Long-Range Linkage Disequilibrium in the Human Genome

Presented by Evan M. Koch

May 9, 2012

In partial fulfillment of the requirements for graduation with the Dean’s Scholars Honors Degree in Biology

Prof. Mark Kirkpatrick, Supervising Professor

Prof. Ruth Buskirk, Honors Advisor in Biology
Abstract

In the study of genomic variation, the nonrandom association of alleles has been a valuable tool for understanding population history, correlating disease phenotypes, and mapping genetic structural variation. These associations, termed linkage disequilibrium, have been quantified to great extent in humans over short distances. Linkage disequilibrium between genetically distant sites has garnered far less attention. These long-range patterns are quantified using a statistic $p_D$ that can distinguish true associations from chance occurrences. Using null distributions generated by randomization we detect statistically significant long-range linkage disequilibrium on 19 of 22 human chromosomes. Given that recombination tends to break down associations in relatively few generations, this finding indicates that countervailing forces are at work. The factors that could have created the observed long-range linkage disequilibrium are epistasis, recent population admixture, genetic structural variation, and artifacts in the data collection. This analysis is unable to rule on any particular cause, but it seems admixture is the most likely to have produced the following results.
1 Introduction

Much work since the sequencing of the first human genome has gone into characterizing the genetic differences between individuals. This has included everything from single nucleotide polymorphisms (SNP) to chemical and chromosomal alterations. SNPs in particular have had the advantage of being relatively simple to discover and an easy target for statistics. An obvious feature to consider when studying the amount and distribution of variation is the covariation between variable nucleotides. This notion was introduced into population genetics by Lewontin and Kojima in 1960 as linkage disequilibrium (LD), or the nonrandom association of genotypes between two loci on a chromosome. The processes capable of creating LD that are focused on here are epistasis, recent population admixture, genetic structural variation, and artifacts in data collection. LD is expected and has been well-characterized in the human genome between loci that are tightly linked (physically close on a chromosome). Such association over short distances is due to random genetic drift and the fact that mutations will arise on different genetic backgrounds. Without frequent recombination to break up these associations, they can appear as large chromosomal areas over which only a few variants are present in a population.

Despite an intense focus on LD in both medical and evolutionary studies, much less attention has been paid to LD between genetically distant sites, and that is what we investigate here. Since recombination tends to break down disequilibria rapidly between those, finding significant long-range LD suggests that countervailing forces may be at work. Potential candidates are epistatic selection, previously undetected population admixture, and genetic structural variation. We aim to characterize patterns of long-range LD across the genome while removing any influence from the well-characterized and explained LD over short distances.

Shortly after the sequencing of the first human genome, a map of single nucleotide polymorphisms (SNPs) in the human genome was published (Sachidanandam et al., 2001). The map identified 1.42 million SNPs, roughly half of the candidates for which came from reduced representation and complete shotgun sequencing of 24 individuals. The other candidates were identified by the Human Genome Project using hierarchical shotgun sequencing. From the beginning, this study of genetic variation had the explicit purpose of aiding the identification of mutations contributing to disease. The general strategy was to characterize the variation most common in human populations and then correlate disease phenotypes with areas of variation. Although it is unlikely that any of the common variants contribute to the disease, the culprit mutations should be closely linked to those identified. This information provides researchers with a place in the genome to look in greater depth. A feature characterized extensively in the first SNP map but known previously to exist (Li and Sadler, 1991) is that at closely linked sites the diversity is comprised of a small number of specific allele combinations. These combinations of alleles inherited together are called haplotypes.

The study of LD that used the SNP map from Sachidanandam et al. looked in greater detail at distribution of haplotype blocks (Reich et al., 2001). The SNP data from the map itself was not sufficient to make any conclusions about LD, so the researchers choose 3,000 SNPs whose minor variant was present at at least 35% frequency and sequenced these in 44 individuals from Utah, 48 from Sweden, and 96 Yorubans.
from Nigeria. This resulted in the characterization of haplotypes in 19 regions of 160 kb. With an estimated recombination rate of 0.5 - 5.9 cM Mb\(^{-1}\), the distal ends of the regions sequenced by Reich et al. have the possibility of being what is considered long-range by this study. However, in a genome of 3000 Mb this is not enough to perform anything as broad as a genome-wide association study. The goal of the study was to determine how far blocks of LD extended. The greater the extent of LD the more useful information about common haplotypes can be for association studies. The finding of a typical length around 60 kb prompted the International Haplotype Map Project (HapMap) to do what Reich et al. had several over small genomic sections for the whole human genome (The International Haplotype Map Consortium, 2003).

The populations included in phases I and II were a US Utah population of Northern and Western European descent, Yoruban in Ibadan, Nigeria, Japanese in Tokyo, Japan and Han Chinese in Beijing, China. The diversity of ancestries was chosen in part because of the medical focus. For instance, when sequencing an individual and trying to understand a complex disease phenotype, it is helpful to have a comprehensive set of common genetic backgrounds to compare to, and those backgrounds could vary considerably between ancestry groups. This study focuses only on the sample from the Yoruban in Ibadan, Nigeria (YRI). In addition to making the LD computations more tractable, the YRI sample has been found to have (as a results of higher ancestral population size) shorter, more diverse haplotype blocks(The International Haplotype Map Consortium, 2005; The International Haplotype Map Consortium, 2007). Both of these factors should contribute to lower the amount of long-range LD than we would expect to see by chance. For this reason we decided to focus our analysis on the YRI sample.

Unlike Sachidanandam et al. who brought together data from multiple projects, the HapMap picked candidate SNPs from the public repository dbSNP (Sherry et al., 2001) with the intention of having one SNP every 5kb. In phase I this resulted in a dense map of 920,102 SNPs with only 3.3% of inter-SNP distances greater than 10kb and an average block length for YRI was 7.3kb. One reason for trying to achieve this high density and number was to determine the full extent to which the human genome consisted of haplotype blocks similar to those found by Reich et al. and others (provide refs) The boundaries of haplotype blocks had also been shown to be associated with recombination hot-spots, although some have disputed this as causative (Phillips et al. 2003). A problem with haplotype blocks that do not associate with recombination hotspots is that they might have arisen by chance and may not persist given that there isn’t a strong mechanism keeping them around. This would limit the blocks usefulness in disease association studies. Phase I found many cases when blocks failed to observe recombination rate peaks as boundaries but was optimistic for disease association based on the number of perfect correlations with neighboring SNPs that many sites showed. Our study aims to correct for these so-called perfect proxies and other high short-range associations by using the sample with half as many perfect proxy SNPs (YRI) and developing a procedure to account for their influence.

The second phase of the HapMap increased the number of SNPs to 2,935,635. The overall resolution of the map increased due to this the fact that SNPs with a lower minor allele frequency were included (The International Haplotype Map Consortium, 2007). Increased resolution allowed recombination hotspots to be placed more precisely and the struc-
ture of LD to be examined in different ways. By examining the amount of similarity that extended across recombination hotspots, it was possible to determine the average identity-by-descent in each sample. Further studies performed using the HapMap data have included genome-wide association (GWA) studies for complex diseases (Manolio et al., 2008), scans for loci under positive selection (Akey, 2009), studying the effect of geography on human evolution (Coop et al. 2009), and searching for different types of genomic structural variation (Myers et al., 2008; Cáceres et al., 2012). Others have been less optimistic about the HapMap’s utility and the project has been criticized on the grounds that the assumptions made in LD-based association studies can lead to failure to identify disease variants (Terwilliger and Hiekkalinnna, 2006).

Previous studies have used the term long-range linkage disequilibrium to refer to regions with longer than expected LD. These are essentially the right tail of the distribution of haplotype block sizes. A finding going back to Reich et al. has been that, although LD generally drops off with distance between SNPs, this is highly variable. The authors attributed this observation to a bottleneck in the European population. Other pre-HapMap studies have focused on specific regions showing LRLD. One study examined a region surrounding the NF1 gene on chromosome 17 and supported the bottleneck hypothesis (Schmegner et al., 2005), while another related LRLD on the X to subdivision of ancestral African populations followed by recent admixture (Garrigan et al., 2005). Post-HapMap studies in humans have no longer needed to focus on specific areas of extended LD, but the patterns across the genome have been used in more complex analyses that have investigated a variety of evolutionary factors (Coop et al., 2009).

In studies of demography, the existence of LRLD and extended haplotypes is attributed to the influence of genetic drift during periods of low population size (Slatkin, 2008). For this study we attempt to consider only regions of LRLD where this has not played a significant role. Aside from choosing an African population so that fewer bottlenecks are expected, we did this by looking only at regions of LD a certain distance apart and not part of a continuous haplotype block. Theoretical work has been done that gives us a base expectation about the degree of LD to expect between distal regions with no intervening association. Before the first sequencing of a human genome, simulations had shown that little LD was to be expected at distances over 10 kb (Kruglyak, 1999). Once empirical work had shown this not to be the case, the models underlying the theory and simulations were updated to accommodate the difference. Changes that were made included population expansion, structure, and bottlenecks, in addition to the explicit modeling of recombination hotspots (Schaffner et al., 2005). Both simulation and analytical studies showed higher than otherwise expected LD at linked and unlinked markers could be created by strong population structure (Pritchard and Pzreworski, 2001; Wakeley and Lessard, 2003).

A fundamental difference between this study and previous studies is that instead of describing the extent of LD across a continuous chromosomal portion, we look for two such associated portions at a time. One consequence of this is a lack of the notion of decaying LD away from some principal SNP. A second consequence is that by examining each pair of SNPs we are making a number of independent associations that vastly outnumbers that from studies focused solely on sites within some minimal distance. The statistical considerations are thus different from conventional studies of LD and novel
approaches have been developed to account for the increased number and independence of calculations.

1.1 Statistical measures of linkage disequilibrium

Several statistics have been developed to measure LD between pairs of sites, each with its own advantages and drawbacks (Hedrick, 1987). Other measures have also been devised for comparing more than two sites, but these are not considered here and have been rarely used in other studies (Slatkin, 2008). Given the large number of sites we are testing and the low expected LD between any two distant pairs, we need a way to distinguish between chance associations due to sampling and true disequilibria. Most common measures of LD are not well suited to this purpose (Slatkin, 1994). We here review these measures and explain why a different statistic, $p_D$, is appropriate for detecting nonrandom disequilibria.

The association metrics for allele frequencies at two sites use the following contingency table for sites $A$ and $B$. The bases are labeled 0 and 1.

<table>
<thead>
<tr>
<th></th>
<th>$a_{0,0}$</th>
<th>$a_{0,1}$</th>
<th>$r_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{1,0}$</td>
<td>$a_{1,1}$</td>
<td>$r_1$</td>
<td></td>
</tr>
<tr>
<td>$s_0$</td>
<td>$s_1$</td>
<td>$n$</td>
<td></td>
</tr>
</tbody>
</table>

In the above table $a_{i,j}$ is the number of haplotypes in the sample that have base $i$ at site $A$ and base $j$ at site $B$. The marginal totals for base $i$ at sites $A$ and $B$ are respectively $r_i = a_{i,0} + a_{i,1}$ and $s_i = a_{0,i} + a_{1,i}$. The total number of samples is given by $n = r_0 + r_1 = s_0 + s_1$. The frequencies of base 1 at sites $A$ and $B$ are $p_A = r_1/n$ and $p_B = s_1/n$.

1.1.1 $D$ and $D'$

$D$ was introduced by Lewontin and Kojima in 1960. It is the classical measure of linkage disequilibrium and can be interpreted easily as the covariance between bases 0 and 1 at sites $A$ and $B$. $D$ can be calculated as:

$$D = \frac{a_{0,0}a_{1,1} - a_{0,1}a_{1,0}}{n^2} = \frac{a_{0,0}}{n} - \frac{r_0 s_0}{n n}$$

This is equivalent to the difference between the observed and expected frequency of the $a_{0,0}$ haplotype so that when $D = 0$ there is linkage equilibrium.

Aside from being easily interpreted as a covariance, the property of $D$ that has made it especially useful in population genetics studies is that under conditions of Hardy-Weinberg equilibrium it decreases exponentially at a rate given by $1 - c$ where $c$ is the recombination rate between the two loci in question. A drawback of $D$ is that its range of possible values is dependent on the allele frequencies at the two loci, and a large magnitude does not indicate statistical significance. To avoid this problem it is common
to normalize $D$ by dividing by the maximum obtainable value at the observed allele frequencies to get:

$$D' = \frac{D}{D_{max}}$$  \hspace{1cm} (3)

In this expression $D_{max} = \min(p_A(1-p_B), (1-p_A)p_B)$ so $D'$ now ranges from $-1$ to $1$ (Lewontin, 1964). This normalization makes for an easier comparison of LD between pairs of SNPs, but it still is not informative about how likely the association was to arise by chance sampling error. Values of $D'$ near 1 are more likely if marginal frequencies are near 0 or 1 than if they are near 1/2.

### 1.1.2 Log Odds Ratio

The odds ratio is simply the ratio of the conditional odds for each possible outcome from our contingency table. It can be calculated as:

$$OD = \frac{odds(A = 0|B = 0)}{odds(A = 0|B = 1)} = \frac{a_{0,0}a_{1,1}}{a_{0,1}a_{1,0}}$$  \hspace{1cm} (4)

This is the probability of finding base 0 at site A given that we find base 0 at site B divided by the probability of finding base 0 at site A given that we find base 1 at site B. The more similar these probabilities are, the less closely associated the two sites are. The odds ratio reduces to the same value regardless of which particular probabilities are chosen. $OD$ is then scaled by taking the natural log to give

$$LOD = \ln\left(\frac{a_{0,0}a_{1,1}}{a_{0,1}a_{1,0}}\right)$$  \hspace{1cm} (5)

which has a range of $(-\infty, \infty)$ where a value of 0 indicates no association. $LOD$ has the benefit that, unlike $D$ and $D'$, its range is not constrained by the marginal frequencies of $p_A$ and $p_B$. Unfortunately, values of $\infty$ are difficult to interpret and compare. $LOD$ still faces the problem like $D$ and $D'$ that it does not convey how likely the observed value is to have arose by chance.

### 1.1.3 $r^2$

This statistic is simply the correlation coefficient between the alleles found at A and B. It can be calculated as

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$$  \hspace{1cm} (6)

This measure has the advantage, like $D'$, that it ranges from 0 to 1 and thus allows for easy comparison between SNP pairs. $r^2$ is also a natural choice for association mapping studies and allows for assessment of how likely it is to have arose by chance due to the fact that it is equivalent to the $\chi^2$-test statistic divided by the number of chromosomes (Weir 1990, Ch. 3). However, it has been shown that the $\chi^2$ does not perform well under many reasonable population genetics conditions (Slatkin 1994).
2 Methods

2.1 The \( p_D \) Statistic

The drawbacks of the discussed LD measures lead us to consider a statistic based on the probability that the association found in a sample would be observed under an appropriate null model. Even small values of \( D \) can be statistically significant under certain situations, and we are interested not in the strength of LD but rather how much of it cannot be attributed to chance. We therefore use \( p_D \), the probability that a value of the disequilibrium \( D \) as large or larger than that in the sample would be observed if there is no association in the population from which the sample is drawn, conditioned on the sampled allele frequencies at the two loci. This probability is given by the tail of Fisher’s exact test (Weir 1990, Ch. 3; Slatkin, 1994; Lewontin, 1995). Slatkin showed that completely linked sites evolving under a neutral model will often show significant association as the result of their shared ancestry. Golding (1984) and Hudson (2001) developed methods to calculate the probabilities of samples of two locus haplotypes with association as the result of their shared ancestry.

\( p_D \) is intuitively the sum of the individual probabilities of getting the exact \( a_{i,j} \) values that will give a \( D \) as or more extreme than observed.

\[
p_D = \sum_{i=0}^{L} p(a_{0,0} + i, a_{0,1} - i, a_{1,0} - i, a_{1,1} + i)
\]

(7)

Where \( L = \begin{cases} \min (a_{0,1}, a_{1,0}) & \text{if } a_{0,0}a_{1,1} > a_{0,1}a_{1,0} \\ -\min (a_{0,0}, a_{1,1}) & \text{if } a_{0,0}a_{1,1} < a_{0,1}a_{1,0} \end{cases} \)

We can calculate \( p_D \) here by counting the number of ways of getting the particular set of \( a_{i,j} \) values divided by the total number of configurations allowed by the observed allele frequencies. This probability is obtained as:

\[
p(a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1}) = \frac{(a_{0,0} + a_{0,1})(a_{0,0} + a_{1,1})(a_{0,1} + a_{1,1})(a_{1,0} + a_{1,1})!}{a_{0,0}a_{0,1}a_{1,0}a_{1,1}!} (8)
\]

Substituting this result into 7 gives a way to calculate \( p_D \) from Fisher’s exact test.

\[
p_D = \sum_{i=L_1}^{L_2} \frac{(a_{0,0} + a_{0,1})(a_{0,0} + a_{1,0})(a_{0,1} + a_{1,1})(a_{1,0} + a_{1,1})!}{(a_{0,0} + i)(a_{0,1} - i)(a_{1,0} + i)(a_{1,1} + i)(a_{0,0} + a_{1,0} + a_{0,1} + a_{1,1})!} (9)
\]

The values of \( L_1 \) and \( L_2 \) are given by:

\( L_1 = 0 \) and \( L_2 = \min(a_{0,1}, a_{1,0}) \) if \( a_{0,0}a_{1,1} > a_{0,1}a_{1,0} \)

\( L_1 = -\min(a_{0,0}, a_{1,1}) \) and \( L_2 = 0 \) otherwise.

7
2.2 From Pairs to Patches of $p_D$

After calculating $p_D$ for all pairs of sites on each chromosome it is necessary to determine how many areas of high $p_D$ exist and where they are located. We developed a routine to calculate the number of high $p_D$ patches given a threshold value. To decide how many patches of $p_D$ to allow we used an ad-hoc criteria based on the number of SNPs the HapMap found on that chromosome. The formula used to calculate the number of patches on chromosome $i$ is:

$$P_i = \frac{n_i^2}{2 \times 10^8}$$

(10)

This is based off of the fact that $n_i/2$ is the number of SNPs on chromosome $i$, and the fraction of these we wanted to see represented by patches is $1/10^8$.

When breaking up pairs of sites into patches it is important to have some measure of the distance between them. One way to do this is as the probability of recombination. A problem with using this as a distance measure is that it is not additive. Simply using the number of bases between two sites is also not effective because the recombination rate per base pair is not uniform across the genome. Instead, what is used is a genetic map distance that takes into account the fact that the probability of recombination asymptotically approaches .5 as the physical distance between sites grows large. Using a marker set of genetic map distances from estimated recombination rates, it was possible to interpolate the position of each SNP. When it was necessary to convert between the map distance and recombination probability the Kosambi map function was used (Kosambi, 1944). The genetic map distance is denoted as $r$.

In the patch-finding algorithm we refer to a pair of sites as a peak. Each peak thus has two coordinates (one for each site) and a $p_D$ value. We define a patch as a group of peaks where both coordinates of that peak are within at least some distance $r_{min}$ from the coordinates of at least one other peak. The number of patches found will then be determined by how the vast number of peaks are filtered. The first step is to mark all the peaks whose coordinates are less than $r_{min}$ apart. After marking all these peaks the routine removes all non-diagonal peaks with a $p_D$ value below a the given cutoff. The remaining peaks are aggregated into patches if both of their sites were within $r_{min}$ of each other. Patches which contained a diagonal peak were then removed because this association with closely linked sites could confound the results by allowing drift to play a major role.

Since we already know how many patches should be found on each chromosome, the above routine was run for varying cutoff values of $p_D$ to determine which resulted in the desired patch number. Since values of $p_D$ are not independent within a patch, we took the peak with the largest $p_D$ value as $p_D^{max}$ and used this as a measure for the entire patch. The mean $p_D^{max}$ for a chromosome was used as an overall signal for the significance of long-range LD. This routine was applied to the randomized data in order to provide a comparison between the two. The cutoff $p_D$ values were used to determine how many patches existed for a randomized set.
2.3 Generation of Null Distributions

In order to determine if the number of $p_D$ patches (as defined previously) and the signal of $p_D$ in each patch is due to a genuine biological or demographic phenomenon it was necessary to quantify the degree of signal and patch structure one would expect if all patterns seen were a product of random chance. Randomly generating $N$ haplotypes and then performing all the above mentioned analyses will not accomplish this. Since our statistic of $p_D$ is conditioned on the observed allele frequencies, these should be kept constant for each site. However, shuffling SNPs between all haplotypes does not generate an appropriate null distribution either. The reason for this is that over short map distances we expect LD that is not due to selection, demography, or random association but to the fact that these areas are physically linked in the genome. Any randomizations not containing this local structure would confound the fact that the subsequent analysis directly accounts for it when generating the null distribution.

The way we performed randomizations was to take each chromosome and assign it at random to another such that each chromosome is pointed to (note that the assignments need not go both ways). This gives two sets of haplotypes: shuffled and real. The shuffled haplotype has the property that the same shuffling is performed at each site. Viewed from the perspective of a pair of sites, one in each set, this is equivalent to shuffling the haplotypes at one of those sites while keeping the other the same. By calculating $p_D$ between the shuffled and real sets of haplotypes we are able to randomize $p_D$ while keeping local structure intact. We move along the real haplotype set from left to right and calculate $p_D$ with every site to the left using the shuffled haplotypes at those sites. This randomization process and its benefits are illustrated below:

<table>
<thead>
<tr>
<th>$s_1$</th>
<th>$s_2$</th>
<th>$s_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Calculating $p_D$ for possible combinations we get:

\[
p_{D_{s1s2}} = 0.004662; p_{D_{s1s3}} = 0.587413; p_{D_{s2s3}} = 0.61655
\]

Here sites 1 and 2 are in LD while site 3 is in equilibrium with both of them. We now shuffle site 1 such that it is in LD with site 3 and see that after this site 2 will be as well, while the $p_D$ value between them is preserved exactly.
Calculating $p_D$ for the new combinations we get:

\[
p_{D_{s_1's_2'}} = 0.004662; \quad p_{D_{s_1's_3}} = 0.004662; \quad p_{D_{s_2's_3}} = 0.0250583
\]

Viewed from the opposite direction: calculating $p_D$ between a shuffled site and two real-ordered ones, the same situation would be expected if the real sites were tightly linked. This would be equivalent to shuffling site 3 so that it has a high $p_D$ with 1 and 2 instead of the other way around. Another property of the randomization procedure that should be mentioned is that, as we move from left to right across each chromosome we will not tend to calculate high $p_D$ values between sites that are close together. This is because one of the sites will always be shuffled. That is to say, $p_{D_{s_1's_2'}}$ is never actually calculated because $p_{D_{s_1's_2}}$ is instead. Whether $p_{D_{s_1's_2}}$ is high or not turns out not to matter due to the fact that in the patch-finding algorithm close sites are removed anyway. Rather, the effect of $p_{D_{s_1's_2'}}$ contributes to the calculation of $p_{D_{s_1's_3}}$ and $p_{D_{s_2's_3}}$.

The randomization routine was performed on each chromosome and the results were processed in the same way as the real data, with the exception that the cutoff $p_D$ values were those taken from the real data to ensure a certain number of patches. In all runs of the processing routine we chose $r_{\text{min}} = .25 \text{ cM}$. This value is greater than 90% of the haplotype blocks (The International HapMap Consortium, 2005 Suppl. Fig. 9b). The effective population size of the YRI population is estimated to be 7,500 (Tenesa et al., 2007), so our minimum distance corresponds to $\rho = 4N_e r = 75$. That is expected on theoretical grounds to provide enough recombination to very largely eliminate disequilibrium from shared ancestry (Hudson, 2001).

3 Results

3.1 Examination of $p_D$ Patterns

The long-range patterns of $p_D$ varied widely between chromosomes. We show this using two graph styles that highlight different aspects of the distribution. The first is a triangle that plots a chromosome against itself so that both axes represent the same chromosome.
Patches are shown as circles, the radii of which are scaled to the $p_D^{max}$ value for that patch. Patches closer to the diagonal are those with shorter genetic distances between their regions, while those closer to the origin have regions at more distal ends of the chromosome. This plot style has been used before to show haplotype structure over short sections of chromosome, and we have adapted it to visualize larger scale patterns. To provide an example of the distribution differences between chromosomes we present triangle plots of chromosomes 7 and 10 (Figure 1).

One feature of this triangle plot is that chromosome 7 appears to have multiple patches that are near to each other or even overlap. There is the possibility for overlap in the sense that sites can be in LD with multiple others, but it should be noted that no SNP pairs are shared between patches, and their apparent overlap is due to the fact that size was made proportional to $p_D^{max}$. Yet there is clearly a clustered nature to the distribution and this is even more pronounced in 7 than 10. In particular, some regions around 77, 8, 25, and 163 cM seem to make an inordinately high number of associations. Both tend not to have patches between the extreme ends of the chromosome, while associations between regions that are relatively close together are common. This is likely a product of the overall greater number of site pairs within close distances of each other. The scarcity of patches in some areas of the plot may be partially attributed to the fact that, although SNPs were evenly spaced over physical distances, the recombination rate could vary enough that this even spacing does not carry over to map distances.

Figure 1: Triangle plot: distribution of patches on chromosomes 7 and 10. The circles marking the patches are placed at the location of the peaks with the greatest value of $-\ln(p_D)$ from that patch. This is refered to as $p_D^{max}$. The diameters, not the area of the circles corresponds to $p_D^{max}$. Chromosome 7 contains 121 patches while 10 has 124.
Though informative, the triangle plots do make it difficult to see certain features of LRLD. One of these is the number and strength of patches formed by a particular chromosomal region, especially when those patches might have a high $r$. To visualize better the distribution of long-range interactions we present the same results shown in Figure 1 in an alternate form. The circle plots shown in Figure 2 bend the chromosome into a semicircle and connect patches with straight lines. The color of these lines is scaled to the $p_D^{max}$ for the patch.

Using this method we can take another look at the interesting areas identified in the triangle plot (77, 8, 25, and 163 cM on chromosome 7). It is much easier to see which other sections the 77 cM region of chromosome 7 forms patches with. However, it is more difficult to see how many and where the region at 163 cM forms patches because these have a smaller $r$. In this plot we can again see clearly chromosome 7 has more patches with high $p_D^{max}$ than 10 and also that these are clustered in certain areas. All save a few patches on chromosome 10 are within a narrow $p_D^{max}$ range. In chromosome 7 we can now see a three-way interaction with sites around 63, 13 and 132 cM. These patches also have some of the strongest interactions on the chromosome. Whether there is a significant number of three-way patches relative to chance expectation was not investigated, but this feature does show up on chromosomes 1, 2, 6, 8, 10, and 20.

![Circle plot: distribution of patches on chromosomes 7 and 10. Lines connect top peaks and the color scales to $p_D^{max}$.](image)

Figure 2: Circle plot: distribution of patches on chromosomes 7 and 10. Lines connect top peaks and the color scales to $p_D^{max}$. 

12
To investigate whether the distance had an effect on the number and amount of LD between patches we calculated the correlation between the $r$ between sites and the number of patches or $p_{D}^{\text{max}}$. After normalizing for the number of possible SNP pairs at each interval genetic distance, most chromosomes show a negative correlation between distance and number of patches. On the larger chromosomes this association was usually significant. As the size of a chromosome grows, the number of possible pairs at long genetic distances ($0.4 < r < 0.5$) increases greatly relative to shorter distances. This finding indicates that we draw disproportionately from the short distances for patches. There is only a significant negative correlation between $p_{D}^{\text{max}}$ and $r$ for chromosomes 1 and 2. Other chromosomes also show negative correlations but fail to reach significance.

### 3.2 Comparison to Randomizations

Randomizations show that the majority of chromosomes show a greater amount of LRLD both in the number of patches present and the strength of LD than one would expect by chance alone. Based on 1000 randomizations we found a significant number of LRLD patches in all chromosomes except 14, 19, and 21 ($\alpha = 0.05$). Additionally, a significantly higher average $p_{D}^{\text{max}}$ than expected by chance was found in all chromosomes except 19. Table ?? shows these results for each chromosome.
Table 1: Comparison of disequilibria patches to null distributions and potentially selected regions. Number of patches for each chromosome was determined by equation 10. Column Akey 2 denotes patches where both sites fell within selected region, while Akey 1 denotes patches where only one site does. Bold values indicate $p < .05$.

3.3 Comparison to Potentially Selected Sites

To evaluate the potential importance of the chromosomal patches we have identified, we compared them to a list of genomic regions that have been identified by multiple studies as potential targets for positive selection (Akey, 2009). We asked whether, from each patch, one or both of the top sites fell within one of these regions. To assess the probability of identifying sites with selected regions by chance alone, we performed 100,000 randomizations for each chromosome by dropping the selected regions at random and counting where our top peaks fell. These results are shown in (Table 1) where patches with both of the top sites in a potentially selected region are under Akey 2 and those with one are under Akey 1. In general, although we often find one or both sites within a candidate regions, this association does not appear to be significant. This is because...
the cumulative size of the identified regions is large, with individual regions as large as 10 cM.

4 Discussion

Previous studies of LD have examined genome-wide patterns but have focused primarily on distances shorter than 300kb. When long-range LD has been found it has been as extended haplotype blocks. LD over these short distances has been extremely useful in a number of different contexts including population structure, natural selection, disease association, and fine scale mapping of recombination rate. Much less attention has been paid to patterns of long-range LD, but these could conceivably be used to complement such work. For instance, when mapping a disease or phenotype to several genomic regions it could help to know beforehand which distant sites are in LD and for what reason. Knowing when to assume independence is important.

This study aimed to characterize the extent and distribution of long-range LD in the human genome. We wanted to find covarying regions that were both statistically significant and not confounded by the short-range LD that exists due to tight linkage and genetic drift. To account for the fact that we were searching a large number of pairs of sites we, quantified our results with the statistic $p_D$ which gives the probability of finding by random chance a disequilibrium $D$ as high or higher than observed, conditioned on the sampled allele frequencies. Patterns were further refined by the organization of linked pairs into patches and these were compared to null distributions generated by randomization. Significant regions were detected in 19 of 22 chromosomes, while general patterns differed greatly between chromosomes in the distribution of $p_D$ and the genetic distance between sites (YRI HapMap phase II). With a few exceptions neither the number of patches or $p_D$ showed a significant correlation with the distance $r$ between sites, indicating that distance did not have a great effect on the significance level. The association of number of patches with $r$ was a different story, indicating the way we identified patches was biased towards small $r$. When comparing SNPs with the highest level of $p_D$ to a list of putative selected regions (Akey 2009), we found that on most chromosomes large number of our identified sites fell within these areas. However, the associations were generally not statistically significant. This is attributed to the fact that the regions can span a large proportion of each chromosome (up to 10%).

The factors that could have created the observed patterns of LD are epistasis, recent population admixture, genetic structural variation, and artifacts in the data collection. Epistasis can create long-range LD if selection acts on particular pairings of alleles, as this would cause favored pairs to be overrepresented relative to allele frequencies. Recent admixture could create these patterns if allele pairs are present in different frequencies in a group that has begun mixing with the sampled population. Some of these pairs would not yet had time for their association to be broken up by recombination. Genetic structural variation can feasibly cause long-range LD if these features cause the recombination rate to be far less than expected for between alleles known to be physically distant. Data artifacts, though we have no good reason to believe they exist to this extent, could conceivably explain the patterns. A potential scenario is one where on different days
when different chromosomal portions were sequenced, the machinery showed biases for particular nucleotides. These biases would then show up in our analysis as long-range LD.

Our study provides no mechanism to formally distinguish between these potential causes of long-range LD. However, we can evaluate generally how reasonable they each are. Epistasis in particular is an unlikely candidate because the degree of selection needed to produce a pattern such as that we observe would be extreme. That is, the number of selective deaths (failures to reproduce) that would have had to occurred to maintain this level of association is unrealistic for a human population. Structural variation is also difficult to justify. Such features would have to exist over long distances, be present in a large portion of the population, and have escaped detected by human geneticists. It has been shown that LD can be used to detect inversions on human chromosomes (Cáceres et al., 2012), but there is little justification that inversions would produce the discontinuous pattern observed. We are thus left with recent admixture as the most likely explanation.

Given that our analysis cannot differentiate between the discussed ways that long-range LD can be caused, a logical next step would be the development of methods that can. Even if, as we suggest, the patterns are due primarily to admixture, it is possible that part of what we observe can be attributed to the other factors. Now that we have a method for successfully characterizing long-range LD, it might be time to think about how exactly to incorporate this into scans and other studies that comb the genome for information. The presence of nonrandom long-range LD in the human genome certainly exists, but what it’s exact importance is remains to be seen.

References


