Methods for improved reconstruction of haplotypes of highly polymorphic loci

J.B. Ekins[1], L.A.D. Hutchison[1,2], J.E. Ekins[1], N.M. Myres[1], K.Hadley[1], L. Layton[1], M. L. Lunt[1], S.S. Masek[1], A.A. Nelson[1], M.E. Nelson[1], K.L. Pennington[1], U.A. Perego[1], J.L. Peterson[1], A. Sims[1], T. Tolley[1], A. Welch[1], S.R. Woodward[1,3]

[1]Sorenson Molecular Genealogy Foundation, 2511 S. West Temple, Salt Lake City, Utah [2]Department of Computer Science, Brigham Young University, Provo, Utah [3]Department of Microbiology and Molecular Biology, Brigham Young University, Provo, Utah

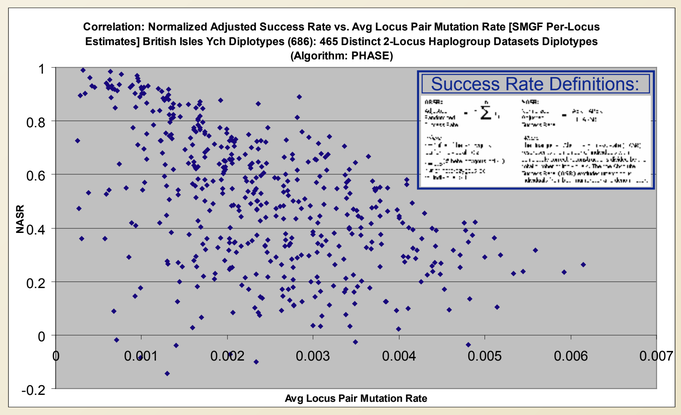
Abstract:

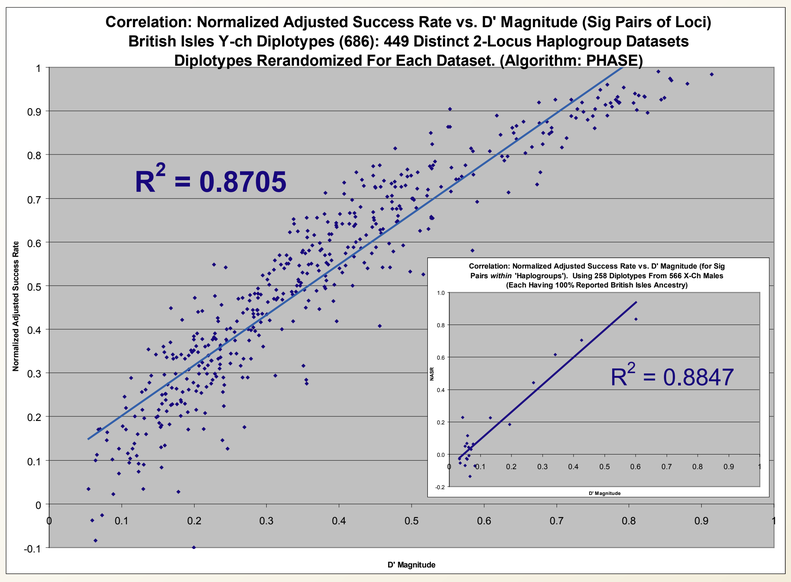
Derivation of haplotype information from unphased genotypic data is vital to many genetic and genealogical questions, including the reconstruction of recent population history. Microsatellite loci exhibit higher mutation rates than biallelic loci and thereby become good candidates for asking questions about recent population history. These rates contribute to higher levels of identity-by-state haplotypes and offer a reasonable explanation for the lower absolute success rates obtained by haplotype reconstruction algorithms like PHASE2.0.2 (when compared to SNP datasets). This difficulty is exacerbated by the low average density of STR loci in the genome (as compared with available SNP loci) which, generally speaking, increases the noise introduced by recombination events between any two loci. Some have successfully applied selection parameters to SNP datasets such as measures of disequilibrium with encouraging results. Herein, selection parameters (e.g. LD) and informational parameters (e.g. the extent of population admixture) are shown to improve haplotyping success rates in datasets containing linked STR loci. In addition to the use of simulated datasets (such as phase known 'diplotypes' produced by combining haploid data), trials using large, 'real' STR datasets (phase-known diploid individuals) are performed for the first time--answering a need to establish success rates within 'real' datasets containing microsatellite loci. We present a new haplotype reconstruction algorithm (Reconstrux) that iteratively infers population haplotype prevalences and haplotype phase probabilities for each sample. In addition to producing reconstructions of STR datasets, Reconstrux operates about $6x10^3$ times faster than PHASE and accepts information regarding admixture while achieving comparable levels of accuracy when equivalent datasets are used.

Parameteric Improvements:

A Strong Correlation Exists Between Success Rate and Magnitude of Disequilibrium Between Microsatellite Markers

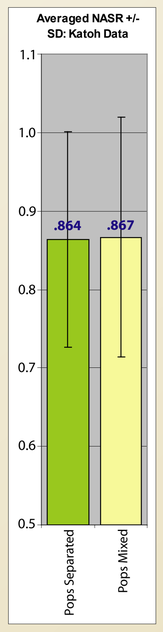
X-chromosome datasets containing haploid males (see Dataset Legend) with 100% reported ancestry in the British Isles were constructed by combining haplotypes at random to form simulated diploid genotypes and then phased using PHASE 2.0.2. The 13 X-chromosome markers separate into three presumptive haplogroups--defined as having the same genetic map distance[4]. D' significance and D' magnitude for all pairs of loci were calculated using ARLEQUIN. Each dataset of diplotypes consisted of two loci. Normalized Adjusted Success Rates (NASR--see Success Rate Definitions) for datasets representing every possible two-locus combination were produced. Pairs of loci which were not found to be in significant disequilibrium with each other were excluded from analysis. There is a strong correlation between the magnitude of D' between pairs of loci and the corresponding NASR for a dataset consisting of a particular pair of loci. This correlation was predicted by a similar experiment using Y-chromosome data (see data below). By combining loci into groups in which individual pairs of loci show large disequilibrium values (<.75), long STR haplogroups (~7 markers) can be constructed which attain absolute success rates (UASRs) of > 90% (based on simulated data--not shown).





No ameliorative effect observed for phasing success rate when reported ancestry used to subdivide a mixed population into several geographically distinct populations:

Twelve separate trials were analyzed using a pair of loci found to be in significant disequilibrium in each of the six populations sampled by Katoh (see Dataset Legend). Six of the trials corresponded to the six populations [Pops Separated]. These trials contained an average of ~32 diplotypes. Six more trials consisting of 64 haplotypes choosen at random (with replacement) from the entire set to create 32 diplotypes [Pops Mixed]. All twelve were phased separately using Phase 2.0.2. Generally, success rates were high, each category experienced a low NASR anomaly, and the averages are not statistically different.



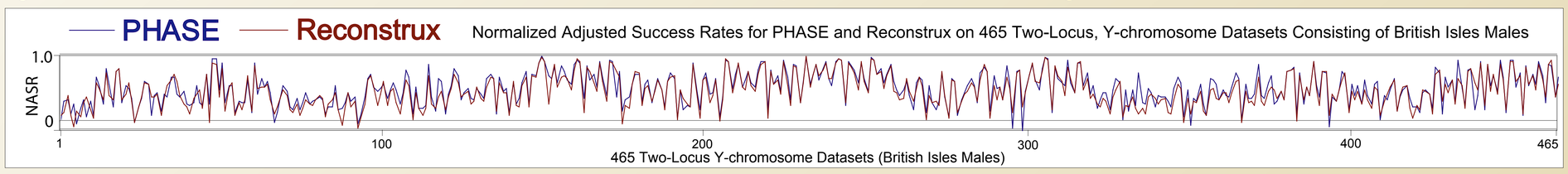
A Low Correlation Exists Between Average Mutation Rate and Success Rate:

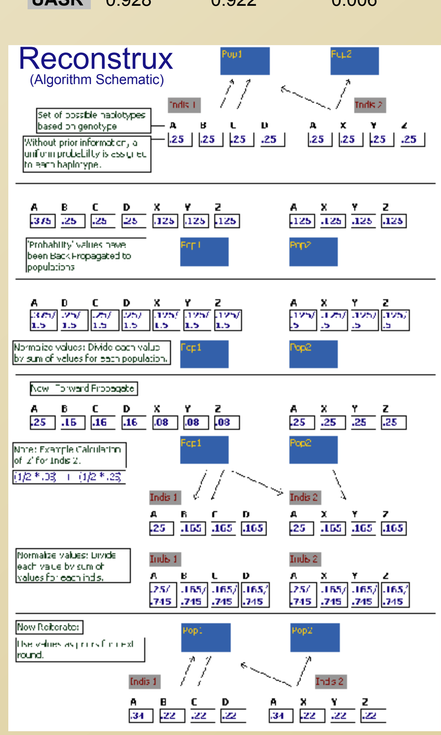
Datasets representing every possible pair of a set of 31 Y-chromosome loci containing the combined haplotypes of 1372 British Isles males were run using PHASE 2.0.2. Success rates (NASR) were calculated. Mutation rates (SMGF estimates) were averaged for each pair of loci in order to provide a summary value reflecting the mutability of a particular pair of loci. The correlation between NASR and the averaged locus-pair mutation rates does not appear to be linear. It can be inferred that when two loci have a high averaged mutation rate, they are likely to be difficult to correctly phase. When two loci share a low averaged mutation rate, success rates tends to be higher.

PHASE vs. Reconstrux:

Using the above-described Y-ch dataset of 1372 British Isles males, 465 two-locus datasets (representing every combination of 31 loci) were analyzed using PHASE 2.0.2 and Reconstrux. Although PHASE 2.0.2 does perform better on average than Reconstrux, the success rates are comparable. Reconstrux outperformed PHASE on 33% of the trials--both algorithms being run on default settings. The standard deviations of the averages for each algorithm were quite similar to one another. See schematic for a description of the Reconstrux algorithm.

	PHASE	Reconstrux	Difference
NASR	0.501	0.466	0.035
ASR	0.750	0.733	0.018
UASR	0.928	0.922	0.006





A Real Dataset:

A novel feature of this research is the use of a 'Phase Known' female X-chromosome dataset. As the X-chromosome is haploid in males, X-chromosome phase may be inferred for a category of female relations (e.g. mother/son, father/daughter). A set of 129 unrelated females were phased based on phase-informative males. Datasets corresponding to every possible pair of the X-chromosome loci herein described were generated for these phase known females. Success rates were calculated after being run on Phase 2.0.2. The NASR (see Success Rate Definitions) for each trials was compared with equivalent trials produced for the 566 Male X-chromosome British Isles data presented above. As might be predicted, a two-tailed Student's t-test fails to reject the null hypothesis that there is no difference between the trials (p=0.278). Female X-chromosome genotypes are roughly analogous to male X-chromosome diplotypes. A more substantial question is whether it is appropriate to use Y-chromosome diplotypes to model the effects of selection parameters (like D') on real genotypes. In an transitive fashion, (a) the similarity between the success rates in the above mentioned male and female X-chromosome datasets and (b) the fact that the strong correlation between NASR and D' Magnitude for the said male dataset was predicted by the correlative analysis first performed on the Y-chromosome dataset of 1372 British Isles males (c) provide evidence that suggests a legitimacy in use of simulated Y-chromosome diplotypes to predict the importance of disequilibirum on real genotypes.

Dataset Legend

Dataset	Loci	Ancestry & Size
SMGF Y-ch:	DYS388, DYS389A, DYS389B, DYS390, DYS391, DYS392, DYS426, DYS438, DYS439, DYS447, DYS448, DYS460, YGATAH4, DYS437, DYS444, DYS446, DYS452, DYS462, GGAATB07, YGATAA10, YGATAC4, DYS393, DYS394, DYS449, DYS455, DYS458, DYS461, DYS463, DYS445, DYS454, DYS456	1372 Males: British Isles
SMGF X-ch:	SMGF_X37.35: DXS989, DXS8099, DXS8027 DXS1048, DXS8047 SMGF_X44.98: DXS1214, DXS1036, DXS1067, DXS997 SMGF_X144.78: DXS8071, DXS1041, DXS994, DXS1062	566 Males: British Isles 129 Females: Mixed Ancestry (Predominantly European)
Katoh X-ch:	DXS983, DXS986, DXS8092, DXS8082, DXS1225, DXS8037, DXS995	24 Males: European; 50 Males: Japanese; 41 Males: Khalkh; 20 Males: Khoton; 27 Males: Uriankhai; 29 Males: Zakhchin;

References:

[1]Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111-122.
[2]Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927.
[3]Katoh T (2002) Genetic Isolates in East Asia: A Study of Linkage Disequilibrium in the X Chromosome . Am J. Hum. Genet. 71:395-400.
[4]Marshfield Clinic. http://research.marshfieldclinic.org/genetics/Physical_Maps/Chromosome23.htm.
[5]Moore D (2003) The Basic Practice of Statistics, 2:367.
[6]Pritchard J, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945-959.
[7]Pritchard JK, Wen W (2002) Documentation for structure software: version 2. http://pritch.bsd.uchicago.edu.-Rosenberg N, Burke T, Elo K, Feldman M, Freidlin P, Groenen M, Hillel J, Mäki-Tanila A, Tixier-Boichard M, Vignal A, Wimmers K, Weigend S (2002) Genetic Structure of Human Populations. Science 298:2381-2385.
[8]Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, Feldman M (2002) Genetic Structure of Human Populations. Science 298:2381-2385.
[9]Schneider S, Roessli D, Excoffier L (2000) Arlequin: A software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.
[10]Stephens M, Smith N, Donnelly P (2001) A New Statistical Method for Haplotype Reconstruction from Population Data. Am J. Hum. Genet. 68:978-989.

Presented at the American Society of Human Genetics 2004 Annual Meeting, October 2004, Toronto, Ontario, Canada