

# Characterization of family-specific LD patterns in Sorenson Molecular Genealogy Foundation dataset

K. L. Pennington<sup>1\*</sup>, K. H. Ritchie<sup>1</sup>, J. E. Ekins<sup>1</sup>, J. L. Peterson<sup>1</sup>, N. M. Myres<sup>1</sup>, J. B. Ekins<sup>1</sup>, L. A. D. Hutchison<sup>1,2</sup>, L. Layton<sup>1</sup>, M. Lunt<sup>1</sup>, S. Masek<sup>1</sup>, A. A. Nelson<sup>1</sup>, M. E. Nelson<sup>1</sup>, U. A. Perego<sup>1</sup>, T.A. Tolley<sup>1</sup>, S.R. Woodward<sup>1</sup> 1) Sorenson Molecular Genealogy Foundation (SMGF), Salt Lake City, UT, USA; 2) Computational Biology Research Group, CSAIL, Massachusetts Institute of Technology, MA, USA

## Abstract

Inference of familial relationships is a useful tool in a variety of contexts, such as disease linkage and association studies, forensic applications, and inference of personal histories. To date, Y-chromosomal and mtDNA testing have been used extensively for these purposes. However, little has been done using autosomal DNA due to the complex inheritance pattern involved. Herein characterize family-specific LD patterns and haplotype frequencies as a tool for inferring family group assignments of unknown individuals through relationship probabilities. Six extended family groups ranging from 28 to 142 samples and one dataset of 1095 randomly selected samples were analyzed for significant linkage disequilibrium (LD) within twelve autosomal 3-locus STR haplogroups. Additionally, one locus was chosen from each of these haplogroups and pairwise LD calculations were run among these unlinked loci. Each of the families showed a unique LD pattern with significantly more LD both within and between haplogroups than did the random dataset. Haplogroups and unlinked locus pairs (HLPs) demonstrating significant LD were further analyzed to identify family-selective haplotypes and allele pairs (HAPs). Several HAPs were found within each family set with high frequencies characteristic of that particular family group relative to the other datasets. In the future, these HAPs may be used to assign probabilities of membership of an unknown individual into these family groups.

## Datasets

We compiled eight datasets consisting of one 1095-member random set and six family datasets collected from the same region as the Random dataset, ranging from 28 to 172 members. The eight family datasets include: KBI, KBII, LM, LN, SMI, SMII, and SN. See Table 1 for information on size and depths of each family. The datasets reflect a multigenerational sampling of related individuals with a common ancestor born in the year indicated in Table 1. Some familial overlap exists between two of the families sampled, and thus 17 individuals are found in each of two different sets. KBII and SMI contain these overlapping individuals. KBI and SMII are the same sets, respectively, but differ by having the 17 overlapping individuals removed. The use of these four closely related datasets will suggest the power of our method to distinguish between similar family groupings. To analyze this data, twelve autosomal three-locus haplogroups were chosen as seen in Table 2. These loci were chosen for proximity as well as completeness in our database. Calculations and comparisons were performed within these twelve haplogroups and also among representative loci from each haplogroup, as indicated in the table.

## Linkage Disequilibrium

Alleles in close physical proximity on a chromosome tend to be inherited as a bundle due to the rarity of recombination between the loci and are said to be in linkage disequilibrium.<sup>1</sup> Many factors influence the observation of linkage disequilibrium in a population, including population stratification,<sup>2</sup> admixture, and non-random mating.<sup>3</sup> In a randomly mating population, physically unlinked loci will sort independently, have haplotype frequencies equal to the product of allele frequencies, and will exhibit no significant linkage disequilibrium.<sup>4</sup> In our Random dataset, the levels of LD among both physically linked and unlinked loci gives a measure of the background levels within the study population. In the family datasets, the excess of LD among these loci is further indicative of their commonly shared ancestry and level of relatedness. Within physically linked haplogroups, the difference in levels of LD among the Random and the family datasets is significant for most families at the 0.01 confidence level. However, the high levels of LD in all datasets makes differentiation based on these patterns difficult. By also using the physically unlinked allele pairs, differences in LD patterns are more pronounced. Also, the absence of physical linkage eliminates proximity<sup>5</sup> as a cause of co-inheritance and therefore the observed levels of LD in random datasets represent the non-ideality of the study population and gives a measure of the background levels of LD in that population. Also, the increased levels of LD found in the family datasets represents the effects of relatedness within the datasets.

We ran calculations within the twelve 3-locus haplogroups and between the sixty-six unlinked allele pairs to determine the presence or absence of significant linkage disequilibrium at the 0.05 significance level. Linkage disequilibrium calculations were run using Arlequin<sup>6</sup> Software. The presence or absence of significant linkage disequilibrium is summarized in Figures 1-2 and Table 3. Figure 1 shows results for the physically linked haplogroup data. Figure 2 shows the results for the physically unlinked data. Table 3 shows the num odds ratios for the amount of LD in the families versus the random dataset. The z-test was used to test for the significance of the increase in the number of locus pairs exhibiting significant LD levels within each family dataset compared to that in the random dataset, using equivalence as the null hypothesis. All but one of the families exhibited significant deviation from Random at the 0.01 significance level for the haplogroups. All of the families exhibited significantly increased LD for the physically unlinked locus pairs.

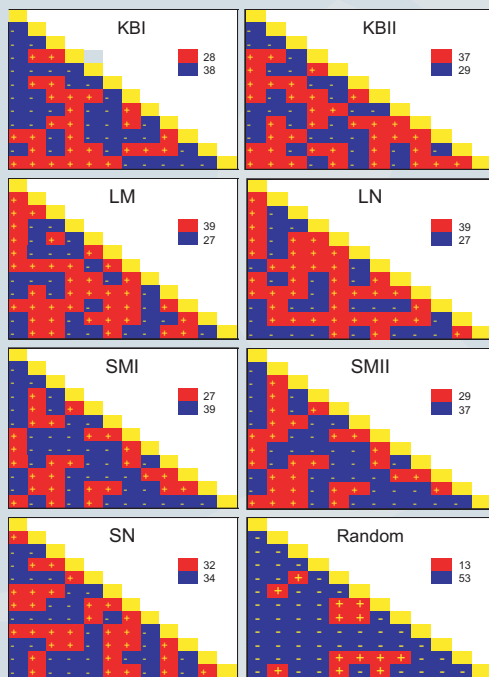


Figure 2. LD patterns among physically unlinked loci. Calculated and represented as in Figure 1.

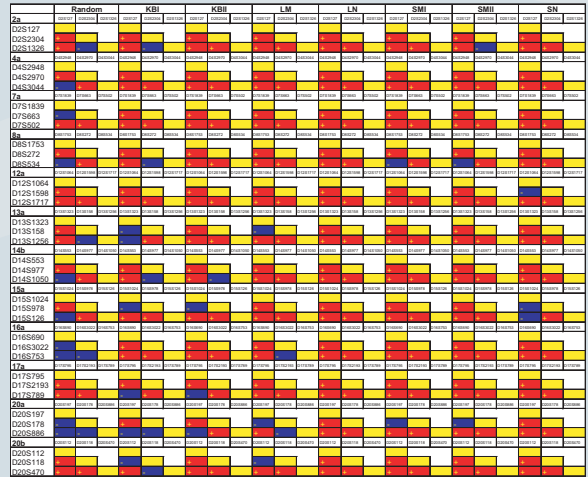


Figure 1. LD patterns within haplogroups. The presence or absence of significant linkage disequilibrium is indicated as a + or -, respectively. Linkage disequilibrium calculations were run using Arlequin<sup>6</sup> Software, with 10000 permutations and 10 initial conditions. Other parameters include: deletion weight = 1; transition weight = 1; transversion weight = 1; value = 1e-07; significant digits for output = 5; allowed level of missing data = 0.05.

Table 3. Odds ratios for total LD.

|        | Haplogroup |    | Interhaplogroup |    | OR = P(+JF) / P(+JR) |       | P (p <sub>2</sub> > p <sub>1</sub> ) |        |
|--------|------------|----|-----------------|----|----------------------|-------|--------------------------------------|--------|
|        | +          | -  | +               | -  | Hap                  | Inter | Hap                                  | Inter  |
| Random | 23         | 13 | 36              | 13 | 53                   | 66    | 1                                    | 1      |
| KBI    | 25         | 11 | 36              | 28 | 38                   | 66    | 1.087                                | 2.1538 |
| KBII   | 32         | 4  | 36              | 37 | 29                   | 66    | 1.3913                               | 2.8462 |
| LM     | 31         | 5  | 36              | 39 | 27                   | 66    | 1.3478                               | 3      |
| LN     | 36         | 0  | 36              | 39 | 27                   | 66    | 1.5852                               | 3      |
| SMI    | 34         | 2  | 36              | 37 | 39                   | 66    | 1.4783                               | 2.0769 |
| SMII   | 33         | 3  | 36              | 29 | 37                   | 66    | 1.4348                               | 2.2308 |
| SN     | 33         | 3  | 36              | 32 | 34                   | 66    | 1.4348                               | 2.4615 |

Table 1. Family datasets.

| Family | # Indiv | MRCAs born |
|--------|---------|------------|
| KBI    | 28      | 1801       |
| KBII   | 45      | 1801       |
| LM     | 142     | 1678       |
| LN     | 104     | 1818       |
| SMI    | 113     | 1744       |
| SMII   | 96      | 1744       |
| SN     | 61      | 1814       |

| Hap | Locus 1  | Locus 2  | Locus 3  | Location   |
|-----|----------|----------|----------|------------|
| 2a  | D2S2304  | D2S127   | D2S1326  | A02_149.89 |
| 4a  | D4S2948  | D4S2970  | D4S3044  | A04_38.77  |
| 7a  | D7S1839  | D7S1902  | D7S463   | A07_78.65  |
| 8a  | D8S272   | D8S534   | D8S1753  | A08_154.02 |
| 12a | D12S1064 | D12S1598 | D12S1717 | A12_95.03  |
| 13a | D13S1323 | D13S158  | D13S1256 | A13_84.87  |
| 14b | D14S553  | D14S977  | D14S1050 | A14_107.13 |
| 15a | D15S1024 | D15S978  | D15S126  | A15_45.62  |
| 16a | D16S690  | D16S3022 | D16S753  | A16_57.79  |
| 17a | D17S789  | D17S795  | D17S2193 | A17_89.32  |
| 20a | D20S197  | D20S178  | D20S886  | A20_66.16  |
| 20b | D20S112  | D20S118  | D20S470  | A20_39.25  |

\*Indicates locus used for inter-haplogroup analyses

## Haplotype and Allele Pair Frequency Comparisons

Another result of relatedness among a group of individuals is that they will tend to share similar alleles due to their common ancestry. The coinheritance of alleles that created the increase of linkage disequilibrium in the family datasets can also be seen by looking at haplotypes and allele pairs shared within the families at greater than random frequencies. These family-selective haplotypes may later be used to assist in probabilistically assigning an individual of unknown ancestry to a family group within the database. We calculated allele frequencies, haplotype and allele pair (HAP) frequencies, and 'expected' HAP frequencies for each of the datasets studied.<sup>7</sup> The expected HAP frequencies were calculated as the product of the allele frequencies in the Random dataset. Arlequin<sup>6</sup> Software was used to produce an EM estimate of HAP frequencies in all datasets. We compared HAP frequencies generated by Arlequin<sup>6</sup> in family datasets to the expected frequencies calculated from the random dataset. Odds ratios were calculated and the HAPs in the families with familial frequencies of at least 0.01 and > 100x expected frequencies are shown in the attachment. These frequencies were compared among the families and the random dataset to identify any family-selective haplotypes that may be further used to probabilistically assign membership into a family group. Allele pair data is only given for locus pairs in each family where significant LD was found.

## Conclusions

The characterization of the family groups using LD patterns serves as a proof of concept for further analysis of relatedness within the SMGF database. By identifying clusters within the database which exhibit high within-cluster LD, we hope to identify closely related groups which will assist in future relationship assignments of unknown individuals to these family-level groupings. In addition to describing general characteristics of family groups, we also wanted to find specific haplotypes within each family that may serve to selectively describe that particular family group. Table 5 shows many haplotypes with potential to serve as family-descriptive haplotypes for assignment of unknown individuals, when taken together with other family-selective haplotypes. Further analysis needs to be performed to investigate the use of these haplotypes to probabilistically assign membership into family groups.

## Acknowledgements

Special thanks to everyone who contributed their DNA and pedigrees to the project and to all those who assisted with sample collections. Thanks to Charles Jensen and Robert Hughes for assistance in automating the calculations and also to Edgar Gomez for poster design and formatting assistance. This work was funded by the Sorenson Molecular Genealogy Foundation and its contributors.

## References

- Lewontin, R. C. On Measures of Gametic Disequilibrium. *Genetics* 1988, 120, 849-852.
- Nei, M.; Li, W.-H. Linkage Disequilibrium in Subdivided Populations. *Genetics* 1973, 75, 213-219.
- Ardlie, K.G.; Kruglyak, L.; Seielstad, M. Patterns of Linkage Disequilibrium in the Human Genome. *Nat. Rev. Genet.* 2002, 3(Apr), 299-309.
- Vitart, V.; Carothers, A.D.; Hayward, C.; Teague, P.; Hastie, N.D.; Campbell, H.; Wright, A.F. Increased Level of Linkage Disequilibrium in Rural Compared with Urban Communities: A Factor to Consider in Association-Study Design. *Am. J. Hum. Genet.* 2005, 76, 763-772.
- Pritchard, J. K.; Przeworski, M. Linkage Disequilibrium in Humans: Models and Data. *Am. J. Hum. Genet.* 2001, 69, 1-14.
- Schneider, S.; Roessli, D., and Excoffier, L. (2000) Arlequin: A software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.
- Schouten, M.T.; Williams, C.K.I.; Haley, C.S. The impact of using related individuals for haplotype reconstruction in population studies. *Genetics* 2005, doi: 10.1534/genetics.105.042762.