

Proceedings

Open Access

Haplotypic structure of the X chromosome in the COGA population sample and the quality of its reconstruction by extant software packages

Fabio Marroni¹, Chiara Toni², Benedetto Pennato¹, Ya-Yu Tsai³,
Pryia Duggal³, Joan E Bailey-Wilson³ and Silvano Presciuttini*^{1,3}

Address: ¹Center of Statistical Genetics, c/o Centro Retrovirus, SS Abetone e Brennero 2, 56127 Pisa, Italy, ²Unit of Legal Medicine, University of Pisa, Pisa, Italy and ³Center for Inherited Disease Research, NHGRI, NIH, Baltimore, USA

Email: Fabio Marroni - fabio.marroni@biomed.unipi.it; Chiara Toni - chiara@hint.it; Benedetto Pennato - b.pennato@libero.it; Ya-Yu Tsai - ytsai@cidr.jhmi.edu; Pryia Duggal - pduggal@nhgri.nih.gov; Joan E Bailey-Wilson - jebw@nhgri.nih.gov; Silvano Presciuttini* - sprex@biomed.unipi.it

* Corresponding author

from Genetic Analysis Workshop 14: Microsatellite and single-nucleotide polymorphism Noordwijkerhout, The Netherlands, 7-10 September 2004

Published: 30 December 2005

BMC Genetics 2005, 6(Suppl 1):S77 doi:10.1186/1471-2156-6-S1-S77

Abstract

Background: The haplotypes of the X chromosome are accessible to direct count in males, whereas the diplotypes of the females may be inferred knowing the haplotype of their sons or fathers. Here, we investigated: 1) the possible large-scale haplotypic structure of the X chromosome in a Caucasian population sample, given the single-nucleotide polymorphism (SNP) maps and genotypes provided by Illumina and Affimetrix for Genetic Analysis Workshop 14, and, 2) the performances of widely used programs in reconstructing haplotypes from population genotypic data, given their known distribution in a sample of unrelated individuals.

Results: All possible unrelated mother-son pairs of Caucasian ancestry ($N = 104$) were selected from the 143 families of the Collaborative Study on the Genetics of Alcoholism pedigree files, and the diplotypes of the mothers were inferred from the X chromosomes of their sons. The marker set included 313 SNPs at an average density of 0.47 Mb. Linkage disequilibrium between pairs of markers was computed by the parameter D' , whereas for measuring multilocus disequilibrium, we developed here an index called D^* , and applied it to all possible sliding windows of 5 markers each. Results showed a complex pattern of haplotypic structure, with regions of low linkage disequilibrium separated by regions of high values of D^* . The following programs were evaluated for their accuracy in inferring population haplotype frequencies: 1) ARLEQUIN 2.001; 2) PHASE 2.1.1; 3) SNPHAP 1.1; 4) HAPLOBLOCK 1.2; 5) HAPLOTYPER 1.0. Performances were evaluated by Pearson correlation (r) coefficient between the true and the inferred distribution of haplotype frequencies.

Conclusion: The SNP haplotypic structure of the X chromosome is complex, with regions of high haplotype conservation interspersed among regions of higher haplotype diversity. All the tested programs were accurate ($r = 1$) in reconstructing the distribution of haplotype frequencies in case of high D^* values. However, only the program PHASE realized a high correlation coefficient ($r > 0.7$) in conditions of low linkage disequilibrium.

Background

With the advent of dense genetic maps of single-nucleotide polymorphisms (SNPs), large population samples of diallelic multilocus genotypes are increasingly available for studies in the fields of population genetics, marker-disease association, and evolutionary genetics. However, current genotyping methods do not provide information on the individual diplotypes (the haplotype pair composing a genotype). This information would nonetheless increase the power of any genetic analysis substantially. Several methods for estimating haplotype frequencies from a sample of genotyped but unphased diploid individuals have been developed. These include a sequential haplotype inference algorithm [1], several expectation-maximization based algorithms [2-4], a coalescent-based algorithm based on a Markov chain Monte Carlo approach [5], a Bayesian approach that uses a Dirichlet prior distribution for the haplotype frequencies [6], and a recent method based on Bayesian networks that takes account of recombination hotspots, bottlenecks, genetic drift, and mutation [7]. The X chromosome is unique as a population genetics tool because of its diploidy in females and haploidy in males, a characteristic that, among other things, renders its haplotypes accessible to direct count. The potential of the X chromosome to contribute to fine-scale microevolutionary studies (which are dominated by mtDNA and the Y chromosome) has probably been underused [8].

The purposes of the present work were dual. First, we wanted to ascertain whether a haplotype structure could be detected on the X chromosome in a Caucasian population sample, given the genotypes provided by Illumina and Affimetrix for Genetic Analysis Workshop 14. Second, we wanted to test and compare the capability of widely used programs in reconstructing haplotypes, given their distributions in a sample of individuals with known diplotypes; for this purpose, the haplotypes in a sample of unrelated mothers (treated as independent individuals) were determined using the data on their sons, so that the accuracy of different methods of inferring haplotypes from genotype data could be evaluated by comparing the true and the inferred distributions of haplotype frequencies.

Methods

All possible unrelated mother-son pairs of Caucasian ancestry were selected from the 143 families of the Collaborative Study on the Genetics of Alcoholism pedigree files. The diplotypes of the mothers were inferred from the X chromosomes of their sons. An integrated map of the X chromosome for the Illumina and Affimetrix SNP datasets was obtained by querying the NCBI Human Genome (Build 34) for marker position. The Affimetrix dataset was first cleared of the markers with heterozygosity <0.2 in

order to render the 2 datasets more homogeneous. The final map included 313 markers (121 Illumina, 192 Affimetrix). It spanned 146.5 Mb at an average density of 0.47 Mb. A large gap (7.8 Mb) was located at 56.5 Mb. Linkage disequilibrium (LD) between pairs of markers was computed by the parameter D' . For multilocus disequilibrium, we defined here an index called D^* . This is computed as $D^* = 1 - (H_d - H_{\min}) / (H_{\text{eq}} - H_{\min})$, where the haplotype diversity H_d is computed as $H_d = 1 - \sum p_i^2$ (p_i being the frequency of haplotype i , analogous to the gene diversity of a single locus), the expected haplotype diversity under no LD, H_{eq} , is calculated as $H_{\text{eq}} = 1 - \sum E\{p_i\}^2$ ($E\{p_i\}$ being the expected frequency of each possible haplotype, i.e., the product of the frequencies of the alleles composing that haplotype), and the minimum possible value of haplotype diversity H_{\min} is obtained computationally. Specifically, if n haplotypes typed for s SNPs are arranged in a n by s matrix and the alleles are coded consistently (e.g., 0 = low frequency allele at all loci), H_{\min} is obtained by computing H_d with the above equation, after the matrix is rearranged by sorting iteratively each column independently of the others. Another measure of multilocus disequilibrium (the normalized entropy difference, ϵ) was published previously [9]. We applied both D^* and ϵ to all possible sliding windows of 5 SNPs each.

The following programs were evaluated for their accuracy in inferring population haplotype frequencies: 1) ARLEQUIN 2.001; 2) PHASE 2.1.1; 3) SNPHAP 1.1; 4) HAPLOBLOCK 1.2; 5) HAPLOTYPYPER 1.0. In inferring haplotypes, ARLEQUIN and SNPHAP use an expectation-maximization algorithm, HAPLOTYPYPER uses a Bayesian approach assuming a prior Dirichlet distribution of haplotype frequencies, PHASE uses a coalescence-based algorithm for inferring the prior distribution of haplotype frequencies coupled with a Markov chain Monte Carlo approach to approximate the posterior distribution, and HAPLOBLOCK uses a Bayesian network method. ARLEQUIN and SNPHAP ignore the missing data, whereas PHASE and HAPLOTYPYPER make informed guesses; in HAPLOBLOCK, users can choose between these 2 options. Among all programs, only PHASE include the possibility of specifying a genetic map and modeling the process of recombination. Genotypes at 5 or 10 consecutive markers were selected from the Illumina dataset based on varying levels of D^* , and were submitted to all programs. The accuracy of each program was measured using the Pearson correlation coefficient between the true and the inferred haplotype frequencies.

Results

The final dataset analyzed in the present work consisted of 104 unrelated Caucasian females with known diplotypes at 313 SNPs on the X chromosome. Figure 1 shows the parameter D' between all adjacent markers and between

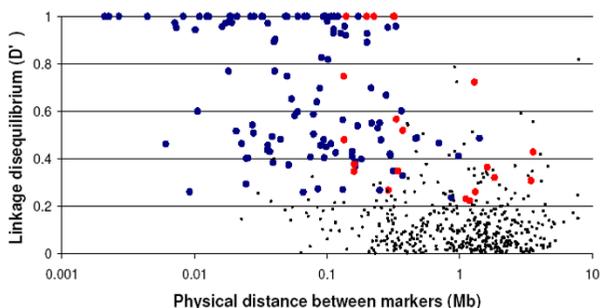


Figure 1
Standardized linkage disequilibrium as a function of intermarker distance. Standardized linkage disequilibrium (D') between markers of the X chromosome as a function of the intermarker distance. Large symbols: D' values with nominal $p \leq 0.01$ (blue: adjacent markers; red: LD computed at 5-marker intervals). Dots: D' values with $p > 0.01$. Marker pairs with distance < 1 kb have been omitted.

each marker and the fifth marker downstream; 84% of marker pairs closer than 100 kb showed high levels of LD, with p -values < 0.01 , in comparison with 34% of the pairs 100 to 500 kb apart, and 4.4% of the pairs 500 kb to 2 Mb apart. Two marker pairs with an intermarker distance > 3 Mb showed highly significant LD. Then, the multilocus haplotype structure of the X chromosome was investigated by considering sliding windows of 5 markers and calculating both the parameters D^* and ϵ . The 2 measures were highly correlated ($r = 0.945$). Because of the uneven marker distribution in the maps, the length of the 5-marker segments was highly variable, from 93 kb to 7.94 Mb; in the present analysis, segments longer than 5 Mb were not considered. Several regions of high values of D^* (low haplotype diversity) were separated by segments with similar values of H_d and H_{eq} (no LD, Figure 2). One instance of $D^* = 1.0$ (i.e., in which $H_d = H_{min}$) was located at about 56 Mb, near the large gap in the chromosome map. These 5 markers were part of a chromosome segment of 10 markers spanning 1.33 Mb for which only 7 haplotypes, out of 1,024 theoretically possible, were observed. The value of D^* for this segment of 10 markers was 0.74. This is consistent with previous reports of a substantial recombination decrease in the centromere of the X chromosome [10].

Fourteen series of unphased genotypes with different values of D^* (10 series consisting of 5 consecutive markers and 4 of 10 markers) were submitted to each of the 5 programs. Table 1 shows Pearson correlation coefficients between the observed and the inferred haplotype frequencies. For 5-marker haplotypes, the correlation coefficients were high even in situations of moderate LD for all programs. In the case of 10-marker haplotypes (last 4 rows in

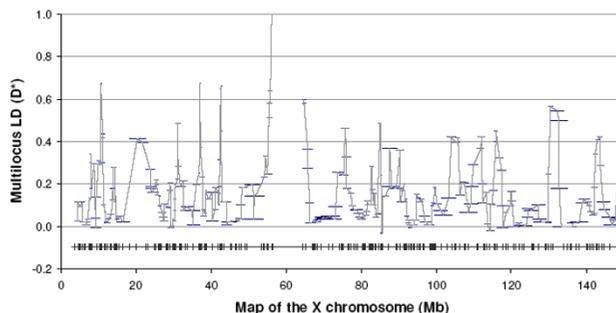


Figure 2
Multilocus LD of the X chromosome. Bars represent sliding windows of 5 markers each, whose D^* value is plotted. The line under the chart shows the marker location; a large gap centered at 60 Mb may be noted.

Table 1), all the programs reconstructed perfectly well the true haplotype distribution when the number of different haplotypes in the sample was small in comparison with the total number of possible haplotypes. With the increase of haplotype diversity (series 11 in Table 1), the performance of the programs started to decrease and differentiate, though the correlation between the true and the estimated haplotype frequencies was still high; PHASE realized the best performance ($r = 0.996$). In the opposite situation, when the haplotype diversity was high (130 different haplotypes in a sample of 104 individuals) the performances were generally poor; only PHASE realized a high correlation coefficient (0.737). When the majority of the haplotypes is unique (last row in Table 1), the inferred haplotypes are clearly unreliable.

Discussion

We investigated the large-scale haplotypic structure of the X chromosome in a Caucasian population sample by computing D' for all adjacent markers and any fifth marker; high levels of LD were detected even at distances > 1 Mb. We then applied to all possible segments of 5 consecutive markers a measure of multilocus LD, here called D^* . This parameter is easily computed and is based on the standard definition of heterozygosity; D^* reaches its maximum possible value of 1.0 when the haplotype diversity is at a minimum, i.e., when LD is complete. Thus, D^* appeared to be a suitable measure in studies of large-scale multilocus linkage disequilibrium. In addition, we wanted to test the capability of widely used programs in reconstructing the haplotypes of population samples. All investigated programs perform well when the number of markers is small (5) even in situations of low values of D^* . With a higher number of markers (10), high correlation values between true and inferred haplotype frequencies are attained only in conditions of high D^* . PHASE is an exception, in that it has reconstructed the true distribu-

Table 1: Correlation between the true haplotype frequencies and those estimated by five programs

	Series	Location ^a	Length ^b	#Hap ^c	D [*]	Arlequin (EM) ^d	Phase (CP) ^e	SnP Hap (EM)	Haplo Block (BN) ^f	Haplo typer (D) ^g
5-marker haplotypes	1	55,992,117	0.14	3	1	1	1	1	1	1
	2	10,464,050	0.14	10	0.6	1	1	1	1	1
	3	54,499,735	1.42	5	0.59	1	1	1	1	1
	4	71,732,319	2.62	8	0.46	1	1	1	1	1
	5	36,765,524	0.13	11	0.36	1	1	1	1	1
	6	84,938,832	1.46	9	0.28	0.97	0.99	0.97	0.98	0.89
	7	42,260,770	4.25	20	0.19	0.95	0.99	0.96	0.95	0.92
	8	90,806,576	2.23	16	0.18	0.94	0.96	0.94	0.83	0.92
	9	63,668,482	3.84	30	0.05	0.84	0.95	0.82	0.75	0.77
	10	3,562,850	3.86	31	0.02	0.75	0.91	0.75	0.67	0.65
10-marker haplotypes	11	54,499,735	1.63	6	0.74	1	1	1	1	1
	12	55,920,880	10.51	15	0.44	0.99	1	0.98	0.93	0.97
	13	74,309,896	10.79	130	0.02	0.51	0.78	0.50	0.07	0.50
	14	14,437,620	13.36	165	0.02	0.14	0.35	0.19	0.25	0.17

^anucleotide position of the first marker of the haplotype

^blength of the haplotype (Mb)

^cnumber of different haplotypes in the sample

^dEM: expectation-maximization algorithm

^eCP: coalescent prior

^fBN: Bayesian networks

^gD: Dirichlet prior

tion of haplotype frequency with good accuracy even in a difficult situation. This program employed significantly more computing time than the others (10–20 minutes in comparison with less than a second using in the same machine), with the exception of HAPLOBLOCK, which ran for more than 30 hours.

Conclusion

The SNP haplotypic structure of the X chromosome is complex, with regions of high haplotype conservation (most notably, around the centromere) interspersed among regions of higher haplotype diversity. A more detailed definition of this structure, to be accomplished in further studies, could be useful in evolutionary analyses and in disease association studies.

All the tested programs were accurate ($r = 1$) in reconstructing the true distribution of haplotype frequencies in case of high LD. Only the program PHASE realized a high correlation coefficient ($r > 0.7$) in case of low linkage disequilibrium.

Abbreviations

LD: Linkage disequilibrium

SNP: Single-nucleotide polymorphism

Authors' contributions

FM participated in all phases of statistical analyses and drafted the manuscript. CT participated in the analysis of 2-locus and multilocus linkage disequilibrium. BP participated in the analyses of software performances. Y-T integrated the SNP maps. PD selected the families to be used in the study. JEB-W participated in the study conception, and provided critical revision of the manuscript for important intellectual content. SP conceived of the study, participated in its design and coordination, and helped to draft the manuscript. All authors read and approved the final manuscript.

References

1. Clark AG: **Inference of haplotypes from PCR-amplified samples of diploid populations.** *Mol Biol Evol* 1990, **7**:111-122.
2. Excoffier L, Slatkin M: **Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.** *Mol Biol Evol* 1995, **12**:921-927.
3. Hawley ME, Kidd KK: **HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes.** *J Hered* 1995, **86**:409-411.
4. Long JC, Williams RC, Urbanek M: **An E-M algorithm and testing strategy for multiple-locus haplotypes.** *Am J Hum Genet* 1995, **56**:799-810.
5. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68**:978-989.
6. Niu T, Qin ZS, Xu X, Liu JS: **Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms.** *Am J Hum Genet* 2002, **70**:157-169.

7. Greenspan G, Geiger D: **High density linkage disequilibrium mapping using models of haplotype block variation.** *Bioinformatics* 2004, **20(Suppl 1)**:1137-1144.
8. Schaffner SF: **The X chromosome in population genetics.** *Nat Rev Genet* 2004, **5**:43-51.
9. Nothnagel M, Furst R, Rohde K: **Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks.** *Hum Hered* 2002, **54**:186-198.
10. Mahtani MM, Willard HF: **Physical and genetic mapping of the human X chromosome centromere: repression of recombination.** *Genome Res* 1998, **8**:100-110.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

