Growing the Family Tree: The Power of DNA in Reconstructing Family Relationships *

Luke A. D. Hutchison† Natalie M. Myres‡ Scott R. Woodward§

Sorenson Molecular Genealogy Foundation (www.smgf.org)

2511 South West Temple, Salt Lake City, Utah 84115, USA

Abstract

The Sorenson Molecular Genealogy Foundation is building the world's largest database of correlated genetic and genealogical information to enable genealogical research to be performed using DNA analysis techniques. DNA samples with associated 4-generation pedigree charts have so far been collected from approximately 40,000 volunteers. Up to 170 regions of DNA are currently analyzed for each individual, and the corresponding pedigree chart is extended as far as genealogical databases allow, to currently include over 700,000 ancestral records. By combining these two sets of correlated data on an unprecedented scale, we are enabling progress for the first time into the new field of "molecular genealogy."

Molecular genealogy is the application of DNA analysis techniques and statistical population genetics to the task of reconstructing unknown genealogies from the genetic and genealogical information of living individuals. We address aspects of using DNA for genealogical research, including those of identification and differentiation of populations (with population boundaries defined not just by factors of demographic separation, but also by time periods), differences in inheritance models of the various types of genetic data, clustering, statistical reconstruction of ancestral trees, inference of ancestral genetic signatures, and inference of surname based on paternal-line DNA.

1 Introduction

Every living individual carries within themselves a combination of the genetic signatures of their ancestors. This unique combination of signatures forms the individual's unique genetic identity, which is subsequently passed on to become a constituent part of succeeding generations. We are thus intrinsically linked to, and part of, our forebears and our descendants. Truly, in the words of Donne, "*No man is an island, entire of itself; every man is a piece of the continent, a part of the main . . . any man's death diminishes me, because I am involved in mankind.*" [1].

The vast majority of our DNA is identical to that of all others in the human race. It is this pattern, common to all human life, that identifies us as human. And yet, almost paradoxically, the small differences between genetic signatures give us identity as individuals. The number of genetic markers that differ between humans is disproportionately small compared to the total size of the human genome (the genetic "blueprint" of each human being). However, the total number of differences between any two humans is numerically large enough that each individual is unique among all other individuals who have ever lived. The regions of DNA that differ between individuals, known as *polymorphic sites*, give us a unique identity and a place in the human family tree.

Molecular (or genetic) genealogy is the application of DNA analysis techniques and statistical population genetics to the task of reconstructing unknown genealogies from the genetic and genealogical information of living individuals. The purpose of molecular genealogy is to supplement, not supplant, traditional techniques for genealogical research. The types of answers that may be provided by molecular genealogy include the derivation of populations of origin of unknown ancestors at genealogical "walls;" the reconstruction of ancestral *genotypes* (genetic signatures) from the genotypes of living descendants; the quantification of relatedness and possible kinship of two individuals; the inference of surnames on paternal lines in patronymic lineages; the investigation of possible non-paternities or adoptions; and, ultimately, the reconstruction of unknown pedigrees, or the tying of living individuals to specific previously-unknown ancestors.

The Sorenson Molecular Genealogy Foundation ("SMGF," `www.smgf.org`) is building the world's

*Originally published in: *Proceedings of the First Symposium on Bioinformatics and Biotechnology (BIOT-04, Colorado Springs)*, p. 42–49, Sept. 2004. Republished in: *Family History Technology Workshop*, BYU Provo, March 2005. © 2005 Sorenson Molecular Genealogy Foundation.

†Director of Bioinformatics, SMGF – `luke@smgf.org`

‡Co-Principal Investigator, SMGF – `natalie@smgf.org`

§Principal Investigator, SMGF – `scott@smgf.org`

largest database of correlated genetic and genealogical information, to enable genealogical research to be performed using DNA analysis techniques. Currently, DNA samples with corresponding 4-generation pedigree charts have been collected from approximately 40,000 volunteers. The DNA for each sample is analyzed at up to 170 locations across the genome, and the corresponding pedigree chart is extended as far as genealogical databases allow, to include over 700,000 ancestral records. The combination of these two types of correlated data on an unprecedented scale presents rich opportunities for analysis, and uncovers new, challenging problems by enabling the first real large-scale exploratory research into the field of molecular genealogy.

2 Types of Genetic Data

2.1 Sequence Data

DNA sequences are the most fundamental form of genetic information. The four nucleotides, abbreviated A, G, C and T, are the atomic units of a DNA sequence. Cells in the body contain four billion pairs of nucleotides (referred to as bases) that uniquely identify the individual, and that completely specify the structure and function of the entire organism. DNA sequences differ between individuals, predominantly because of the genetic processes of mutation and recombination. Algorithms exist for finding the "genetic distance" between the sequences of two individuals, or the number of edit operations (insertions, deletions and substitutions) needed to convert one sequence into the other [11]. While complete DNA sequence data can be used to derive all other genetic data, currently it is prohibitively expensive and time-consuming to obtain substantial sequence data for large numbers of individuals.

2.2 SNPs

Single Nucleotide Polymorphisms (SNPs, pronounced "snips") are single-base mutations in a DNA sequence where one base changes to another (Figure 1). These tend to be rare events (in some cases, unique events in the history of the human race), with mutation rates estimated at around 175 total SNP mutations per individual per generation, or .000002% per base per generation [8]. SNPs thus allow for the tracing of extremely deep-rooted pedigrees. SNPs are more useful for anthropological studies than genealogical studies because of their typically low mutation rate. Considering multiple SNPs together provides the ability to more accurately pinpoint the actual time of divergeance of two ancient lineages, and allows for non-unique-event SNPs to be identified.

The Y Chromosome Consortium [13] has identified a set of SNPs useful in classifying males into populations of origin. They present a *decision tree* for hierarchically classifying individuals into major *clades* or lineage forks, then into specific *haplogroups*, or subgroups of more closely-related individuals within each clade. The hierarchical designation appears to map reasonably closely to the demographics of known ancestral populations of tested individuals. It is worth noting though that these are paternal-lineage populations, because the SNPs used are all on the Y chromosome. Paternal-lineage populations have different properties than do traditional populations, as will be explained in Section 4.1.

2.3 STRs / Microsatellite Loci

A *short tandem repeat* (STR) or *microsatellite* locus is a region (or *locus*) of DNA in which a *repeat unit*, in the form of a specific sequence of bases, is repeated a number of times (Figure 2). The repeat region is amplified (copied millions of times) using *PCR* (Polymerase Chain Reaction), and is then genotyped to determine the number of repeat units at each locus for each DNA sample. The number of repeats, or *allele value*, at a particular marker or locus on a chromosome is passed down from parent(s) to child unchanged, unless there is a mutation, which will usually make the region longer or shorter by one complete repeat unit. STRs tend to have much higher mutation rates than SNPs (estimated at around 0.3% per locus per generation [7]), meaning they are much more useful on a genealogical timescale.

Of the different types of genetic data, the most cost-effective to obtain in large quantities is currently STR data. (Techniques for detection of SNPs and sequencing of DNA on a large scale are rapidly improving however.) Consequently, most current research in molecular genealogy primarily employs STR data.

3 Genetic Mechanisms Affecting Molecular Genealogy

3.1 Mutation Models and Mutation Rates

In general, genetic variation between generations results from the genetic processes of *recombination* and *mutation*, and may happen at the individual-base level, or may affect multiple bases (for example, in the case of entire STR repeat units being inserted or deleted).

While the occurrence of mutation is taken for granted, exact models that describe the mutation process are not known, particularly in the case of STRs. Some of this difficulty arises from the low probability of actually observing a mutation at a specific locus in any given generation, and from the size of the average generation gap in humans. Models have thus been proposed to approximate

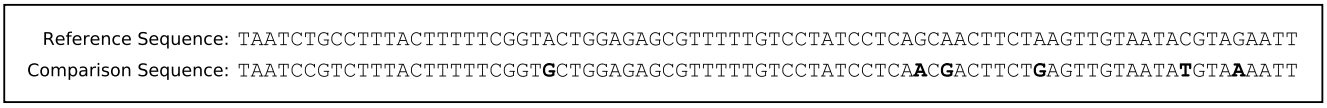

Figure 1: Single Nucleotide Polymorphisms (SNPs) are single-nucleotide mutations in the DNA sequence. Typically, SNPs have extremely low mutation rate probabilites and are therefore treated as single-event mutations, or analyzed together with other SNPs to detect IBS matches (Section 3.2).

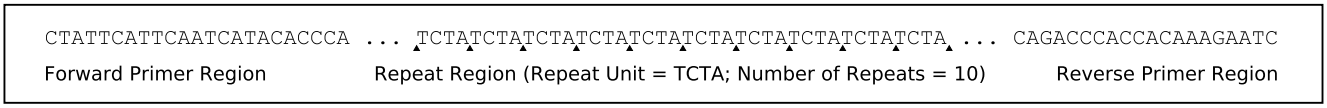

Figure 2: A Short Tandem Repeat (STR) locus is a region of DNA in which a *repeat unit* (here TCTA) is repeated a number of times. The repeat region is enclosed by two primer regions, which are used as start and end points for PCR amplification, which is the process of making millions of copies of the STR locus. The locus is then genotyped to determine the number of repeat units for the individual (the *allele* for the locus).

the actual behavior of a locus over many subsequent generations, including the *infinite alleles* and *stepwise mutation* models [6]. It is generally agreed that these models oversimplify the actual process of mutation, although they do provide useful tools for analysis of patterns of mutation under certain limited conditions.

Mutation rates have only been estimated approximately, and for small numbers of loci, due to these difficulties in actually observing mutations [5, 7], and because of the time and cost currently involved in determining the genotypes of large numbers of individuals. The rates which have been determined by observational studies appear to vary significantly between loci, meaning that a single average mutation rate cannot meaningfully be applied to all STR loci for most purposes.

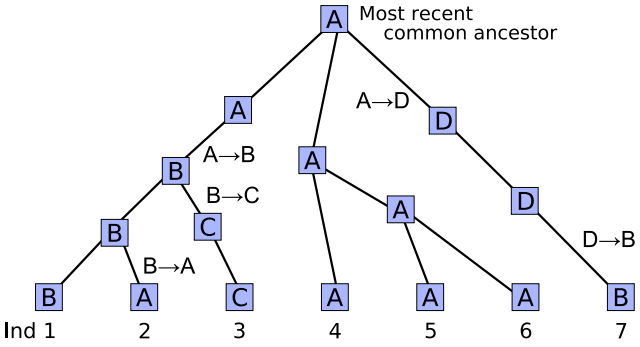

Figure 3: Mutation and back (or recurrent) mutation in haploid (non-paired) DNA. Letters represent genotypes; mutations are labeled on lines of descent. Individuals 4 and 5 are identical by descent (IBD), 4 and 7 are different by state (DBS), and 1 and 7 are identical by state (IBS).

3.2 IBD vs. IBS

If an allele (genetic marker value) at a specific locus is passed down from an ancestor to two descendants unchanged, it is said that the two descendants are *identical by descent* (IBD). If the two descendants mismatch due to one or more mutations, then the descendants are said to be *different by state* (DBS). However the two lineages may separately mutate away from the original allele, and then eventually randomly mutate again to a matching configuration. This is known as *identical by state* (IBS) (see Figure 3). IBD matches occur over relatively short timescales (as no mutation has been observed on either lineage); DBS mismatches typically occur over longer timescales; and IBS matches typically occur over much longer timescales (because multiple mutations are observed). IBS matches can be problematic, because if treated as IBD matches, they would imply a much shorter *time to most recent common ancestor* (TMRCA) than a true IBD match. Analyzing several loci together can help discern IBS matches, because if a large proportion of loci match between two individuals, it is much more likely that the matches are IBD than IBS.

The *infinite alleles model*, mentioned above, assumes that every mutation produces a new, globally-unique allele, disallowing IBS matches. This serves to simplify many mathematical analyses, but does not capture the reality that IBS matches occur a great deal in nature. IBS matches are particularly a problem when the mutation rates of loci under consideration are very different.

4 Genetic Inheritance Models

4.1 Y Chromosome (Ycs) DNA

The Y chromosome, possessed only by males, is passed down from father to son mostly unchanged. The majority of the Y chromosome is formed of non-recombining, *haploid* (non-paired) DNA, meaning the changes that arise in the Y

chromosome are primarily due to mutation. Typically, the Ycs markers that are used for molecular genealogy are STR loci in the non-recombining (NRY) region, with an average mutation rate of approximately 0.3% per locus per generation [7].

The inheritance model of the Y chromosome is immediately useful to genealogists, because it follows the same inheritance pattern as that of surnames in many western (and even non-western) societies. Thus, there is a correlation between observed Y chromosome genotypes and surnames. This is not a 1-to-1 correspondence, because of adoption, non-paternity, multiple origins for the same name, mutation, etc., but a fuzzy search against a database of surname-labeled Y chromosome genotypes nevertheless provides a useful way of finding possible family names beyond these events on paternal lineages. It also provides a means to identify others who share common *biological* ancestors on the paternal line where there was an unknown biological relationship, helping genealogical researchers who were unaware that they were biologically related to find each other. On a coarser scale, there is a correspondence betwen DNA patterns found in the Ycs and various world populations, which can allow researchers to trace the population of origin of a paternal-line ancestor.

It should be noted that the definition of *population* or *cluster* is somewhat unusual when dealing with the Y chromosome, because we are considering non-recombining paternal-line DNA. The characteristics of paternally-related populations are different from those of populations defined by recombining DNA (which produce the "traditional" definition of a population). For example, multiple unrelated lineages (*paternal populations*) can coexist in a common geographical location for an indefinite period of time without direct genetic interaction. Populations defined by non-recombining DNA are immune to traditional population-genetic forces that are caused by recombination, such as inbreeding. Paternal populations are also not affected by population growth effects in the same way as traditionally-defined populations, such as in the case of a historic contraction in population size. Population contraction affects a haploid population in the same way as a slow population expansion (possibly with genetic *drift* over time), followed by a rapid expansion. From the point of view of present-day genetic analysis, it is as if branches of the Ycs tree that did not make it through the population contraction never existed, because there is no further trace of the Ycs of specific paternal lines that ended at some point in history.

In this light, it makes sense to define a paternally-related population or cluster as a group of individuals who share a common paternal ancestor recently enough to match IBD at a significant number of loci, or as a group of lineages that descended from a common ancestor and whose living descendants are still genetically similar to their common ancestor. This definition of paternally-related populations necessarily impacts any inferences about paternal population structure that are made from the Ycs of living descendants. Paternally-related populations are specifically defined by common ancestry, and thus are only indirectly correlated with geographical origins (due to the physical location of the common ancestor).

Several methods for visualization of relatedness of entire populations have been developed. Many of these methods take a matrix of pairwise relatedness measures as input. Cladograms (branching tree-like diagrams) were extensively used in the past to visualize relatedness among individuals; they have been superceded by *median networks* (diagrams that may posess loops) and other visualization methods, due to the fact that cladograms yield results that are in general not theoretically sound (they do not capture the true nature of relatedness between individuals in different branches). Attempts are also often made to reconstruct the actual Ycs inheritance tree (or set of paternal lines descending from a common ancestor), using *phylogeny* (tree-building) algorithms such as those provided by PHYLIP [4]. These algorithms employ a form of heuristic random search of all possible lineage trees under specific phylogeny criteria, and yield an approximate solution. The best phylogeny for a dataset usually cannot be determined, because the total number of trees that may be reconstructed for a dataset of a given size scales exponentially with the number of individuals, quickly rendering the problem intractable (uncomputable) for moderately-sized datasets. Unfortunately, while the output of a phylogeny algorithm is generally accepted as authoritative, the search space is so large, and the pairwise-distance data often so internally inconsistent (due to IBS matches and limited numbers of actual loci in the genotypes) that different phylogeny algorithms and different runs of the same algorithm almost always give different results. For these reasons, phylogenies generated by current algorithms should be treated as informative but not authoritative.

4.2 Mitochondrial DNA (mtDNA)

Like the Y chromosome, mitochondrial DNA (mtDNA) is haploid (non-paired). It is present in the mitochondria, or energy-producing units of the cell, rather than in the nucleus. There are typically hundreds of mitochondria per cell, and multiple mtDNA molecules per mitochondria. The mother's mitochondria are those present in the zygote, or first cell of a new human being, and thus a mother passes her mtDNA to her sons and daughters. Her sons, however, will not pass their mtDNA to the next generation. Thus the mtDNA may be thought of as almost exclusively *maternal-line* DNA. Most of the observations made above for Ycs DNA and paternally-related populations are also

true for mtDNA and maternally-related populations, because mtDNA is essentially inherited along the maternal line.

Typically, mtDNA data obtained for genealogical purposes consists of SNPs from the mtDNA region known as the *D-loop*. Mitochondrial SNPs (as well as some Ycs SNPs) are often used to trace phylogenies on a deeper (anthropological) scale, because of their lower relative mutation rate compared to STRs in nuclear DNA, and because of the haploid nature of mtDNA.

4.3 Autosomal DNA

Autosomal DNA is the *diploid* (paired-chromosome) DNA that forms the vast majority of the DNA in most human cells. Each of the two alleles at a specific locus on an autosomal chromosome of one parent has a 50% chance of being passed on to each child, meaning that on average, a specific allele is passed on to half of the children. Additionally, autosomal DNA recombines, meaning that at each generation, sections of DNA are exchanged between pairs of corresponding (*homologous*) chromosomes received from the two parents.

If there is a low probability of a recombination between two or more loci, then they have a high probability of being inherited together by successive generations, and the loci are said to be in *linkage disequilibrium* (LD), or simply *linked*. Loci may be linked if they are located physically close together on the chromosome, or because there are few potential recombination sites between the loci. It is even possible to observe statistical correlation or linkage *by association* between distant sites, meaning that specific combinations of alleles at the loci occur together much more frequently than can be accounted for by chance. Groups of alleles on a single chromosome at loci that are in disequilibrium are called *haplotypes*. Haplotypes take a much greater range of possible combined forms than the individual loci they are comprised of, meaning they are more specific than individual loci, and are therefore more useful for genealogical purposes. A single STR locus may have ten possible allele values, meaning everybody in the population falls into one of ten categories, resulting in a moderately high chance of IBS match between two random individuals. A 3-locus haplotype, however, may have over a thousand possible configurations of alleles, resulting in an increase in specificity and a decrease in likelihood of IBS match.

When analyzing multi-locus genotypes, it is impossible to determine which chromosome of a pair a specific allele came from – the data is said to be *unphased*. The problem of determining which alleles at each locus of a set of linked diploid loci are physically located on the same chromosome is known as *haplotyping* or *determining phase* (Figure 4(a)). For example, for a set of three linked autosomal loci, we

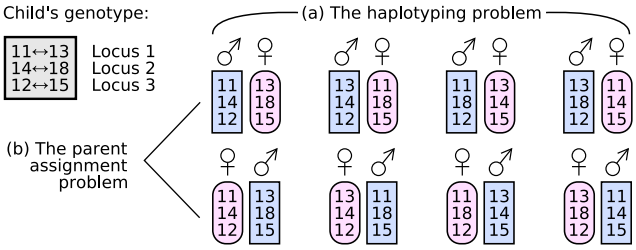

Figure 4: (a) The problem of *haplotyping*, or determining which alleles in a diploid genotype come from the same chromosome; (b) Determining which chromosome came from which parent.

have $3 \times 2 = 6$ alleles in the unphased genotype, yielding a maximum of $2^3 = 8$ possible assignments of alleles to specific chromosomes, or $2^{3-1} = 4$ possible phases when not distinguishing between the chromosomes. Depending on the allele values, some of these alignments or phases may be identical to each other, due to *homozygosity* (where the two alleles at a locus are identical).

If the genotypes of the parents are unknown, it is not possible to determine which allele in the child came from the father and which allele came from the mother. Additionally, only one of the two alleles at each locus are passed down from each parent to each child. When three or more siblings' genotypes are known, it may be possible to reconstruct the two parent genotypes unambiguously, but it is not possible to determine which genotype corresponds to which parent (the mother or father) without genotypes of extended families. Once phase is set in a child, the assignment of the two resulting haplotypes to the correct parent of origin is important, in order to be able to propagate haplotypes back through the genealogy, to infer ancestral types (Figure 4(b)).

Autosomal loci that are unphased and *unlinked* (and therefore not able to be haplotyped) are very difficult to trace genealogically without knowing the genotypes of large numbers of individuals in an extended family group, because a specific allele could have come from either parent at each generation, and only one of each parent's two alleles was passed on to each child. However, unlinked autosomal loci can still be used for genealogical purposes, by clustering together similar individuals, and then looking for patterns in the geographic origin of the known ancestors of the individuals that fell into the same cluster. Populations in general have a distribution of alleles that is distinct from that of other populations.

One algorithm that does a reasonable job of clustering individuals based on unlinked autosomal loci, known as STRUCTURE, uses a Markov Chain Monte Carlo (MCMC) algorithm to iteratively improve cluster membership probabilities until a reasonable solution is found [10]. STRUCTURE has been used to cluster many autosomal datasets.

4.4 X Chromosome (Xcs) DNA

The X chromosome has a very interesting inheritance model: because each male has an X-Y combination of sex chromosomes and each female has an X-X, males of necessity received their Y chromosome from their father, and received one of their mother's two X chromosomes. Females received one X from their mother and one from their father. This is useful in haplotyping the X chromosome in females – as the X chromosome is haploid in males and diploid in females, it is possible to always unambiguously set phase in the genotypes of any mother-son or father-daughter pair. By creating a dataset of phase-known females mixed with phase-unknown females, we can estimate how well any given haplotyping algorithm performs in haplotyping the Xcs in a large female population: the accuracy with which phase was determined for phase-known females gives an estimate of performance on phase-unknown data. This is a good model for estimating performance of haplotyping algorithms on autosomal data, since Xcs STRs are believed to have genetic properties in females that are similar to those of autosomal STR loci.

Haplotyping has so far proven to be a difficult problem, although several researchers have created tools that can successfully reconstruct a large proportion of haplotypes from a set of random simulated genotypes [2, 12]. Determining phase for autosomal loci is difficult when the relationship of individuals is not known, because analysis can only be performed on a population level. It is hard to check the validity of haplotyping results, because the haplotype phase was unknown to start with. In order to test the validity of phase reconstruction, haplotyping algorithms are typically tested with data that is simulated and therefore of known phase. In our experience, these algorithms do not work nearly as well as claimed when they are applied to real, phase-known data, such as the Xcs data we have obtained from known father/son and mother/daughter pairs in our database. Haplotyping algorithms can also be very slow to run. At SMGF, we have created a new haplotyping algorithm that sets phase in a population of haplotypes with an accuracy that is close to that of the current best algorithm, PHASE v2, yet runs several orders of magnitude faster. This algorithm will be described in a future publication. Our dataset of 220 phase-known individuals (combined with several thousand phase-unknown individuals), derived from real data, will also be of interest to those working on the haplotyping problem.

4.5 Comparison of Inheritance Patterns

It is interesting to compare the modes of inheritance of the various human chromosomes in the context of genealogical reconstruction. The chromosomal inheritance patterns have different characteristics depending on whether the inheritance is considered forwards or backwards through time.

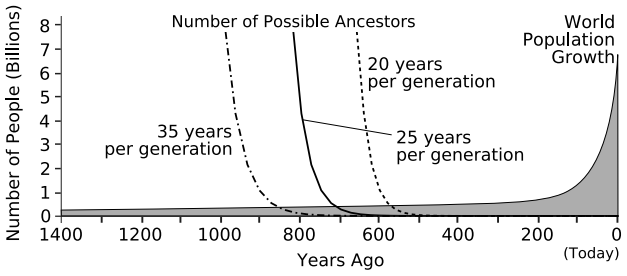

Figure 5: Total possible ancestors (2^n people at the nth generation back) compared to historic growth of world population. At some point in the very recent past, significant proportions of the world's population shared all ancestors. (After Jobling et al.)

The Y chromosome, for example, may be inherited by any number of sons at each generation, yielding a paternal tree relationship when viewed forwards through time. However, each son received his Y chromosome from exactly one father, yielding a single paternal lineage when viewed backwards through time. Mitochondrial DNA has very similar inheritance patterns on the maternal line, producing a maternal tree and maternal lineage if viewed forwards and backwards respectively, except that the "maternal tree" also has male leaf nodes (sons) connected to many of the female nodes in the tree.

Autosomal alleles follow a zig-zag pattern (single-path random walk) back through time, since they could have come from either parent at each generation. The number of possible ancestors that any given autosomal allele could have come from at the nth generation is 2^n. Interestingly, if one traces a pedigree chart far enough back, the same ancestor begins to appear on multiple branches of the pedigree: the pedigree actually *coalesces*. Even further back, the founding ancestors of the human race would appear on every branch of the pedigree – or, if the pedigree chart were drawn such that coalescing ancestors were drawn once, the chart would diverge and then converge again (this is effectively what is known in discrete mathematics as a *lattice*, a specific form of *partial ordering*). Also, at some recent point in human anthropological history, large proportions of those living today shared almost all of their ancestors [9] due to extreme coalescence of ancestral lines (Figure 5).

Looking forward through time, autosomal alleles are potentially inherited by multiple children at any generation, so a single allele follows a path that resembles a lightning bolt (i.e. the forward-inheritance mechanism is a branching random walk).

The X chromosome, however, has the most intriguing mode of inheritance. When viewed forward through time, each male may pass their X chromosome to zero or more females (and exactly zero males), and each female may pass their X chromosome to zero or more children (male

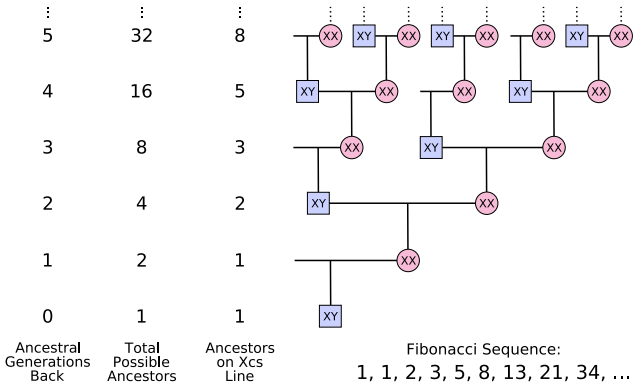

Figure 6: The number of ancestors at generation n from whom a living individual may have received an X chromosome allele is F_n, the nth term of the Fibonacci Sequence. The ratio of successive terms in the Fibonacci sequence converges on the Golden Ratio $\phi = 1.618$.

or female). Looking backward through time, the number of potential ancestors that could have been the source of any given allele on the X chromosome at generation n back grows as the sequence $F_n = 1, 1, 2, 3, 5, 8, 13, \ldots$. This will be familiar to many as the Fibonacci Sequence, whose ratio of successive terms converges upon the Golden Ratio $\phi = 1.618$ (Figure 6).

5 Importance of Genealogical Data

The true importance of the SMGF database for molecular genealogy lies in the genealogical data that accompanies each of the 40,000 genotypes, which currently totals over 700,000 ancestral records. The presence of comprehensive genealogical data, polished by qualified genealogists, for every DNA sample in the database, allows for an entire dimension of analysis that is not possible using the genetic data alone. The combination of genetic and genealogical data present in the SMGF database is unprecedented on this scale.

In particular, it is important that *identity linking* is performed as accurately and thorougly as possible. To statistically reconstruct genotypes of ancestors, we need to know the DNA of as many living descendants of that ancestor as possible. If an ancestor is present in the unlinked pedigrees of several different individuals, then the ancestor has several different identities in the database, and there is significantly less information available to infer the ancestral genotype. Conversely, the more correct identity links that are made for a common ancestor of living individuals, the stronger the inferences that can be made as to the ancestral genotype, since DNA from the ancestor is likely to have made its way to multiple living descendants at a higher relative frequency than that found at random in the population. Without these links, at least for autosomal DNA, each allele is equally likely to have come from any one of the ancestors at a specific generation. Thus the key to verifiable molecular genealogy, particularly for recombining DNA, lies in accurate identity linking. It is very likely that better linking technology would result in the identification of further identity links between many of the 700,000 ancestral records in the SMGF database.

This raises issues of data accuracy and datafield normalization – much genealogical data is incomplete, incorrect, or inconsistent between different sources. The error rate and incompleteness rate increases the further back the genealogy is traced. However, there is a significant percentage of available genealogy that is certainly correct; we minimize initial error as much as possible by employing proficient genealogists and by drawing from the best data sources, and then it is our goal to identify remaining inconsistencies in the genealogical data by consulting the DNA.

It is interesting that genetics can serve as a verification of genealogy, and vice versa. Genealogy can also serve as a prior for genetics-based pedigree reconstruction, with the effect of reducing the total problem space, and of detecting and correcting errors by providing informational redundancy (as with error correcting codes in data transmission).

6 Population Genetics

Statistical population genetics provides many important clues and analysis techniques to achieve the goals of molecular genealogy [3]. In particular, quantification of various genetic and population parameters (such as gene diversity, locus homogeneity, kinship coefficients, linkage disequilibrium measures and average time to most recent common ancestor) can aid in understanding a population's genetic history.

Much of statistical population genetics relies upon a simplification of population dynamics, known as *Hardy-Weinberg Equilibrium* (HWE). A population is in HWE if the population is infinitely large, there is no migration to or from the population, all members of the population reproduce, all mating is random, everyone produces the same number of offspring, and neither mutation nor natural selection occur. There are no real populations that ever (even approximately) satisfy these criteria, yet the criteria are required for legitimate application of many population-analytic formulae. In general, however, it is often too difficult to mathematically model the actual dynamics of a real population, so this simplified model is used.

Factors that can cause a population to violate HWE include mutation, gene migration, genetic drift (where the balance of genes in a population changes over time, particularly in small populations), nonrandom mating, population bottlenecks or founder effects, and natural selection. These

commonly occur in real human populations – in particular, the HWE requirement of random mating is violated due to the existence in any geographic region of numerous demographic barriers such as race, religion, language, and physical barriers such as mountain ranges and oceans. Migration rates have usually been low until recent history, but there have been many sudden large-scale migrations corresponding to events in world history. Thus, even an approximate 1-to-1 mapping between geographic populations and genetic populations may not exist. It is important to observe that a non-HWE population is defined in terms not only of the specific combination of genes present, but also the *time period* that is being considered (since a population changes over time).

Interestingly, it is exactly the differential between HWE and the actual dynamic history of a real population that exposes the intrinsic structure of interrelatedness of a population. Eventually, advances in analysis of these effects will allow for family histories to be reconstructed from descendants' DNA.

7 Conclusion

We have described the field of molecular genealogy, which is the process of using the combination of genetic and genealogical data to reconstruct the unknown genealogies of living individuals. The relationship of common genetic concepts to molecular genealogy was discussed. Progress has already been made in using genetics for genealogical purposes, particularly with the Y chromosome, which is immediately useful to genealogists because of the correlation of its inheritance mechanism to that of surnames in many societies. Current algorithms for approximate reconstruction of haploid (maternal or paternal) lineage trees were covered, as well as clustering of autosomal DNA to determine population membership. Haplotyping of autosomal and X chromosome loci has been shown as a mechanism to increase the specificity of genetic signatures. Issues of "identity linking" and data accuracy in genealogical data were addressed, in light of the importance of genealogical data to molecular genealogy. Relevant concepts from population genetics were covered. Overall, much progress has been made in developing the tools and concepts that are needed for molecular genealogy, and specific DNA analysis techniques for genealogical research exist today, such as the Y chromosome surname search. However, this field is still in its infancy, and much work still needs to be done to enable genealogists to supplement traditional genealogical research with genetic analysis techniques.

References

[1] J. Donne. Meditation XVII. Devotions Upon Emergent Occasions, 1624.

[2] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5):921–927, 1995.

[3] Falconer and Mackay. *Quantitative Genetics*. Prentice Hall, 1996.

[4] J. Felsenstein. Phylip – phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[5] E. Heyer, J. Puymirat, P. Dieltjes, E. Bakker, and P. de Knijff. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human Molecular Genetics*, 6(5):799–803, 1997.

[6] M. A. Jobling, M. Hurles, and C. Tyler-Smith. *Human Evolutionary Genetics*. Garland Science, 2004.

[7] M. Kayser, L. Roewer, M. Hedman, L. Henke, J. Henke, S. Brauer, C. Krüger, M. Krawczak, M. Nagy, T. Dobosz, R. Szibor, P. de Knijff, M. Stoneking, and A. Sajantila. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *American Journal of Human Genetics*, 66:1580–1588, 2000.

[8] M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, 2000.

[9] S. Ohno. The Malthusian parameter of ascents: What prevents the exponential increase of one's ancestors? *Proceedings of the National Academy of Sciences*, 93:15276–15278, December 1996.

[10] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[11] D. Sankoff and J. Kruskal. *Time Warps, String Edits and Macromolecules*. CSLI Publications, 1999.

[12] M. Stephens and P. Donnelly. A comparison of Bayesian methods for haplotype reconstruction from population genotype. *American Journal of Human Genetics*, 73:1162–1169, 2003.

[13] YCC (Y Chromosome Consortium). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Research*, 12(2):339–348, February 2002.