Large Scale DNA Variation as an Aid to Reconstruction of Extended Human Pedigrees

S.R. Woodward[1,3], N. Myres[3], J.B. Ekins[3], J.E. Ekins [3], K. Hadley [3], L. Hutchison [2], L. Layton [3], U. Perego [3], A. Sims [3], A. Nelson [3], M. Nelson [3], A. Welch [3], 1) Microbiology and Molecular Biology, Brigham Young University, Provo UT; 2) Computer Science, Brigham Young University, Provo UT; 3) Sorenson Molecular Genealogy Foundation, Salt Lake City, UT

Abstract

Reconstruction of pedigree structure from limited information is frequently required in studies of inherited disease. To aid in this process, and to develop a genealogical tool, we have developed a database of genetic variation based on autosomal, mitochondrial, and Y chromosome markers. We have sampled a total of 37,000 individuals of diverse ethnic origin, with an initial focus on European and Polynesian lineages verified by extensive written genealogies. From various subsets of these individuals, we have typed 14 X chromosome microsatellites, 120 autosomal microsatellites, and 24 Y chromosome microsatellites totaling more than 2.2 million loci tests. In addition, we have sequenced from a subset of these samples 850 base pairs from the mitochondrial D-loop. Autosomal markers were chosen in sets of 3 to 5 that are in tight disequilibrium and are distributed across the genome. The combination of genetic and genealogical data is maintained and managed in a mySQL database from which input files for analysis programs such as STRUCTURE and ARLEQUIN, are generated. In addition, other analysis programs developed by us, Y-TYPEFINDER and HAPLOTYPER, have been employed to generate modal haplotypes from the Y chromosome data and to reconstruct phase in the X chromosome dataset where haplotyping algorithms can be assessed. We have employed the STRUCTURE program and additional analysis with HAPLOTYPER and Y-TYPEFINDER to measure genetic relationships across living individuals, and correlate these relationships with historical genealogical records. These analyses indicate that in Polynesian populations, clear assignment of living individuals to distinct historical lineages is feasible over the range of 3 to 6 generations, and similar assignments may be possible for living individuals in European populations.

Assignments of Population Clusters

A subsample of the database representing 682 samples collected in the Pacific basin was clustered using STRUCTURE. This algorithm has been successful in identifying clusters at continental scales (Bamshad et. al. 2003). This sampling was used to test the clustering algorithm STRUCTURE to determine whether proximal geographic populations could be delineated using a modest number of microsatellite markers. This data set was clustered based solely on genetic data and compared with known genealogical data to group individuals into populations of origin within the relative near past (i.e. within the last 10 generations).

Individuals were genotyped at 58 autosomal STR loci. STRUCTURE was run at k=2 through 20 with no prior genealogical or population data. The maximum number of samples assigned to a population with inclusion scores of >= 0.80 was obtained with k=8. At this level of k and inclusion score, 446/682 = 0.654 of the individuals in the total dataset were included in a specific population. After samples were assigned to a specific cluster, population assignments were made by extracting birthplace information from the genealogical data collected at the time of sampling. Genealogies varied in depth from 4 to 9 generations.

STRUCTURE output was used to generate TULIP (http://www.tulip-software.org) diagrams as shown in figure 1. Eight specific clusters representing the k=8 STRUCTURE run is depicted. Specific population assignments to the clusters were made by inspection of the genealogies of the individuals in each cluster. Percentages of parentage of each individual were then calculated for inclusion in the specific population assignments. For example, in the population labeled China, 93% of the birthplaces of terminal ancestors in the subset cluster were in mainland China. In the Samoan sample, 98% of the individuals terminal ancestral birthplaces were in Samoa. In the Hawaii cluster, only 79% of the terminal ancestral birthplaces were in Hawaii. Assignments therefore reflect two major components, the depth of the genealogical record, (the genealogical record ends prior to the 'actual' ancestral home of the individual) and the amount of admixture that is present in the Hawaiian sample (collection location).

This dataset demonstrates the utility of STRUCTURE to cluster individuals from both diverse and relatively closely related populations (i.e. Samoa and Tonga) into genealogically meaningful groups. This is an initial step in the reconstruction of shallow genealogies, a primary goal of this database.

Figure 1

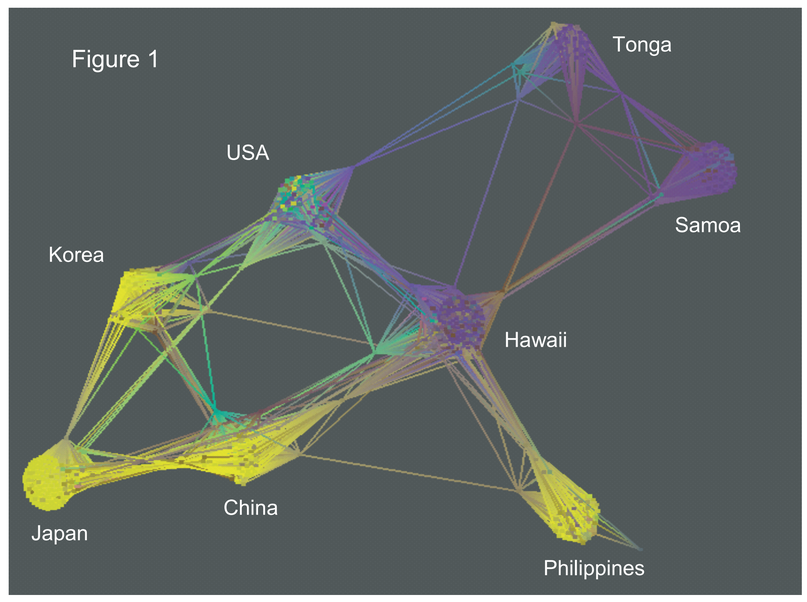
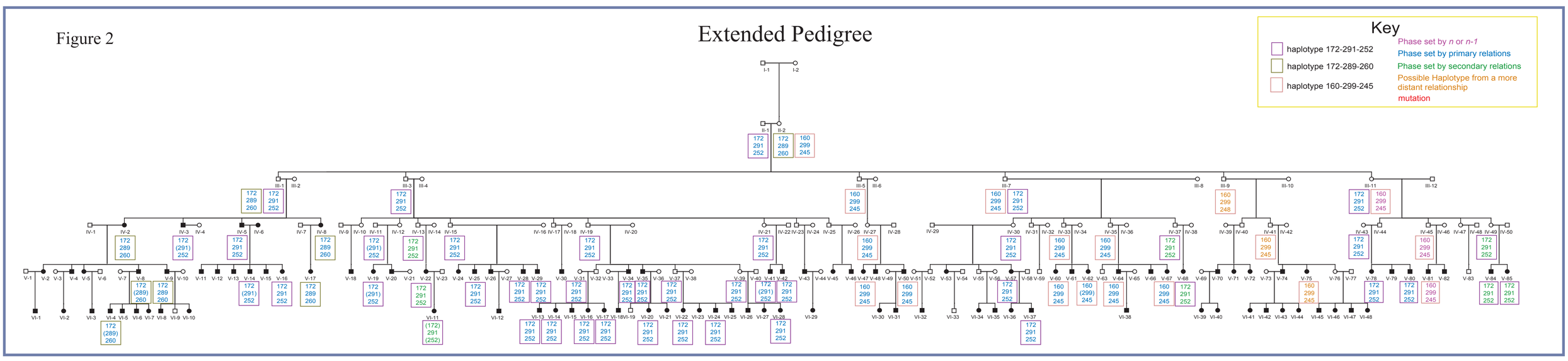


Figure 2

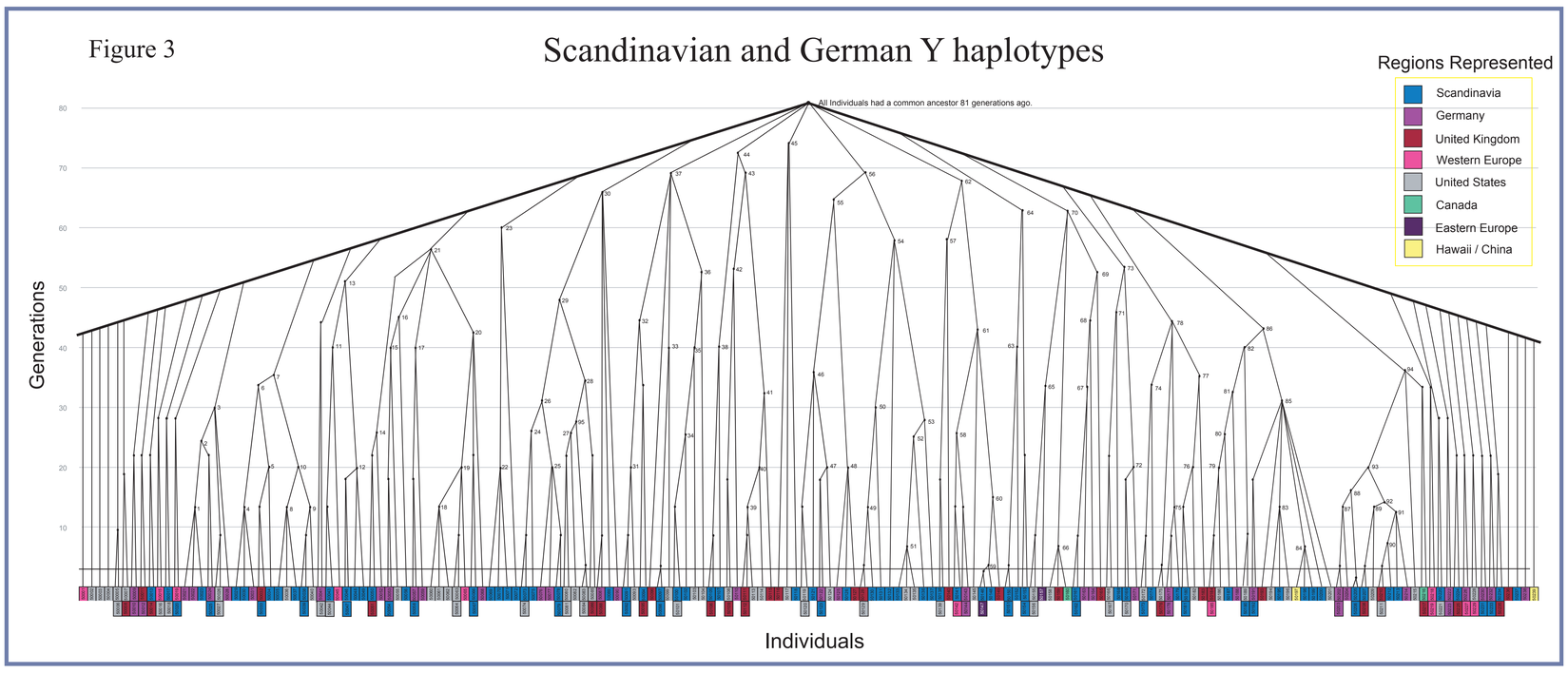


Reconstruction of Ancestral Haplotypes

Central to the reconstruction and verification of recent genealogies is the ability to accurately generate genotypes and haplotypes in probands, and to propagate these haplotypes to ancestors in a pedigree. A preliminary investigation of this process is depicted in figure 2. This pedigree represents six generations of a family test case. The proximal ancestors have six offspring. It is expected that greater than 98% of their genomes are transmitted to the next generation (Bickeboller and Thompson, 1999). The reconstruction of an ancestral path of a chromosome segment requires a set of markers that have very specific characteristics. For the reconstruction of shallow genealogies, highly polymorphic markers with measurable mutation rates were compiled. For this purpose microsatellite STRs are particularly suited for the task. In addition, multiple individual STRs that are in close physical proximity have been chosen. These sets of closely situated markers, expected to be in linkage disequilibrium in extended nuclear families but rapidly approach equilibrium in large populations, were used to construct haplotypes in the nuclear families. In this example three STR loci on chromosome 2 are demonstrated. These loci D2S1326, D2S2304 and D2S127 are grouped closely at the physical map position of 149.89 on chromosome 2. Levels of disequilibrium were measured using the methods of Slatkin 1994, Slatkin and Excoffier 1996, and Lewontin and Kojima 1960 as implemented in ARLEQUIN. Significant levels of disequilibrium between all three markers were observed when measured within five different extended family sets (543 total individuals). However, when the same markers were evaluated in a set of 1095 unrelated individuals from the same general population as the family sets (primarily UK and European immigrants to the US), there was no significant (p=0.05) disequilibrium. This is consistent with expected levels of recombination that would establish equilibrium in panmixic populations but would allow the maintenance of disequilibrium in closely related individuals.

The following criteria were employed to set the phase of known genotypes for individuals and to infer the genetic composition of ancestors with unknown genotypes within the pedigree. A score was given to each haplotype reflecting the strength of the inferred phase. When *n* equals the number of loci in a haplogroup, the phase of a genotype was considered unambiguous when an individual was homozygous at n or *n-1* loci. A score of 1 was given to haplotypes set from *n* or *n-1* homozygous genotypes. A score of 2 was assigned to haplotypes that were inferred by the rules of Mendelian inheritance when known genotypes were available from primary relatives, including parents and siblings. A score of 3 was given to haplotypes determined by inference from known genotypes of secondary relatives, including grandparents, aunts, uncles, and first cousins. A score of 4 was given to haplotypes that were inferred from relationships more distant than secondary relatives. Once the probabilistic phase of genotypes was determined and ranked for individuals, the haplotypes were propagated back to ancestors by correlating haplotypes from primary and secondary relatives at each generation. This process was performed at each generation until haplotypes were propagated back to the patriarch or matriarch of interest or until inference could not be accomplished with the haplotypes present. Although haplotypes of all scores were used, preference was given to haplotypes of lower score in reconstructing the genetic composition of ancestors. We have identified approximately 300 STR markers arranged in 100 haplogroups that meet the criteria outlined above for use in the reconstruction of ancestral haplotypes. These reconstructions will be used to verify, link and extend existing genealogical family trees.

Figure 3



Y Haplotype Clusters

There have been no previous attempts to reconstruct recent genetic histories on a continental scale using the Y chromosome. Here we define a genetic pedigree relating individuals of unknown paternal relationship. We implemented Structure v2.0 to cluster individuals into 'biologically significant' groups (Falush et al. 2003). The most frequently occurring allele at each locus was used to define a modal type (the assumed ancestral type) for each external cluster. The time to the most recent common ancestor was calculated from the modal type for each cluster (Stumpf and Goldstein 2001). A pair-wise comparison of haplotypes was considered between individuals within external clusters to determine relationships with the least mismatches between haplotypes and the time to most recent common ancestor between pairs (Walsh 2001). Deeper genetic relationships were reconstructed by using the modal types as input for analysis into Structure v2.0 (Falush et al. 2003). The pedigree, shown in figure 3, is a subset of our worldwide data set. It is comprised of 239 individuals whose known paternal ancestry terminates predominately in Scandinavia and Germany, and demonstrates that the Y chromosome can be used to reconstruct lineage based clusters with genetic relationships between unknown individuals on a genealogically relevant time scale.

Conclusion

The purpose of the Molecular Genealogy Research Project is to create a large correlated genetic and genealogical database for use as a genealogical tool. We propose to use this data to: 1) Verify the accuracy of documented genealogies using DNA marker technology. 2) Reconstruct genealogies on a personal, lineage and population level. 3) These genealogies will be of varying depths and completeness. The rationale for this project include: 1)Personal interest or curiosity. 2)Legal matters including inheritance and probate, paternity, adoptions, missing children 'kidnappings'. 3)Inherited disease and medical concerns and 4) Elucidation of genetic mechanisms, disease and other phenotypes in populations and in specific lineages within populations.

The first steps toward this goal is demonstrated by the ability to reconstruct population origins, specific Y-chromosome lineages and unique lineage specific autosomal haplotypes identifying ancestral individuals.

The current version of the database contains data on approximately 40,000 individuals, it is anticipated that the first operating version of the database will include genetic and genealogical profiles from approximately 100,000 individuals.

References

Bamshad M., S. Wooding, W. S. Watkins, C. Ostler, M. Batzer, L. Jorde. 2003. Human Population genetic structure and inference of group membership. Am. J. Hum. Genet. 72:578-589.

Bickeboller, H. and E.A. Thompson. 1999. The probability distribution of the amount of an individuals genome surviving to the following generation. Genetics 143:1043-1049.

Falush D., M. Stephens, J. Prichard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

Lewontin, R. C., K. Kojima. 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14:450-472.

Slatkin M., 1994. Linkage disequilibrium in growing populations. Genetics 137:331-336.

Slatkin M., l. Excoffier. 1996. Testing for linkage disequilibrium in genotypic data using the EM algorithim. Heridity 76:377-383.

Stumpf M., D. Goldstein. 2001. Genealogical and evolutionary inference with the human Y chromosome. Science 291: 1738-1742.

Walsh B. 2001. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics 158: 897-912.

ASHG 2003