12-2017

This is just a phase : the impact of population structure on haplotype phasing and linkage disequilibrium measures at functional genetic sites.

Roxanne Kaaren Leiter
University of Louisville

THIS IS JUST A PHASE:
THE IMPACT OF POPULATION STRUCTURE ON HAPLOTYPE PHASING AND LINKAGE DISEQUILIBRIUM MEASURES AT FUNCTIONAL GENETIC SITES

By

Roxanne Kaaren Leiter
B.A. University of Louisville, 2014

A Thesis
Submitted to the Faculty of the
College of Arts and Sciences of the University of Louisville
in Partial Fulfillment of the Requirements
of the Degree of

Master of Arts
In Anthropology

Department of Anthropology
University of Louisville
Louisville, Kentucky

December 2017

This is Just a Phase:

The Impact of Population Structure on Haplotype Phasing and Linkage Disequilibrium Measures at Functional Genetic Sites

By

Roxanne Kaaren Leiter
B.A., University of Louisville, 2014

A Thesis Approved on
December 1, 2017

By the following Thesis Committee

Dr. Christopher Tillquist, Thesis Director

Dr. Fabian Crespo

Dr. Michael Perlin

DEDICATION

This thesis is dedicated to my parents

Rebecca Benson-Leiter

And

Philip Leiter

who have given me endless support and encouragement as well as the opportunity to follow my dreams wherever they may take me. They are the reason I can call myself plucky.

ACKNOWLEDGMENTS

I would like to thank my thesis director, Dr. Christopher Tillquist, for his invaluable mentorship and guidance through this thesis. Dr. Tillquist has shown me what it means to be a true scholar and academic in the face of adversity. I would also like to thank my other committee members, Dr. Fabian Crespo and Dr. Michael Perlin for their review of my work over the years it has taken to complete this thesis. I would also like to thank the family members who have been in my corner since the beginning of this endeavor: Rebecca Benson-Leiter, Philip Leiter, Zoran Kyle Benson, Sarah Stika, Dane and Virginia Benson and the rest of my Ohio family.

ABSTRACT

THIS IS JUST A PHASE: THE IMPACT OF POPULATION STRUCTURE ON HAPLOTYPE PHASING AND LINKAGE DISEQUILIBRIUM MEASURES AT FUNCTIONAL GENETIC SITES

Roxanne K. Leiter

December 5, 2017

The block-like structure of the human genome has been the subject of many scientific papers and is of practical significance in large-scale genome-wide association studies. How stringent haplotype block boundaries are within and between populations has been the subject of ongoing debate within human population genetics. This thesis will contribute to the description of universal and population-specific haplotype blocks at functional sites, namely across the IL-10 gene family (including IL-10, IL-19, IL-20 and IL-24), which is involved in a number of immune system processes, and MAPKAP-K2, an adjacent and functionally significant kinase gene. Beyond the description of blocks across these sites in different populations, this thesis will also measure the impact of the haplotype phasing process on downstream applications of linkage disequilibrium analysis, which underlies much of the research on human haplotype blocks.

The five genes in this analysis span just over 200kb on the *q* arm of chromosome 1. A total of 80 samples from the Coriell Institute of Medical Research

are used in this analysis and represent Andean, Basque, Chinese, Iberian, Indo-Pakistani, Middle Eastern, Russian, South African and North African populations.

Some haplotype block boundaries were concordant with gene boundaries with most populations showing a consistent boundary between IL-20 and IL-24 and at least half of the study populations showing consistent boundaries between MAPKAP-K2, IL-10 and IL-20. The only gene boundary lacking a persistent haplotype block boundary was between IL-19 and IL-20. The haplotype phasing programs PHASE and Beagle shared 13 of 15 haplotype block boundaries in common while MDBlocks and Beagle only shared 2 haplotype block boundaries and PHASE and MDBlocks only shared 1 block boundary.

These data indicate that there are indeed population-specific differences in the distribution of LD across these five sites. Despite these differences, there is a general trend of high LD across each gene with a breakdown of LD at gene boundaries across all populations

¢!.[9 hC /hb¢9b¢{

[L{¢ hC ¢! . [9{

TABLE PAGE

[L{¢ hC CLD¦w9{

FIGURE PAGE

[ĿbY!D9 5Ŀ{9v¦Ŀ[Ŀ.wĿ¦a

Linkage disequilibrium, or LD, is defined as a non-random association between two or more genetic loci (Lewontin and Kojima, 1960; Slatkin, 2008). Loci are said to be in LD when there is a statistical association between two alleles at different loci such that they co-occur together more frequently than would be expected if they were independently assorting (Lewontin and Kojima, 1960). In contrast, loci that are independently assorted and not affected by natural selection, gene flow or mutation are said to be in linkage equilibrium (Flint-Garcia et al., 2003). The term linkage disequilibrium was originally coined by Lewontin and Kojima but did not grow in popularity until its usefulness was realized with genotyping and mapping technology that was developed several decades later (Lewontin and Kojima, 1960; Slatkin, 2008). With genotyping becoming more cost effective, reliable and efficient, polymorphism and disease susceptibility data could be more readily understood in concert with each other (Morton, 2005). The ability to establish correlations between disease phenotypes and particular polymorphisms within the genome gave the concept of linkage disequilibrium the practical significance that maintains its popularity in genetics and disease literature.

Single gene diseases such as Huntington's disease, cystic fibrosis and Alzheimer's disease, were major victories for linkage association studies but a finite number of Mendelian disorders and an overwhelming number of disease phenotypes with unexplained genetic correlations led researchers to establish other means of making associations between diseases and underlying genetic causes (Flint-Garcia et al., 2003; Morton, 2005). With the consideration of multifactorial genetic contributors to disease as well as environmental influences, the feasibility of association studies decreased and the cost of performing them consequently increased (Ardlie et al., 2002; Morton, 2005). The later discovery of haplotype blocks, or fairly large regions of strong LD, in several study organisms and eventually humans made association mapping more efficient because a carefully chosen marker could represent larger stretches of untyped DNA (Slatkin, 2008). The population genetic implications of haplotype blocks will be discussed later in this section but their initial utility in drastically reducing the cost and difficulty of complex genetic association studies spurred interest in haplotype block research.

Realizing the practical limitations of LD analysis in association mapping, more specifically in identifying causative alleles that contribute to a high percentage of disease phenotypes, more recent LD studies have focused on applications to the population genetics of past demographic events and natural selection (Slatkin, 2008). Linkage disequilibrium patterns can reflect past selection events, population histories, population structure, gene flow and mutation events but can also be influenced by local recombination rates, sex differences and population characteristics (Ardlie et al., 2002; Slatkin, 2008).

Influences on LD

Different types of natural selection can determine the overall impact of LD on a particular locus. For example, it has been shown in several studies that balancing selection can maintain strong LD leading to the assertion that selection may act specifically on blocks of LD rather than individual loci (Slatkin, 2008). Natural selection can create what is referred to as locus-specific bottlenecks as selected sites increase or decrease in frequency alongside non-selected but neighboring sites (Maynard Smith and Haigh, 1974; Charlesworth et al., 1993; Ardlie et al., 2002; Flint-Garcia et al., 2003). Also referred to as the hitchhiking effect, haplotypes surrounding a selected locus can be carried into high frequency with a selective sweep depending on the strength of selection (Maynard Smith and Haigh, 1974; Ardlie et al., 2002; Nachman et al., 2004; Handley et al., 2007). Alternatively, background selection entails a decrease in frequency of neutral sites immediately surrounding negatively selected loci because of their proximity to those deleterious sites (Charlesworth et al., 1993). Balancing selection, which favors the maintenance of two or more alleles at a certain locus in a population, can create a pattern where each selected allele has a corresponding suite of non-functional associated sites (Charlesworth, 2006). This tendency for non-selected sites to 'travel' with sites under some kind of selection certainly presents a quandary in association studies as it can be unclear which sites are of functional significance if strong LD is maintained in the region. Huff et al. elaborate on methods for identifying signatures of positive selection in LD studies and conclude that some methods are more robust than others because of the ability to account for confounding factors that can have an impact on LD values, such as variable population histories (Huff et al., 2010).

Recombination greatly affects LD and serves to decrease established LD over time as recombination events, or crossovers, occur in the region (Ardlie et al., 2002;

Flint-Garcia et al., 2003). A very simplistic model of the impact of recombination rates and distances given by Ardlie et al. illustrates the interplay of LD and the properties of recombination:

$$D_t = (1 - r)^t D_0$$

D_0 is the extent of LD at the starting point, r is the recombination rate and D_t is the extent of LD after t generations (Ardlie et al., 2002). Ardlie and colleagues caution against a dogmatic reliance on the general rules of LD decay across small regions as neighboring loci can be very distinctly unlinked. Recombination rates are variable across the genome and can also vary on an individual and sex-specific basis (Broman et al., 1998). Recombination hotspots, which will be discussed in greater detail in the following section of this review, can also influence LD patterns across the genome (Ardlie et al., 2002). It has also been reported that recombination patterns are comparatively low around the centromeres of the chromosome (Nachman et al., 2004). Nachman et al. found this to be the case with two genes, Msn and Alas2, on either side of the X chromosome's centromere showing abnormally low rates of recombination. Considering the effects of both the local recombination and the level and type of selection on a particular locus, Nachman et al. point out that selection will have a much stronger impact on the persistence of linked neutral loci if the local recombination rate is low in that region.

Genetic drift can create LD, even where it did not exist previously, by random sampling (Slatkin, 2008). The effect of genetic drift on linked loci is more straightforward as the population size can impact the extent of observed LD; it follows that smaller populations will on average exhibit stronger signals of LD as a function of group size (Ardlie et al., 2002; Slatkin, 2008). In a simulation study by Zhang et al. where recombination sites are randomly and uniformly distributed across a region, block-like patterns of LD emerge as a result of genetic drift alone (Zhang et al., 2003). They found that the observed blocks in the simulation data did not extend beyond 100 kb, a result which is seen in other studies using empirical data (Reich et al., 2001; Zhang et al., 2003). Zhang et al. also offer the caveat that the pattern of LD across the human genome is likely due to the interplay of forces, including recombination hotspot distribution, natural selection, genetic drift and gene conversion.

Changes in population size also have an impact on LD such that a decrease or increase in population size can determine the amount of LD that accumulates or decays over time (Slatkin, 2008). LD tends to decrease in populations experiencing swift population growth as such growth counteracts the effects imposed by genetic drift (Ardlie et al., 2002). Population bottlenecks, which entail a great reduction in population size, generally increase the amount of LD, as haplotypes are usually lost in the process (Flint-Garcia et al., 2003; Slatkin, 2008). For human populations specifically, it has been proposed that a high degree of LD in some populations compared to others is indicative of a past bottleneck event (Slatkin, 2008)

Varying mutation rates, particularly high mutation rates such as those seen in areas high in CpG dinucleotides, can diminish the amount of LD in an area because it obscures any LD that would otherwise accumulate (Ardlie et al., 2002). A signal of LD can be readily maintained until the extended haplotype is disrupted by a novel mutation.

Gene flow, or admixture, can manipulate LD patterns, specifically increasing LD at the time of the admixture event if the populations are sufficiently divergent, though the patterns are not durable if random mating ensues afterward (Pritchard and Przeworski, 2001; Slatkin, 2008). Because of this trend, a population that has very recently experienced gene flow with another population that is genetically dissimilar may show unusually high patterns of LD, which then decays as a function of local recombination rates (Pritchard and Przeworski, 2001).

Beyond selection and demographic forces, the scale at which LD is measured has an impact on the extent of LD encountered. In a review of many papers on LD, Pritchard and Przeworski note that long-range LD studies tend to discover more extensive LD than what is expected while less than expected LD is often shown at smaller scales (Pritchard and Przeworski, 2001). Pritchard and Przeworski note that current demographic and selection models of LD account for the long-range pattern as these forces affect large blocks of the genome. They posit that gene conversion, a mechanism similar to crossing over but distinct in that it involves a one way transfer of genetic material between homologous chromosomes as opposed to an equal exchange, may explain the lower levels of LD at small scales though they do admit that not much is known about the extent of gene conversion in humans (Pritchard and Przeworski, 2001).

Measures of LD

D is the most commonly cited statistic in linkage disequilibrium studies, although its limitations have been expounded upon in several papers (Hedrick, 1987; Slatkin, 2008) . The measurement of D is straightforward and relies on the observed difference between a two-locus haplotype and what would be expected under the conditions of independent assortment between two loci (Ardlie et al., 2002). The simple equation for D is as follows:

$$D = P_{AB} - P_A \, x \, P_B$$

In this equation A and B represent two adjacent loci with four possible alleles (A, a, B, b). The observed haplotype frequency of any combination is represented by P_{AB}. The expected haplotype frequency is represented as the product of the frequency of the two alleles in the observed haplotype frequency, $P_A \, x \, P_B$ (Ardlie et al., 2002).

Nevertheless, several statistics have been developed employing D in various mathematical formulas that describe different aspects and measures of LD. D may be useful between two adjacent loci but is less reliable across several loci or across different populations for comparative LD measures because the statistic is sensitive to varying allele frequencies (Hedrick, 1987; Ardlie et al., 2002).

r^2 is another commonly used and often preferred measure of linkage disequilibrium because it indicates the predictive power of one locus on a separate locus but also doesn't allow for a value of 1 to be reached unless allele frequencies at both sites are equal (Flint-Garcia et al., 2003).

The equation for r^2 is as follows:

$$r^2 = \frac{(\pi_{AB} - \pi_A \pi_B)^2}{\pi_A \pi_a \pi_B \pi_b}$$

where a pair of loci with alleles A and a at locus one, and B and b at locus two, with alleles frequencies indicated by $\pi_A, \pi_{a,} \pi_{B,}$ and π_b and haplotype frequencies indicated by $\pi_{AB}, \pi_{Ab}, \pi_{aB}, \pi_{ab}$ (Pritchard and Przeworski, 2001). r^2 is preferred over traditional measures of D by most population geneticists because it has the ability to control for variable allele frequencies with more precision by building them into the equation (Ardlie et al., 2002).

D' is a statistic based on Lewontin's D that was developed to account for LD across multi-allelic markers to measure long-range LD, although complications are introduced by variable sample sizes and allele frequencies with low values (Ardlie et al., 2002; Flint-Garcia et al., 2003; Zhao et al., 2007).

The equation for D', with the same variable attributes used in the measure for r^2, is as follows:

$$|D'| = \frac{(D_{ab})^2}{\min(\pi_A\pi_{a,}\pi_B\pi_b)} \text{ for } D_{ab} < 0;$$

$$|D'| = \frac{(D_{ab})^2}{\min(\pi_A\pi_{a,}\pi_B\pi_b)} \text{for } D_{ab} > 0$$

Statistically significant correlations of two loci using D' and r^2 are most commonly evaluated using Fisher's exact test and a similar measure, a multifactorial permutation analysis, is used for multi-allelic sites (Flint-Garcia et al., 2003).

Several authors argue that there are benefits in using either D' or r^2 because they measure different aspects of LD and both control for variation in allele frequencies, at least to some degree (Hedrick, 1987; Flint-Garcia et al., 2003). r^2 is more sensitive to recombination and mutational histories of particular loci while D' is more precise in estimating divergences in recombination (Flint-Garcia et al., 2003). Despite this latter strength, Flint-Garcia et al. caution against an overreliance on D' measures due to the introduction of bias when the sample sizes of the study populations are low because of its inability to appropriately account for small allele frequencies.

The Four Gamete test was used to evaluate LD in this thesis. This algorithm measures the frequency of 4 possible two-marker haplotypes for each pair of loci along the region under assessment. Recombination events are established at two sites where all possible combinations of markers are observed at least 1% of the time in the population (Hudson and Kaplan, 1985; Wang et al., 2002). The Four Gamete test assumes no recurrent and/or backward mutation, as that might generate the association between a pair of loci that would otherwise be interpreted as evidence of a historical recombination event. A haplotype block is determined when there is a contiguous set of adjacent SNPs for which there is no evidence of a recombination event (Wang et al., 2002). Wang et al. caution that a recombination event may be overlooked if the sample size is small and variability at adjacent sites is not adequately represented, which may make haplotype blocks appear more extensive than they truly are (Wang et al., 2002).

LD in the Human Genome

The extent of LD in humans has been vigorously debated, although it is generally agreed upon that it has the potential to span large genomic distances ranging from 60-500 kb (Reich et al., 2001; Flint-Garcia et al., 2003). Important sex-specific differences in LD rates have been observed in human populations, with more recombination occurring at telomeres in males and more centromeric recombination in females (Broman et al., 1998; Kong et al., 2002; Lynn et al., 2004). Furthermore, there was more individual variation in autosomal recombination patterns in females where it was largely absent in males (Broman et al., 1998). Support for the Out-of-Africa model of early human expansion comes from several studies that have established more extensive LD in populations outside of Africa compared to those within Africa, which are assumed to be older and more established (Flint-Garcia et al., 2003).

Recombination hotspots, or areas where high levels are LD that are punctuated by regions of low LD, have been described in humans as well (Jeffreys et al., 2001; Flint-Garcia et al., 2003). Specifically, Jeffreys et al. found that areas of LD correlated with recombination hotspots in the major histocompatibility complex (MHC). The team used sperm-typing to establish where the hotspots were distributed due to crossover events and showed a great degree of divergence in recombination rates across the 216 kb region (Jeffreys et al., 2001). From the results of this study, the authors suggest that not only can the extent of LD be greatly influenced by hotspots but that LD patterns may also hold a degree of predictive power in determining these crossover hot spots at certain loci, as was the case with the MHC. It follows that fine-scale recombination maps may be useful in eliminating recombination as a factor in variable degrees of LD in studies seeking to establish other causes of such observations.

Complicating the impact of recombination on the generation and decay of LD patterns are the assertions that recombination patterns are highly variable (Lynn et al., 2004; Katzman et al., 2011). Recombination hotspots are known to vary between individuals of the same population, between males and females and between different chromosomes in the genome (Kong et al., 2002; Lynn et al., 2004). In humans, females tend to have a higher rate of recombination across all chromosomes when compared against males (Lynn et al., 2004).

In the interest of anthropological study, LD analysis can shed light on the potential historical processes and events that produced the current distribution of genetic variation (Wall and Pritchard, 2003). A point of contention within LD research, and a difficulty that will be addressed in this thesis, is how varying degrees of LD across and within human genes are interpreted alongside models of ancient human dispersion and settlement. Through simulations, Slatkin showed how variable population histories involving population growth versus stasis and local recombination rates impact measures of LD (Slatkin, 1994). Slatkin concluded in suggesting that rapidly expanding populations are more likely to exhibit less LD relative to populations with a historically constant size (Slatkin, 1994). He further explains that population size determines the extent to which genetic drift alone can produce significant linkage disequilibrium, which seems to occur more readily in populations with a historically consistent size. It is further noted in the paper that extensive LD can arise in founding populations if non-random associations are present in successful founders (Slatkin, 1994).

The distribution of LD patterns in and between human populations has been described in several studies (Reich et al., 2001; Gabriel et al., 2002). Reich et al. detailed a few general characteristics of LD distribution across human populations, namely that the extent of LD is greater in non-African populations when compared against sub-Saharan populations. The research team also found that the extent of LD was greatest in their Northern European samples, leading them to suggest that the discrepancy is due to historical events such as population bottlenecking or founder events (Reich et al., 2001). A paper by Wall and Pritchard also corroborated the general pattern described by Reich et al. where long-range and more extensive LD patterns were observed more in non-African populations (Wall and Pritchard, 2003). Another research team that studied the extent of LD across 3 entire chromosomes also found that blocks of LD were shorter in African populations relative to non-African populations (De La Vega et al., 2005). Although these explanations are in concordance with other anthropological findings, there are many factors beyond population history that contribute to the pattern and extent of LD observed (Ardlie et al., 2002). As mentioned earlier in this section, the degree of genetic drift imposed on a population and the relative size of that population can generate LD (Slatkin, 1994). Also mentioned previously, different natural selection regimes can entirely remove or amplify regions of neutral sites around selected sites in a population (Maynard Smith and Haigh, 1974; Charlesworth et al., 1993). Other notable influences on LD patterns are variable recombination and mutation rates, population structure and admixture (Ardlie et al., 2002).

¢I9 t9ht[LbD hC ¢I9 h[5 ² hw[5 !b5 tht¦[!¢Lhb {¢w¦/¢¦w9

The origin of modern humans and their subsequent dispersal of the species across the globe commands considerable attention in several fields of study, including population genetics. Modern human fossil evidence occurring outside of Africa as well as archaeological evidence of human activity places a conservative estimate of dispersal outside of Africa at 40 KYA (Jobling et al., 2014). The routes of this exit are still under debate but it is generally agreed upon that modern humans left Africa to colonize the Levant region before spreading further into Asia and Europe (Mellars, 2006).

From the perspective of population genetics, the origin and dispersal of modern humans necessarily entails an original source population and the movement of gene pools away from the source combined with increased isolation and diversification (Slatkin, 2008; Trifonova et al., 2012). Identifying and parsing the similarities of the human genome as well as its population specific differences underlies much of human population genetics and the evidence presented in favor or refutation of demographic models of human evolution.

Sewall Wright first described population structure ontologically and mathematically as the non-random breeding that occurs within a population or a deviation from the assumed state of panmixia (Wright, 1950). Wright suggests several factors including sex and age distribution, population density, and isolation by distance that may determine the degree of subdivision with a population.

The concept of population structure can certainly be applied to human populations, though there is a much greater potential for nuance in the patterns of population structure given differing population histories. Africa, boasting long and rich population histories, has had the longest time and the greatest opportunity to develop fine-scale population structure. This is evident in the higher effective population size and greater complexity of genetic variation within Africa (Tishkoff et al., 2009).

Human genetic variation is often described as clinal, meaning that variability increases or decreases with corresponding geographic distance from a source population (Handley et al., 2007). These clines are observable on two scales: genetic differentiation increasing on a local scale from population to population and an overall decrease in genetic variation with increasing distance from Africa (Handley et al., 2007). This view of human genetic variation is not universally accepted, however, and other research has depicted variation as clustered instead of clinal in nature.

(Rosenberg et al., 2005) argue that variation presents itself in clusters. Furthermore, the research team reanalyzed their original dataset with the intention of elucidating the effects of study design on their conclusions and found that the observed cluster pattern was robust (Rosenberg et al., 2005). Through the reanalysis, Rosenberg et al. paid special attention to the impact of the number of loci studied, the sample size and the number of clusters on the outcome of the study.

This retrospective analysis was prompted by criticisms by Serre and Paabo, who contended that the study design of the initial publication was influencing the observed clustering pattern (Serre and Pääbo, 2004). Although the argument can be misconstrued as an "either or" scenario, Rosenberg et al. implore that both clines and clusters are description of human genetic variation at different levels of inspection (Rosenberg et al., 2005). Perhaps taking the best of both models is yet another, that of isolation by distance, which makes similar predictions on the degree of genetic similarity between populations as distance increases between them but does not necessarily restrict those patterns to clines moving in singular directions (Relethford, 2004). Relethford demonstrates the applicability of the isolation by distance model on several lines of data, including blood cell polymorphisms, genetic data and craniometric features.

A north-south cline of genetic differentiation in European populations is often interpreted as evidence that is in concordance with a Neolithic dispersal from the Middle East throughout the rest of Europe (Chikhi et al., 2002; Auton et al., 2009). The demic diffusion model posits that the initial spread of agriculture was accomplished through movement of people, who transmitted the technology as they migrated into Europe. This model stands in contrast to another model that assumes that the transition was largely cultural and did not necessarily entail large-scale population migrations (Chikhi et al., 2002). There is considerable debate over this topic but several lines of evidence, including data from nuclear, Y chromosome and mitochondrial DNA support the demic diffusion model, though the debate is far from closed (Chikhi et al., 2002; Auton et al., 2009; Fu et al., 2012).

tI¸[hD9hDw!tI¸ !b5 a9!{¦w9{ hC !bL{h¢wht¸

Phylogeographic approaches to population genetics allow for considerations of physical space as a modifying factor in the distribution of genetic variation at the population level. Building on existing fields such as demography, geography and population history, phylogeography combines these research endeavors into a unified discipline that considers both micro- and macro-evolutionary processes in shaping geographical distributions (Avise, 1998). The geographic distribution of genetic attributes has been a topic of study for many decades and continues to reveal more insights into the effects of space on the human genome, both individually and collectively. Spatial patterning in the distribution of traits has been previous described for ABO blood groups and rhesus factor (Falsetti and Sokal, 1993).

Anisotropy reflects the degree to which a variable is directionally dependent (Jay et al., 2013). Spatial autocorrelation in particular measures the dependence of variables like allele frequencies on the values at physical locations nearby (Falsetti and Sokal, 1993). Much like linkage disequilibrium analysis, spatial autocorrelation measures the predictive power of one variable on the value of another variable in close proximity, capitalizing on the assumption that proximity is a predictive tool. Some estimates of spatial autocorrelation can estimate the particular bearing, in terms of cardinal directions into account, which summarizes the degree to which directional clines explain spatial patterns of variability in the data (Falsetti and Sokal, 1993). What is not provided by these analyses is the reason that any cline or a lack of a cline exists; therefore further research is needed to fully explain cline presence or absence.

I!t[h¢¸t9 tI!{LbD

In order for linkage disequilibrium to be assessed, haplotype phase must be estimated. Haplotype phase refers to the particular inherited combination of alleles on a contiguous stretch of chromosome and it must be estimated, either computationally or experimentally, because genotyping only reveals pairs of alleles (Browning, 2008). Underlying all haplotype phasing programs and built within phasing algorithms that estimate phase computationally are unique haplotype comparison processes. The link between this methodology and ancient human origins is the notion that the relative relatedness of individuals within a population has much to do with the geographical location and age of that population. Younger, more dispersed populations such as those that are furthest from Africa are expected to be more genetically homogeneous and have a larger effective population size (Tenesa et al., 2007). In terms of phasing, this should imply that fewer model haplotypes are needed to accurately phase a collection of samples from regions of the world where less genetic variability is expected. The opposite should be true for populations that are nearer to or within Africa. This conclusion is expected given that the shorter population histories of more geographically dispersed populations provide less opportunities for recombination to disturb extended haplotypes and that those geographically dispersed populations underwent recent successive bottleneck events as they moved out of Africa.

Sample population is one factor that is believed to impact the phasing process and may compromise phasing accuracy estimates when population, as well as sample size, are not accounted for carefully (Browning and Browning, 2012). Allele frequencies and the density of SNPs confound accuracy estimates when comparisons across different populations are conducted (Browning and Browning, 2012). There are expectations that populations that are further from Africa are more genetically homogenous and conversely, that populations that are closer to or within Africa harbor more variability, exhibit lower levels of LD and have more diverse haplotypes (Tenesa et al., 2007; Browning and Browning, 2012). If the sample size of a population is so small such that it is not representative of the population and may result in imputation inaccuracy, samples from closely related or neighboring populations can be added to the sample set to increase accuracy. It was also noted that the phasing of historically admixed populations may yield more robust estimates if there are samples from each population that contributed to the admixed population (Browning and Browning, 2012).

Haplotype Phasing Algorithms

There are three distinct statistical methods used in each of the programs used for phasing these data. The latest version of PHASE uses a Bayesian haplotype reconstruction method that combines expected haplotypes as a prior with the observed genetic data as the likelihood to determine the prior distribution. The prior distribution gives the method the predictive power needed to reconstruct haplotypes given the observed genotype data (Stephens and Donnelly, 2003). The novel approach in this latest version of the program combined an approximate coalescent prior approach and a more efficient computational technique to make the estimates conceivably more accurate and faster to obtain. The approximate coalescent approach addresses an issue when haplotype reconstruction models encounter an unobserved haplotype. In another approach outlined in the paper by Stephens and Donnelly, they describe how a Dirichlet prior model breaks up unresolved haplotypes in order to compare them to known haplotypes and then forces the model to choose a haplotype at random from that point. In the approximate coalescent approach, the haplotype comparison process is the same but the model has the ability to put more weight on a reconstruction that involves haplotype combinations that are similar but not identical to observed haplotypes (Stephens and Donnelly, 2003). This model, Stephens and Donnelly argue, allows for an appropriate weighting of the more likely scenario, a haplotype that is a slight deviation from what is already observed, when haplotypes cannot be broken up into smaller, observed haplotype sequences.

The primary objective of the novel haplotype phasing approach employed in the program Beagle was to make imputation more efficient given large influxes of data coming from genome wide association studies. Existing methods were not scalable for large datasets with many individuals and dense sets of SNPs (Browning and Browning, 2007). Beagle uses a localized haplotype-cluster model that determines haplotype phase by joining nearby alleles into a small cluster, comparing that cluster to real data and assessing model fitness accordingly before reclustering the haplotype (Browning and Browning, 2007). Browning and

Browning note that this approach allows for more flexibility in where recombination blocks might fall and how many clusters are appropriate for a given dataset, which more accurately captures the complexity in recombination patterns observed in biological populations.

MDBlocks uses minimum-description-length (MDL) criterion as a novel way to model haplotype phase. Where most measures of phasing algorithms summarize the decay of LD between haplotype blocks or the amount of diversity within blocks, the MDL approach considers both the decay of LD between haplotype blocks as well as the diversity of haplotypes within blocks simultaneously (Anderson and Novembre, 2003). The MDL principle is an advent from the computer sciences and is particularly valuable for summarizing large amounts of data because it takes any regularity to a dataset and compresses the data to summarize it most efficiently (Grünwald, 2000). In a review paper on the subject of the MDL principle, it is stated that the principle is ideal for minimizing regularity but still capturing data because it strikes a delicate balance between under-fitting and over-fitting a model to data such that it adequately summarizes without cluttering up the model with an excess of parameters (Grünwald, 2000). Given this consideration, model selection should favor those that are simple yet give a reasonable fit to the data rather than ones that summarize genetic data to a precise yet cumbersome degree (Grünwald, 2000).

The MDL criterion used in MDBlocks may be better suited for the purposes of this study because the adaptive model fitting approach may better account for population specific differences when phasing the data.

D9b9{ hC Lb¢9w9{¢

Most of the genes used in this study belong to a single gene family that is believed to be a result of gene duplication (Brocker et al., 2010). The IL-10 gene cluster includes IL-10, IL-19, IL-20 and IL-24, all of which are situated next to each other on the *q* arm of chromosome 1 (Kotenko, 2002; Jones and Flavell, 2005; Brocker et al., 2010; Hofmann et al., 2012). The cytokines are classified as part of the same gene family and share nucleotide sequence as well as amino acid homology but have a diverse range of functions. Some of these classification schemas rely on functional and genetic similarities while others are based on structural attributes of the cytokines and other characteristics (Brocker et al., 2010).

Although an entire literature review could be compiled on this subject alone, the function of interleukins and more specifically, the products of the IL-10 gene cluster will be briefly described first. Interleukins are a particular class of cytokines, which participate in immune function by acting as a messenger between immune cells (Brocker et al., 2010). Interleukins are primarily involved in immune cell development, activation and differentiation and can inhibit or promote inflammatory responses. Fifty-five interleukin and IL-related genes have been identified and can be found on nearly every chromosome in the human genome (Brocker et al., 2010).

The IL-10 gene cluster is situated on the *q* arm of chromosome 1 and spans about 200 kilobase pairs (Kotenko, 2002; Brocker et al., 2010). The basic structure of the 4 genes within the IL-10 cluster located on chromosome 1 is similar with most genes containing 5 exons and conservation of intron/exon junctions between the cytokines (Kotenko, 2002). Kotenko points out, however, that there is variation in this general pattern such that some genes may contain greater or fewer exons or vary in the number of bases in introns.

In light of the "crosstalk" between IL-10 gene cluster cytokines and other components of the immune system, some researchers have proposed that each member contributes to an intricate cytokine cascade that manifests itself as a pro- or anti-inflammatory response (Wang and Liang, 2004). Although the members of the IL-10 gene cluster share sequence and amino acid homology, there is a wide array of functions attributed to the cytokines within this family and they each interact with a variety of cells involved in the immune system and immune response.

Evolution of the IL-10 Gene Cluster

Cytokine genes are said to be one of the most rapidly evolving classes of genes in the human genome (Brocker et al., 2010). Surprisingly, however, the amino acid sequences of several cytokines within the IL-10 gene family are well conserved across species. This high degree of conservation is evidenced by studies showing that interleukin 24 cytokines from rats could bind with the appropriate human receptors (Wang et al., 2004). Prior to the study by Wang et al. it was suggested that although there was a high degree of similarity in sequence and layout of the IL-10 gene family between humans and rats, that the genes carried out diverse functions in those unique biological systems. Wang et al., after showing that the rat's equivalent of IL-24, MOB-5, could not only bind to but activate both heterodimeric IL-24 receptor complexes, they concluded that MOB-5 was a true homologue to IL-24 (Wang et al., 2004).

The arrangement of the cytokine genes within the IL-10 gene cluster is identical to the order of genes seen in murine animals and exon composition and length is conserved as well, although there is slight variation in this latter respect (Wang and Liang, 2004; Hofmann et al., 2012). It has been suggested that the IL-10 gene cluster arose through gene duplication and subsequent divergence of the duplicated sequences over time (Xu, 2004).

Interleukin-10

A brief introduction of the IL-10 cluster gene products at the beginning of each cytokine section will cover basic properties and functions of the cytokines but the function of each gene and cytokine will not be covered extensively until later on in this thesis. Instead, this section serves as a primer for a more detailed discussion of the IL-10 gene cluster's structural and functional properties. Figure 1 shows how the gene family is situated on the chromosome as well as the conserved pattern in mice.

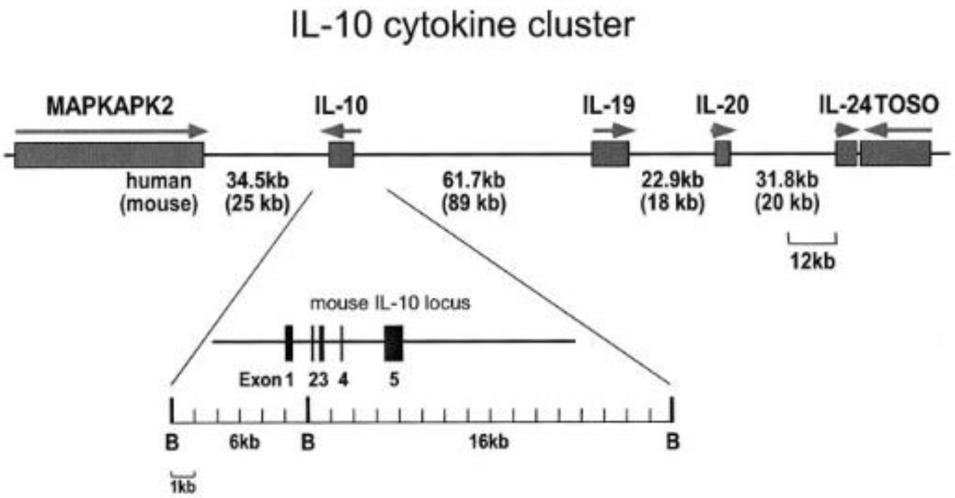


Figure 1: This image shows the IL-10 family genes as they are situated on chromosome 1. There are two non-immune genes, MAPKAPK2 and TOSO, flanking the gene cluster. Also shown is a homologous section of the IL-10 homolog in mice, which has a similar exon structure (Im et al., 2004).

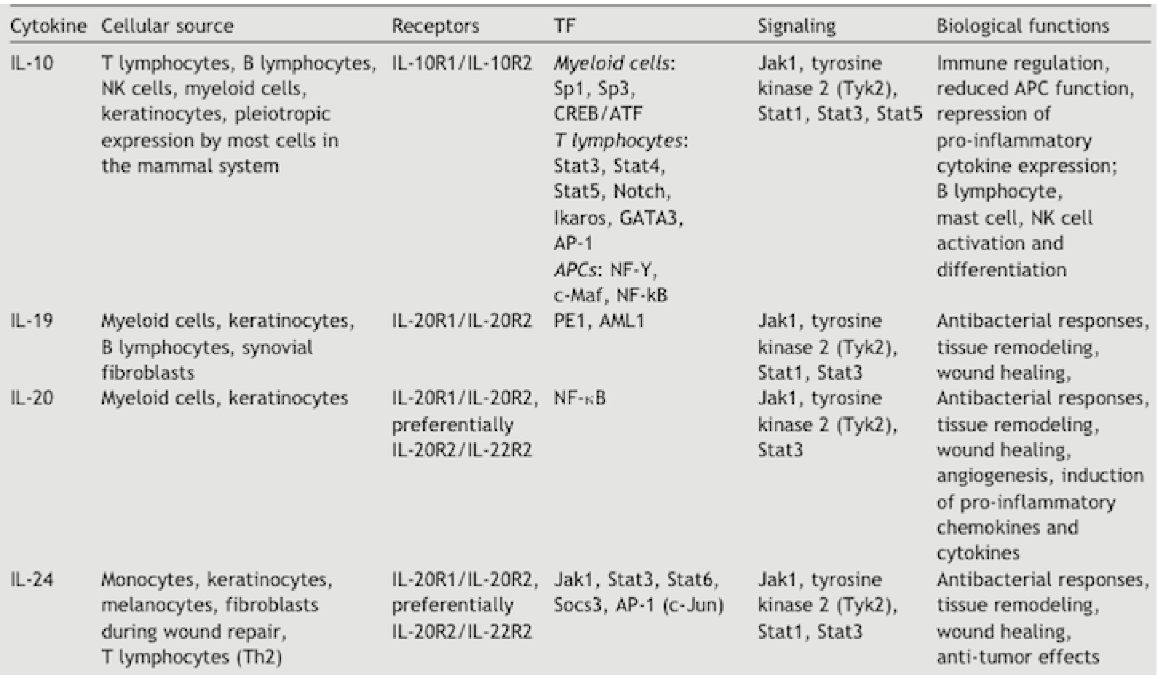
IL-10 Products

Interleukin-10 was the first of this family to be discovered and much more is known about IL-10 relative to its other homologs. IL-10 is considered to be an anti-inflammatory cytokine due to its ability to inhibit other cells, namely antigen presenting cells and macrophages, involved in inflammatory immune responses (Gallagher et al., 2000; Moore et al., 2001; Brocker et al., 2010). It also aids in regulating the responses of a number of immune cells including — but not limited to — B cells, several T cell subtypes, dendritic cells, etc. (Tone et al., 2000; Moore et al., 2001). Experiments in murine models with IL-10 knockouts as well as other association studies showed that several autoimmune disorders occurred when IL-10 was not functioning normally, lending more credence to the anti-inflammatory categorization of this interleukin (Kuhn et al., 1993; Cantó et al., 2014). The effective use of IL-10 as a treatment option for disorders caused by a general hyper-inflammatory state of the body furthers this claim (Cao et al., 2005).

IL-10 Structure & Function

The IL-10 gene has been extensively studied but the exact location of its promoter has not yet been determined. Despite this limitation, several potential regulatory sites have been identified in the IL-10 promoter region (Im et al., 2004; Jones and Flavell, 2005; Hofmann et al., 2012). Jones and Flavell reported the discovery of Th2-specific DNase hypersensitivity sites within the IL-10 promoter, although it should be noted that these experiments were performed in vitro. The specificity of interactions between interleukins and other cells in the immune system makes it difficult to speculate about its precise regulation.

IL-10 stimulation is pleiotropic in nature and thus yields varying functional responses from different immune cell types (Cheng and Sharma, 2015). The cytokine induces the proliferation and differentiation of B cells, T cells and natural killer cells and dampens pro-inflammatory responses by blocking the synthesis of or other otherwise inhibiting pro-inflammatory cytokines (Jordan et al., 2005; Cheng and Sharma, 2015). IL-10 inhibits the activity of mast cells, macrophages and dendritic cells, which are key components of inflammatory responses (Paul et al., 2012). Through the course of an immunological insult, macrophages are able to initially synthesize pro-inflammatory cytokines, including other interleukins, followed by the expression of anti-inflammatory cytokines, including IL-10 (Cao et al., 2005). This ability corresponds with the necessity for the immune system to produce a response that is robust enough to combat physiological insult without reaching a level where the host is harmed. IL-10 may also play an important role in modulating inflammatory responses in pregnant women, where a conservative immune response is necessary for maintaining the pregnancy. This role is evidenced by an association with aberrant IL-10 functioning with poor pregnancy outcomes (Cheng and Sharma, 2015).



Cytokine	Cellular source	Receptors	TF	Signaling	Biological functions
IL-10	T lymphocytes, B lymphocytes, NK cells, myeloid cells, keratinocytes, pleiotropic expression by most cells in the mammal system	IL-10R1/IL-10R2	*Myeloid cells*: Sp1, Sp3, CREB/ATF *T lymphocytes*: Stat3, Stat4, Stat5, Notch, Ikaros, GATA3, AP-1 *APCs*: NF-Y, c-Maf, NF-kB	Jak1, tyrosine kinase 2 (Tyk2), Stat1, Stat3, Stat5	Immune regulation, reduced APC function, repression of pro-inflammatory cytokine expression; B lymphocyte, mast cell, NK cell activation and differentiation
IL-19	Myeloid cells, keratinocytes, B lymphocytes, synovial fibroblasts	IL-20R1/IL-20R2	PE1, AML1	Jak1, tyrosine kinase 2 (Tyk2), Stat1, Stat3	Antibacterial responses, tissue remodeling, wound healing,
IL-20	Myeloid cells, keratinocytes	IL-20R1/IL-20R2, preferentially IL-20R2/IL-22R2	NF-κB	Jak1, tyrosine kinase 2 (Tyk2), Stat3	Antibacterial responses, tissue remodeling, wound healing, angiogenesis, induction of pro-inflammatory chemokines and cytokines
IL-24	Monocytes, keratinocytes, melanocytes, fibroblasts during wound repair, T lymphocytes (Th2)	IL-20R1/IL-20R2, preferentially IL-20R2/IL-22R2	Jak1, Stat3, Stat6, Socs3, AP-1 (c-Jun)	Jak1, tyrosine kinase 2 (Tyk2), Stat1, Stat3	Antibacterial responses, tissue remodeling, wound healing, anti-tumor effects

Table 1: This table shows the source cells, receptors, known transcription factors, signaling pathways and immunological functions of each cytokine (Hofmann et al., 2012).

Interleukin-19

IL-19 Products

The IL-19 gene and the associated cytokine were first identified based on its homology to the IL-10 gene using expressed sequence tags (ESTs) from sequence databases (Gallagher et al., 2000). Tests measuring its release from cells using traditional immunological stimulation techniques confirmed that this newly discovered cytokine responded to signals similar to those that stimulate IL-10 expression (Kotenko, 2002; Chen et al., 2006). Despite in vitro tests confirming its existence and similar stimulatory cues, the function of interleukin-19 is not well understood and cannot be concretely related to that of IL-10 although it is expected to play some role in the orchestration of inflammatory responses. Much like other interleukins discussed in this paper, improper IL-19 functioning has been associated with a variety of autoimmune disorders and diseases associated with immune dysfunction (Cantó et al., 2014).

IL-19 Structure & Function

The IL-19 gene structure is very similar to that of IL-10 with 5 exons and 4 introns in both sequences, with variation in the length of introns but not exons. Despite this, IL-19 is more closely homologous to IL-20 in terms of amino acid sequence (Hofmann et al., 2012). The 3' UTR of IL-19 is truncated compared to IL-10 and harbors only one mRNA destabilizing region where IL-10 has multiple (Gallagher et al., 2000).

The sources of IL-19, at least to the extent that is has been studied, are epithelial, myeloid, and B cells and it has been suggested that the cytokine up-regulates Th2 responses, which are generally anti-inflammatory in nature. Interestingly, the cytokine does not up-regulate Th1 type cells as it does Th2 cells when stimulated (Jordan et al., 2005). Giving more credence to this association, asthma and uremia, two disorders linked to the overexpression of IL-19, are characterized by excessive Th2 responses (Cantó et al., 2014). Cantó et al. caution against using preliminary explanations for the function of IL-19, which have sometimes been generated using murine models, because IL-19 interacts with other cytokines differently when compared with humans.

IL-19 stimulation causes the up-regulation of IL-10 and the production of IL-10 mRNAs, although it is not been confirmed that this high level of mRNA translates to a high level of expression of the anti-inflammatory cytokine (Jordan et al., 2005).

Interleukin-20

IL-20 Products

IL-20 was first discovered through algorithm predictions based on the suspected structure of the cytokine (Kotenko, 2002; Xu, 2004). The role of IL-20 in immune function has not been fully elucidated yet but experiments in murine models and in vitro using human cells has given clues to its function. In both human and mice experiments, aberrant skin conditions are observed when IL-20 is not functioning properly, suggesting a potential role in epidermal maintenance (Blumberg et al., 2001; Wang and Liang, 2004; Xu, 2004). Some authors have suggested that the presence of psoriasis and psoriasis-like symptoms in human and mice models respectively indicates that proper IL-20 functioning is essential for healthy skin development (Kotenko, 2002; Xu, 2004).

IL-20 Structure & Function

Six potential promoter sites for IL-20 have been identified based on the presence of TATA boxes and all exhibited promoter function though at different degrees (Chen and Chang, 2009). Chen and Chang suggested that regulatory elements at the IL-20 locus could explain the variable results of promoter activity for the cytokine but this assertion remains to be fully investigated.

The binding of IL-20 to its associated receptor complex activates a Stat1 and Stat3 cascade (Hofmann et al., 2012). Although not much is known about how IL-20 exerts its effects, the association of skin disorders with aberrant regulation or polymorphisms of IL-20 suggests that it plays an important role in maintaining skin integrity. Other reports of high expression of IL-20 during the course of infection also suggest that the cytokine plays a role in pathogen defense mechanisms (Hofmann et al., 2012).

Interleukin-24

IL-24 Products

IL-24, initially named MDA-7 or melanoma differentiation association gene-7, was first discovered due to its abundance in melanoma cells relative to normally functioning melanocytes (Kotenko, 2002; Wang and Liang, 2004). It was not initially categorized with IL-10 or even recognized as a cytokine until later discoveries confirmed homology with other cytokines and similarity to IL-10 despite it not exerting anti-inflammatory effects (Kotenko, 2002). Despite the initial discovery of elevated IL-24 in melanoma cells, the cytokine has been the subject of increased attention and research funding following publication of results indicating that it inhibits the growth of tumors by triggering apoptosis in cancerous cells (Zheng et al., 2007). Zheng et al. also contend that the ability to trigger apoptosis is unique to IL-24 and is not a feature of other IL-10 cytokines.

IL-24 Structure & Function

IL-24 shares the most amino acid sequence homology with IL-20 at 45%, followed closely by IL-19 (Brocker et al., 2010). The promoter region has been identified and transcription factor recognition sites, such as Jak1, Stat3 and Stat6 sites, have also been found within the promoter (Hofmann et al., 2012).

Once IL-24 is bound to a cell surface receptor, it activates the JAK/STAT signaling pathway (Wang et al., 2004). Interestingly, IL-24 seems to exhibit different cellular responses that are dependent on the presence of a corresponding receptor. While binding to surface receptors generally engenders a proliferative response in affected cells, the intracellular presence of IL-24 seems to promote apoptosis (Wang and Liang, 2004; Wang et al., 2004). Because of this latter observation, the use of IL-24 as a cancer treatment option has been the subject of much of the latest IL-24 research.

LD Across Members of IL-10 Gene Family

One study established that IL-19 and IL-20 are in significant LD with each other and that the polymorphisms studied in IL-20 are in almost complete LD in an Estonian population of unaffected and psoriatic individuals (Kõks et al., 2004). In a second study by the same research team, two blocks of LD were established across IL-19, 20 and 24 with one block extending across IL-19 and most of IL-20 and the second block spanning one SNP from IL-20 and across most of the SNPs included under IL-24 (Kõks et al., 2005). The latter paper by Koks et al. also reported a recombination site within the IL-20 gene, which seems to correspond with the break between blocks of LD between rs2232360 and rs1518108 (Kõks et al., 2005).

MAPKAP-K2

Mitogen-activated protein kinase-activated protein kinases, which are encoded by the MAPKAP family of genes, are intricately involved in the regulation of a variety of different cell signaling pathways. Technologies that allowed for the perturbation of kinase genes in cells at first and entire model organisms later, led to the discovery of novel kinases and the pathways in which they were involved (Gaestel, 2006). MAPKAP kinase-2 is part of a phosphorylation cascade and is activated when it is phosphorylated by a MAP kinase and is known to further activate downstream residues in several distinct pathways (Stokoe et al., 1993; Ben-Levy et al., 1998). Two other enzymes, MAPKAP-K3, which is found on chromosome 3 and MAPKAP-K5, along with MAPKAP-K2 make up the MAPKAP-K subfamily (Gaestel, 2006). MAPKAP-K2 involvement has been implicated in the up-regulation of inflammatory responses via post-transcriptional control of cytokine mRNA stability, the regulation of certain cell cycle checkpoints and other cellular activities (Gaestel, 2006).

MAPKAP-K2 Structure and Function

MAPKAP-K2's catalytic domain, which contains a conserved phosphorylation site, and C-terminal regulatory portion are highly conserved across distinct species, such as fruit flies, worms and mammals although the sequence features diverge in different species beyond those central components (Gaestel, 2006). Other members of the MAPKAP-K subfamily, namely MAPKAP-K3 and MAPKAP-K5 are not as well conserved across species, which Gaestel suggests may be the result of a gene-duplication event after the divergence of higher order species. Other important sequence features of all Map kinases are in the C-terminal domain where two amino acid sequences, nuclear export signal (NES) and nuclear localization signal (NLS), allow MKs to be imported into the nucleus and exit from it respectively (Gaestel, 2006)..

MAPKAP-K2, in particular, seems to play a unique role in the regulation of the inflammatory response and has been proposed as a more promising treatment for chronic inflammatory diseases (Gaestel, 2006). A paper by Moens et al. outlines the myriad ways in which MAPKAP-K2 modulates the inflammatory response, such as activating the transcription factors of immune-involved genes, mRNA stabilization of several immune genes, and inhibiting the synthesis of other immune cells and modulators (Moens et al., 2013).

t¦wth{9 hC {¢¦5¸

In light of the landmark LD studies briefly alluded to in the literature review section of this proposal, particularly those concerned with LD in human populations as an indicator of past human population histories, this thesis will contribute new data to that conversation. Firstly, it will provide new data to the subject of human LD study, which tests the robustness of accepted conclusions while incorporating novel SNPs to the long list of loci studied under LD models. Secondly, this thesis will consider the observed LD patterns and possible influences from multiple alternative explanations, such as those outlined in the Ardlie et al. paper (genetic drift, population growth, admixture, etc.). Thirdly, the collection of SNPs used in this study are closely spaced and densely packed within the regions of study and offer a more fine scale dataset compared to the sparse but widely distributed SNP markers that are generally employed in LD research. Evans and Cardon call for more fine-scaled SNP samplings across populations to determine the extent to which LD patterns are conserved in different populations (Evans and Cardon, 2005). Weiss and Clark similarly suggest that a dense snapshot of SNPs over a smaller area reduces what they call the internal heterogeneity problem, which has come to light after novel SNPs were found in re-sequenced haplotypes (Weiss and Clark, 2002). Finally, the impact of different phasing algorithms, namely those employed by Beagle and PHASE, on the downstream applications of LD analysis will be briefly explored. Also noteworthy in this study, is how functional genes, especially those that have been repeatedly implicated as crucial to normal functioning, might also be mediated at the population level by demographic processes, which have been traditionally studied using neutral markers.

RESEARCH QUESTIONS

The research questions in this thesis are as follows:

- To what degree are linkage disequilibrium patterns at the loci under study shared between the populations?
- Does LD at these loci increase in a clinal fashion across Eurasia similar to other studies that have established this pattern?
- Are there differences in linkage disequilibrium patterns between haplotype phasing programs?
- Are there differences in haplotype block boundaries between different programs?
- Does quality control trimming improve estimates of linkage disequilibrium?

a !¢9wL![{ !b5 a9¢Ih5{

Samples

Samples for this thesis were obtained from the Human Variation Panel of the Coriell Institute for Medical Research database. A total of 74 individuals were sampled across 57 sites from males in the following populations: Andes (7), Basque (7), Chinese (8), Iberia (9), Indo-Pakistan (9), the Middle East (9), Russia (9), North Africa (7), and South Africa (9) in the untrimmed dataset. A total of 63 individuals were sampled across 40 sites from males in the following populations: Andes (7), Basque (7), Chinese (7), Iberia (8), Indo-Pakistan (6), the Middle East (9), Russia (8), North Africa (5), and South Africa (6) in the untrimmed dataset.

Haplotype Phasing

Haplotype phasing was completed using three haplotype phasing programs Beagle, PHASE and MDBlocks (Anderson and Novembre, 2003; Stephens et al., 2004; Browning and Browning, 2007). For haplotype phasing, input files were made for each population separately, each gene separately, all sites in each population, and all populations for each gene in order to assess the impact of population structure on phase resolution. The data were also trimmed based on the robusticity of each genotype call as recorded in the original data file. The data were trimmed on the basis of how conservative the genotype call was from the electropherogram, as specified in the original dataset. Samples with genotype calls that were labeled as "non-conservative" were omitted from the trimmed dataset, as were samples that had less than 40 genotype calls. The untrimmed datasets have all individuals and all sites from the original dataset. The trimmed dataset removed the following individuals and sites (sites are specified by the physical location of the marker in the GRCh38.p2 Assembly):

> Samples: 16689 (Basque), 17100 (Iberia), 17022, 17026, 17027 (3 from IndoPak), 17333 (MidEast), 17378, 17383 (2 from North Africa), 13912 (Russia), 17342, 17344 and 17349 (3 from South Africa)
>
> Sites: 204928558, 204932826, 204944397, 204945745, 204951961, 204956171, 204963245, 204974646, 204976620 (9 sites from MAPKAP-K2), 205011268 (IL-10), 205051213, 205053168, 205061027, 205069997, 205072837 (5 sites from IL-19), 205138774, 205139138 (2 sites from IL-24)

Beagle required input files in Variant Call Format, which were constructed from the original data files using the text-editing program, TextWrangler (Bare Bones Software, Inc. 2009). Genotypes were encoded according to their reference (ancestral) and alternative (derived) alleles, with 0 indicating the reference allele and 1 indicating the alternative. Missing data were input as '.'. The input file contains three header lines that specify what type of data is used in the file, the version used and the date the file was constructed. The sample ID, chromosome number, position ID, rs number, reference allele, alternative allele, data format and genotypes were all specified in each input file. Beagle version 4.1, the latest version of the program, was used to complete haplotype phasing (Browning and Browning, 2007). The following command was used to run the program for each file:

```
java –Xmx[memory allotment]m -jar beagle[program version].jar
gt=[input file name].vcf out=[output file name]
```

The command 'gt' specifies the type of input file and the nature of the data being imputed. No other parameters were used in Beagle as the small dataset used in this study did not require the use of any arguments that would make the program run more efficiently. Input files were subsetted in different ways to assess whether or not population structure impacted the phasing process and downstream applications of LD analysis. Files were subsetted according to the following variables: each individual gene for each population, each population for all genes together, each gene for all populations together and finally, all populations and all genes. Input files according to these rules were constructed for a trimmed and an untrimmed dataset in order to assess the impact of sample call quality on the phasing process and downstream applications.

PHASE version 2.1 requires text files with the extension .inp as input files. In each file, the number of individuals, the number of sites, the locations of each site, individual IDs and genotypes are specified. The same subsetting approach employed in Beagle was also used for PHASE. The following command and parameters were used to run PHASE:

```
./PHASE -MR PHASE_[input file name].inp [output file name].out [number of
iterations] [thinning interval] [burn-in]
```

The initial command './PHASE' calls up the program while the –MR function specifies use of a model that incorporates assumptions about recombination. This function slows the run time of the program compared to the more efficient model that does not account for recombination but this inefficiency was not an issue for this study because of the smaller sample size. 10 iterations were completed for each input file, the thinning interval was left at 1 and the burn-in was set at 1000. It was suggested that a minimum of 5 iterations should be performed in order to obtain reliable phasing results so 10 iterations were completed for each file (Stephens et al., 2004). Since the input files were not prohibitively large, an increase in the number of iterations for the sake of accuracy could be completed without compromising computation time.

MDBlocks also uses text files constructed in a format similar to PHASE where the number of sequences, the number of markers and genotype data are specified. The number of sequences corresponds to the total number of haplotypes in the input file, which in the case of diploid individuals is the number of samples multiplied by two. The genotype data are coded as either 1s or 0s depending on which allele is the ancestral or derived. Missing data was indicated with a -1. Because Beagle used a similar format for genotype data, the input files for MDBlocks were constructed by making the necessary modifications to the existing Beagle input files. The following command for MDBlocks was simple and the nature of the dataset did not require any modification to the parameters:

```
MDBlocks [file name].txt
```

There is an option to use a different algorithm, which makes the program run more efficiently, but since these commands were not used in the other haplotype phasing programs and the dataset is small, it was omitted for MDBlocks. Input files were subsetted by population and with all populations together and were not subsetted by gene or by trim status.

LD Analysis

LD analysis, specifically the Four Gamete test, was performed using the program Haploview, which allows data input, LD analysis, and triangular heat map generation (Wang et al., 2002; Barrett et al., 2005). The triangular heat map allows LD across the IL-10 family and MAPKAP-K2 to be visualized within and across populations with the degree of LD indicated by shaded boxes between sites and haplotype blocks outlined in black triangles.

Haploview requires two files to run: a .ped file with sample IDs and genotype data and a corresponding text file with the SNPs of the gene included in the .ped file and the corresponding physical locations of each site. Once the input files are loaded, the program prompts the selection of an algorithm to assess LD and the Four Gamete test was chosen for all analyses.

Bearing Analysis

A bearing analysis was run on these data across MAPKAP-K2 and the IL-10 family in order to assess the bearing at which most variability is summarized and to detect a potential cline in the physical distribution of allele frequencies (Falsetti and Sokal, 1993). Angular correlation was also assessed in order to determine maximum correlation of the variability in these data (Simon, 1997). The analyses was accomplished with the program PASSaGE and incorporates the methods developed by Simon and Falsetti and Sokal (Falsetti and Sokal, 1993; Rosenberg and Anderson, 2011, Simon 1997). PASSaGE requires two files in order to run the bearing analysis: a file with allele frequency data for the sites, referred to as a data distance matrix, and a coordinates file, or a geographic distance matrix, which supplies the longitude and latitude coordinates of each population in the study. The data distance matrix was generated from the original data file using Excel to calculate allele frequencies at each site.

w9{¦[¢{

IL-10

PHASE IL-10 Trimmed

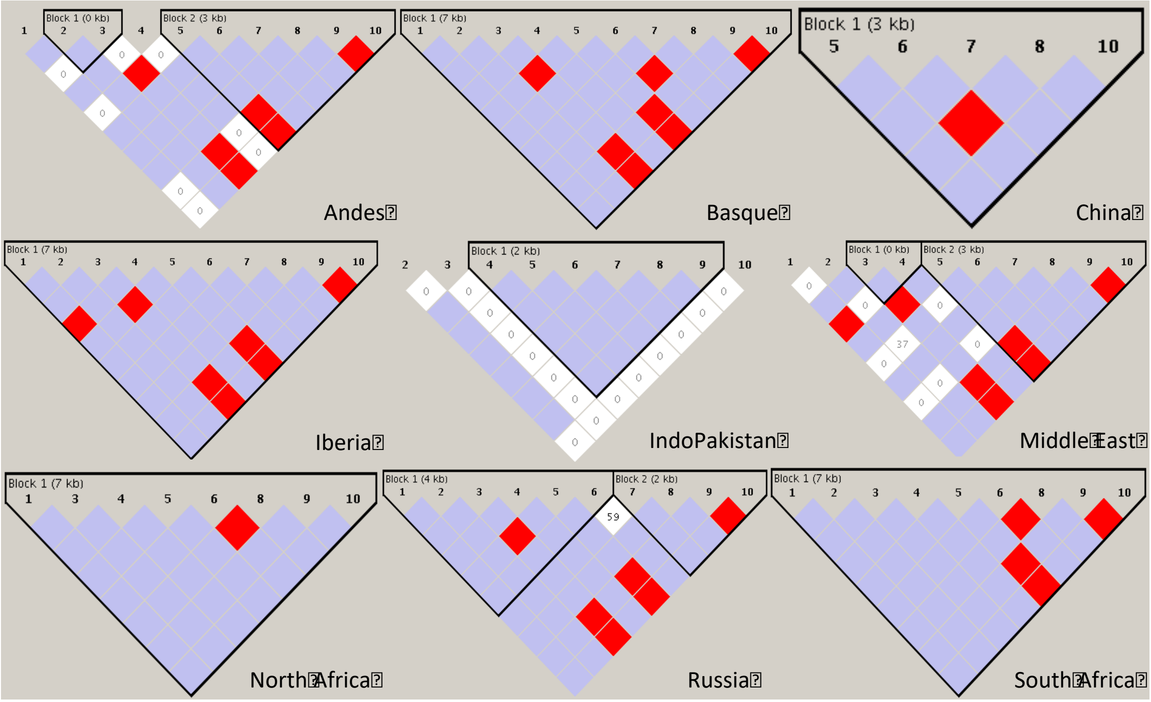


Figure 2: This figure shows the LD plots generated with Haploview for the trimmed IL-10 dataset with results from PHASE.

Figure 2 shows the LD heat maps generated in Haploview for the phased trimmed IL-10 data from PHASE. Each heat map shows the sites in high LD in red, sites in low LD in purple and pairs of sites with no LD between them in white. Haplotype blocks are outlined with bold black lines with block boundaries falling between sites. The Basque, Iberian, North African and South African populations show only one haplotype block and none of the populations show any extensive LD across the gene though there are signals of deeper LD across most populations with the exception of IndoPakistan and North Africa. The Andes and Middle East populations are similar in that they show a short first haplotype block and an extended second haplotype block while Russia shows a longer initial block and shorter second block.

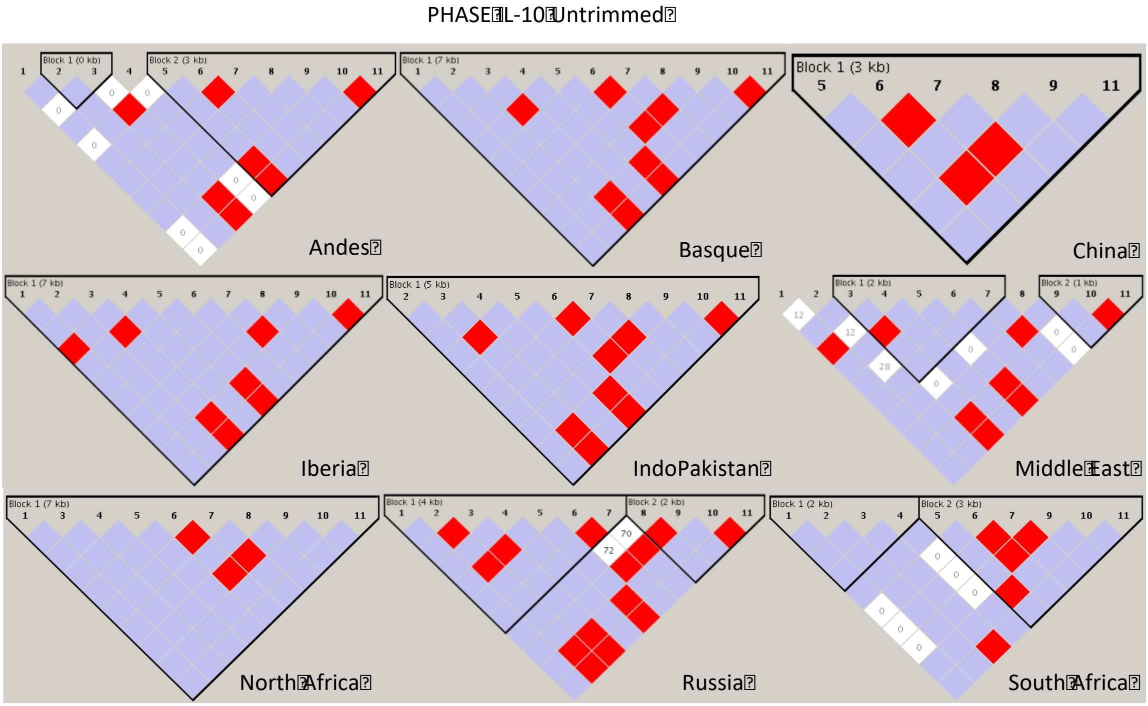


Figure 3: This figure shows the LD plots generated with Haploview for the untrimmed IL-10 dataset with results from PHASE.

Figure 3 shows the heat maps for each population at the same site as Figure 2 except that Figure 3 shows the results when the dataset is left untrimmed. Only a few more pairs show high LD where they did not in the trimmed dataset but some haplotype block boundaries have shifted when trim status is altered. All populations, save North Africa, show the pattern of deep LD. Deep LD between distant sites is indicated by the red boxes that signify high LD at the bottom of the heat map. IndoPakistan samples lost all high LD designations and have a smaller haplotype block across the site in the trimmed dataset compared to the untrimmed set. The Middle East has very similar LD designations, save for one pair of SNPs, but has different block boundaries. South Africa shows a different LD score pattern and also experienced a block shift, with the trimmed dataset having one haplotype block across all of IL-10 to two haplotype blocks in the untrimmed set.

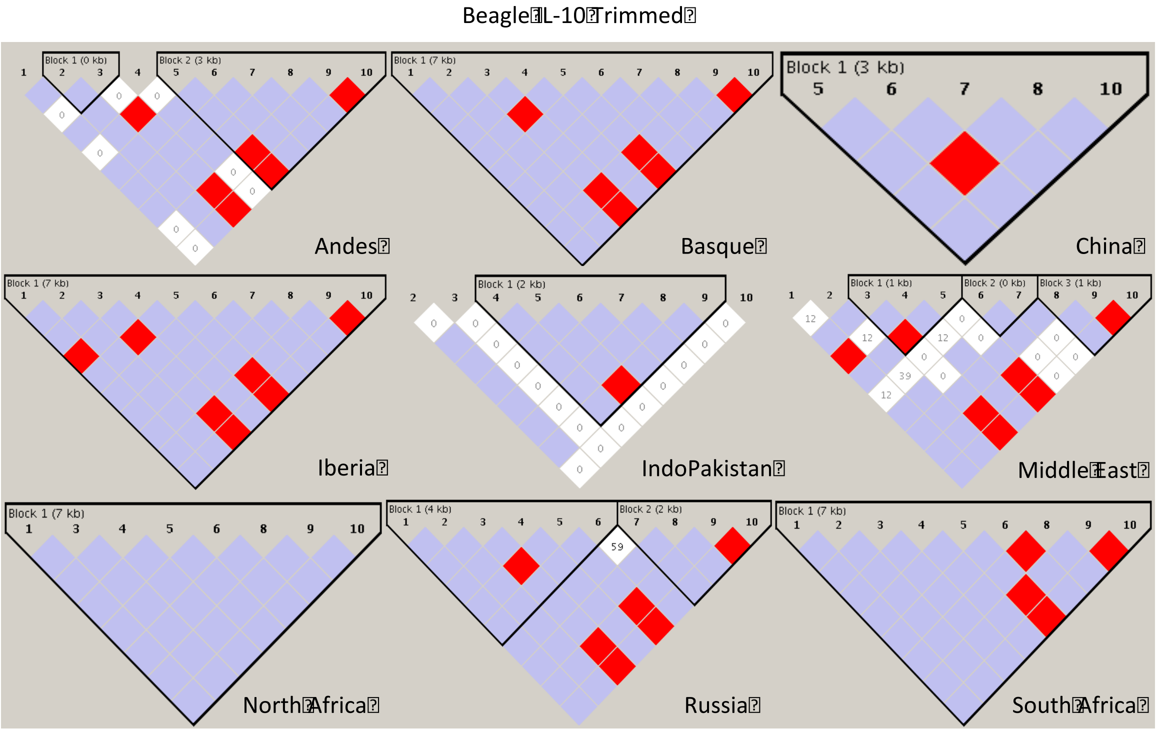


Figure 4: This figure shows the LD plots generated with Haploview for the trimmed IL-10 dataset with results from Beagle.

Figure 4 shows the LD heat maps generated from the trimmed IL-10 dataset that was run through the phasing program Beagle. The high LD designations were almost the same when comparing the results from the trimmed datasets between PHASE and Beagle, except for one high LD designation in the Basques, one in IndoPakistan, and one in South Africa. Haplotype block designations were also very similar between the two programs except that Beagle assigned three blocks to the Middle East while PHASE only assigned two even though the LD designations between the Middle East in PHASE and Beagle were very similar. In North Africa there were no high LD designations nor was there more than one haplotype block.

Beagle IL-10 Untrimmed

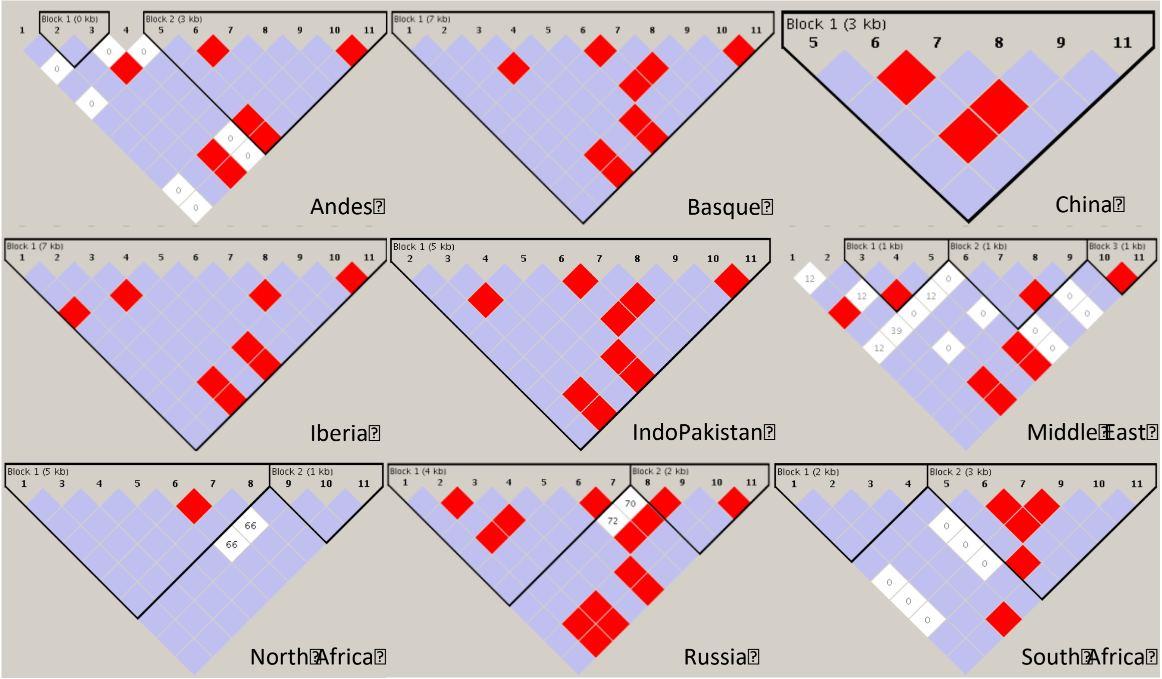


Figure 5: This figure shows the LD plots generated with Haploview for the untrimmed IL-10 dataset with results from Beagle.

Figure 5 shows the LD heat maps generated using the untrimmed IL-10 dataset that was run through Beagle for phasing. The Basques showed three new high LD designations in the untrimmed dataset compared to the trimmed dataset. China showed two new high LD designations given the same comparison. IndoPakistan showed the highest increase with eight new pairs of sites having the high LD classification and it also showed a haplotype block that covered all sites in the untrimmed set. The Middle East and North Africa showed only one new high LD designation in the untrimmed dataset. Russia showed six new high LD designations in the untrimmed set and finally, South Africa showed one new designation though, the particular pairs in high LD varied as well. The Middle East showed three haplotype blocks in comparing the trimmed and untrimmed dataset but the markers that were covered by those blocks changed slightly. North Africa and South Africa showed two haplotype blocks in the untrimmed set where it only showed one haplotype block across the whole gene in the trimmed dataset. When comparing the untrimmed dataset from Beagle to that of PHASE, there are only two high LD designations in the results from PHASE. In the Middle East samples, there is an increase from two haplotype blocks for PHASE to three blocks for Beagle and in North Africa, the results from Beagle show two haplotype blocks where PHASE results show only one block.

IL-19

PHASE IL-19 Trimmed

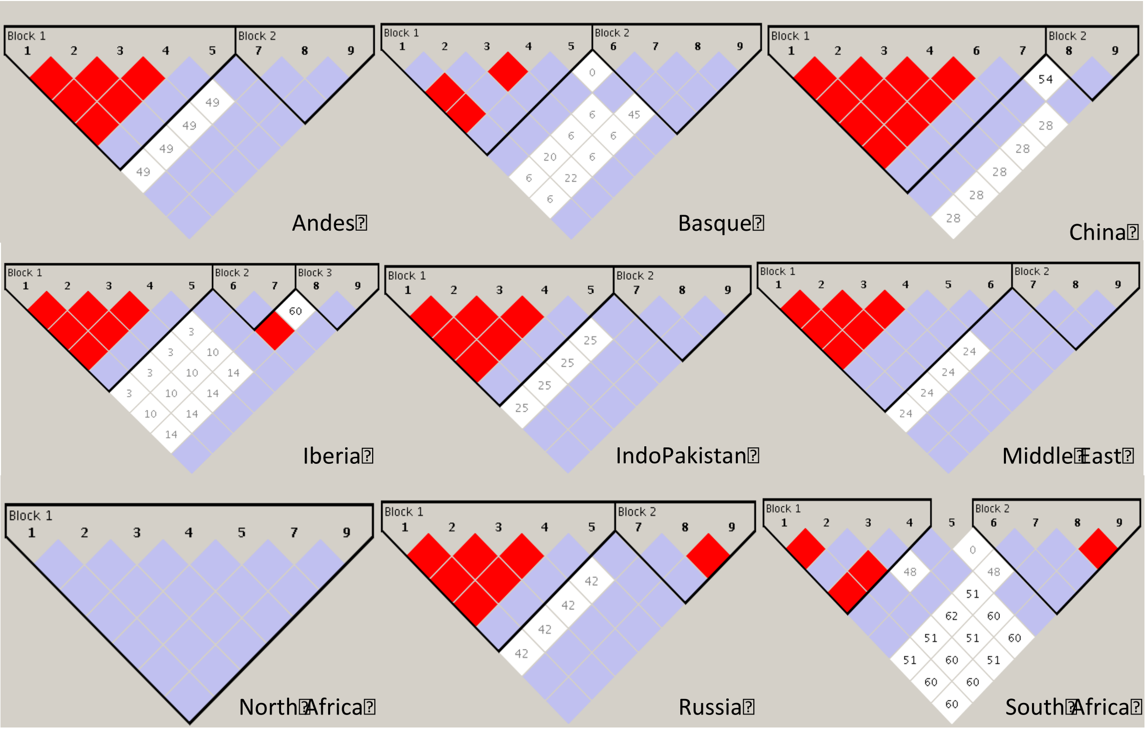


Figure 6: This figure shows the LD plots generated with Haploview for the trimmed IL-19 dataset

with results from PHASE.

Several populations — Andes, China, Iberia, IndoPakistan, Middle East and Russia — show a consistent LD block at the beginning of IL-19. Basque and South African populations show high LD designations and have similar haplotype block boundaries as the aforementioned populations but the signal of LD is not as strong nor as deep across the first four to five markers in IL-19. North Africa lacks any high LD designations and only has one haplotype block across the locus.

PHASE IL-19 Untrimmed

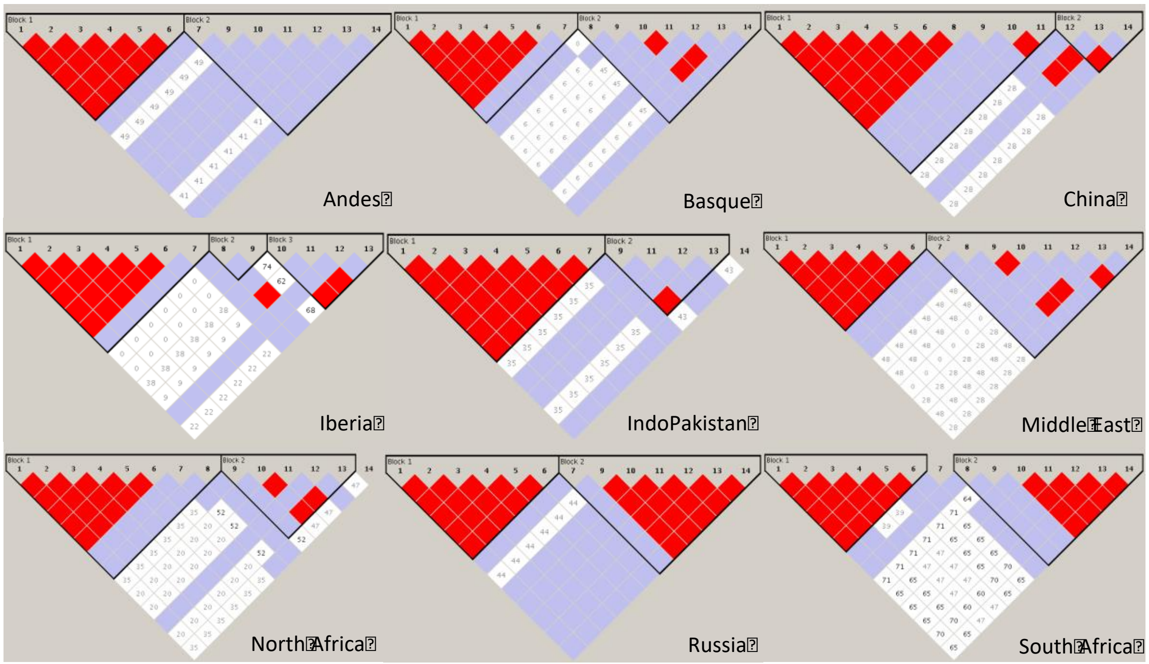


Figure 7: This figure shows the LD plots generated with Haploview for the untrimmed IL-19 dataset with results from PHASE.

In contrast to the trimmed dataset, Figure 7 shows a persistent pattern of extended and deep LD at the first five markers of IL-19. Furthermore, some populations — Russia and South Africa — show a strong signal of LD at the end of IL-19. All populations, save Iberia, show two haplotype blocks across the site and the coverage of those haplotype blocks is similar in all populations except for China and North Africa, which show an extended haplotype at the beginning of the gene and a shorter haplotype toward the end.

Beagle IL-19 Trimmed

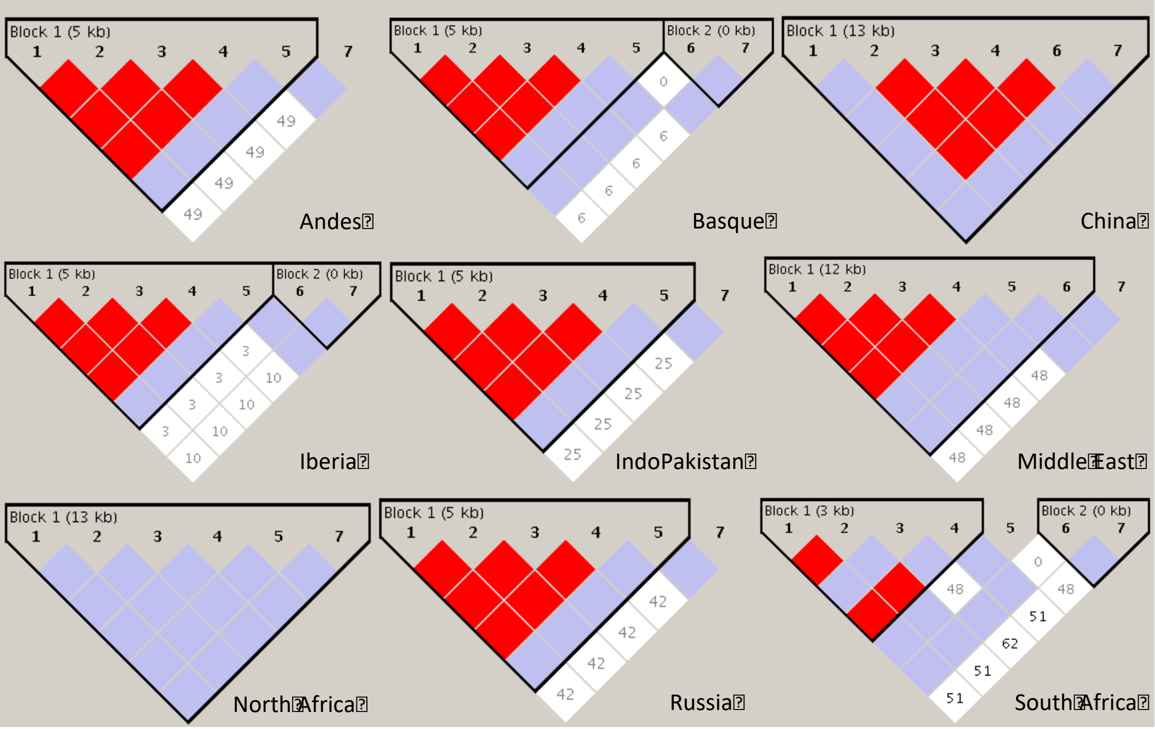


Figure 8: This figure shows the LD plots generated with Haploview for the trimmed IL-19 dataset with results from Beagle.

Much like the trimmed IL-19 dataset from PHASE, Beagle shows the North African population without any high LD designations. There are fewer haplotype block boundaries when comparing the trimmed results from Beagle to PHASE with Beagle showing less. The consistent pattern of strong LD at the beginning of IL-19 is maintained in this dataset, though the signal is not as strong as the Basque samples. In both PHASE and Beagle in the trimmed dataset, South Africa does not show strong LD between each adjacent marker though it does show strong and deep signal of LD in haplotype block one.

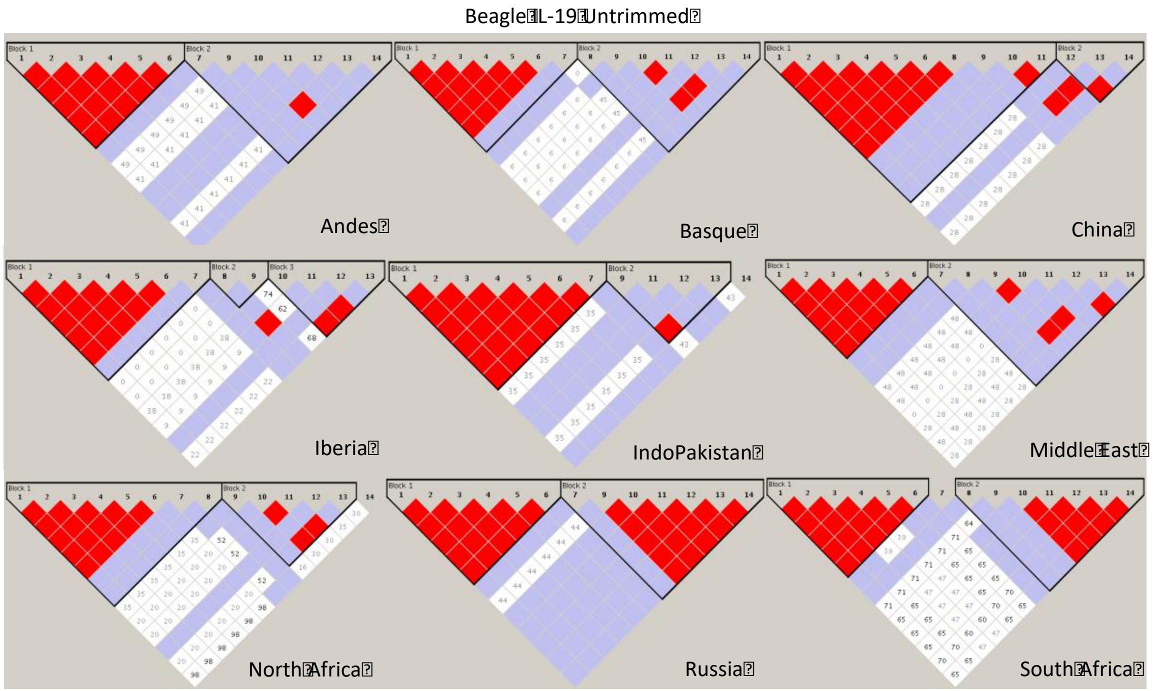


Figure 9: This figure shows the LD plots generated with Haploview for the untrimmed IL-19 dataset with results from Beagle.

Similarly to Figure 7, Figure 9 shows strong and deep blocks of LD across the first haplotype block in IL-19. Russia and South Africa also show a second strong block of LD in the second and last haplotype block for IL-19. The haplotype block patterns are very similar between the two programs and trim statuses as well, with Iberia showing three haplotype blocks and China and North Africa showing a much longer first block and a shorter second block when compared against the remaining populations.

IL-20

IL-20, because there were only three sites available for analysis in both the trimmed and untrimmed datasets, did not yield any heat maps that were of value.

IL-24

As was the case for IL-20, the trimmed datasets for IL-24 were reduced from five sites down to three, making those results difficult to interpret or draw any conclusions from. Therefore, in this section for IL-24 results, I will only report untrimmed results.

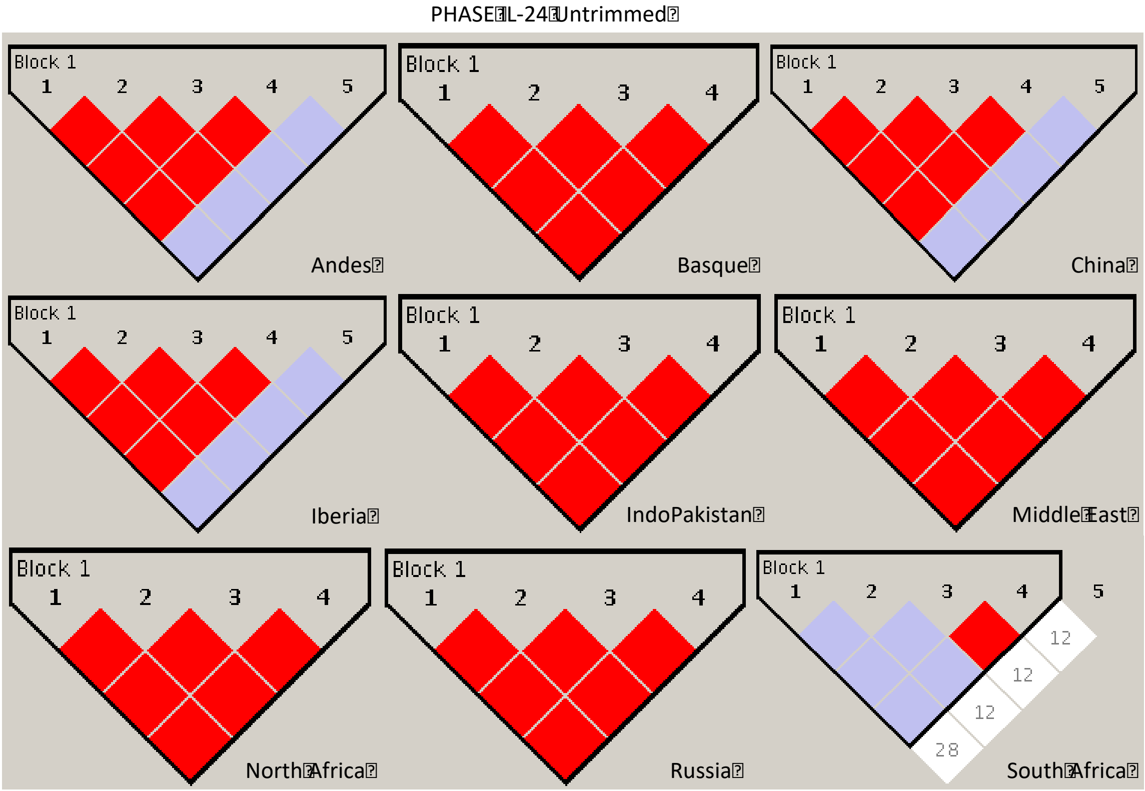


Figure 10: This figure shows the LD plots generated with Haploview for the untrimmed IL-24 dataset with results from PHASE.

Figure 10 shows the results from the untrimmed IL-24 dataset where haplotype phase was determined with the program PHASE. In all 9 populations, the fifth site of IL-24 is not in high LD with the rest of the site, though it is included in the same haplotype block in the Andean, Chinese, Iberian and South African samples. South Africa is rather distinct from the rest of the populations in that there are only two adjacent sites that are in strong LD with each other, whereas the rest of the populations have all four sites in strong LD.

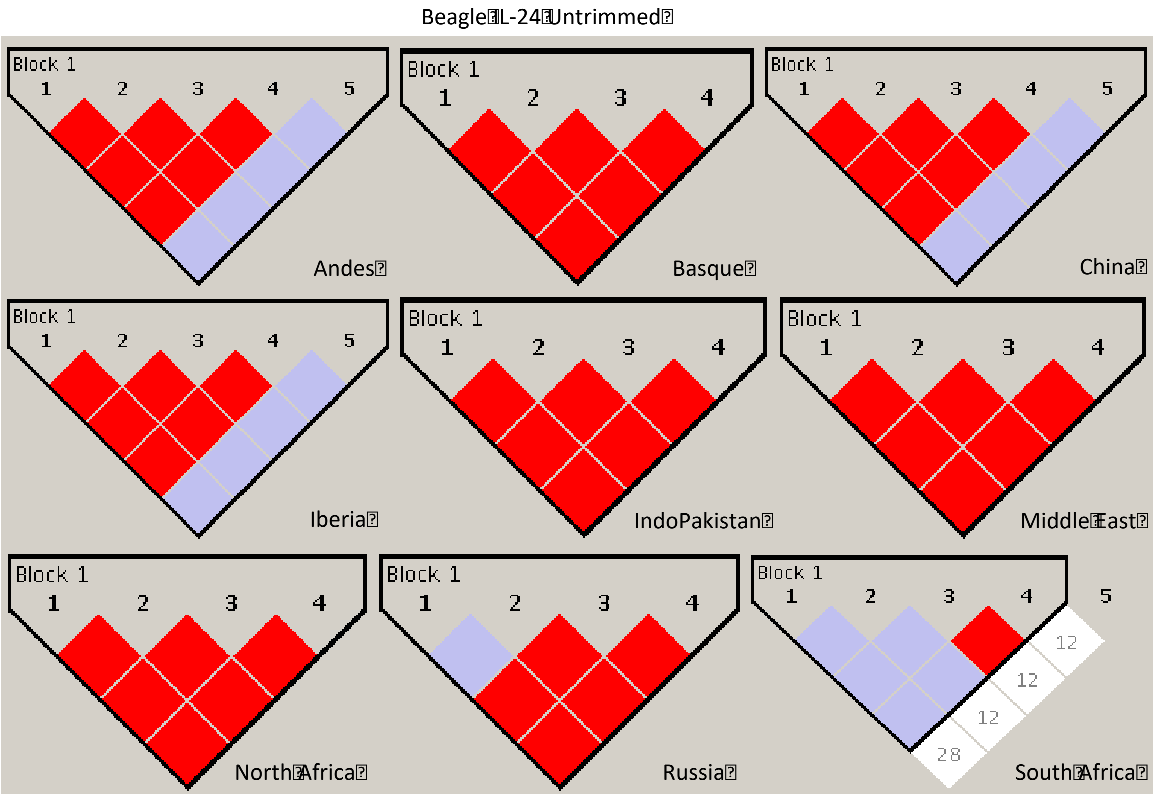


Figure 11: This figure shows the LD plots generated with Haploview for the untrimmed IL-24 dataset with results from Beagle.

Figure 11 shows almost the exact same signals as Figure 10 with the same trim status except that the haplotype phase was resolved using Beagle. All of the heat maps were identical across all of the populations except for a small difference in Russia between the two programs; sites 1 and 2 were not in LD in the data generated by Beagle while they are in the data from PHASE.

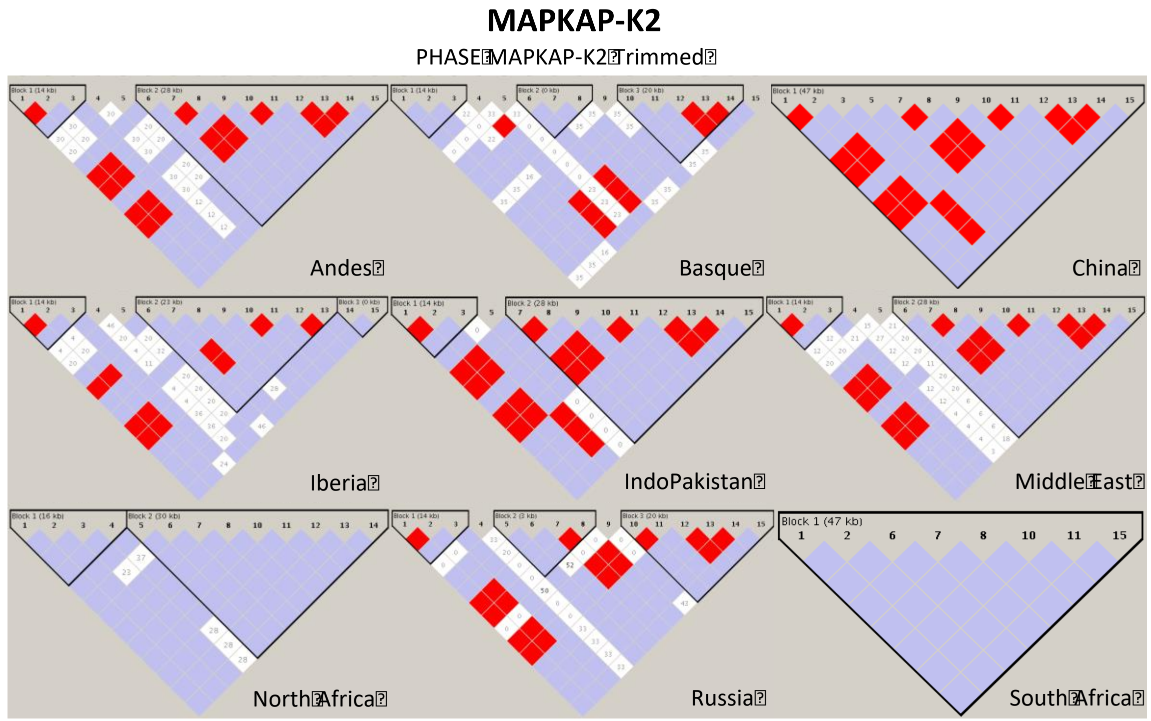


Figure 12: This figure shows the LD plots generated with Haploview for the trimmed MAPKAP-K2 dataset with results from PHASE.

Figure 12 shows the heat map results of the trimmed MAPKAP-K2 dataset that had haplotype phase determined by PHASE. There are shared sites with high LD in the Andean, Chinese, IndoPakistani, Middle East and Russian samples with Andes, IndoPakistan and China and Russia and the Middle East having nearly identical or identical high LD designations. The Basque, Iberian, North African and South African samples all show lower levels of LD across MAPKAP-K2 with North and South Africa showing now high LD designations at any of the sites across MAPKAP-K2.

PHASE MAPKAP-K2 Untrimmed

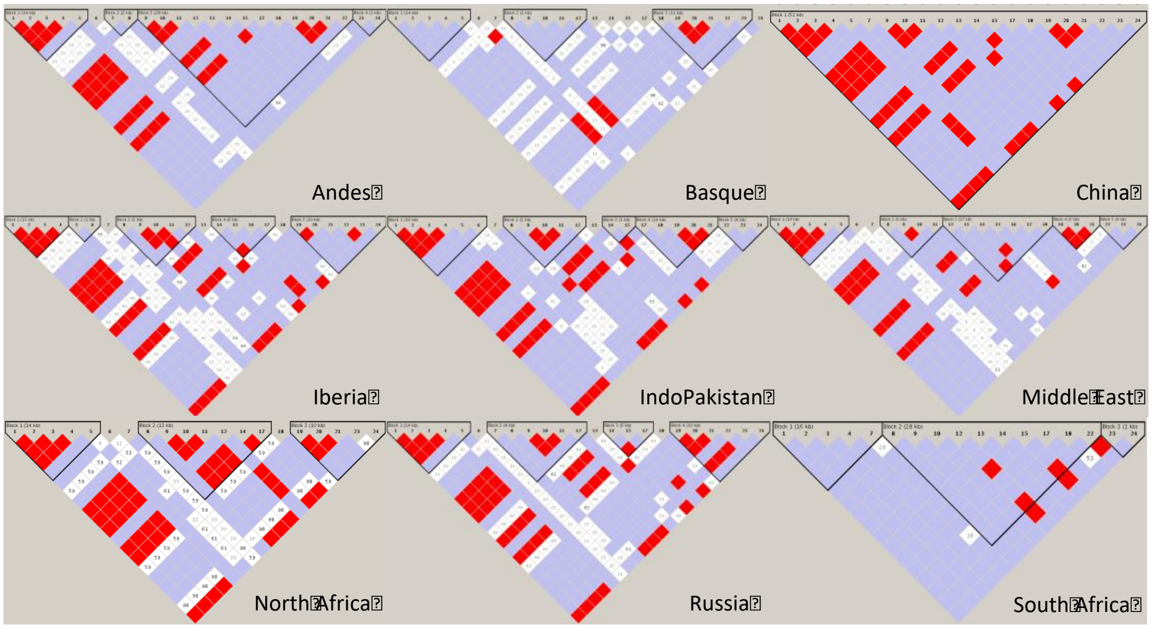


Figure 13: This figure shows the LD plots generated with Haploview for the untrimmed MAPKAP-K2 dataset with results from PHASE.

Much like Figure 12, Figure 13 shows that China and IndoPakistan are nearly identical, except for one high LD designation. Unlike Figure 12, however, Russia and the Middle East do not share as many LD designations in common. Every population, aside from China and the Basques, exhibits more haplotype blocks in the untrimmed dataset compared against the trimmed dataset for MAPKAP-K2. The haplotype block coverage is not exactly the same for MAPKAP-K2, but it is fairly equivalent given a difference in the number of sites between the trimmed and untrimmed data.

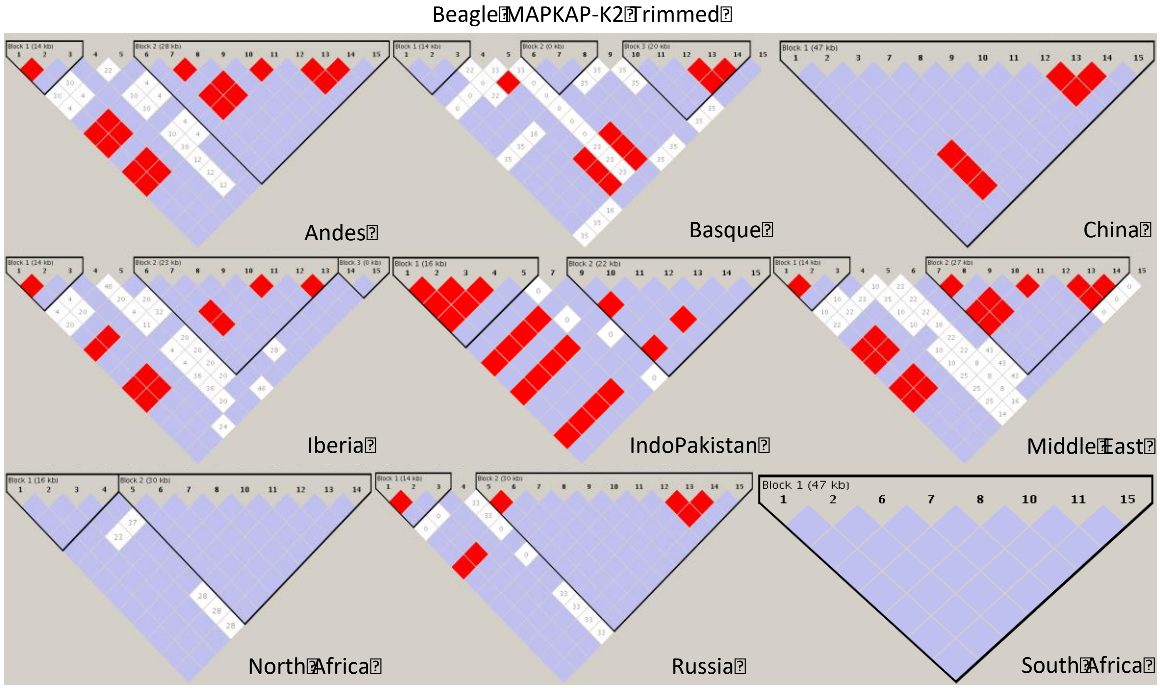


Figure 14: This figure shows the LD plots generated with Haploview for the trimmed MAPKAP-K2 dataset with results from Beagle.

In Figure 14, it is apparent that the high LD designations and the boundaries of the haplotype blocks are the same between Beagle trimmed data and PHASE trimmed data at MAPKAP-K2 across the following populations: Andes, Basque, Iberia, Middle East, North Africa and South Africa. Russia has 11 more high LD designations in the PHASE heat map when compared to the Beagle heat map. IndoPakistan has the same number of high LD designations but they are arranged differently across the gene. China shares some high LD designations in common between programs but PHASE shows much more LD across the entirety of the gene compared to Beagle.

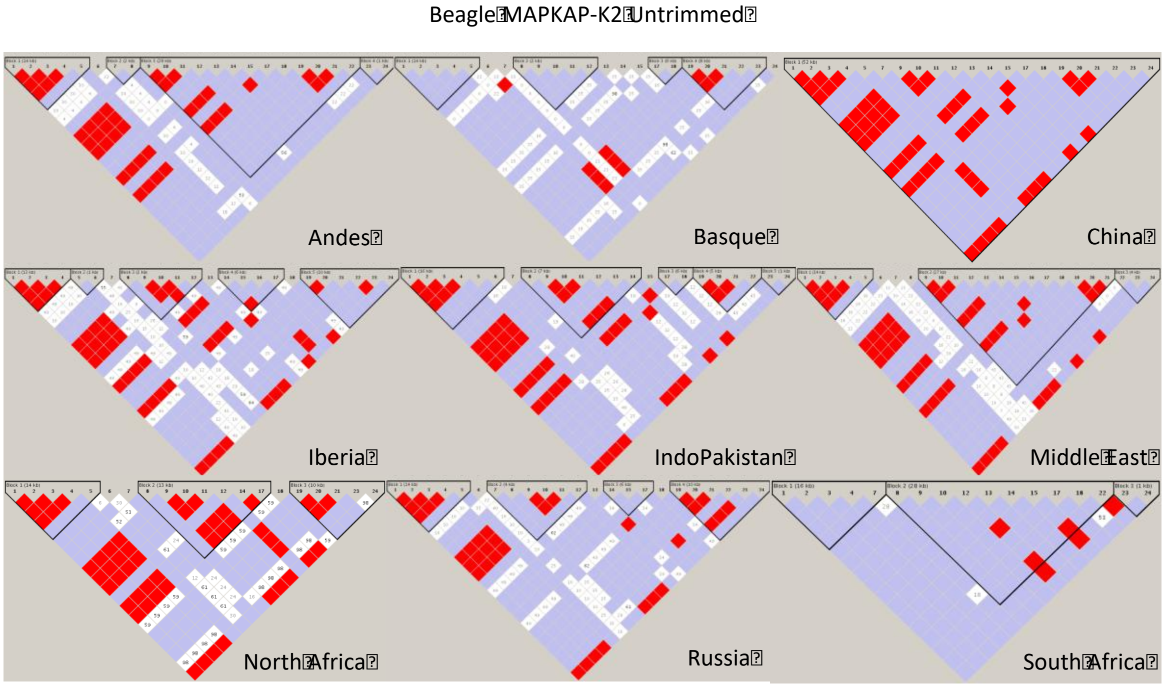


Figure 15: This figure shows the LD plots generated with Haploview for the untrimmed MAPKAP-K2 dataset with results from Beagle.

Figure 15 shows the untrimmed MAPKAP-K2 heat maps after phasing through Beagle. The Basque population has showed a consistent LD profile across the two haplotype phasing programs and the trim status differences. North Africa and China show much more extensive LD across the entirety of MAPKAP-K2 in the untrimmed dataset when compared against the trimmed dataset for Beagle. The Andes population is also consistent across the two different phasing programs but is less consistent when it comes to trim status of the datasets. South Africa shows six high LD designations compared to zero in the trimmed Beagle dataset, and the same pattern of six designations is shown in the untrimmed PHASE dataset heat map. In the untrimmed datasets of both programs, it is apparent that South African and the Basque populations show considerably less extensive LD compared to other populations in this study. When the trimmed datasets are incorporated, North Africa shows little LD as well.

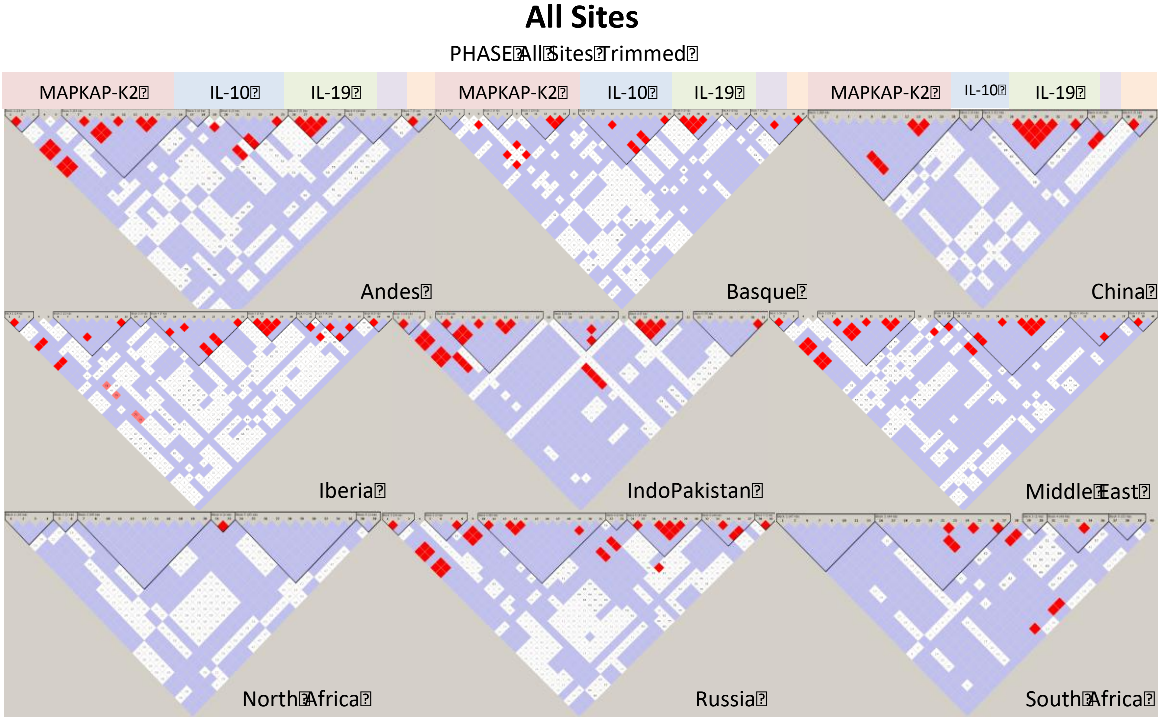


Figure 16: This figure shows triangular heat maps generated in Haploview using trimmed output from PHASE across all sites. The pink box indicates the boundaries of MAPKAP-K2, the blue box indicates the boundaries of IL-10, the green box indicates IL-19, the purple indicates IL-20 and finally, the orange box indicates IL-24.

Figure 16 shows trimmed output from PHASE across all sites in this study across all nine populations. There is a consistent block of high LD at the very beginning of IL-19 that is shared across all populations except for North and South Africa. North and South Africa are also the two populations that show less extensive LD in general though there are a few signals of long-range LD in South Africa.

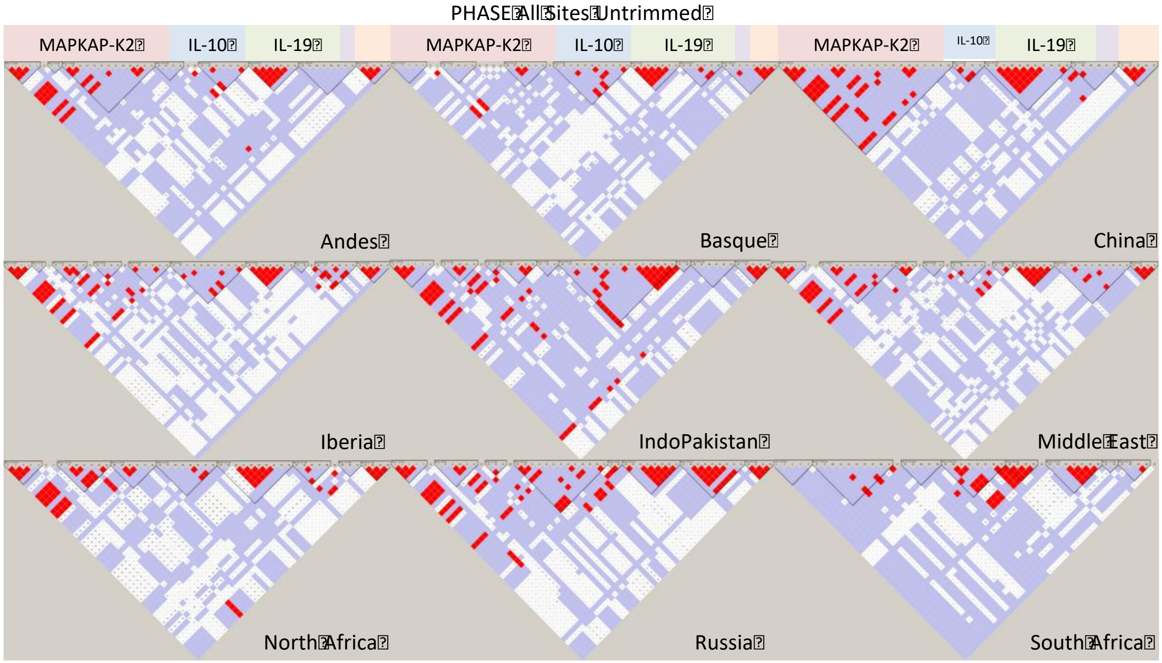


Figure 17: This figure shows triangular heat maps generated in Haploview using untrimmed output from PHASE across all sites. The pink box indicates the boundaries of MAPKAP-K2, the blue box indicates the boundaries of IL-10, the green box indicates IL-19, the purple indicates IL-20 and finally, the orange box indicates IL-24.

Figure 17 shows the untrimmed output from PHASE across all sites in the study. Interestingly, we can see here that the SNPs of IL-19 that were in strong LD are now shared across all populations including North and South Africa. More generally, signals of LD are more prevalent when the untrimmed datasets are used when generating heat maps. High LD scores are more extensive and dip below toward the bottom of the heat map, indicating that LD is much more extensive across distant sites in the untrimmed datasets. South Africa is still the population with fewer high LD designations. IL-24 has several SNPs in high LD within that gene and much like the individual gene heat maps showed, that pattern is consistent across different populations but also missing from South Africa.

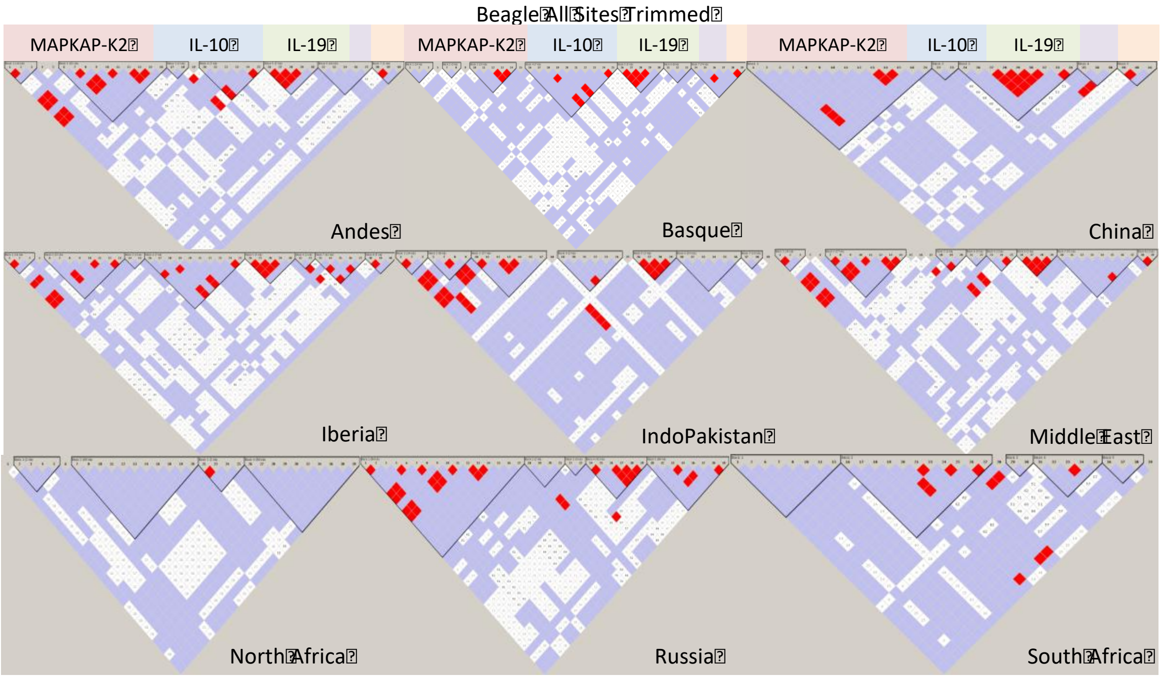


Figure 18: This figure shows triangular heat maps generated in Haploview using trimmed output from Beagle across all sites. The pink box indicates the boundaries of MAPKAP-K2, the blue box indicates the boundaries of IL-10, the green box indicates IL-19, the purple indicates IL-20 and finally, the orange box indicates IL-24.

Figure 18 shows the distribution of LD across all sites from the trimmed dataset in Beagle. There is a somewhat consistent pattern of strong LD at the beginning of IL-19, though that signal is mostly absent in North and South African populations. There is a somewhat strong, albeit diffuse, pattern of LD across MAPKAP-K2 that is evident in the Andes, IndoPakistani, Middle Eastern, Russian and to some extent, Iberian, populations. Compared with Figure 18, which shows the untrimmed counterpart to this dataset from Beagle, high LD designations does not go as deep in the trimmed dataset.

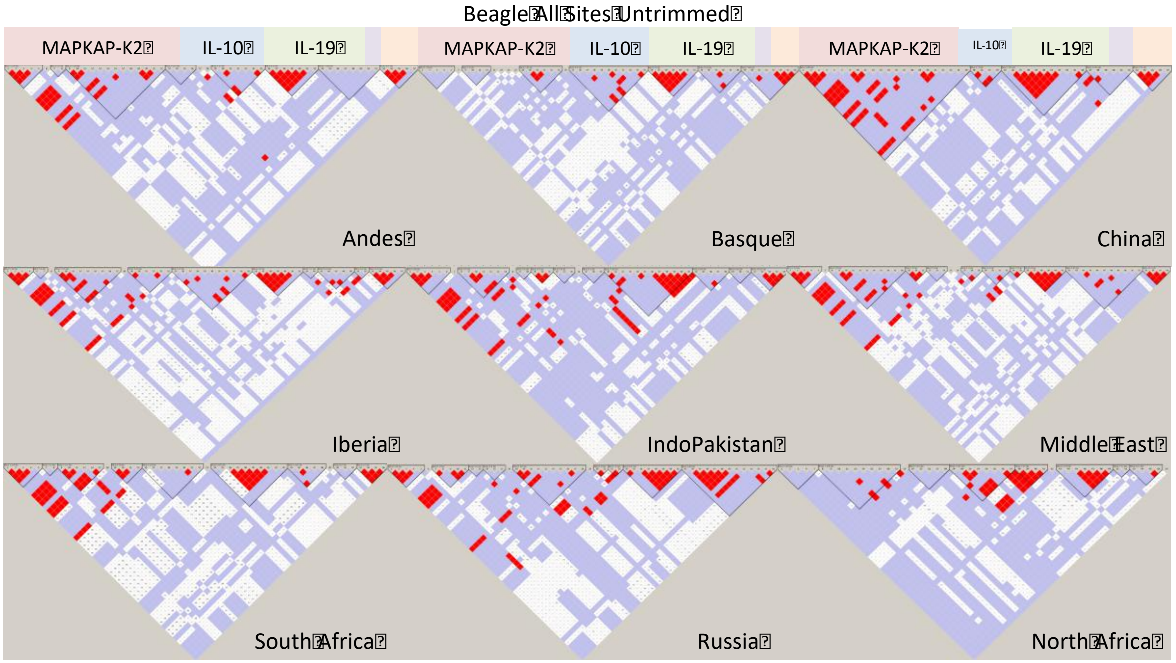
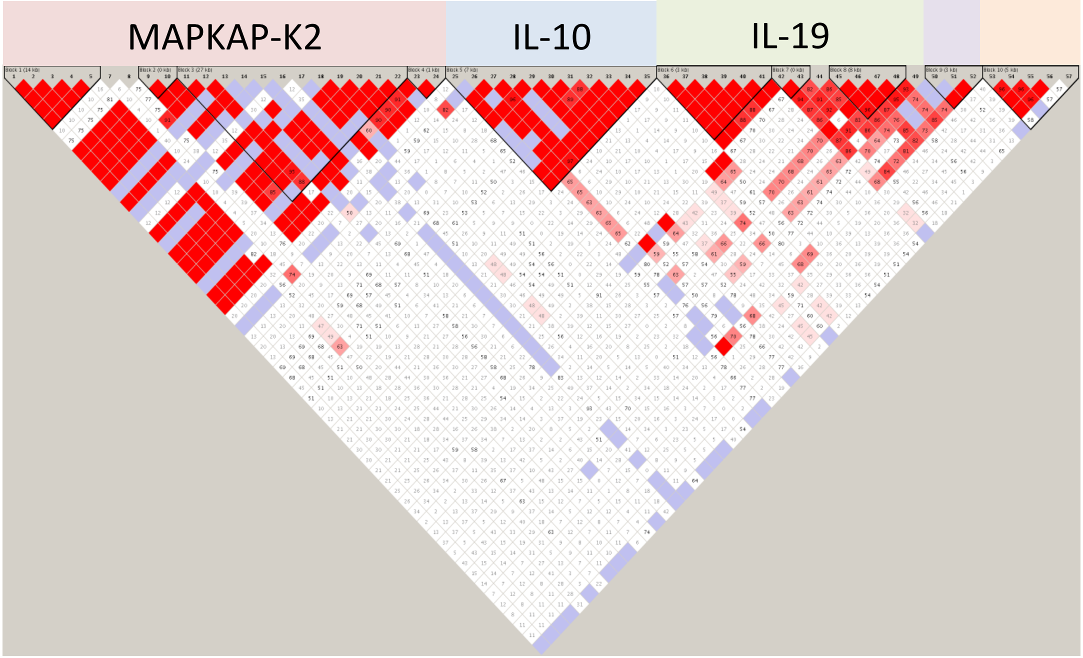


Figure 19: This figure shows triangular heat maps generated in Haploview using untrimmed output from Beagle across all sites. The pink box indicates the boundaries of MAPKAP-K2, the blue box indicates the boundaries of IL-10, the green box indicates IL-19, the purple indicates IL-20 and finally, the orange box indicates IL-24.

Figure 19 generally reflects the untrimmed data from PHASE, although the high LD across IL-24 is less apparent in North Africa and Russia. There is also a group of SNPs at the beginning of MAPKAP–K2 that are in high LD in the Andes, China, Iberia, IndoPakisan, the Middle East, South Africa and Russia but they are absent from North Africa and the Basque populations. In this group of heat maps, it is not the Basque population that shows the least extensive LD across the region.

Beagle All Sites All Populations Untrimmed



PHASE All Sites All Populations Untrimmed

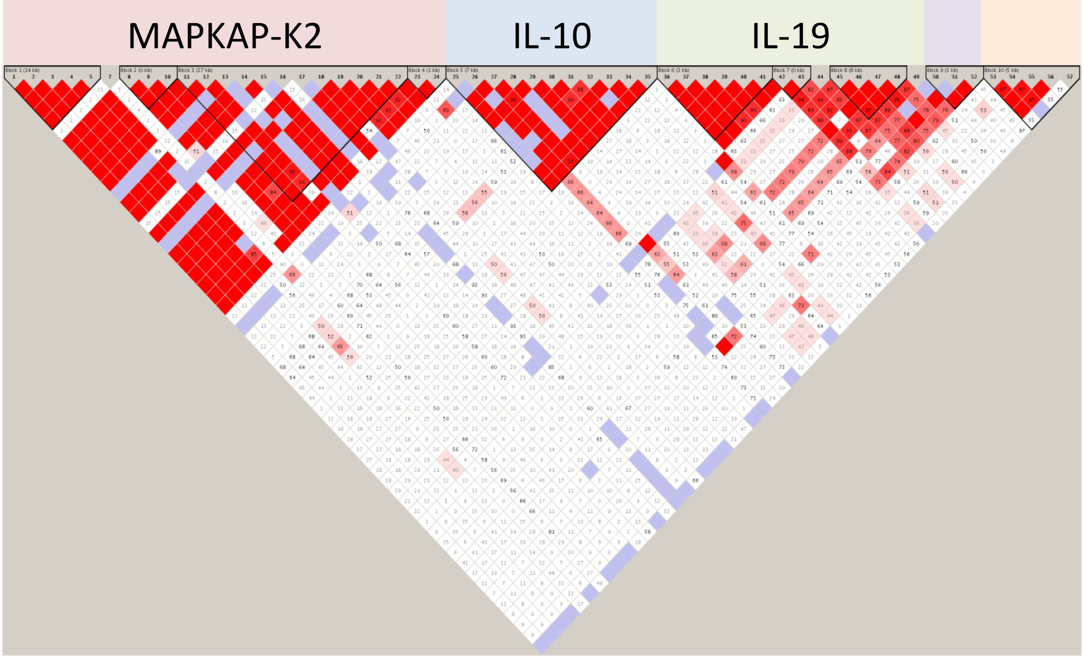


Figure 20 shows heat maps of all sites across all populations in the untrimmed dataset where haplotype phased was resolved using Beagle (top) and PHASE (bottom). The pink box indicates the boundaries of MAPKAP-K2, the blue box indicates the boundaries of IL-10, the green box indicates IL-19, the purple indicates IL-20 and finally, the orange box indicates IL-24.

Figure 20 shows that phasing outcomes, when it comes to applications to LD analysis, are consistent between the programs PHASE and Beagle. The designations of high LD are concordant between programs but the haplotype block boundaries drawn around those extended haplotypes is variable. PHASE called nine haplotype blocks while Beagle only called six in the same region. The most extensive block of high LD appears to occur across IL-10 with MAPKAP-K2 and the beginning of IL-19 showing high LD as well.

PHASE Sites (Untrimmed)

	All Sites	**IL-10**	**IL-19**	**IL-20**	**IL-24**	**MAPKAP**
All Populations	5.3	11	4	3	5	5.5
Andes	6.375	4.5	6.2	2	5	5.5
Basque	6.875	11	7	3	4	5.33
China	9.6	6	6.5	3	5	21
Iberia	5.2	11	4.333	3	5	4.2
IndoPak	5.555	10	5.5	3	4	4.2
MidEast	5.667	4	7	3	4	4.4
NorthAF	6	10	6.5	3	4	5.667
Russia	6.25	5.5	6.5	2	4	5
SouthAF	5.625	5	6.5	2	4	5.667

PHASE Sites (Trimmed)

	All Sites	**IL-10**	**IL-19**	**IL-20**	**IL-24**	**MAPKAP**
All Populations	4	10	3	3	3	4.333
Andes	5	4	4	2	3	6.5
Basque	5	10	4.5	3	2	3.667
China	6	5	4	3	3	13
Iberia	4.75	10	3	3	3	4.333
IndoPak	6.2	6	4	2	2	6
MidEast	6	4	4.5	3	2	6.5
NorthAF	5.166	8	7	NA	2	6.5
Russia	5	5	4	2	2	4.333
SouthAF	5.8	9	4	2	2	8

Beagle Sites (Untrimmed)

	All Sites	**IL-10**	**IL-19**	**IL-20**	**IL-24**	**MAPKAP**
All Populations	5.2 (-)	11 (/)	4.33 (+)	3 (/)	5 (/)	5.5 (/)
Andes	6.375 (/)	4.5 (/)	6.5 (+)	2 (/)	5 (/)	5.5 (/)
Basque	6.875 (/)	11 (/)	7 (/)	3 (/)	4 (/)	4.25 (-)
China	9.6 (/)	6 (/)	6.5 (/)	3 (/)	5 (/)	21 (/)
Iberia	5.3 (+)	11 (/)	4.333 (/)	3 (/)	5 (/)	4.2 (/)
IndoPak	4.9 (-)	10 (/)	5.5 (/)	3 (/)	4 (/)	4.2 (/)
MidEast	5.2 (-)	3 (-)	7 (/)	3 (/)	4 (/)	**7.333 (+)**
NorthAF	5.22 (/)	**5 (-)**	6.5 (/)	3 (/)	4 (/)	5.667 (/)
Russia	5.555 (-)	5.5 (/)	6.5 (/)	2 (/)	4 (/)	5 (/)
SouthAF	5.625 (/)	5 (/)	6.5 (/)	2 (/)	4 (/)	5.667 (/)

Beagle Sites (Trimmed)

	All Sites	**IL-10**	**IL-19**	**IL-20**	**IL-24**	**MAPKAP**
All Populations	3.8 (-)	10 (/)	3 (/)	3 (/)	3 (/)	4.33 (/)
Andes	4.857 (-)	4.5 (+)	4 (/)	2 (/)	3 (/)	6.5 (/)
Basque	5 (/)	10 (/)	4.5 (/)	3 (/)	2 (/)	3.667 (/)
China	6.2 (+)	5 (/)	**7 (+)**	2 (-)	3 (/)	12 (/)
Iberia	5 (+)	10 (/)	3 (/)	3 (/)	3 (/)	4.333 (/)
IndoPak	4.57 (-)	6 (/)	4 (/)	3 (+)	3 (+)	6.667 (+)
MidEast	**4.125 (-)**	**2.667 (-)**	4.5 (/)	3 (/)	2 (/)	5.5 (-)
NorthAF	**7.25 (+)**	8 (/)	7 (/)	NA	2 (/)	6.5 (/)
Russia	6 (+)	5 (/)	3.5 (-)	2 (/)	2 (/)	**6.5 (+)**
SouthAF	5.8 (/)	9 (/)	4 (/)	2 (/)	2 (/)	8 (/)

Table 2: This table shows the average haplotype block length for each population, at each site, across all populations. The two tables on top are from PHASE and the two on the bottom are from Beagle. Each program also has averages for the untrimmed and trimmed datasets. Cells that are highlighted

in purple are those where the Beagle average was lower and cells that are highlighted in green are those where the Beagle average was higher than that of PHASE. Given these differences, deviances of more than 1 average marker per block are bolded.

There is not a consistent increase in the average haplotype block length by population across the data as evidenced by Table 2. There are some populations closer to Africa that exhibit shorter average haplotype block lengths compared to those outside of Africa but there are few discernible clinal block length patterns. IL-24 is the only gene that shows a consistent signal of decreasing average block length in populations that are close to or within Africa across both programs and trim statuses, though the decrease is very slight. Several populations show aberrant averages, which is likely due to the small numbers of samples per population used in this study.

Comparison of Haplotype Block Boundaries Across Populations

		1		2		3		4		5		6		7		8		9		10		11		12		13		14		15		16		17		18		19		20		21		22		23		24		25		26		27		28		29		30	
All Pops	█										█	▨		▨	█						█																										█				█											
Andes	█										█	▨	█				█															▨													█								█	▨		▨	█				
Basque	█										█	▨		▨	█										█	▨		▨		▨		▨		▨	█												█	▨	█												
China	█											▨				▨																▨																		▨		▨		▨		▨			█		
Iberia	█								█				█	▨	█										█	▨	█								█	▨	█												█												
IndoPak	█												█	▨	█										█	▨	█				█	▨	█										█						█	▨	█										
Middle East	█										█	▨		▨	█																						█						█	▨	█						█	▨	█								
North Africa	█										█	▨		▨	█											▨				▨		▨			█									▨					█			▨									
South Africa	█									▨		▨			█							▨										▨						▨		▨		▨			█	▨	█										█				
Russia	█										█	▨	█												█	▨	█					▨			█	▨	█																								
	MAPKAP																																																	IL-10											

		31		32		33		34		35		36		37		38		39		40		41		42		43		44		45		46		47		48		49		50		51		52		53		54		55		56		57	
All Pops											█												█				█	▨	█								█	▨	█						█										█
Andes											█														█	▨	█													▨					█										█
Basque											█														█												█																█		
China							█	▨	█															▨									█												█										█
Iberia											█														█				█	▨	█							▨							█										█
IndoPak										▨													█			▨			█	▨	█												█										█		
Middle East			█	▨	█						█												█																						█								█		
North Africa							█																				█										█	▨	█						█								█		
South Africa				▨											█								█	▨	█														█	▨	█				█								█		
Russia			█				█																█			▨														▨					█								█		
	IL-10											IL-19																												IL-20						IL-24									

Table 3: This table shows haplotype block boundaries by population given by Haploview with untrimmed output from PHASE. Black bars indicate the block boundary. White boxes indicate markers that are listed within haplotype blocks and grey boxes are markers that were not listed. The gene boundaries are indicated along the bottom of the table.

Comparison of Haplotype Block Boundaries Across Three Phasing Programs

		1		2		3		4		5		6		7		8		9		10		11		12		13		14		15		16		17		18		19		20		21		22		23		24		25		26		27		28		29		30	
Beagle	█										█	▨		▨		▨	█				█																										█				█											
PHASE	█										█	▨		▨	█						█																										█				█											
MDBlocks	█																█																						█										█				█									
	MAPKAP																																																	IL-10											

		31		32		33		34		35		36		37		38		39		40		41		42		43		44		45		46		47		48		49		50		51		52		53		54		55		56		57	
Beagle											█												█				█	▨	█								█	▨	█						█										█
PHASE											█		█										█				█	▨	█								█	▨	█						█										█
MDBlocks									█												█				█										█								█										█		
	IL-10											IL-19																												IL-20						IL-24									

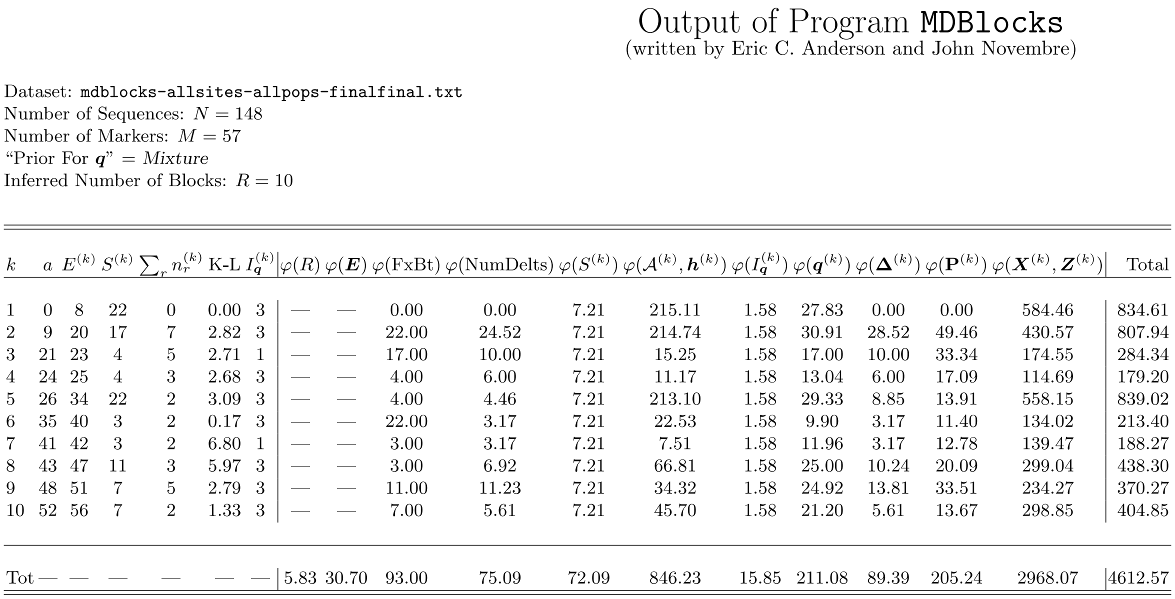
Table 4: This table shows haplotype block boundaries across all sites for each haplotype phasing program: PHASE, Beagle and MDBlocks. Black bars indicate the block boundary. White boxes indicate markers that are listed within haplotype blocks and grey boxes are markers that were not listed. The gene boundaries are indicated along the bottom of the table.

Table 3, which shows haplotype block boundaries by population, shows that there are some salient block boundaries that are concordant with gene boundaries. This general trend of haplotype block boundaries falling at the junction between genes is evident in the row that shows block boundaries across all populations. It also shows that there are shared block boundaries within genes that are also shared across several populations. There is also evidence of recombination sites within genes, namely within IL-19, which has block boundaries between markers 41 and 46 across all populations in this study. Both of these tables were inspired by a figure in a paper by Liu et al. that investigated population-specific variation in haplotype block structures (Liu et al., 2004).

Table 4 shows the haplotype block boundaries that were determined by the three haplotype phasing programs used in this thesis. No single block boundary was shared by all three phasing programs. PHASE and Beagle shared block boundaries much more frequently than either shared with MDBlocks. Out of 15 identified haplotype block boundaries, PHASE and Beagle shared 13 but MDBlocks and Beagle only shared 2 common block boundaries, showing the discordance between PHASE and Beagle with MDBlocks. The same general trend of block boundaries around genes and some boundaries within genes such as IL-19 is shown across phasing programs as it was across different populations in Table 3.

MDBlocks

MDBlocks, which uses an adaptive model fitting approach in determining haplotype blocks, found 10 blocks across the 57 markers included in this study when all populations were included. Haplotype blocks range from 12 markers to only 2 markers in length. Generally, the number of unique haplotypes in a given block increases with block length, though there are exceptions such as block 8, which is only 5 markers in length but has 11 unique haplotypes (see Table 5). Only three individual populations were run through MDBlocks. For an unknown reason and likely due to the nature of the data, the files for several populations would not run through the program.



Output of Program `MDBlocks`

(written by Eric C. Anderson and John Novembre)

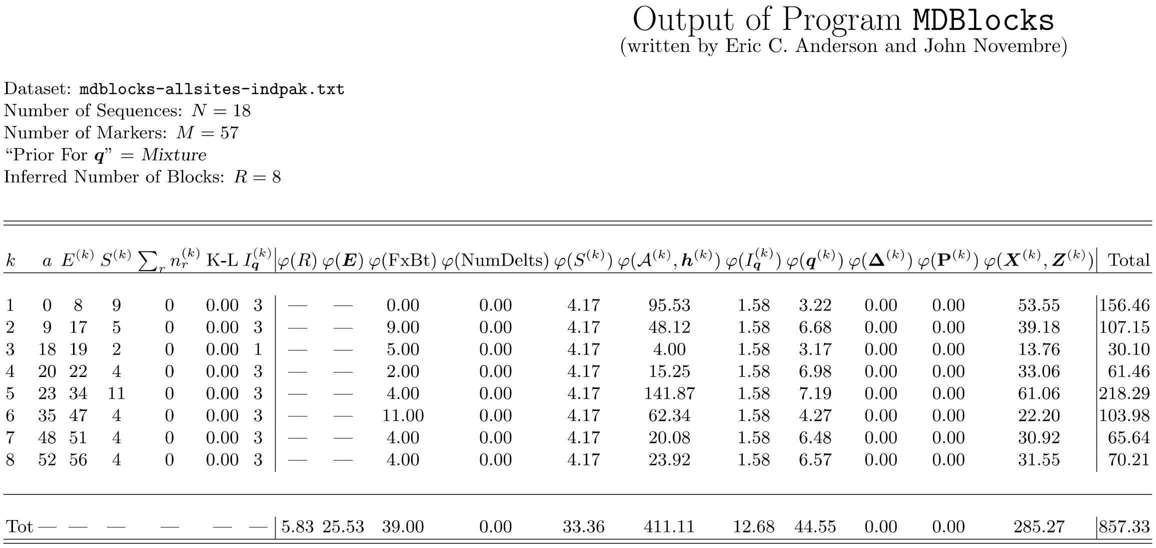
Dataset: `mdblocks-allsites-allpops-finalfinal.txt`
Number of Sequences: $N = 148$
Number of Markers: $M = 57$
"Prior For $\boldsymbol{q}$" $=$ *Mixture*
Inferred Number of Blocks: $R = 10$

k	a	$E^{(k)}$	$S^{(k)}$	$\sum_r n_r^{(k)}$	K-L	$I_{\boldsymbol{q}}^{(k)}$	$\varphi(R)$	$\varphi(\boldsymbol{E})$	φ(FxBt)	φ(NumDelts)	$\varphi(S^{(k)})$	$\varphi(\mathcal{A}^{(k)}, \boldsymbol{h}^{(k)})$	$\varphi(I_{\boldsymbol{q}}^{(k)})$	$\varphi(\boldsymbol{q}^{(k)})$	$\varphi(\boldsymbol{\Delta}^{(k)})$	$\varphi(\mathbf{P}^{(k)})$	$\varphi(\boldsymbol{X}^{(k)}, \boldsymbol{Z}^{(k)})$	Total
1	0	8	22	0	0.00	3	—	—	0.00	0.00	7.21	215.11	1.58	27.83	0.00	0.00	584.46	834.61
2	9	20	17	7	2.82	3	—	—	22.00	24.52	7.21	214.74	1.58	30.91	28.52	49.46	430.57	807.94
3	21	23	4	5	2.71	1	—	—	17.00	10.00	7.21	15.25	1.58	17.00	10.00	33.34	174.55	284.34
4	24	25	4	3	2.68	3	—	—	4.00	6.00	7.21	11.17	1.58	13.04	6.00	17.09	114.69	179.20
5	26	34	22	2	3.09	3	—	—	4.00	4.46	7.21	213.10	1.58	29.33	8.85	13.91	558.15	839.02
6	35	40	3	2	0.17	3	—	—	22.00	3.17	7.21	22.53	1.58	9.90	3.17	11.40	134.02	213.40
7	41	42	3	2	6.80	1	—	—	3.00	3.17	7.21	7.51	1.58	11.96	3.17	12.78	139.47	188.27
8	43	47	11	3	5.97	3	—	—	3.00	6.92	7.21	66.81	1.58	25.00	10.24	20.09	299.04	438.30
9	48	51	7	5	2.79	3	—	—	11.00	11.23	7.21	34.32	1.58	24.92	13.81	33.51	234.27	370.27
10	52	56	7	2	1.33	3	—	—	7.00	5.61	7.21	45.70	1.58	21.20	5.61	13.67	298.85	404.85
Tot	—	—	—	—	—	—	5.83	30.70	93.00	75.09	72.09	846.23	15.85	211.08	89.39	205.24	2968.07	4612.57

Table 5: This table shows the output file from MDBlocks when all populations using all sites were run through the program. The first column indicates the number of blocks found. The second and third columns indicate the boundaries for each block. Ex. Block 1 contains markers 1-8 while block 2 contains markers 9-20. The fourth column in the table indicates the number of unique haplotypes in each block.

In the IndoPakistani population, only 8 haplotype blocks were found and the variance of haplotype block length was from 2 to 12 markers in length (see Table 6).

Unlike when all populations are considered simultaneously, there is not a tendency for shorter haplotype blocks to exhibit higher than anticipated haplotype diversity. In this population, there is a trend of increasing haplotype diversity with block length except for the longest haplotype block, which had only four unique haplotypes representing it.



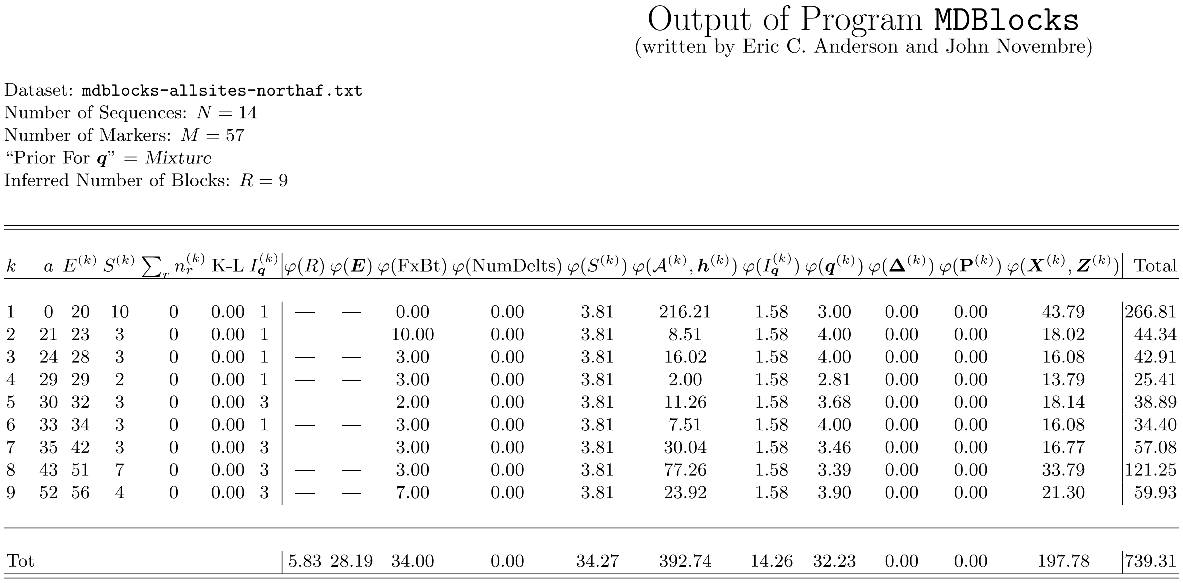
Output of Program MDBlocks
(written by Eric C. Anderson and John Novembre)

Dataset: `mdblocks-allsites-indpak.txt`
Number of Sequences: $N = 18$
Number of Markers: $M = 57$
"Prior For $\boldsymbol{q}$" $=$ *Mixture*
Inferred Number of Blocks: $R = 8$

k	a	$E^{(k)}$	$S^{(k)}$	$\sum_r n_r^{(k)}$	K-L	$I_{\boldsymbol{q}}^{(k)}$	$\varphi(R)$	$\varphi(\boldsymbol{E})$	$\varphi(\text{FxBt})$	$\varphi(\text{NumDelts})$	$\varphi(S^{(k)})$	$\varphi(\mathcal{A}^{(k)}, \boldsymbol{h}^{(k)})$	$\varphi(I_{\boldsymbol{q}}^{(k)})$	$\varphi(\boldsymbol{q}^{(k)})$	$\varphi(\boldsymbol{\Delta}^{(k)})$	$\varphi(\mathbf{P}^{(k)})$	$\varphi(\boldsymbol{X}^{(k)}, \boldsymbol{Z}^{(k)})$	Total
1	0	8	9	0	0.00	3	—	—	0.00	0.00	4.17	95.53	1.58	3.22	0.00	0.00	53.55	156.46
2	9	17	5	0	0.00	3	—	—	9.00	0.00	4.17	48.12	1.58	6.68	0.00	0.00	39.18	107.15
3	18	19	2	0	0.00	1	—	—	5.00	0.00	4.17	4.00	1.58	3.17	0.00	0.00	13.76	30.10
4	20	22	4	0	0.00	3	—	—	2.00	0.00	4.17	15.25	1.58	6.98	0.00	0.00	33.06	61.46
5	23	34	11	0	0.00	3	—	—	4.00	0.00	4.17	141.87	1.58	7.19	0.00	0.00	61.06	218.29
6	35	47	4	0	0.00	3	—	—	11.00	0.00	4.17	62.34	1.58	4.27	0.00	0.00	22.20	103.98
7	48	51	4	0	0.00	3	—	—	4.00	0.00	4.17	20.08	1.58	6.48	0.00	0.00	30.92	65.64
8	52	56	4	0	0.00	3	—	—	4.00	0.00	4.17	23.92	1.58	6.57	0.00	0.00	31.55	70.21
Tot	—	—	—	—	—	—	5.83	25.53	39.00	0.00	33.36	411.11	12.68	44.55	0.00	0.00	285.27	857.33

Table 6: This table shows the output file from MDBlocks when all sites within the IndoPakistan population were run through the program. The first column indicates the number of blocks found. The second and third columns indicate the boundaries for each block. The fourth column in the table indicates the number of unique haplotypes in each block.

In the North African population, 9 haplotype blocks were found and the variance of block size was from 1 to 20 markers in length. There was also a trend in the North African samples where longer haplotype blocks begat higher haplotype diversity with the exception of one block composed of seven markers only harboring 3 unique haplotypes (see Table 7).



Output of Program `MDBlocks`
(written by Eric C. Anderson and John Novembre)

Dataset: `mdblocks-allsites-northaf.txt`
Number of Sequences: $N = 14$
Number of Markers: $M = 57$
"Prior For $\boldsymbol{q}$" = *Mixture*
Inferred Number of Blocks: $R = 9$

k	a	$E^{(k)}$	$S^{(k)}$	$\sum_r n_r^{(k)}$	K-L	$I_{\boldsymbol{q}}^{(k)}$	$\varphi(R)$	$\varphi(\boldsymbol{E})$	φ(FxBt)	φ(NumDelts)	$\varphi(S^{(k)})$	$\varphi(\mathcal{A}^{(k)}, \boldsymbol{h}^{(k)})$	$\varphi(I_{\boldsymbol{q}}^{(k)})$	$\varphi(\boldsymbol{q}^{(k)})$	$\varphi(\boldsymbol{\Delta}^{(k)})$	$\varphi(\mathbf{P}^{(k)})$	$\varphi(\boldsymbol{X}^{(k)}, \boldsymbol{Z}^{(k)})$	Total
1	0	20	10	0	0.00	1	—	—	0.00	0.00	3.81	216.21	1.58	3.00	0.00	0.00	43.79	266.81
2	21	23	3	0	0.00	1	—	—	10.00	0.00	3.81	8.51	1.58	4.00	0.00	0.00	18.02	44.34
3	24	28	3	0	0.00	1	—	—	3.00	0.00	3.81	16.02	1.58	4.00	0.00	0.00	16.08	42.91
4	29	29	2	0	0.00	1	—	—	3.00	0.00	3.81	2.00	1.58	2.81	0.00	0.00	13.79	25.41
5	30	32	3	0	0.00	3	—	—	2.00	0.00	3.81	11.26	1.58	3.68	0.00	0.00	18.14	38.89
6	33	34	3	0	0.00	1	—	—	3.00	0.00	3.81	7.51	1.58	4.00	0.00	0.00	16.08	34.40
7	35	42	3	0	0.00	3	—	—	3.00	0.00	3.81	30.04	1.58	3.46	0.00	0.00	16.77	57.08
8	43	51	7	0	0.00	3	—	—	3.00	0.00	3.81	77.26	1.58	3.39	0.00	0.00	33.79	121.25
9	52	56	4	0	0.00	3	—	—	7.00	0.00	3.81	23.92	1.58	3.90	0.00	0.00	21.30	59.93
Tot	—	—	—	—	—	—	5.83	28.19	34.00	0.00	34.27	392.74	14.26	32.23	0.00	0.00	197.78	739.31

Table 7: This table shows the output file from MDBlocks when all sites within the North African population were run through the program The first column indicates the number of blocks found. The second and third columns indicate the boundaries for each block. The fourth column in the table indicates the number of unique haplotypes in each block.

Russian samples showed 8 haplotype blocks in total and the variance of block length was from 2 to 14 markers. Per the usual trend, the longest haplotype blocks harbored the most haplotype variability with one exception of a long block that had only three unique haplotypes. The longer blocks with unexpectedly low haplotype variability occurred across roughly the same markers (35-42) in all of the individual populations (see Table 8).

Output of Program MDBlocks

(written by Eric C. Anderson and John Novembre)

Dataset: `mdblocks-allsites-russia.txt`
Number of Sequences: $N = 18$
Number of Markers: $M = 57$
"Prior For $\boldsymbol{q}$" $=$ *Mixture*
Inferred Number of Blocks: $R = 8$

k	a	$E^{(k)}$	$S^{(k)}$	$\sum_r n_r^{(k)}$	K-L	$I_{\boldsymbol{q}}^{(k)}$	$\varphi(R)$	$\varphi(\boldsymbol{E})$	$\varphi(\text{FxBt})$	$\varphi(\text{NumDelts})$	$\varphi(S^{(k)})$	$\varphi(\mathcal{A}^{(k)}, \boldsymbol{h}^{(k)})$	$\varphi(I_{\boldsymbol{q}}^{(k)})$	$\varphi(\boldsymbol{q}^{(k)})$	$\varphi(\boldsymbol{\Delta}^{(k)})$	$\varphi(\mathbf{P}^{(k)})$	$\varphi(\boldsymbol{X}^{(k)}, \boldsymbol{Z}^{(k)})$	Total
1	0	8	9	0	0.00	3	—	—	0.00	0.00	4.17	88.36	1.58	3.54	0.00	0.00	52.69	148.77
2	9	10	2	0	0.00	3	—	—	9.00	0.00	4.17	4.00	1.58	2.61	0.00	0.00	9.06	28.84
3	11	17	5	0	0.00	3	—	—	2.00	0.00	4.17	39.39	1.58	4.77	0.00	0.00	36.30	86.64
4	18	19	2	0	0.00	1	—	—	5.00	0.00	4.17	4.00	1.58	3.17	0.00	0.00	17.84	34.18
5	20	34	12	0	0.00	3	—	—	2.00	0.00	4.17	195.33	1.58	2.72	0.00	0.00	60.30	264.53
6	35	42	3	0	0.00	1	—	—	12.00	0.00	4.17	27.28	1.58	5.46	0.00	0.00	25.84	74.75
7	43	51	5	0	0.00	3	—	—	3.00	0.00	4.17	55.59	1.58	5.68	0.00	0.00	32.33	100.77
8	52	56	4	1	0.70	3	—	—	5.00	2.00	4.17	23.92	1.58	5.74	2.00	3.81	25.26	71.91
Tot	—	—	—	—	—	—	5.83	25.53	38.00	2.00	33.36	437.89	12.68	33.69	2.00	3.81	259.63	854.43

Table 8: This table shows the output file from MDBlocks when all sites within the Russian population were run through the program. The first column indicates the number of blocks found. The second and third columns indicate the boundaries for each block. The fourth column in the table indicates the number of unique haplotypes in each block.

Bearing Analysis and Angular Correlation

Bearing analysis across the IL-10 gene family and MAPKAP-K2, which was run with the program PASSaGE largely showed an east-west clinal distribution of variation but none of the estimates reached the set significance threshold. One SNP in MAPKAP-K2, rs7515374, nearly reached significance but fell just short. A closer inspection of the bearing analysis, which is more suitable for populations of smaller size, for that particular SNP showed that they were indeed some significant values.

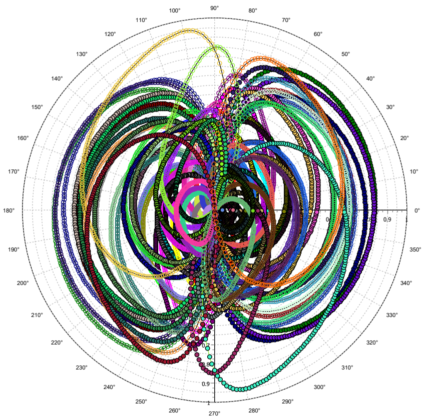


Figure 21: This figure shows the angular correlation plot for each locus in MAPKAP-K2 and the IL-10 gene family. The direction of the colored circles represents the direction that each locus is mostly directly correlated with the allele frequencies observed in the data.

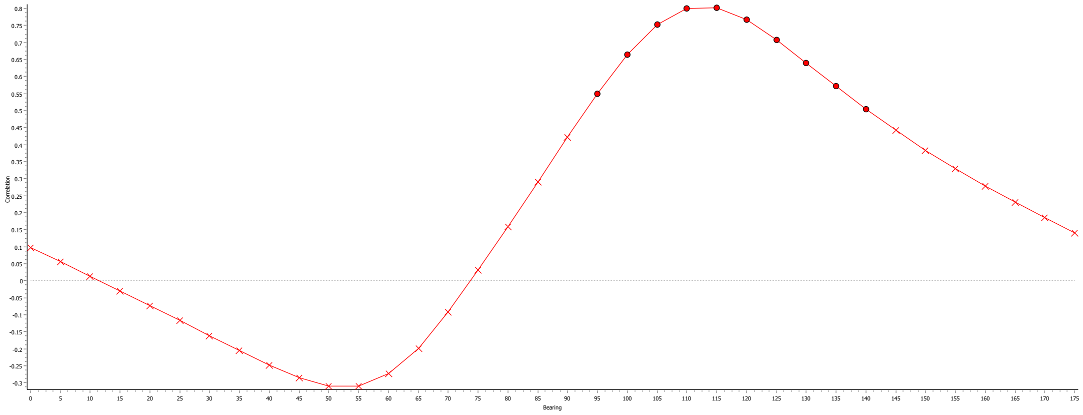


Figure 22: This is a bearing analysis plot for the SNP, rs7515374, in MAPKAP-K2. Circles along the line indicate significant values while crosses signify non-significant values.

The bearing analysis in Figure 22 indicates that the greatest positive correlation trends in a NNW direction. The cline, which goes in the direction of greatest negative correlation and is the opposite of the greatest positive correlation, goes in the NNE direction. As far as allele frequencies are concerned, which are the data that spatial analysis is being performed on, this means that frequencies vary along the SSW-NNE axis.

5L{/¦{{Lhb

Population-Specific Differences in LD Patterns

Within the general pattern of extensive LD signals spanning entire genes, slight population-specific block boundaries do exist, which is in concordance with authors that have described how haplotype block structure varies across populations (Liu et al., 2004). The results of this study corroborate that claim, with only seven of the 57 sites showing a shared haplotype block boundary. For the purposes of this comparison, a shared block boundary is defined by a boundary that is common amongst at least five of the nine populations in this study (see Table 3). Four of the seven shared block boundaries occur at the junction between genes.

South Africa exhibits less extensive LD more generally and especially where all other populations in the study possess strong signals of LD at IL-24. This observation is expected given the Out of Africa model of human dispersal and the expectation that older, more established populations have had sufficient time to accumulate mutations and accrue the effects of other demographic forces that serve to break up LD patterns over time.

Unlike the conclusions drawn by Gabriel et al., the population specific differences in the extent of LD across the sites does not seem to reflect the demographic histories of the populations aside from the relaxation of LD across IL-

24 in the South African samples (Gabriel et al., 2002; Sawyer et al., 2005). High LD at the beginning of IL-19 across populations and a relaxation of LD at IL-24 in South African populations may be a reflection that IL-19 is under stronger selection with certain haplotypes being driven to higher frequencies due to positive selection or alternative haplotypes being eliminated through stronger negative selection whereas IL-24 variability may be primarily determined by stochastic demographic forces. Kõks et al. noted LD across IL-19 and IL-20 and suggested that there may be protective haplotypes that play a role in psoriasis susceptibility at these loci due to associations between IL-19 and IL-20 with psoriatic phenotypes (Kõks et al., 2004). It could also be the case that IL-19 and IL-24 exhibit stronger LD signals because MAPKAP-K2 and part of IL-10 are in an area of relatively high recombination (1.9 cM/Mb) where the rest of the sequence under study is in an area of low recombination (.5 cM/Mb) (Kent et al., 2002).

The occurrence of haplotype block boundaries around rather than within the boundaries of genes is in concordance with the conclusion by McVean et al. that recombination is less likely to occur within genes, presumptively because recombination within exons is deleterious (McVean et al., 2004). Despite this assertion, all three programs found haplotype block boundaries within gene boundaries, most notably within IL-19, but also within genes with a greater density of SNPs.

Population-specific differences in the number of haplotype blocks could indicate different population histories. Given models of human dispersal and demographic expectations of LD, populations outside of Africa are expected to have extended haplotypes, more extensive LD and thus, fewer haplotype blocks with low haplotype diversity. Populations within or close to Africa are expected to exhibit less extensive LD, shorter haplotypes, shorter blocks of LD and should, therefore, have more numerous and shorter haplotype blocks (Tishkoff et al., 1996). A consistent increase of average haplotype block length with increasing distance from Africa was not observed in these data (see Table 2). IL-24 showed a somewhat consistent decrease of average haplotype block length that is in concordance with expectations under a population history model, though it was slight. This suggests that demographic forces are likely not the only forces constraining or creating patterns of LD in this region of chromosome 1.

LD heat maps did not show a consistent pattern of LD across IL-19 and IL-20, which was reported by Kõks et al. in their study on haplotypes spanning the two genes, though there was strong LD across the beginning of IL-19 (Kõks et al., 2004). Only the Russian population showed a pattern of strong LD that links IL-19 with IL-20 and even in that case, the block did not encapsulate the entirety of IL-19 in the same block as IL-20 (Figure 23). However, when considering the population specific block boundaries across each gene boundary, the junction between IL-19 and IL-20 has fewer shared block boundaries across all populations compared to any other junction in this study. Notably, the only individual populations that exhibit a block boundary between IL-19 and IL-20 are from the North and South African population samples.

Extent of LD in MAPKAP-K2 and the IL-10 Gene Family

In the output from MDBlocks, there were haplotype blocks from marker 35-42 that were comparatively long but had unexpectedly low haplotype diversity given that the general trend was increasing haplotype block length led to increased haplotype diversity. This pattern was seen across the individual populations that were run through MDBlocks (IndoPak, Russia and North Africa) but not all populations together. This region is at the beginning of IL-19 and there are several other pieces of data outlined in this thesis that capture this low haplotype diversity. Table 1 shows almost no haplotype block boundaries in this region of IL-19. Heat maps for IL-19 (see Figures 6-9) also show extended and extensive LD across the first half of IL-19, which would lower haplotype diversity.

PHASE Untrimmed Data across All Sites in Russian Population

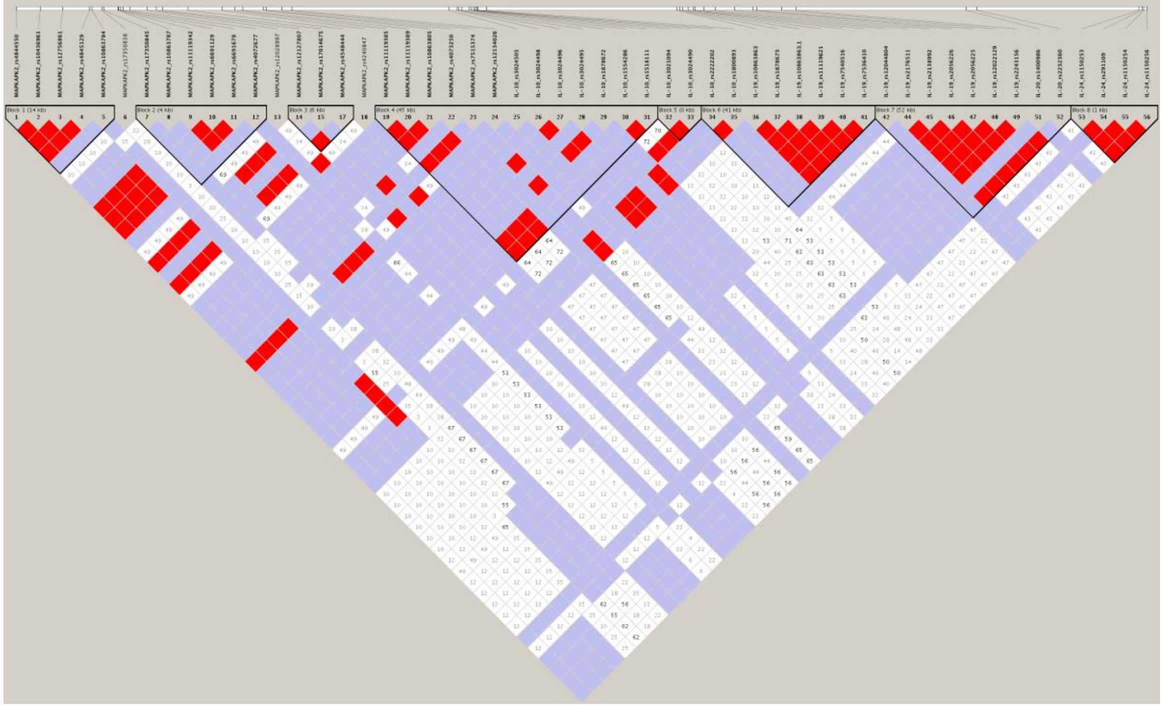


Figure 23: This figure shows the heat map across all sites in the untrimmed Russian dataset. This heat map was generated using Haploview with haplotype phasing done with the PHASE program.

There were some consistent signals of LD across all or most of the study populations, namely in the region of high LD in a portion of IL-19, however, there were also many population- specific and program-specific regions of elevated LD as well. The Basque population showed a notable relaxation of LD across MAPKAP-K2 where other populations showed at least sporadic instances of LD between markers.

Reconciling Population-Specific Differences and General Patterns of LD

One of the goals in this thesis is to evaluate the effect of population structure on LD estimates as compared to LD evaluation across all populations. Most of the heat maps and haplotype block boundary charts shown in the Results section show that there is ample evidence for nuance in LD estimates within populations. Some regions within this study showed stronger signals of LD that transcended the population level view, such as IL-19, but other strong signals of LD are not apparent unless all populations are evaluated together.

Figure 19 is an LD heat map of all populations together and given this view, it is apparent that there are salient block boundaries around each of the genes, particularly MAPKAP-K2 and IL-10, with the exception of the block boundary that occurs within IL-19. The significance of a general pattern of LD with population-specific differences in both the extent of LD at any given locus at the exact location of haplotype block boundaries has different implications given the application. Pritchard and Przeworski elaborate on how a nuanced understanding of LD at different scales can be used to glean unique pieces of information about complex disease causation. In complex disease association studies, where the goal is to find a general region where a disease locus may lie along the human genome, large-scale views of LD patterns such as those given in Figure 19 can summarize large amounts of genetic data and narrow the field of search (Pritchard and Przeworski, 2001). On the other hand, when exact localization of a disease allele is the goal as it often is after association studies have found an associated disease region, the smaller scale nuances afforded by gene-by-gene and population-specific analysis allow for the pinpointing of the causative locus.

The Effect of Trim Status on LD

Certain individuals and loci were trimmed from the original dataset due to quality control concerns. To evaluate the effect of said trimming, input files were subsetted on this basis, as well as population-by-population and gene-by-gene. Figures 15 through 18 show the effect of trim status on LD across two programs: PHASE and Beagle. For the untrimmed PHASE dataset, IndoPakistan, Russia and North Africa showed high LD designations deep within the heat map when compared against the trimmed dataset, indicating that sites that are farther apart on this region of the chromosome are in strong LD even though local LE has been generated within those larger blocks of disequilibrium. This pattern of long-range LD deep within the heat map is also seen in the untrimmed dataset from Beagle with Iberia, South Africa and Russia showing deep levels of LD compared to their counterparts in the trimmed dataset.

The consequence of evaluating trim status on LD measures in populations with small population sample sizes may have inadvertently revealed the limitations of the algorithms that assess LD. In Figures 6 through 9, which show heat maps for LD in IL-19 between trimmed and untrimmed datasets inputted into Beagle and PHASE for haplotype imputation, the effect of small sample size on LD estimates is apparent. Both trimmed heat maps show no LD across all of IL-19 in North African samples, even though the signal of LD at the beginning of this region is well represented in all other study populations. Because of trimming, this population was only represented by four haplotypes. In the untrimmed heat maps, the signal of LD in North Africa is almost exactly identical to all other populations that maintained that strong signal of LD across the first half of IL-19, which is even apparent at the level of all populations in Figure 19. For the purposes of this study and given the nature of these data, the untrimmed datasets tended to be more revealing of patterns of LD. The trimming of data based on conservative genotyping calls seemed to only reduce the number of samples and data points, not necessarily any error in the dataset.

Accurate Predictions of Haplotype Phase

A concern of this study is the accurate prediction of missing genotype data in the imputation process, which is resolved by the programs employed, PHASE, Beagle and MDBlocks, through different statistical approaches. PHASE relies on a coalescent model, which infers missing genotypic data given the assumption that imputed haplotypes are likely to be similar to other observed haplotypes in the dataset (Browning and Browning, 2007). Conversely, Beagle uses a Hidden Markov Model approach in imputing missing haplotypes given the assumption that the observables, in this case background haplotypes and local levels of LD, will accurately predict haplotypes with missing data (Pei et al., 2008). In evaluating the effectiveness of different inference methods, Pei et al. note that Beagle's particular treatment of the HMM approach suffers in accurate predictive power, at least compared to MACH and IMPUTE, by not considering reference haplotypes such as those from the HapMap project as hidden states. While the approach used in PHASE is computationally intensive, it is reported to be a reliable method for predicting missing data during imputation (Browning and Browning, 2007). While the goal of this study is to evaluate the imputation accuracy between the programs Beagle and PHASE, the small size of the sample set biases the results in favor of PHASE.

While no methods were employed to specifically assess the accuracy of phasing methods given this dataset, the results given by each program were fairly consistent and in concordance with each other. Furthermore, downstream applications on LD analysis also showed little variation in results between PHASE and Beagle. Other variables, such as assessing LD in each population separately versus all populations together or trimming the datasets, seemed to have a greater impact on the results of LD analyses than did the program used for haplotype phasing.

Spatial Autocorrelation Analyses

The results of the bearing analysis showed an East-West distribution across the loci on chromosome 1, which is concordant with the observations of ABO blood groups and rhesus factor highlighted in the paper by Falsetti and Sokal (Falsetti and Sokal, 1993). Unfortunately, none of the data points were significant so the conclusions drawn from the bearing analysis are necessarily limited. Likely, the lack of significance across the loci could be that the sampling distribution from an east-west direction is not as uniform as it is going from south to north.

/!±9!¢{

A caveat to these data is the degree to which admixture is influencing the resolution of haplotype phase and downstream estimates of LD. The samples were purchased from Coriell and the exact location of sampling is unknown. If, for example, Chinese samples come from the countryside, which is likely to be less admixed than city centers such as Beijing or Hong Kong, then LD estimates may differ accordingly. This uncertainty likely has implications for the influence of other stochastic demographic forces, like genetic drift, on the results obtained. Another restriction on this study given these data is that in order to assess the impact of stochastic demographic forces on the impact of LD and haplotype block boundary determination, a single region of the genome cannot be used as a proxy for a population history. While conclusions may be cautiously drawn about the impact of natural selection, a force that is site- specific in its impact, one must be even more conservative when conjecturing about the implications for forces that are genome-wide (Slatkin, 2008).

Another caveat speaks to the limitations of LD as a measure of summarizing genetic variability and the necessary caution one must take when extrapolating conclusions. LD analysis speaks only to the association of alleles, not to the cause of their association. Trifonova et al. elaborate on the limited applicability of LD analysis results given the research questions they are proposed to answer

(Trifonova et al., 2012). While it would certainly be useful to efficiently summarize large swaths of genetic data for the purposes of dealing with the burden of large-scale genome wide association data and the task of teasing apart the intricate genetic influences in complex disorders, there are likely meaningful population-specific and site-specific differences that make such a silver bullet solution miss the mark.

The study population and the numbers of each site for each locus were small in this study, perhaps so much so that some of the signals may be an artifact of small sample size. A small sample size can lead to an overestimation of the extent of LD because points of recombination are not accurately detected (Want et al., 2002). Furthermore, this study was limited to male samples, which may reveal a sex-specific bias in regard to recombination rate and its impact on LD results.

/hb/[¦{Lhb{

This thesis sought to examine the methodological implications of different haplotype phasing and subsetting approaches to downstream applications of LD analysis. Two different and widely employed haplotype phasing programs, PHASE and Beagle, were used to phase genotype data from MAPKAP-K2 and the IL-10 gene family. A third haplotype phasing program, MDBlocks, was used to compare phase results against the aforementioned programs as well. Results from PHASE and Beagle were used as input into Haploview for LD analysis. At this point, several subsetting approaches were used to partition the data such that the impact of these differences on the final results could be assessed. Input files were constructed for each population individually as well as all populations together. The original dataset was trimmed based on the confidence of genotyping calls and the trimmed and untrimmed sets were run to assess the impact of that variable as well.

All three haplotype phasing methods performed similarly in determining how many haplotype blocks occurred across the region and where haplotype block boundaries fell. Subsetting the data by population versus inputting all populations together in LD analysis revealed population-specific differences in haplotype block length and block boundary positions. The input of all populations into Haploview for LD analysis revealed general patterns of block boundaries that flank gene boundaries with the exception of some block boundaries that fall within genes, most notably within IL-19.

w9C9w9b/9{

"The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: 13820 13838 13852 13876 13877 13911 13912 13913 13914 15883 15884 15885 15886 15887 16185 16188 16654 16688 16689 17014 17015 17016 17017 17018 17021 17022 17023 17024 17026 17027 17028 17029 17030 17091 17092 17093 17094 17095 17096 17097 17099 17100 17301 17302 17303 17307 17308 17309 17310 17331 17332 17333 17334 17335 17336 17337 17339 17340 17341 17342 17343 17344 17345 17346 17347 17348 17349 17378 17379 17380 17381 17382 17383 17384 17385 17386 17387 17388 17390 17391."

Anderson EC, Novembre J. 2003. Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336–54.

Ardlie KG, Kruglyak L, Seielstad M. 2002. Patterns of Linkage Disequilibrium in the Human Genome. Nat Rev Genet 3:299–309.

Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN, King KS, Nelson MR, Bustamante CD. 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res 19:795–803.

Avise JC. 1998. The History and Purview of Phylogeography : a Personal Reflection. Mol Ecol 7:371–379.

Bare Bones Software, Inc. 2009. TextWrangler 5.5. Barebones.com

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

Ben-Levy R, Hooper S, Wilson R, Paterson HF, Marshall CJ. 1998. Nuclear export of the stress-activated protein kinase p38 mediated by its substrate MAPKAP kinase-2. Curr Biol 8:1049–1057.

Blumberg H, Conklin D, Xu W, Grossmann A, Brender T, Carollo S, Eagan M, Foster D, Haldeman BA, Hammond A, Haugen H, Jelinek L, Kelly JD, Madden K, Maurer MF, Parrish-novak J, Prunkard D, Sexson S, Sprecher C, Waggie K, West J, Whitmore TE, Yao L, Kuechle MK, Dale BA, Chandrasekher YA. 2001. Interleukin 20: Discovery, Receptor Identification, and Role in Epidermal Function. Cell Press 104:9–19.

Brocker C, Thompson D, Matsumoto A, Nebert DW, Vasiliou V. 2010. Evolutionary divergence and functions of the human interleukin (IL) gene family. Hum

Genomics 5:30.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination. Am Hournal Hum Genet 63:861–869.

Browning SR. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124:439–450.

Browning SR, Browning BL. 2007. Rapid and Accurate Haplotype Phasing and Missing Data Inference for Whole Genome Association Studies by Use of Localized Haplotype Clustering. Am J Hum Genet 81:1084–1097.

Browning SR, Browning BL. 2012. Haplotype phasing: Existing Methods and New Developments. Nat Rev Genet 12:703–714.

Cantó E, Garcia Planella E, Zamora-Atenza C, Nieto JC, Gordillo J, Ortiz MA, Metón I, Serrano E, Vegas E, García-Bosch O, Juárez C, Vidal S. 2014. Interleukin-19 impairment in active Crohn's disease patients. PLoS One 9:e93910.

Cao S, Liu J, Song L, Ma X. 2005. The Protooncogene c-Maf Is an Essential Transcription Factor for IL-10 Gene Expression in Macrophages. J Immunol 174:3484–3492.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effects of deleterious mutations on neutral molecular variation. Genet Soc Am 134:1289–1303.

Charlesworth D. 2006. Balancing Selection and its Effects on Sequences in Nearby Genome Regions. PloS Genet 2:e64.

Chen P-J, Wei C-C, Wang C, Chen F-W, Hsu Y-H, Chang M-S. 2006. Promoter analysis of interleukin-19. Biochem Biophys Res Commun 344:713–20.

Chen W-Y, Chang M-S. 2009. IL-20 is regulated by hypoxia-inducible factor and up-regulated after experimental ischemic stroke. J Immunol 182:5,003-5,012.

Cheng S Bin, Sharma S. 2015. Interleukin-10: A pleiotropic regulator in pregnancy. Am J Reprod Immunol 73:487–500.

Chikhi L, Nichols R a, Barbujani G, Beaumont M a. 2002. Y genetic data support the Neolithic demic diffusion model. Proc Natl Acad Sci U S A 99:11008–11013.

Evans DM, Cardon LR. 2005. A Comparison of Linkage Disequilibrium Patterns and Estimated Population Recombination Rates across Multiple Populations. Am J Hum Genet 76:681–687.

Falsetti AB, Sokal RR. 1993. Genetic structure of human populations in the British Isles. Ann Hum Biol 20:215–229.

Flint-Garcia SA, Thornsberry JM, S E, IV B. 2003. Structure of Linkage Disequilibrium in Plants. Annu Rev Plant Biol 54:357–374.

Fu Q, Rudan P, Paabo S, Krause J. 2012. Complete Mitochondrial Genomes Reveal Neolithic Expansion into Europe. PLoS One 7:e32473.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The Structure of Haplotype Blocks in the Human Genome. Science (80-) 296:2225–2229.

Gaestel M. 2006. MAPKAP kinases — MKs — two's company, three's a crowd. Nat Rev Mol Cell Biol 7:120–130.

Gallagher G, Dickensheets H, Eskdale J, Vazquez N, Pestka S. 2000. Cloning, expression and initial characterisation of interleukin-19 (IL-19), a novel homologue of human interleukin 10 (IL-10). Genes Immunol 19:442–450.

Grünwald P. 2000. Model Selection Based on Minimum Description Length. J Math Psychol 44:133–152.

Handley LJL, Manica A, Goudet J, Balloux F. 2007. Going the distance: human population genetics in a clinal world. Trends Genet 23:432–9.

Hedrick PW. 1987. Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341.

Hofmann SR, Rösen-Wolff A, Tsokos GC, Hedrich CM. 2012. Biological properties and regulation of IL-10 related cytokines and their contribution to autoimmune disease and tissue injury. Clin Immunol 143:116–27.

Hudson, RR, Kaplan, NL 1985. Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences. The Genetics Society of America 111: 147-164.

Huff CD, Harpending HC, Rogers AR. 2010. Detecting Positive Selection from Genome Scans of Linkage Disequilibrium. BMC Genomics 11:1–9.

Im S-H, Hueber A, Monticelli S, Kang K-H, Rao A. 2004. Chromatin-level Regulation of the IL-10 Gene in T cells. J Biol Chem 279:46,818-46,825.

Jay F, Sjödin P, Jakobsson M, Blum MGB. 2013. Anisotropic isolation by distance: the main orientations of human genetic differentiation. Mol Biol Evol 30:513–25.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222.

Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. 2014. Human Evolutionary Genetics. Garland Science.

Jones EA, Flavell RA. 2005. Distal Enhancer Elements Transcribe Intergenic RNA in the IL-10 Family Gene Cluster. J Immunol 175:7,437-7,446.

Jordan WJ, Eskdale J, Boniotto M, Lennon GP, Peat J, Campbell JDM, Gallagher G. 2005. Human IL-19 regulates immunity through auto-induction of IL-19 and production of IL-10. Eur J Immunol 35:1576–82.

Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. Genome Biol Evol 3:614–26.

Kent W, Sugnet C, Furey T, Roskin K, Pringle T, Zahler A, Haussler D. 2002. UCSC Genome Browser. Genome Res 12:996–1006.

Kõks S, Kingo K, Rätsep R, Karelson M, Silm H, Vasar E. 2004. Combined haplotype analysis of the interleukin-19 and -20 genes: relationship to plaque-type psoriasis. Genes Immun 5:662–667.

Kõks S, Kingo K, Vabrit K, Rätsep R, Karelson M, Silm H, Vasar E. 2005. Possible relations between the polymorphisms of the cytokines IL-19, IL-20 and IL-24 and plaque-type psoriasis. Genes Immun 6:407–415.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. 2002. A high-resolution recombination map of the human genome. Nat Genet 31:241–247.

Kotenko S V. 2002. The family of IL-10-related cytokines and their receptors: related, but to what extent? Cytokine Growth Factor Rev 13:223–240.

Kuhn R, Lohler J, Rennick D, Rajewsky K, Muller W. 1993. Interleukin-10-Deficient Mice Develop Chronic Enterocolitis. Cell Press 75:263–274.

De La Vega FM, Isaac H, Collins A, Scafe CR, Halldórsson B V., Su X, Lippert R, Wang Y, Laig-Webster M, Koehler RT, Ziegle JS, Wogan LT, Stevens JF, Leinen KM, Olson SJ, Guegler KJ, You X, Xu LH, Hemken HG, Kalush F, Itakura M, Zheng Y, de Thé G, O'Brien SJ, Clark AG, Istrail S, Hunkapiller MW, Spier EG, Gilbert D. 2005. The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. Genome Res 15:454–462.

Lewontin R., Kojima K. 1960. Evolutionary Dynamics of Complex Polymorphisms. Evolution (N Y) 14:458–472.

Liu N, Sawyer SL, Mukherjee N, Pakstis AJ, Kidd JR, Kidd KK, Brookes AJ, Zhao H. 2004. Haplotype Block Structures Show Significant Variation among Populations. Genet Epidemiol 27:385–400.

Lynn A, Ashley T, Hassold T. 2004. Variation in Human Meiotic Recombination. Annu Rev Genomics Hum Genet 5:317–349.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favorable gene. Genet Res 23:23–35.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Science (80-) 304:581–4.

Mellars P. 2006. Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonizations of Eurasia. Science (80-) 313:796–800.

Moens U, Kostenko S, Sveinbjørnsson B. 2013. The role of mitogen-activated protein kinase-activated protein kinases (MAPKAPKs) in inflammation. Genes (Basel) 4:101–133.

Moore KW, Malefyt RDW, Robert L, Garra AO. 2001. Interleukin-10 and the interleukin-10 receptor. Annu Rev Immunol 1:683–765.

Morton NE. 2005. Linkage disequilibrium maps and association mapping. J Clin Invest 115:1425–1430.

Nachman MW, D'agostino SL, Tillquist CR, Mobasher Z, Hammer MF. 2004. Nucleotide Variation at Msn and Alas2, Two Genes Flanking the Centromere of the X Chromosome in Humans. Genet Soc Am 167:423–437.

Paul G, Khare V, Gasche C. 2012. Inflamed gut mucosa: downstream of interleukin-10. Eur J Clin Invest 42:95–109.

Pei YF, Li J, Zhang L, Papasian CJ, Deng HW. 2008. Analyses and comparison of accuracy of different genotype imputation methods. PLoS One 3.

Pritchard JK, Przeworski M. 2001. Linkage Disequilibrium in Humans: Models and Data. Am J Hum Genet 69:1–14.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. 2001. Linkage Disequilibrium in the Human Genome. Nature 411:199–204.

Relethford JH. 2004. Global Patterns of Isolation by Distance Based on Genetic and Morphological Data. 76:499–513.

Rosenberg MS, Anderson CD. 2011. PASSaGE: Pattern Analysis, Spatial Statistics and

Geographic Exegesis. Methods Ecol Evol 2:229–232.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLoS Genet 1:e70.

Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, Kidd KK. 2005. Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet 13:677–686.

Serre D, Pääbo S. 2004. Evidence for gradients of human genetic diversity within and among continents. Genome Res:1679–1685.

Simon, G. 1997. An Angular Version of Spatial Correlations, with Exact Significance Tests. Geographic Analysis 29: 267-278.

Slatkin M. 1994. Linkage Disequilibrium in Growing and Stable Populations. Genet Soc Am 137:331–336.

Slatkin M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485.

Slatkin M, Excoffier L. 2012. Serial founder effects during range expansion: A spatial analog of genetic drift. Genetics 191:171–181.

Stephens M, Donnelly P. 2003. A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. Am J Hum Genet 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2004. Documentation for PHASE, version 2.1. Seattle.

Stokoe D, Caudwell B, Cohen PT, Cohen P. 1993. The substrate specificity and structure of mitogen-activated protein (MAP) kinase-activated protein kinase-2. Biochem J 296 (Pt 3:843–849.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent Human Effective Population Size Estimated from Linkage Disequilibrium. Cold Spring Harb Lab Press 17:520–526.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, James L, Williams SM, Tishkoff SA, Reed FA, Francoise R, Lema G, Moore JH, Mortensen H, Wambebe C, Weber JL, Williams SM. 2009. The Genetic Structure and History of Africans and African Americans. Science (80-) 324.

Tone M, Powell MJ, Tone Y, Thompson S a. J, Waldmann H. 2000. IL-10 Gene Expression Is Controlled by the Transcription Factors Sp1 and Sp3. J Immunol 165:286–291.

Trifonova EA, Eremina ER, Urnov FD, Stepanov VA. 2012. The Genetic Diversity and Structure of Linkage Disequilibrium of the MTHFR Gene in Populations of Northern Eurasia. Acta Aaturae 4:53–69.

Wall JD, Pritchard JK. 2003. Haplotype Blocks and Linkage Disequilibrium in the Human Genome. Nat Rev Genet 4:587–597.

Wang M, Liang P. 2004. Interleukin-24 and its receptors. Immunology 114:166–170.

Wang M, Tan Z, Thomas EK, Liang P. 2004. Conservation of the genomic structure and receptor-mediated signaling between human and rat IL-24. Genes Immun

5:363–70.

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. 2002. Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. Am J Hum Genet 71:1227–1234.

Weiss KM, Clark AG. 2002. Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18:19–24.

Wright S. 1950. The Genetical Structure of Populations. Ann Hum Genet.

Xu W. 2004. Interleukin-20. Int Immunopharmacol 4:627–633.

Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L. 2003. Randomly Distributed Crossovers may Generate Block-like Patterns of Linkage Disequilibrium: an Act of Genetic Drift. Hum Genet 113:51–9.

Zhao H, Nettleton D, Dekkers JCM. 2007. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. Genet Res 89:1–6.

Zheng M, Bocangel D, Doneske B, Mhashilkar A, Ramesh R, Hunt KK, Ekmekcioglu S, Sutton RB, Poindexter N, Grimm E a, Chada S. 2007. Human interleukin 24 (MDA-7/IL-24) protein kills breast cancer cells via the IL-20 receptor and is antagonized by IL-10. Cancer Immunol 56:205–215.

CURRICULUM VITAE

Roxanne Kaaren Leiter

1019 English Avenue, Louisville, KY 40217
Personal Phone: (216) 904-8202
Office Location: Strickler Hall room 126K
Office Phone: (502) 852-0028
E-mail: roxanne.leiter@louisville.edu

EDUCATION

University of Louisville
M.A. in Anthropology (in progress) **2014 – present**
Thesis: Impact of the Phasing Process on Linkage
Disequilibrium Measures across an Immune Gene Family
Advisor: Dr. Christopher Tillquist

University of Louisville
BA. In Anthropology
Minor in Biology **2010 – 2014**

TEACHING AND LAB EXPERIENCE

University of Louisville
Instructor for GEN 105 – Supplemented College Reading **2014 – present**
Developed syllabi, lesson plans, and collaborated in developing and amending curriculum with other instructors, met with students and administered grades.

Guest Lecturer for ANTH 510 – Methods in Biological Anthropology **2017**
Presented the topics of linkage disequilibrium and haplotype phasing and also introduced thesis research

Learning Assistant for ANTH 202 – Introduction to Biological Anthropology **2013 - 2014**
Tutored students for several sessions per week, held exam reviews and met with student as needed

Peer Supervisor for REACH program **2014**
Oversaw tutor instruction and provided feedback for improvement

Lab Trainee in the Molecular Anthropology Population Studies Lab **2013 – 2015**
Trained in wet lab techniques including DNA extraction, DNA
polymerase reactions, gel electrophoresis and visualization

CONFERENCES AND SEMINARS

Celebration of Teaching and Learning
University of Louisville Delphi Center **2015 – 2017**

Annual American Association of Physical Anthropologists Meeting
American Association of Physical Anthropologists **2014 – 2016**

University of Louisville Undergraduate Research Symposium 2013

POSTERS AND RESEARCH PROJECTS

R. K. Leiter and C. T. Tillquist "*Linkage Disequilibrium at Functional Sites and Global Populations*"
Poster presentation at the Annual American Association of Physical Anthropologists Conference, Atlanta, GA. 2016

R.K. Leiter, A.E. Mann, M.F. Casanova and C.T. Tillquist "*Apolipoprotein E Polymorphisms in Austism Families*"
Poster Presentation at the Undergraduate Research Symposium 2013

A.E. Mann, R. K. Leiter and C.R. Tillquist, "*Founder Effects Impacts APOE Variability in Northern Europe*" 2013

AWARDS

Anthropology Department Annual Merit Award	2013 – 2014
College Reading and Learning Association Tutor Certification Levels I and II	2013 – 2014
Anthropology Department Graduate Merit Award	2017
Graduate Dean's Citation Award	2017

RELEVANT SKILLS

Wet lab experience in both genetics and microbiology
Experience with programs including R, Beagle, PHASE, fastPHASE, and Haploview
Presentation in both research and education-oriented seminars

MEMBERSHIPS

American Association of Physical Anthropologist 2014 – 2017
University of Louisville Anthropology Graduate Student Association
Lambda Alpha Association

REFERENCES

Dr. Christopher Tillquist
Department of Anthropology, Director of Graduate Studies
Professor
Phone: (502) 852-2422
Email: Christopher.tillquist@louisville.edu
Dr. Fabián Crespo
Department of Anthropology
Assistant Professor
Phone: (502) 852-2427
Email: facres01@louisville.edu

Dr. Geoff Bailey
REACH Program
Executive Director
Phone: (502) 852-8105
Email: gkbail01@louisville.edu

Mark Woolwine
REACH Program
Coordinator of Supplemental College Reading Programs and Student Success Seminars
Phone: (502) 852-2320
Email: mawool01@louisville.edu