RC25010 (W1006-042) June 11, 2010
Computer Science

IBM Research Report

Recombinomics: Population Genomics from a Recombination Perspective

Asif Javed, Laxmi Parida
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Recombinomics [*] : Population Genomics from a Recombination Perspective

Asif Javed
Computational Genomics
IBM T J Watson Research
Yorktown, USA
asijaved@us.ibm.com

Laxmi Parida
Computational Genomics
IBM T J Watson Research
Yorktown, USA
parida@us.ibm.com

ABSTRACT

As biotechnologies improve, coupled with falling costs, more and more genomic data becomes available fostering a renewed interest in understanding recombinational dynamics at unprecedented levels. In this paper, we survey the field for established as well as exploratory methods for handling the challenging task of untangling recombinations in observed data. These include ways to measure the effects of recombinations as well as infer the recombination events (crossover) themselves. We next track the progress made in model (or phylogeny) based approaches as they have the potential for ushering in the *recombinome*, the set of all the crossover points along the chromosome in a given organism.

General Terms

Theory

Keywords

genomics, recombinations, algorithms, phylogeny

1. INTRODUCTION

As next generation sequencing and other related biotechnologies, including robotic technologies, improve in leaps and bounds, more and more data become available, paving the way for interesting genome-wide analysis. The improvements in the last decade have made high throughput sequencing very inexpensive, increasing accessibility to the technologies which in turn lead to building of extensive data repositories [19]. Not only does the volume of data increase, but so does its resolution. i.e., the genomic data of not just a generic member of the species, but that of a specific member is becoming available at unprecedent levels. Thus the trend is towards understanding the variations/relationships *within* a species at increasing levels of detail.

The relevant genetic events that are responsible for these genomic variations can be broadly classified into two: (a) duplication and (b) genetic exchange events. Single nucleotide polymorphisms (SNP), short tandem repeats (STR) and copy number variations (CNV) are examples of effects of the genetic processes of the first category. Recombinations and gene exchange events are in the second category of genetic processes. Since the second category has the effect of shuffling the DNA material through generations, the detection of these events in a given set of extant units is quite challenging. However, almost the entire genome (at least more than 98%) of the organisms are subjected to these confounding shuffling genetic events. Thus there is little respite from these processes in any genome-wide analysis.

Any approach to understanding recombinations studies the footprints that past genetic events have left on the genome. Curiously enough, we must presuppose mutation events in the chromosome to study, understand or measure recombination events. As an extreme case, if there were no mutations (and no other such detectable duplication events), all recombiantions would go undiscovered for all practical purposes. In literature, mutations are assumed to occur under an *infinite sites model*. The simple implication is that no two mutations occur at the same site (assuming that the genome strand is of infinite length). This is a considerable simplification but this model is widely accepted, applicable to most of the genome (highly mutative sites like CpG islands being an exception). An important implication is that there are no back or parallel mutations: this implies that if a site mutates say from A to a, then across subsequent generations, only the descendants of this site (and none others) will display the mutated value of a. For this discussion we assume SNP datasets: as mentioned earlier SNPs are genetic variations of a single nucleotide base. SNPs are the most abundant and well studied form of genetic variation which have gained immense popularity in genetic studies due to the continuing improvements in the underlying biotechnology and decreasing costs. A further simplification is usually made: all SNP loci harbor only two allelic (biallelic) variations. Triallelic human SNPs are known but much less common [14].

Recombination processes play a vital role in shaping the genome. Although every genetic events actually occurs at the level on an individual, the focus of the studies is on the impact of the genetic recombination, as well as its detection, at a population level. Thus it is helpful to understand the processes in the context of the evolution of a population. For this, we observe the units (or members) at each generation and study the flow of genetic material from a unit in one generation to another in the next generation. Thus there is a need to characterize the populations. For simplicity in an ideal population the generations are non-overlapping and of constant size. Further the units display random mating and no selection (see Section 2.2). Such a population is called a Wright Fisher population. While these characteristics of a population may appear non-realistic at first glance, these assumptions are reasonable

[*] The term was coined by Jaume Bertranpetit.

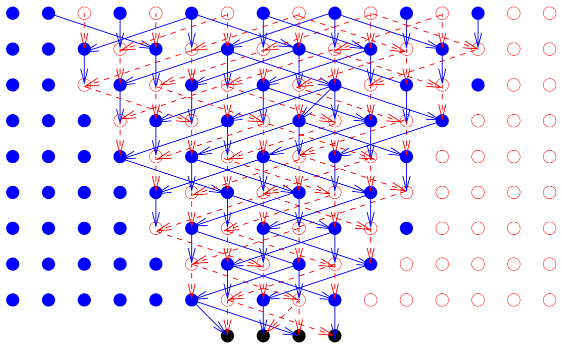

Figure 1: The solid (blue) dots represent one gender, say males and the hollow (red) dots represent the other gender (females) in a population. Each row is a generation with the direction on edges indicating the flow of the genetic material from one generation to another and the four extant units (solid black dots) are at the bottommost row. Only the genetic flow that contributes to the four extant units, in the last 10 generations, is shown here.

for the purposes of the study of the genetic variations at the population level. In fact, models with varying population size and/or overlapping generations can be reparameterized for an equivalent Wright-Fisher Model. Also, diploids (organisms with two copies of a chromosome) can be treated as haploids (one copy) for the purposes of studies at population levels (see texts such as [13, 21]). A simple example of genetic flow in populations across generations is shown in Fig 1 (taken from [32]). A recombination event occurs during the meiotic process, hence this is captured in this picture, when two parents give forth to an offspring. With even further simplification, we can assume that the genome of a unit is the result of the recombination of the two parent units. A mutation event may occur during the transmission of the genetic material from a parent to the offspring. Now given, say the SNP data of only the four extant units of Fig 1, the task is to unravel their recombinational history.

We categorize the various approaches to studying this genetic history into two: (1) recombination profiles and (2) recombination landscapes. In both the input is the SNP data of $n(> 1)$ distinct extant units. Clearly, nothing of interest can be inferred when $n = 1$. The first category studies the telltale impression of recombination events left on the n chromosomal segments as a recombination rate profile. Its central idea is based on the correlation of a pair (or more) of SNPs on the n units: this is quantified as *linkage disequilibrium*. This measure gives the relative extent of recombinations along the chromosomal segment, as can be inferred from the n units. This has been the traditional approach to studying recombinations in practice and still continues to hold the attention of researchers.

The second category of approaches, attempts to identify the individual recombination events. We call the sum total of each recombination event's location on the chromosomal segment (the *where*) as well as the participating ancestral lineages (the *who*) as the *recombinational landscape*. It is indeed the network of Fig 1: a relevant substructure is called the Ancestral Recombinations Graph (ARG) in literature. As an illustrative example, Fig 2 shows an ARG of Fig 1. However, an oddity that may not escape the astute reader is regarding the size of the ARG. In other words, does there exist a single common ancestor to all the extant units? In most cases there exists a common ancestor (termed the Grand Most Common Recent Ancestor or GMRCA). when this structure is studied as a mathematical object [11, 32]. It is unrealistic to reconstruct the fully resolved ARG and this has also been demonstrated through mathematical modelings [32]. However, any subset of recombinational landscape is of immense interest and this is currently an area of intense research.

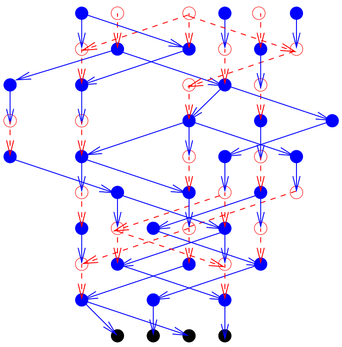

Figure 2: An ancestral recombinations graph (ARG) as a much less resolved graph of the entire ancestry information of Fig 1.

Towards recombinomics.

An important consequence of the quest for the recombinational landscape is a step in the direction of "El Dorado" of recombination (or crossover) events in the chromosome in a population. We mean by *recombinome* the total set of all the recombination events on the chromosome, possibly with the additional information about parental segments as well. The challenge however is in accurate inferencing, at the algorithmic level, of the crossovers.

2. EVOLUTIONARY PERSPECTIVE

Although most of the discussions here apply to all diploids, we focus our attention on humans. The human genome consists of twenty three pairs of chromosomes which contain (nearly) all of our genetic code, with one member of each pair inherited from each parent. Of these, twenty two pairs are *autosomes* and the members of the twenty third pair are *sex chromosomes*. The gender in humans is determined by the male Y chromosome. Thus a male has non-homologous XY pair and the female has homologous XX pair. Further, each paired chromosome, i.e., autosomes and X chromosomes in females, is a *recombination* of the parent's corresponding homologous pair. The male Y chromosome does not recombine outside two tiny pseudoautosomal regions, where some X-Y recombinations still occur. The pseudoautosomal regions are located at the tips of the chromosome and together comprise about 5% of the Y chromosome. The male X chromosome is inherited entirely from the female parent but is still a recombining chromosome. All together the haploid human genome comprises of more than 3 billion base pairs, more than 99% of which is recombining in males. The female nuclear genome is longer than the male, and is all recombining. The only exception in both males and females is the mitochondrial DNA, which is primarily all non-recombining.

Loosely speaking, recombination is nature's mechanism of defining the offspring genome, of fixed size as typical of the species, by shuffling that of both the parents. Genetically it is the process by which a strand of DNA breaks and then joins a different molecule. Without digressing any deeper into cell biology, it is interesting

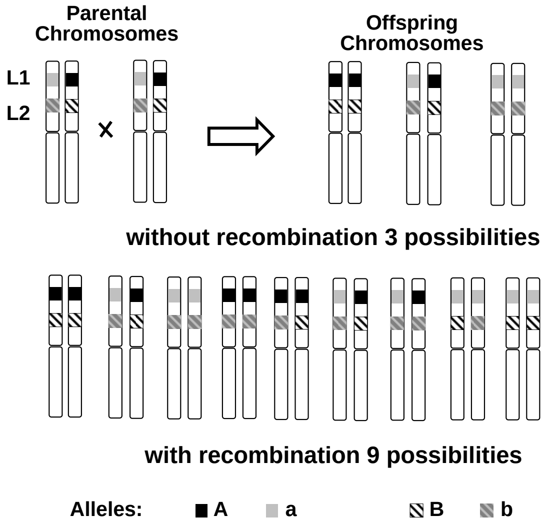
Parental Chromosomes
Offspring Chromosomes
L1
L2
x
without recombination 3 possibilities
with recombination 9 possibilities
Alleles: ■ A ■ a ▨ B ▨ b

Figure 3: Recombination increase diversity by allowing offsprings to inherit different allele combinations from those in their parents.

to note that the underlying biological processes are vital for the integrity of the genome (via their segregation during gamete formation). A breakdown in this process could lead to aneuploidy [1], the gamete gaining or losing one or more chromosome: this is often fatal at the embryonic stage.

2.1 Genetic diversity

Genetic diversity is important for the adaptability and survival of a population in response to new environment factors, as well as its resistance to threats such as new diseases. Recombination plays a key role in increasing the genetic variability within a population. Without this, the next generation would inherit a near replica of the previous generation chromosomes For example, consider Fig 3, L1 and L2 are two loci on the same chromosome. L1 harbors two alleles A and a; similarly L2 has B and b. In the parental chromosomes 'A' always co-occurs with 'B', and 'a' with 'b'. For simplicity, assume that this pattern is reflective of the entire population. Then, in the absence of recombination between these loci, the offsprings can only inherit the parental allelic combinations (either ab or AB) from either parent. Hence there are only three possible chromosome pairs (the order of chromosomes does not matter in a pair). *Genetic recombination leads to offsprings inheriting different allele combinations than those in their parents*. Thus it is vital to not only increasing the genetic diversity within a population but also help evolve the progenies faster to adapt to favorable combination of allelic variations.

2.2 Natural selection

An individual's genetic makeup plays an important role in determining his or her physiological response to various environmental factors and susceptibility to different diseases. Individuals winning the genetic lottery, and inheriting the favored allele, will tend to reproduce at a higher rate. The reasons could be epidemiological, resistance to an early onset disease; biological, improved fertility rate; or simply sublime, higher success in selecting a mate. But the population impact is that the favored allele will pass on to the next generation at a higher rate. Eventually, after many generations, it may achieve a complete sweep and the other allele at this locus will disappear.

A recently mutated allele under strong positive selection rapidly increases in frequency and may quickly reach fixation (frequency of 100% in the population). A frequent derived allele can be old and neutral, or young and selected. In the former case the increase in frequency can be attributed to genetic drift, and both the ancestral and derived allele will have similar haplotype background. In the later case, the haplotype around the selected derived allele would be passed on to the next generations at a faster rate than incidence of recombination. There would be a decrease in variability in the region. The increased homozygosity can thus be used as evidence of recent positive selection [37]. Nagylaki showed that unless a pair of loci are strongly linked or alleles are under strong selection, at an evolutionary level recombination prevails in destroying the linkage disequilibrium (see Section 3 for definition of linkage disequilibrium) [31].

Consider the case when there are two genetically neighboring loci undergoing selection with the respective favored alleles present in a disjoint set of haplotypes. In case of the previous example Fig 3 assume A and b are the advantageous alleles at the two loci. Under conditions of no or low recombination, these favored alleles would be competing against one other, until one prevails and achieves fixation, while the other despite its fitness advantage disappears from the population. This is known as the *Hill Robertson effect*. It allows deleterious alleles to hitchhike a neighboring favored allele and gain fixation in low recombination regions. In general, with a reasonable frequency of recombination between the the two loci, a combination of favored allele (in our example Ab) would appear in the population. This haplotype blend the selective advantage at both sites and hence has an advantage over all other combinations. Felsenstein was among the first to show through a simulation study the negative impact of Hill Robertson effect on the fitness of the population [8].

2.3 Genome Wide Aggregate Analysis

While the advantage of an evolutionary model based analysis is the obvious one of offering a biological explanation for the observed data, sometimes ignoring the annoying details of reality to simplify analysis can also be useful. The autosomal regions undergo crossing over in all generations, yet since the recombinant is still homologous to the two parental ones, it is not outrageous to ignore the genetic shuffling events.

At any one locus most of the variability stems from intra population differences. However, aggregating across multiple loci may and sometimes does reveal broader inter population structures [5]. Principal Component Analysis is a popular unsupervised dimensionality reduction method which can be used straight out of the box [35]. It summarizes the data by projecting it onto a lower dimensional subspace of maximal variability. The first few principal components have been shown to correlate with coalescence time between the underlying populations [25]. Another approach is to directly estimate the allele frequencies at every locus, in each hypothetical population, using Bayesian methods[6]. Both these methodologies ignore the finer role of recombination by painting the genome in broader strokes.

3. RECOMBINATION PROFILE

Linkage disequilibrium.

Non-random association between alleles at different loci within the same chromosome is termed as linkage disequilibrium (LD). A mutation at a site is perfectly correlated with alleles at neigh-

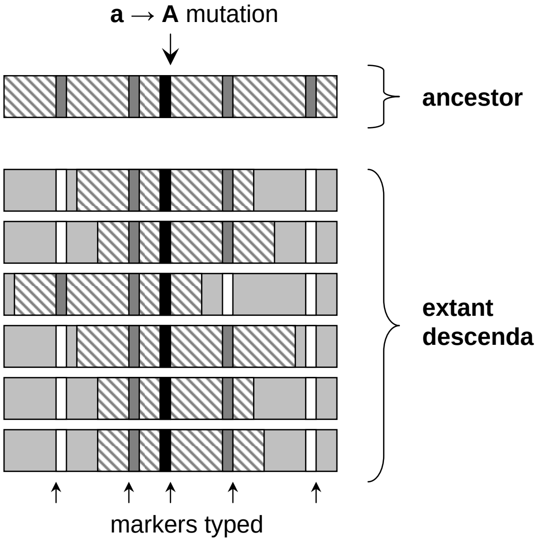

Figure 4: Recombination over successive generations break down the length of the haplotype inherited from the ancestral sequence, thereby introducing new alleles at the neighboring loci.

boring loci co-inherited from the same ancestor, until successive recombination over multiple generations introduce other alleles at these loci and reduce the correlation (see Fig 4). The speed of this breakdown will vary with the physical distance from the new allele. By the very nature of the dependence of the linkage disequilibrium value on allele frequencies, in general, *mutation increases while recombination decreases linkage disequilibrium*.

Linkage disequilibrium between a pair of loci can be quantified using various measures. The most basic definition computes deviation from independence.

$$D = p_{AB} - p_A p_B,$$

where p_{AB} is the probability of alleles A and B co-occurring on the same chromosome, and p_A and p_B are probabilities of occurrence of the respective allele. It can be shown with some algebraic manipulation that D decreases with each successive generation as

$$D(t+1) = (1-c)D(t),$$

where c is the recombination frequency between the two loci. Lewontin suggests normalization of D by its maximum absolute value [23]. Another commonly used statistic is r^2 which relates to D as

$$r^2 = \frac{D}{p_A p_a p_B p_b}.$$

r^2 has gained popularity because it relates to the Chi-square statistic as $\chi^2 = Nr^2$, where N is the number of samples used to estimate the probabilities. This helps compute the significance of the correlation and is very useful in studying indirect correlation while using one marker as a proxy for another in Genome Wide Association Studies (GWAS). In fact a key focus on linkage disequilibrium studies in the past decade can be attributed to the success and interest in GWAS. Linkage disequilibrium allows scientists to identify a representative subset of known genetic variation which can be efficiently genotyped using microarrays [4]. Without it, all known variants would have to be typed, making these association studies cost prohibitive.

Linkage disequilibrium carries evidence of the interplay of various evolutionary forces. The major factors impacting LD in any segment of the genome include mutation rate, recombination rate, natural selection and demography.

Again recent mutations tend to increase the linkage disequilibrium values while recombinations reduce it. Strong selection leaves its imprint on the haplotype surrounding the favored allele. But selection affects only a small number of loci. Changing in demography, on the other hand, impacts the whole genome. Individuals migrating from a genetically different gene pool, inject new alleles and haplotypes; thus increasing the diversity. The new alleles (and haplotypes) increase LD in a manner similar to a recent mutation. Newly founding populations, such as the initial migration of humans out of Africa, carry only a subsample of the source population haplotypes. Population bottlenecks such as natural disasters or epidemics act in a similar manner reducing the set of haplotypes. Drastic reduction in population size tends to increase LD [39].

3.1 Recombination rate estimates

The frequency of recombination varies across the genome [2]. Most recombination events occur within narrow regions (of about 1-2kbp) known as recombination hot spots. Although the numeric definition varies across studies, the general idea of a recombination hot spot is a narrow region where frequency of recombination is significantly higher than the surrounding haplotype. Certain sequence motifs have been identified which tend to co-occur at a higher rate within hot spots [29]. But the molecular mechanisms leading to meiotic crossovers is poorly understood. Hot spot positions and strength change at an evolutionary time scale. At a species level, no commonality has been found in their locations between human and chimpanzee genomes [44]. But how much of this shift happened in the more recent expansion of humans across the globe is open to debate. Graffelman et al found significant variability in recombination rate estimates among a diverse set of populations [10]. The differences tend to be lower among populations within the same continent, suggesting that variation in recombination frequency may be contributing to the diversity of the populations.

Despite their *hotness*, hot spots account for about 60% of meiotic recombination; the remaining 40% are spread across the genome. The rate of incidence of recombination across the genome can be inferred by (a) Pedigree analysis of familial data, (b) Statistical methods using population data and (c) Sperm typing.

3.1.1 Pedigree Analysis

The classical way to study co-inheritance of alleles is through pedigree analysis of familial data. Knowledge of parental and offspring genotypes allow inference of the switch in haplotype patterns to identify individual meiotic recombination events. However, unlike plant species, humans do not allow the liberty of mating appropriate genotypes at will, and the number of offsprings is severely limited. This limits the resolution of recombination placement. Furthermore statistical methods, presented in the next section, may be tweaked to incorporate available progeny information. Thus the use of pedigree analysis for fine scale human recombination rate estimates is very restricted.

3.1.2 Statistical estimates

Rapidly improving microarray technology has driven down the genotyping costs, making SNP datasets more abundant as well as popular. Many statistical methods have been developed to manipulate high density genome wide SNP datasets to estimate recombination rates at an unprecedented resolution. These coalescent based methods compute the population recombination rate between every consecutive pair of SNPs along a chromosome. Population recombination rate ρ related to per generation recombination rate c as

$$\rho = 4N_ec.$$

where N_e is the effective population size of a Wright Fisher population which exhibits the same genetic drift. Wright Fisher reproduction model is a basic population genetics model with simplistic assumptions including constant population size, non-overlapping generations, and random mating. Despite these unrealistic assumption, it generates a fairly representative structure of population sequences suitable for most population genetics problems. Methods for estimating ρ can be broadly binned into three categories [41].

Moment Estimators.

Moment estimators first quantify complex sequence variation using a few summary statistics. Pairwise difference between sequences[16], number of distinct haplotypes, and estimated minimum number of recombination [42] have proven to be useful candidates. The value of ρ which maximizes the likelihood of the statistic(s) is then estimated. The computationally expensive maximum likelihood step hence deals only with the summary statistics, allowing moment estimators to scale to a large number of haplotypes. However the ease of computation comes at the cost of high variance and strong reliance on coalescence model assumptions.

Full likelihood methods.

Full likelihood approaches on the other hand use all available data. They compute the likelihood of the observed data under an assumed coalescence model. Computationally intensive simulations are conducted to infer the probability space of the model parameters, including mutation and recombination rate, which best fit the observed data [7]. The computational cost of the numeric simulations restricts the scalability of these methods to even moderate sized datasets.

Approximate likelihood approaches.

Approximate likelihood methods tradeoff some of the high accuracy by approximating the parameter likelihood surface. This is done by either ignoring the low frequency less informative markers or taking a subset of markers at a time. Hudson conducted the seminal study to utilize this *composite likelihood* approach [17]. He first calculated the pairwise likelihood between all pairs of markers using simulations assuming an infinite site model. These likelihood are then multiplied to compute the composite likelihood. McVean et al. extended this idea by allowing recurrent mutations [26]. Li and Stephens introduced a distinctly different approach [24]. They study all loci simultaneously by approximating the conditional probability of haplotypes. Many other coalescence based methods have also been proposed which share the commonality of the coalescence, yet differ in model assumptions and/or the numeric method to estimate the likelihood. In general, Markov Chain Monte Carlo (MCMC) algorithms which churn their way to parameter estimates, and Importance Sampling methods which recurse over potential ancestral states, have proven useful. Recombination estimates computed by different methods, using the same data, tend to agree. But there is some discrepancy. All likelihood estimates are based on simulations. Theoretically results computed using the same method should always converge to the same value. Practically this does not always happen due to the limited number of iterations.

The statistical recombination rate estimates have gained immense popularity. They use readily available SNP data; and the approximate methods scale reasonably well to chromosome wide analysis. Despite the variability in estimates, they provide a useful tool to compare recombination rates across different genomic regions in the same population; and also to compare genetic variability across populations. Deviation from coalescence model helps identify *interesting* genomic regions, which may reflect complex demographic history or carry evidence of strong selection. But these estimates must be taken with a grain of salt; they have some well known limitations. Firstly, they are sex averaged. It is known that recombination occurs 1.6 times more frequently during female meiosis. Coalescence models ignore the role of sex in recombination. Secondly, population recombination rate does not translate directly to individual recombination rate. Effective population size N_e is reflective of a complex demographic history. Its estimate would vary across the genome, specially in admixed populations. Thirdly, the results of these methods is a statistical average which provides no information on individual recombination events. And finally, recombination rate estimates would suffer from any ascertainment bias present in the selection of SNPs.

3.1.3 Sperm Typing

Sperm typing is the gold standard in identifying meiotic recombination events. In principle the idea is very simple. First males sufficiently heterozygotic in markers in region of interest are identified. Then individual sperms from their semen sample are isolated and polymorphic markers typed in each sperm independently. The results are statistically analyzed to exclude potential contamination, and to compute average recombination rate across the samples. Since each sample contains more than three hundred million sperms, the resolution is potentially infinite. In practise however isolating and typing each sperm is an extremely laborious manual procedure. Studies using sperm typing have thus restricted to narrow regions of intense interest [3].

The results of sperm typing tend to agree with likelihood based estimates [20]. However linkage disequilibrium reflect the historic imprint of recombination on the genome and sperm typing represents the current recombination rate. If a recombination hot spot has evolved recently, its evidence will vary between the past and present, and hence there will be a discrepancy in the estimates. A key limitation of sperm typing is that it is obviously male specific.

4. RECOMBINATIONAL LANDSCAPE

Population recombination rate estimates provide an invaluable tool to analyze the comparative behavior of recombination across the genome. However they reveal little to no direct information on individual recombination events. Recall that we call the sum total of each recombination event's location on the chromosomal segment (the *where*) as well as the participating ancestral lineages (the *who*) as the recombinational landscape. It is difficult to infer an individual phylogeny based on a statistical population averages. Sperm typing does identify individual events, but the phylogeny of these events terminates in the petri dish.

In human phylogenetic studies recombination have generally been treated as a nuisance to be avoided because of the complexity it brings to the genetic sequences. Genetic crossovers bring together sequences with different phylogenetic histories. And the task of untwining these phylogenies is nontrivial. Thus classically the analysis of human phylogeny is done using non-recombining loci, Y chromosome and mitochondrial DNA. These loci undergo uniparental inheritance and harbor higher inter-population variance compared to the rest of the genome [22]. But Y chromosome consists of 58Mbp and mitochondrial DNA only 16kbp. Together they comprise about 2% of the haploid human genome. Can these loci alone paint the complete human phylogenetic landscape? Most autosomal population structure studies circumvent recombination by using a statistical aggregate of allele variations [6, 35]. Based on the role of mutation and recombination, phylogeny models can be divided into three categories.

4.1 Uniparental Model

The uniparental models consider mutations as the lone source of genetic variation. In the absence of recombination, relatedness of two sequences is estimated by their similarity. Since the sequences have evolved uniparentally the underlying phylogeny can be represented by a tree. The tree may be rooted, in which case the root of the tree represents the most recent common ancestor (MRCA) of the extant sequences being studied. Each leaf represents a unique extant sequence. The internal nodes may represent an extant sequence or a hypothetical ancestor.

Algorithms to infer uniparental phylogeny can be broadly classified as clustering based or search based methods. *Clustering* based methods recursively group sequences, a pair at a time. The criteria for selection of the pair varies. The simplest of these methods, unweighted pair-group method with arithmetic mean selects the most similar sequences in each iteration. *Search* based methods consider all possible trees that fit the data and select the one which best fits their optimization criteria; maximum parsimony and maxmimum likelihood are two popular criterion. Phylogeny estimation using non-recombining loci is a very well studied and developed field [15]. Despite the plethora of literature and variation in technical and theoretical details, there is a broad consensus on the results inferred by these methods.

4.2 Mosaic Model

Mosaic model assumes that the individuals being studied are sampled from a recently founded population. This population initially comprised of a small number of founding individuals and the founding event is so recent that no mutation has occurred in the loci being studied. Thus each extant sequence can be represented as a *mosaic* of the unknown founding sequences. The minimum mosaic problem is defined as, given a set of set of extant aligned sequences in a population, and the number of founders k_f, find the set of founders and the mosaic with minimal number of breakpoints [45]. The complexity class of the problem has not been resolved, but no known algorithm solves it in polynomial-time.

The mosaic model is the polar opposite of the uniparental model. It assumes no role of mutation, with recombination being the lone source of variability among the sequences. But mutations do occur at a variable rate across the human genome (2.3×10^{-8} per base per generation on average). Absence of mutation makes the minimum mosaic problem very restrictive, limiting it's application to shallow ancestry in low mutation regions of the genome only.

4.3 Ancestral Recombination Graph Model

Ancestral Recombination Graphs (ARGs) incorporate the best of both worlds. It allows for the interplay of mutation and recombination events [11]. Phylogeny is depicted by a network (Fig 5). The leaves of the network represent extant sequences and the internal nodes constitute hypothetical ancestors. The root of the ARG represents MRCA. The directed edges show the direction of flow of genetic material. There are three types of events represented in the network. Coalescence events occur, when going backwards in time, the phylogenies of two or more distinct sequences converge. Mutation events occur when genetic material, while being passed on to the descendants, undergo point mutation at one more loci. And recombination events occur when the inherited sequences is a concatenation of a prefix and suffix of the breakpoint, inherited from two different ancestors.

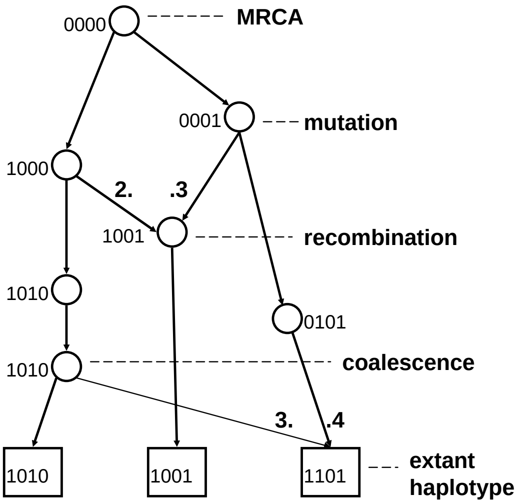

Figure 5: The root of the Ancestral Recombination Graph represents the Most Recent Common Ancestor. The indicated *mutation* passes the ancestral haplotype with a point mutation at the fourth locus. The breakpoint of the highlighted *recombination* is between the second and third loci. The haplotype is a concatenation of the prefix and suffix of the breakpoint inherited from the left and right parent respectively. The *coalescence* event combines phylogenies of its descendants.

If the true ARG is known recombination rates can be calculated by simply counting the recombination events in each interval. The role of different sequence motifs in triggering or facilitating recombination can be directly analyzed. Disease genes can be efficiently identified by tracing back the common ancestry of case samples to causal mutations.

5. INFERRING ANCESTRAL RECOMBINATION GRAPHS

All past recombination events do not leave a trace in the extant sequences. Recombination between near identical sequences can never be detected. Furthermore, newer recombination overwrite the imprint left by the past ones; making it impossible to reconstruct the *complete* true phylogeny [32]. However it is possible to infer a phylogeny consistent with the underlying sequence data under an assumed simplistic model of evolution.

5.1 Minimal ARGs

The minimal ARG problem is defined as, given a set of n samples genotyped at m loci, infer the ancestral recombination graph which minimizes the number of recombination. Wang et al showed the generalized problem to be NP complete under the infinite sites model [43]. They proposed an efficient solution to a restrictive version of the problem that allows for only node disjoint recombination cycles. Gusfield et al solved this limited problem in polynomial time [12]. Song and Hein solved the generalized problem by expressing the data as a sequence of trees along the chromosome [40]. They efficiently searched the subspace of all compatible trees using the so-called subtree prune and redraft operations to construct the ARG with minimal number of recombination. The algorithm takes super-exponential time and the published implementation can analyze up to 9 samples only.

5.2 Bounds on number of recombinations

Since the minimal ARG cannot be inferred for a reasonably sized data, practical methods have been developed to estimate a bound on the minimum number of recombination required. This helps gauge the efficacy of a heuristic solution. Evidence of recombination between any two loci can be gleaned, under the infinite site model, using the four gamete test (FGT). FGT simply states that given a pair of biallelic loci, all four allele combinations cannot occur without recombination.

Hudson and Kaplan developed the first method to compute a lower bound on minimal number of recombination [18]. Myers and Griffiths improved this bound by introducing a composite method which integrates local recombination bounds via dynamic programming [30]. The local bounds were computed as $R \geq H - S - 1$, where H and S are the number of unique haplotypes and loci respectively. Song et al improved the local lower bounds by mapping the problem to set cover and solving it via Integer Linear programming [40]. They further developed a heuristic to create compatible ARGs which attempt to minimize the number of recombination. The algorithm is repeated multiple times and the ARG with the minimum of number of recombination chosen. This provides an upper bound on the minimum number of recombination. If the two agree minimal ARG has been identified.

5.3 Plausible ARGs

Computing the exact minimal ARG is cost prohibitive. Heuristic solutions make random choices and may not always converge to the same solution. Even worse, the minimal ARG may not reflect the true phylogeny. Minichiello and Durbin addressed these limitations while identifying mendelian disease causal loci [28]. They proposed a statistical analysis of multiple plausible ARGs. They used a heuristic approach, with a preference for coalescence over recombination, to infer multiple ARGs compatible with the underlying data. Marginal trees are extracted for each ARG at each locus, and the correlation between descendants of each edge and the disease status computed. The maximum correlation at each locus is averaged over all ARGs and the statistical significance computed using a permutation test. The impact of the bias for coalescence is not analyzed. Furthermore the number of compatible trees for a reasonable sized data is very large and computational restrictions limit the number of sampled ARGs.

5.4 Approximate minimal ARG

Parida et al proposed a model IRiS which exploits SNP patterns to construct compatible phytogenetic networks [34]. The algorithm starts off with dividing the SNPs into blocks. Consecutive blocks with no evidence of recombination within them are merged into segments. The phylogeny in each segment can thus be represented by tree. The trees are then merged pairwise using a bottom up approach to construct a consensus network which is compatible with phylogenies in each underlying tree. The proposed DSR (dominant-subdominant-recombinant) algorithm is generalized to allow for networks containing both mutation and recombination events. The computational time is polynomial in the size of the input networks. Furthermore the algorithm guarantees that the number of recombination introduced during network merger are within an approximation factor ϵ; which is a well behaved function of the size and topologies of the input networks [33].

6. TOWARDS RECOMBINOMICS: IDENTIFYING PAST RECOMBINATIONS

The concept of recombinational junction was first introduced by Fisher [9]. He proposed that meiotic crossover between dissimilar haplotypes creates a unique junction. This junction may be inherited by the descendants of the individual and bears witness to their shared ancestry. Thus it can be used as a phylogeny marker. This novel concept has not been utilized in phylogenetics because of the computational difficulty in identifying past recombinations. Although many different methods have been proposed to detect recombination from DNA sequences [36]. Most of these methods are aimed at placing possible breakpoints or detecting single recombinant sequences. Furthermore these methods do not scale well to modern day data sets.

Mele et al used IRiS to detect historical recombination events by estimating local ARGs along the chromosome[27]. The results were aggregated using a sliding window approach, with varying block sizes, to identify placement and descendants of past recombination with high degree of confidence. Extensive simulations were conducted using COSI [38] to validate this methodology. Moreover, the inferred recombination placement correlated strongly with recombination rate estimates observed by sperm typing and estimated by statistical analysis. The recombination information for each sample was combined in a manner similar to point mutations in the so called *recotypes*. They showed that recotypes provide unique information on the shared history of populations. Since newer recombination are continuously overwriting the evidence of the past, older recombination are harder to detect; and sensitivity was higher for the recent past (last ten thousand years). Furthermore analysis of real data indicated that recombinational diversity estimated using recotypes is robust to ascertainment bias present in the underlying SNP selection.

7. CONCLUSION

The study of recombinations is an active field. There is a plethora of literature available on the wet-lab, computational and statistical analysis of recombination. Recombination bring together sequences with their own unique pasts and the task of extracting the unknown phylogeny from its trace in extant sequences is nontrivial. The computability of even simplistic models involving recombination is very challenging; which paves the way for approximate and heuristic solutions. The advent of cheap high density microarray SNP technology has improved our knowledge of the human genome tremendously. There is renewed interest in understanding recombinational dynamics at a much deeper level. However the size of these genomic databases is growing faster than Moore's law [19] challenging the limits of current methodologies. With the thousand dollar personalized genome on the near horizon, the availability and unprecedented resolution of this deluge of data will not only bring the scalability issue to the fore, but it will also improve our understanding of the relationships between sequence variability and recombination incidences. Nearly all recombination analysis challenges are still open problems on the lookout for better solutions. Nevertheless, the march towards the elusive *recombinome* still looks promising.

8. ACKNOWLEDGMENTS

We would like to thank Jaume Bertranpetit, Marta Melé, Marc Pybus and Francesc Calafell for many insightful discussions.

9. REFERENCES

[1] G. Coop and M. Przeworski. An evolutionary view of human recombination. *Nat. Rev. Genet.*, 8:23–34, Jan 2007.

[2] G. Coop, X. Wen, C. Ober, J. K. Pritchard, and M. Przeworski. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319:1395–1398, Mar 2008.

[3] X. F. Cui, H. H. Li, T. M. Goradia, K. Lange, H. H. Kazazian, D. Galas, and N. Arnheim. Single-sperm typing: determination of genetic distance between the G gamma-globin and parathyroid hormone loci by using the polymerase chain reaction and allele-specific oligomers. *Proc. Natl. Acad. Sci. U.S.A.*, 86:9389–9393, Dec 1989.

[4] P. I. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nat. Genet.*, 37:1217–1223, Nov 2005.

[5] A. W. Edwards. Human genetic diversity: Lewontin's fallacy. *Bioessays*, 25:798–801, Aug 2003.

[6] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, Aug 2003.

[7] P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, Nov 2001.

[8] J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78:737–756, Oct 1974.

[9] R. A. Fisher. A fuller theory of "Junctions" in inbreeding. *Heredity*, 8:187–197, Aug 1954.

[10] J. Graffelman, D. J. Balding, A. Gonzalez-Neira, and J. Bertranpetit. Variation in estimated recombination rates across human populations. *Hum. Genet.*, 122:301–310, Nov 2007.

[11] R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3:479–502, 1996.

[12] D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J Bioinform Comput Biol*, 2:173–213, Mar 2004.

[13] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford Press, 2005.

[14] A. Hodgkinson and A. Eyre-Walker. Human triallelic sites: evidence for a new mutational mechanism? *Genetics*, 184:233–241, Jan 2010.

[15] M. Holder and P. O. Lewis. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.*, 4:275–284, Apr 2003.

[16] R. R. Hudson. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.*, 50:245–250, Dec 1987.

[17] R. R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159:1805–1817, Dec 2001.

[18] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, Sep 1985.

[19] C. A. Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.*, 35:6227–6237, 2007.

[20] A. J. Jeffreys, R. Neumann, M. Panayi, S. Myers, and P. Donnelly. Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.*, 37:601–606, Jun 2005.

[21] M. Jobling, M. Hurles, and C. Tyler-Smith. *Human Evolutionary Genetics: Origins, Peoples and Disease*. Mathematical and Computaional Biology Series. Garland Publishing, 2004.

[22] M. A. Jobling, M. E. Hurles, and C. Tyler-Smith. *Human Evolutionary Genetics: Origins, Peoples and Disease*. Garland Science, 2003.

[23] R. C. Lewontin. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49:49–67, Jan 1964.

[24] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213–2233, Dec 2003.

[25] G. McVean. A genealogical interpretation of principal components analysis. *PLoS Genet.*, 5:e1000686, Oct 2009.

[26] G. McVean, P. Awadalla, and P. Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160:1231–1241, Mar 2002.

[27] M. Melé, A. Javed, M. Pybus, F. Calafell, L. Parida, and J. Bertranpetit. A new method to reconstruct recombination events at a genomic scale. *under submission*.

[28] M. J. Minichiello and R. Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.*, 79:910–922, Nov 2006.

[29] S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.*, 40:1124–1129, Sep 2008.

[30] S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163:375–394, Jan 2003.

[31] T. Nagylaki. Quasilinkage equilibrium and the evolution of two-locus systems. *Proc. Natl. Acad. Sci. U.S.A.*, 71:526–530, Feb 1974.

[32] L. Parida. Ancestral recombinations graph: A reconstructability perspective using random-graphs framework. 2009.

[33] L. Parida, A. Javed, M. Mele, F. Calafell, and J. Bertranpetit. Minimizing recombinations in consensus networks for phylogeographic studies. *BMC Bioinformatics*, 10 Suppl 1:S72, 2009.

[34] L. Parida, M. Mele, F. Calafell, and J. Bertranpetit. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. *J. Comput. Biol.*, 15:1133–1154, Nov 2008.

[35] P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.*, 3:1672–1686, Sep 2007.

[36] D. Posada and K. A. Crandall. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U.S.A.*, 98:13757–13762, Nov 2001.

[37] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837, Oct 2002.

[38] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, 15:1576–1583, Nov 2005.

[39] M. Slatkin. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, 9:477–485, Jun 2008.

[40] Y. S. Song and J. Hein. Constructing minimal ancestral recombination graphs. *J. Comput. Biol.*, 12:147–169, Mar 2005.

[41] M. P. Stumpf and G. A. McVean. Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.*, 4:959–968, Dec 2003.

[42] J. D. Wall. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.*, 17:156–163, Jan 2000.

[43] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *J. Comput. Biol.*, 8:69–78, 2001.

[44] W. Winckler, S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald, R. E. Bontrop, G. A. McVean, S. B. Gabriel, D. Reich, P. Donnelly, and D. Altshuler. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308:107–111, Apr 2005.

[45] Y. Wu and D. Gusfield. D.: Improved algorithms for inferring the minimum mosaic of a set of recombinants. In *Proc. CPM 2007*. Springer, 2007.