 1998) which has a putative hotspot identified by Templeton et al (2000) [2, 10, 12]. Their general model for recombination rate gives results that likewise suggest a hotspot.

3 Broader Context

As discussed, there are a good deal of methods dealing with linkage disequilibrium prior to this Li and Stephens model. In their paper, Li and Stephens discuss coalescent theory, which was generalized to include recombination. Coalescent theory uses a Markov chain with a finite set of states to describe family relationships, and was extended to include recombination in 1983 by Hudson [5]. Hudson describes an infinite-site neutral allele model which focused on genetic drift of neutral mutant alleles, alleles that don't affect an organism's ability to survive and reproduce allowing for crossover at any of an infinite number of sites[5]. This model appears to be extremely generalizable, and makes some wide-reaching assumptions that may be impractical in terms of computation for many loci. In coalescent theory, the population samples contain information on the value of the compound parameter $\rho = 4Nc$ where N is the population size and c is the recombination rate. Li and Stephens discuss that even for moderate-sized autosomal regions, many methods for estimating ρ become intractable. At the time, the most accurate of methods existing at the time used "composite-likelihood" methods for estimating ρ over large genomic regions.

Although Li and Stephens present their method in the context of linkage disequilibrium, it can easily be viewed as a model for haplotype phasing, with the haplotypes modeled in the HMM being viewed as "template haplotypes", as is done in an assessment of haplotype phasing by Browning and Browning (2011)[1]. Two such methods are MACH and IMPUTE2 which estimate an individual's haplotypes using the previous estimated template haplotypes. MACH randomly selects a random subset of sample haplotypes for templates and IMPUTE2 selects a subset of haplotypes similar to the haplotypes for the individual being estimated. Another method, PHASE, uses a similar approach to the Li and Stephens model, but with extra parameters: the coalescent times between a given haplotype and the underlying template haplotype. Coalescent times are obtained by tracing the lineage back to the most recent common ancestor. Certainly it seems that including these coalescent times as parameters could help identify which template haplotypes would make good contributions to the current target haplotype. An alternative method to the Li and Stephens framework is BEAGLE, which forms an HMM by representing locally clustered haplotypes at each marker position as hidden states. They then learn the transition probabilities as an ordered probabilistic traversal across the chromosome in terms of the haplotypes in the sample. Overall, the Li and Stephens model has many applications in genetics. Another such application is HAPMIX which is used for Local Ancestry Inference.

4 Conclusion

In their paper, Li and Stephens introduced a novel way to model linkage disequilibrium at multiple loci. This approach incorporated the underlying recombination rate and examined its effectiveness for estimating the recombination rate itself. Li and Stephens suggest a potential use of their LD model in "case-control" studies where the case chromosomes share some identical by descent region about a causal mutation and thus should be more similar than would be expected by chance. In this scenario, modeling linkage disequilibrium helps to provide information about the causal locus.

As far as limitations, Li and Stephens discuss how there are biological aspects of real data with impacts on LD that haven't been accounted for in this model. Two of which are gene conversion and population structure. While their general model for recombination rates is somewhat robust to these other factors for determining linkage disequilibrium, a more complex model would presumably increase the accuracy of recombination rates (perhaps at the cost of model complexity and tractability). When compared to composite-likelihood methods of Hudson (2001) and Fearnhead and Donnelly (2002) [6, 4], Li and Stephens finds their model to be more computationally efficient at estimating likelihood than Fearnhead and Donnelly and on par with Hudson. At the time this paper was written, Hudson's method assumes a constant recombination rate (although it could be extended to include varying recombination rates in principle). Li and Stephens end by concluding that both their approach and Hudson's offer considerable advantages over other available methods for modeling linkage disequilibrium and inferring patterns of heterogeneous recombination rates. Overall, the Li and Stephens model has proven to be useful for estimating recombination rates, determining linkage disequilibrium, haplotype phasing, and local ancestry inference. Based on the interconnected nature of linkage disequilibrium it is imaginable that there are many other potential applications for this framework.

References

[1] Sharon R. Browning and Brian L. Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[2] Andrew G. Clark, Kenneth M. Weiss, Deborah A. Nickerson, Scott L. Taylor, Anne Buchanan, Jari Stengård, Veikko Salomaa, Erkki Vartiainen, Markus Perola, Eric Boerwinkle, and Charles F. Sing. Haplotype structure and population genetic inferences from nucleotide- sequence variation in human lipoprotein lipase. *American Journal of Human Genetics*, 63(2):595–612, 8 1998.

[3] Paul Fearnhead and Peter Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–1318, 2001.

[4] Paul Fearnhead and Peter Donnelly. Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):657–680, 2002.

[5] Richard R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

[6] Richard R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817, 2001.

[7] A. J. Jeffreys. High resolution analysis of haplotype diversity and meiotic crossover in the human tap2 recombination hotspot. *Human Molecular Genetics*, 9(5):725–733, 2000.

[8] J.f.c. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

[9] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[10] Deborah A. Nickerson, Scott L. Taylor, Kenneth M. Weiss, Andrew G. Clark, Richard G. Hutchinson, Jari Stengård, Veikko Salomaa, Erkki Vartiainen, Eric Boerwinkle, and Charles F. Sing. Dna sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genetics*, 19:233–240, 2000.

[11] Montgomery Slatkin. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, Jun 2008.

[12] Alan R. Templeton, Andrew G. Clark, Kenneth M. Weiss, Deborah A. Nickerson, Eric Boerwinkle, and Charles F. Sing. Recombinational and mutational hotspots within the human lipoprotein lipase gene. *The American Journal of Human Genetics*, 66(1):69 – 83, 2000.