Original Paper

Human Heredity

Hum Hered 2006;62:175–189
DOI: 10.1159/000096599

Received: May 3, 2006
Accepted after revision: August 7, 2006
Published online: October 30, 2006

Linkage Disequilibrium and Haplotype Homozygosity in Population Samples Genotyped at a High Marker Density

Hui Wang[a] Chia-Ho Lin[a] Susan Service[b] Yuguo Chen[d] Nelson Freimer[b]
Chiara Sabatti[a, c] The international collaborative group on isolated populations

[a]Department of Statistics, [b]Center for Neurobehavioral Genetics, and [c]Department of Human Genetics, UCLA, Los Angeles, Calif., and [d]Department of Statistics, University of Illinois at Urbana-Champain, Champain, Ill., USA

Key Words
Linkage disequilibrium measures · Inbreeding · Copy number variation · Population genetics · Genomic loss

Abstract
Objective: Analyze the information contained in homozygous haplotypes detected with high density genotyping. ***Methods:*** We analyze the genotypes of ~2,500 markers on chr 22 in 12 population samples, each including 200 individuals. We develop a measure of disequilibrium based on haplotype homozygosity and an algorithm to identify genomic segments characterized by non-random homozygosity (NRH), taking into account allele frequencies, missing data, genotyping error, and linkage disequilibrium. ***Results:*** We show how our measure of linkage disequilibrium based on homozygosity leads to results comparable to those of R^2, as well as the importance of correcting for small sample variation when evaluating D'. We observe that the regions that harbor NRH segments tend to be consistent across populations, are gene rich, and are characterized by lower recombination. ***Conclusions:*** It is crucial to take into account LD patterns when interpreting long stretches of homozygous markers.
Copyright © 2006 S. Karger AG, Basel

The international collaborative group on isolated populations members that are not listed separately as authors of this manuscript are: Maria Karayiorgou[a], J. Louw Roos[b], Herman Pretorious[b], Gabriel Bedoya[c], Jorge Ospina[d], Andres Ruiz-Linares[c, e], António Macedo[f], Joana Almeida Palha[g], Peter Heutink[h, i], Yurii Aulchenko[j], Ben Oostra[j], Cornelia van Duijn[j], Marjo-Riitta Jarvelin[k, l], Teppo Varilo[m, n], Lynette Peddle[o], Proton Rahman[p], Giovanna Piras[q], Maria Monne[q], Leena Peltonen[m, n], and the affiliations:
[a]Rockefeller University, New York, N.Y, USA; [b]University of Pretoria Weskoppies Hospital, Pretoria, Republic of South Africa; [c]Laboratorio de Genetica Molecular, Universidad de Antioquia, and [d]Departamento de Psiquiatria, Universidad de Antioquia, Medellin, Colombia; [e]The Galton Laboratory, Department of Biology (Wolfson House), University College London, London, UK; [f]Instituto de Psicologia Médica, Faculdade de Medicina, Coimbra, and [g]Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal; [h]Department of Human Genetics, Section of Medical Genomics, VU University Medical Center, and [i]Center for Neurogenomics and Cognitive Research, VU University and VU University Medical Center, Amsterdam and [j]Genetic Epidemiology Unit, Departments of Epidemiology, Biostatistics and Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands; [k]Department of Public Health Science and General Practice, University of Oulu, Oulu, Finland; [l]Department of Epidemiology and Public Health, Imperial College London, London, UK; [m]Department of Medical Genetics, University of Helsinki, and [n]Department of Molecular Medicine, National Public Health Institute, Biomedicum Helsinki, Helsinki, Finland; [o]Newfound Genomics, and [p]Memorial University of Newfoundland, Newfoundland, Canada; [q]Division of Haematology, San Francesco Hospital, Nuoro, Italy.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2006 S. Karger AG, Basel
0001–5652/06/0624–0175$23.50/0

Accessible online at:
www.karger.com/hhe

Chiara Sabatti
Department of Human Genetics, UCLA School of Medicine
695 Charles E. Young Drive South
Los Angeles, CA 90095-7088 (USA)
Tel. +1 310 794 9567, Fax +1 310 794 5446, E-Mail csabatti@mednet.ucla.edu

Identification of excess homozygosity in genotype data may provide valuable information for a wide range of population genetic investigations and also may indicate genomic variations that are associated with diseases. The level of homozygosity at a single marker reflects the age of the polymorphism, population structure and history, and possibly the effect of selection. Excess joint homozygosity of contiguous markers (haplotype homozygosity) can be a signature of selective pressure, linkage disequilibrium, inbreeding, and variations in copy numbers.

Until recently it was not feasible to assess fine-scale homozygosity in the human genome in a systematic manner or on a large scale. It is now evident that high resolution genotype data provide the information needed for such assessment. For example, several datasets of high density genotypes have shown that long stretches of homozygosity are unexpectedly common, and possibly reflect the effects of either inbreeding or selection. Estimates of homozygosity in genotype datasets can also be used for a variety of statistical analyses, for example, as a measure of linkage disequilibrium. One of the beauties of homozygosity-based measures is that, often, they do not require phasing of the data. This is a particularly appealing feature for analyzing large genotype datasets that are now available (see for example references [1–3] for a review of undergoing studies).

To explore several questions involved in identifying and interpreting homozygosity patterns in large genotype datasets, we analyzed a dataset consisting of about 2,500 SNPs on human chromosome 22 in 12 different population samples, each with about 200 individuals [4]. In particular, we were interested in exploring the relationship between homozygosity and linkage disequilibrium (LD) (see, for a review, our previous work in [5]). With this goal in mind, we conducted a two-fold analysis.

On one front, we define a new measure, *Hvol*, which is constructed by normalizing the levels of homozygosity using a volume test approach [6, 7] and is robust to small sample variation. For comparison purposes, we also consider a volume-test version of D'. We find out that measures based on homozygosity are successful in capturing the pattern of LD detected by D' and, especially, by R^2. Additionally, we are able to verify that volume measures avoid the inflation due to small sample sizes that characterizes D'.

On another front, we propose an algorithm to identify genomic regions that, on the basis of their homozygosity levels, are likely candidates for IBD, genomic loss, and/or selection. Our approach is based on a model that leaves a certain degree of ambiguity on the origins of homozygosity (in particular, IBD or genomic loss can be equivalent interpretations). This approach allows us to identify segments of 'non random homozygosity' (NRH) whose genetic origin will be investigated on the basis of data other than genotypes.

The paper is organized as follows. In section 1, we briefly describe the dataset with particular reference to the levels of homozygosity. In section 2, we introduce the measure *Hvol* and describe the results of its evaluation on this dataset. Section 3 introduces a model for the identification of NRH segments, and discusses the possible interpretations for NRH segments together with the results of its application to our dataset.

1 The Genotyped Samples

In this analysis, we study eleven isolated populations and one European-derived (CAU) population. The eleven isolated populations are Antioquia, Colombia (ANT), Ashkenazi (ASH), Azores (AZO), Costa Rican Central Valley (CR), Southwestern Netherlands (ERF), Finland – mainly early settlement – (FIP), Finland mixture of early and late settlement (FIC), late settlement Finland, Kuusamo (FIK), Newfoundland (NFL), Afrikaner (SAF), and Sardinia, province of Nuoro (SAR). The Caucasian outbred sample consists of 60 parents from CEPH trios and 140 Caucasians from the Coriell Institute Human Variation Panel (CIHVP). The sample consists of 200 individuals from each population. Samples were collected through ongoing genetic studies of various common disorders; subjects included in this analysis were either controls or parents of probands. In most populations, genealogy of the subjects was assessed, and it was documented that they were unrelated for at least several generations.

A total of 2,486 SNPs were successfully genotyped, covering 34.2 Mb of chromosome 22, with an average (median) spacing of one marker every 13.8 (8.5) kb. Specifically, 78% of the gaps between markers were less than 20 kb, and only 3.5% of gaps were greater than 50 kb. These markers were in Hardy-Weinberg equilibrium in all populations (after correcting for multiple testing) and were not monomorphic in any population. The twelve populations had similar numbers of markers with minor allele frequency (MAF) $\leq$10%, with percentages ranging from 10% in Antioquia to 14% in the Sardinian isolate. A detailed description of the genotype data, comparisons in

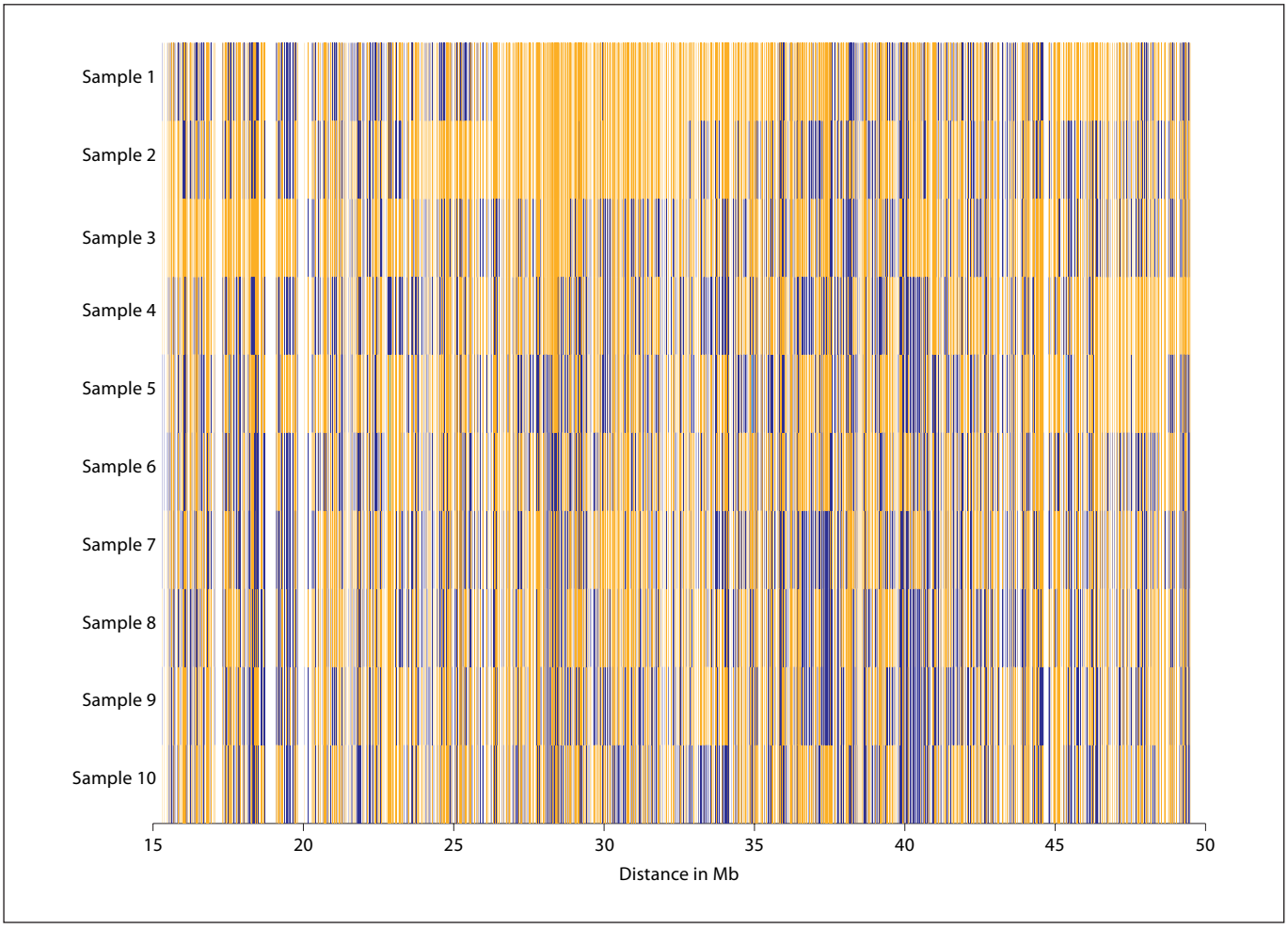

Fig. 1. Representation of genotypes in 10 individuals from Costa Rica. Each horizontal line corresponds to an individual, and each vertical line represents a SNP – whose position in Mb is given on the x axis. The white spaces represent the gaps between SNPs. An orange segment is used to denote a homozygous genotype, i.e. either AA or aa; blue and cyan correspond, respectively, to heterozygous and missing genotypes. Samples 1 to 3 have unusually long homozygous stretches; sample 4 and 5 have suspiciously long homozygous stretches; and samples 6 to 10 are representatives of the overall population. Note that in these samples, the number of ungenotyped SNPs is so small, that practically no cyan can be detected.

allele frequencies between populations, and evaluation of linkage disequilibrium patterns using LD maps can be found in Service et al. [4].

Analysis of overall heterozygosity did not reveal any interesting pattern. The mean heterozygosity of the markers was similar in different populations and ranged from 0.359 (Sardinia) to 0.373 (Antioquia). A moving average of homozygosity values across the chromosomal region also appeared fluctuating randomly around the mean value. Once, however, we started considering joint homozygosity of near-by markers, we were able to observe some interesting features. Firstly, we noticed that when one inspects the homozygosity status of markers along the chromosome in one individual, homozygous markers tend to cluster (see fig. 1). To some degree, this is to be expected because of linkage disequilibrium, as described by Sabatti and Risch [5]. Furthermore, we noticed that in a number of individuals we were able to observe rather long stretches of homozygous markers (see fig. 1). Such long haplotype homozygosity can be due to long range disequilibrium, can be the signature of selection [8], or indicate genomic regions that are identical by descent [9], or again be due to an imprecise genotype call that assigns the status of homozygous to a portion of the genome that

are really monozygous due to genomic loss [10]. We were naturally interested in the reason behind the homozygosity observed in our sample and decided to investigate it further. The next two sections will illustrate first the analysis based on pairwise disequilibrium measures and then the study of stretches of homozygous markers, including a discussion of their origins.

2 Measuring Linkage Disequilibrium

The connection between haplotype homozygosity and linkage disequilibrium has long been noted [11], and has been described in detail by Sabatti and Risch [5]. Let us consider the case of two markers. The alleles at each of these markers can be used to define a partition of the haplotypes. An element of the partition induced by one marker is defined by all the haplotypes that have the same allele at that marker. The level of haplotype homozygosity can be used to define a measure of agreement between the partitions defined by markers one and two, with excess homozygosity (heterozygosity) corresponding to the case of more (less) agreement than expected by chance. Different levels of agreement among these partitions translate into different prediction power, given the allele at one marker, of the allele at the other marker in the haplotype. When agreement is high, we have good prediction power, while low agreement makes prediction more challenging. Excess of haplotype homozygosity, then, corresponds to cases where the R^2 between the two SNPs is high. The relation between haplotype homozygosity values and D' is, instead, more complex, as D' does not really track 'predictability' of one SNP given the other.

While Sabatti and Risch [5] introduced measures of LD based on haplotype homozygosity with values in the range [–1,1], their applicability was hindered by the lack of explicit formulas for the required maximum and minimum values of haplotype homozygosity given the allele frequencies. A solution to this impasse is provided by the use of 'volume measures' of linkage disequilibrium. Sabatti [6] discusses these measures and their precedents in statistical literature [12, 13]. Volume measures can be defined directly on the table of observed haplotype counts, which results in their robustness to the effects of small sample sizes [7]. For simplicity, we consider only the case of two biallelic markers (which is the relevant one here). Let A and B be two SNPs with alleles A_1, A_2 and B_1, B_2, and let n be the number of observed haplotypes. We can summarize these in the table $\mathcal{N}$:

$$\mathcal{N} = \begin{array}{ccccc} & B_1 & B_2 & \\ A_1 & n_{11} & n_{12} & n_{1\cdot} \\ A_2 & n_{21} & n_{22} & n_{2\cdot} \\ & n_{\cdot 1} & n_{\cdot 2} & n \end{array}$$

where $n_{i\cdot}$ ($n_{\cdot j}$) indicated the sum of the elements in row (column) i (j). To define a volume measure of disequilibrium, one firstly needs to select a dissimilarity function, with which compare any table $\mathcal{N}$ against E, the one expected under equilibrium, with entries $e_{ij} = n_{i\cdot}n_{\cdot j}/n$. We consider two such dissimilarities. One is defined only for biallelic markers and is the same evaluated in D'; the other is excess of homozygosity. Volume measures are, then, calculated considering the number of tables T, with the same margins $n_{i\cdot}$, $n_{\cdot j}$ as $\mathcal{N}$, and lower dissimilarity with E.

Specifically, let Ω_1 denote the set of all contingency tables with the same row and column sums as the observed table. Looking at the excess of homozygosity H

$$H(\mathcal{N}) = \sum_{i,j} n_{ij}^2 - \sum_i n_{i\cdot}^2 \sum_j n_{\cdot j}^2 / n^2,$$

we can define the measure *Hvol*:

$$Hvol = \operatorname{sign}\left(H(\mathcal{N})\right) \frac{\sum_{\mathcal{N}' \in \Omega_1} 1_{\left\{|H(\mathcal{N}')| < |H(\mathcal{N})|\right\}} 1_{\{H(\mathcal{N})H(\mathcal{N}') \geq 0\}}}{\sum_{\mathcal{N}' \in \Omega_1} 1_{\{H(\mathcal{N})H(\mathcal{N}') \geq 0\}}}. \quad (1)$$

When $c = r = 2$, let Ω_2 denote the set of all contingency tables with the same row and column sums as the observed table and the same sign of $(n_{11} - n_{1\cdot}n_{\cdot 1}/n)$ as in the observed table. Let

$$M(\mathcal{N}) = \sum_{i,j} \frac{\left(n_{ij} - n_{i\cdot}n_{\cdot j}/n\right)^2}{n_{i\cdot}n_{\cdot j}/n}.$$

We then define *Dvol* as

$$Dvol = \frac{1}{|\Omega_2|} \sum_{\mathcal{N}' \in \Omega_2} 1_{\{M(\mathcal{N}') < M(\mathcal{N})\}}. \quad (2)$$

Note that the above definitions use the strict inequality sign. The choice of $<$ over $\leq$ is irrelevant for large n, but it makes a difference in the case of small n, where strict inequality allows us to better discriminate against apparent associations due to small samples.

Since LD measures based on homozygosity are probably not very familiar to the reader, we illustrate the relation between known measures such as D', $|R|$, and *Hvol*, by calculating them on a subset of our data. We considered the haplotypes defined by consecutive SNPs in one of our population of interest and evaluated the three mea-

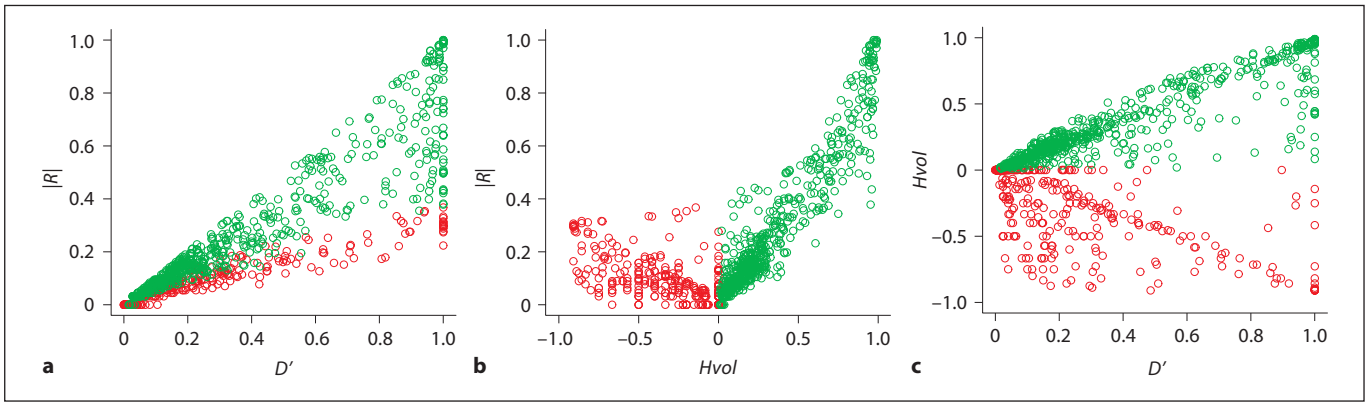

Fig. 2. Illustration of relationships among D', $|R|$, and *Hvol*. Panels **a–c** respectively show the relationships between D' and $|R|$, *Hvol* and $|R|$, D' and *Hvol*. The points are 1,000 random samples from marker pairs with a distance less than 500 kb. The red circles represent tables with excess heterozygosity, i.e. $Hvol \leq 0$, and green circles represent tables with excess homozygosity, i.e. $Hvol > 0$.

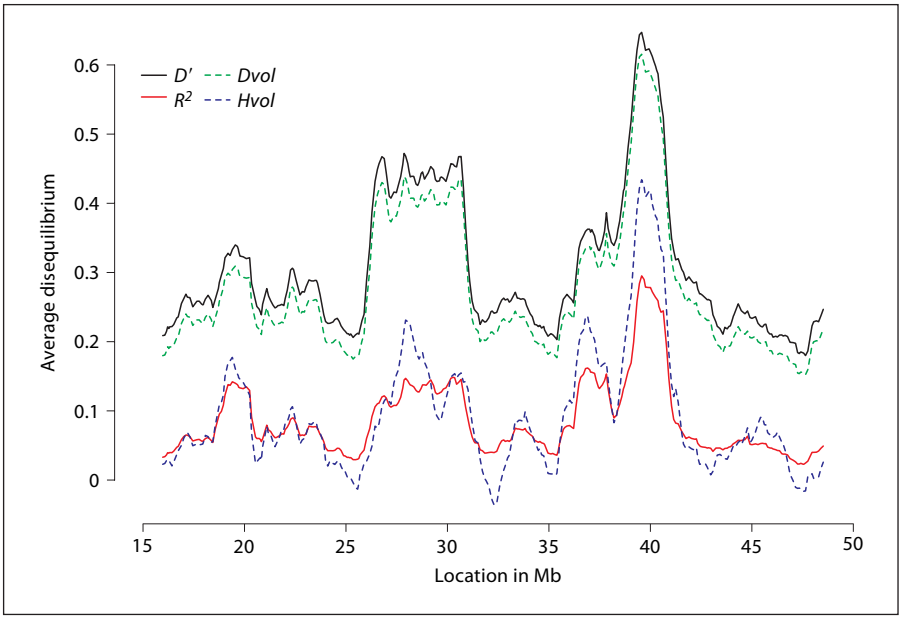

Fig. 3. Linkage disequilibrium on chromosome 22 in the Costa Rican population according to four different measures. D', R^2, *Dvol* and *Hvol* are represented, respectively, with a solid black, a solid red, a broken red, and a broken blue line. The average value of the measures, between markers that are within a 1.7 Mb window, is plotted against the middle point of the window, with the x axis representing the length of chromosome 22.

sures on the sample haplotype frequencies. Figure 2 presents the results. In this display we use $|R|$ rather than R^2, as its values are more directly comparable with D'. In the data analysis that follows, instead, we will use R^2, as it has a more direct interpretation in terms of prediction power. In panel (a) the relation between $|R|$ and D' is illustrated: there are a number of tables with high D' that have low $|R|$. In this display red circles correspond to tables with excess heterozygosity and green circles to tables with excess homozygosity: clearly excess heterozygosity correlates with low values of $|R|$. The positive correlation between $|R|$ and *Hvol* is explicitly illustrated in panel (b): high values of $|R|$ translate in high values of *Hvol*; negative values of *Hvol* correspond to low values of $|R|$; and there are a number of cases where *Hvol* is equal or close to 1, while $|R|$ does not have a very high value. These empirical relations between the values of *Hvol* and $|R|$ mirror the similarity and differences between the notions of prediction and agreement. Panel (c) indicates that the information in *Hvol* is largely orthogonal to the one contained in

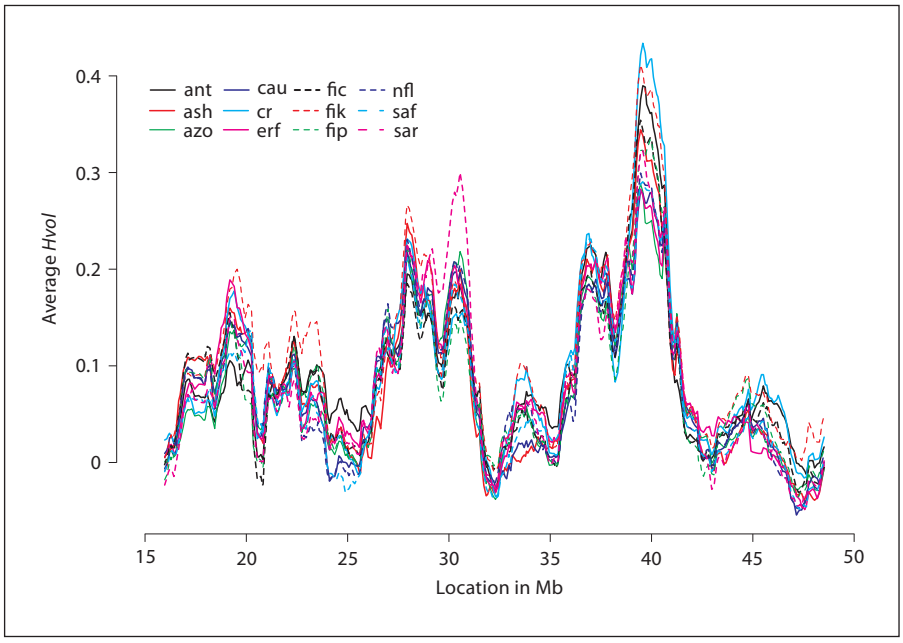

Fig. 4. Linkage disequilibrium on chromosome 22 in the 12 populations as measured with *Hvol*. The figure displays *Hvol* for all 12 populations. The average value of the measure, between markers that are within a 1.7 Mb window, is plotted against the middle point of the window, with the x axis representing the length of chromosome 22.

D′. Overall, the fact that *Hvol* is quite close to $|R|$ suggests that one could use it as a substitute for $|R|$ with multiallelic markers.

To conduct a complete analysis of the linkage disequilibrium patterns in the 12 population samples, we restricted our attention to the SNPs with sample minor allele frequencies larger than 0.1. We did so for uniformity with previous studies, such as the one conducted by Hinds [2], and to make sure that our results were not strongly influenced by the rare markers with exceptionally high homozygosity. This leads us to work with 1920 SNPs. Four measures, *D′*, *Dvol*, R^2 and *Hvol* are calculated for each of the 1,842,240 pairs of SNPs. The results were summarized by averaging the measured disequilibrium within windows of 1.7 Mb sliding along chromosome 22.

Figure 3 reports the values of the four measures in the Costa Rican population. The observed relation between the measures is consistent across populations. In particular it can be noted that the average values of *Dvol* are lower than the ones of *D′*, while clearly exhibiting very similar patterns. This testifies that even if the sample size is moderately large (200 individuals) and only markers with MAF >0.1 are considered, *D′* is inflated due to unobserved rare haplotypes. Turning now our attention to *Hvol*, one can note that its values are closer to the ones of R^2 than to any other measure – not surprisingly, given that both measures are related to predictability. *Hvol* appears to have a higher dynamic range, with larger fluctuations than R^2: this can be explained considering that *Hvol* takes negative values when the agreement is less than expected under independence, while R^2 has its minimum value at zero, in correspondence with independence.

Figure 4 presents the pattern of *Hvol* across all the 12 populations. The most striking aspect is the consistency, across populations, of the LD patterns, suggesting a fundamental role of recombination frequencies. Regions where disequilibrium is higher, allow to better identify differences across populations. In terms of mean values, FIK exhibits the highest level of disequilibrium and AZO the lowest. All of the observations above are consistent with the patterns detected for *D′* and R^2, described in Service et al. [4].

3 Long Stretches of Homozygous Genotypes

As mentioned in the introduction, stretches of adjacent homozygous markers can be signatures of a number of interesting genetic phenomena. In the previous section we have already underscored how linkage disequilibrium between markers in a region can significantly increase their joint homozygosity. An analysis of extended haplotype homozygosities in different populations can be found in Sabatti and Risch [5]. In addition to this, it is

important to keep in mind that selective sweeps substantially reduce diversity in the neighborhood of the selected gene, and thus increase the frequency and length of homozygous segments [8]. Furthermore, the genome of inbred individuals contains regions that are identical by descent and that result in stretches of homozygous genotypes [9, 14]. It is important to recover the boundaries of IBD segments, for example, for gene mapping purposes. Yet another phenomenon that gives rises to stretches of homozygous genotypes is genomic loss, and, more generally, copy number variation. Genotyping technology is unable to accurately determine the number of copies of a given genomic region, so that a haploid segment will be classified as a homozygous diploid [15], and segments that have high copy numbers are also likely to be scored as homozygous as the signal coming from the multiple copies on one chromosome overpowers the one from the normal chromosome. Recently a number of studies [16, 17] have underscored the higher than expected prevalence of variation in copy number in the genome. Their identification is important to correctly interpret results highlighting linkage and/or association between a phenotype of interest and genomic regions.

As all of these different genetic processes result in homozygous segments, it is difficult to distinguish them through the analysis of genotype data alone. Fortunately other data are usually available. Information on raw values recorded during the genotyping reaction can help identify copy number variations. A comparison of the location of homozygous segments across individuals can help separate subject specific effects (likely due to inbreeding) from population ones (that may reflect variation in recombination rate and/or selection). In order to engage in this more complex analysis, however, one needs to initially identify homozygous stretches that are warranted the researcher's attention, as adjacent markers may happen to be jointly homozygous by no more exciting reasons than random chance. Indeed, a quick look at figure 1 makes quite evident that, within each individual, homozygous and heterozygous markers tend to cluster. In order to identify homozygous segments that are reasonable candidates for regions that are identical by descent or that vary in copy number, for example, one needs a model that guides us in the definition of what is to be considered usual or not, as well as an algorithm that allows efficient scanning of the entire collection of genotypes. The rest of this section is devoted to the description of such model and the summary of the results of its application to our dataset.

Interestingly – and reflecting the ambivalence that we have already outlined – there is one model that can be used both to describe IBD status [9] and genomic loss status [15] along the chromosome of one individual. This, combined with a formalization of the genotype generating mechanism, provides a useful tool to analyze our data and identify interesting homozygous segments. Consider data from one individual and let m be the total number of markers on a chromosome (different chromosomes can be considered as independent realizations of the same process). It is useful to introduce a set of hidden variables $\Pi = \{\pi_i\}_{i=1}^{m}$ that indicate the IBD (or loss) status of the individual at the positions corresponding to each of the m markers on a chromosome: $\pi_i = 0$ indicates no IBD (no genomic alterations) at the location of the i-th marker and $\pi_i = 1$ indicates IBD (or an alteration). Leutenegger et al. [9] described the joint distribution of $\{\pi_i\}_{i=1}^{m}$ for an inbred individual using a Markov model of order one with transition probabilities that depend on the distance d_i in centimorgans (cM) between marker i and $i + 1$. In particular, we have the following transition matrix:

$$\begin{pmatrix} t(\pi_{i+1}=1|\pi_i=1) & t(\pi_{i+1}=0|\pi_i=1) \\ t(\pi_{i+1}=1|\pi_i=0) & t(\pi_{i+1}=0|\pi_i=0) \end{pmatrix} = \begin{pmatrix} 1-(1-\delta)\left(1-e^{-\eta d_i}\right) & (1-\delta)\left(1-e^{-\eta d_i}\right) \\ \delta\left(1-e^{-\eta d_i}\right) & 1-\delta\left(1-e^{-\eta d_i}\right) \end{pmatrix}, \quad (3)$$

where the parameter δ corresponds to inbreeding coefficient and η is related to the number of meiotic steps separating the two chromosomes from their most recent common ancestor. Newton and colleagues [10, 18] used the same model to represent the genomic instability of cancer cells, with δ indicating the sporadic loss rate, and the parameter η modelling the dependency among the π_is.

In order to link the unobserved variables Π with the observed genotypes $X = \{x_i\}_{i=1}^{m}$ both Leutenegger et al. [9] and Wang et al. [15] used a hidden Markov model framework. The genotype x_i has four possible values: AA, Aa, aa, –, with – indicating missing data. In the HMM, conditional on π_i, the probability of the four possible values of x_i is independent of the genotype values at the other markers and of the values of π_j with $i \neq j$. If $\pi_i = 1$, we expect a homozygous genotype (modulo genotyping error), while if $\pi_i = 0$, all the genotypes are observed according to their population frequencies. According to the HMM terminology, then, we have the following emission probabilities:

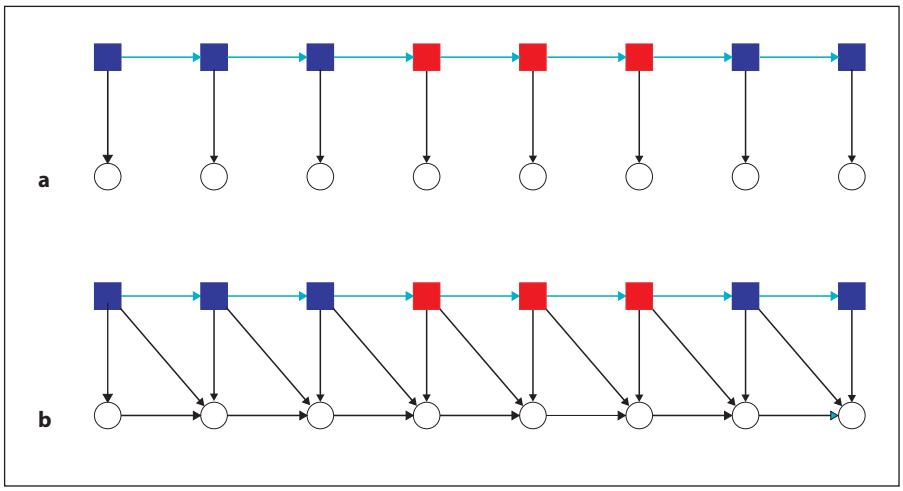

Fig. 5. Hidden Markov Models. Each filled square represents the hidden state π_i at SNP i, red squares correspond to $\pi_i = 1$, blue squares to $\pi_i = 0$, and each circle represents the genotype x_i. The arrows denote the dependence relationships of the HMM. Panel (**a**) gives an illustration of the standard HMM. The next hidden state only depends on the previous hidden state, and the genotype at each SNP only depends on the hidden state of that SNP. Panel (**b**) is an illustration of our generalized HMM. The hidden state still preserves the property of a Markov chain, while the genotype of a SNP depends not only on its hidden state, but also on the hidden state and the genotype of its previous SNP.

$$\begin{pmatrix} e(x_i = AA \mid \pi_i = 1) & e(x_i = Aa \mid \pi_i = 1) & e(x_i = aa \mid \pi_i = 1) & e(x_i = - \mid \pi_i = 1) \\ e(x_i = AA \mid \pi_i = 0) & e(x_i = Aa \mid \pi_i = 0) & e(x_i = aa \mid \pi_i = 0) & e(x_i = - \mid \pi_i = 0) \end{pmatrix}$$
$$= \begin{pmatrix} p_{A^i}(1-\kappa) & 0 & (1-p_{A^i}) & (1-\kappa) & \kappa \\ p_{A^i}^2(1-\kappa) & 2p_{A^i}(1-p_{A^i})(1-\kappa) & (1-p_{A^i})^2(1-\kappa) & \kappa \end{pmatrix} \quad (4)$$

where p_{A^i} is the frequency of allele A for the i-th marker and κ is the missing rate.

While the genotype model in (4) was adequate for the marker density considered in the studies of Leutenegger et al. [9] and Wang et al. [15], it is not for the high density data we collected on chromosome 22. The average inter-marker distance in our data-set is such that we expect nearby markers to be in linkage disequilibrium (and indeed we observed this to be the case in the analysis of the previous section). Model (4) assumes, instead, markers to be in linkage equilibrium, and applying it to our data would result in an excess identification of IBD/loss regions. In order to develop an effective screening tool for homozygous segments, we then modified model (4) to take into account linkage disequilibrium. Figure 5 gives a graphical illustration of the original HMM model (a) and of our generalization (b). For conceptual and computational simplicity, we limited ourselves to a description of linkage disequilibrium that relies on a first order Markov model (see later discussion). The genotype x_i, conditionally on π_i, depends also on the genotype values of the immediate neighboring markers x_{i-1}, x_{i+1}, and actually, the specific form of this link is determined by the values of π_{i-1} and π_{i+1}. Conditionally on (π_i, π_{i-1}, π_{i+1}, x_{i-1}, x_{i+1}), the genotype x_i is independent of every other value of X and Π. We can describe the joint probability of the entire sequence X, by specifying $Pr(x_i \mid x_{i-1}, \pi_{i-1}, \pi_i)$ (which we will denote $e(x_i \mid x_{i-1}, \pi_{i-1}, x_i)$) – as illustrated by the arrows in figure 5 (b). These probabilities are defined combining two elements: the expectation of a homozygous genotype when $\pi_i = 1$, and the dependence between alleles at neighboring markers on the same chromosome according to a Markov model of order 1. To indicate the precise form of these modified emission probabilities, we need to introduce some notation. Let us indicate with $\underline{AA}$ the haplotype consisting of alleles A at marker $i-1$ (succinctly A^{i-1}) and allele A at marker i (succinctly A^i). The population haplotype frequency at the two markers can then be written out as:

$$\begin{array}{ccccc} & & A^i & a^i & \\ & A^{i-1} & p_{\underline{AA}} & p_{\underline{Aa}} & p_{A^{i-1}} \\ & a^{i-1} & p_{\underline{aA}} & p_{\underline{aa}} & p_{a^{i-1}} \\ & & p_{A^i} & p_{a^i} & 1 \end{array} \quad (5)$$

To fix ideas, let's exclude for the moment the possibility of genotyping error and missing data, and consider the case when $\pi_{i-1} = \pi_i = 1$: the genotypes x_{i-1}, x_i are homozygous and represent the alleles at two neighboring markers on the same chromosome, so that given x_{i-1}, the value of x_i is determined by the conditional population frequency of alleles at marker i given the observed allele at marker $i-1$. We then have:

$$e(x_i = AA \mid x_{i-1} = AA,\ \pi_{i-1} = \pi_i = 1) = p_{\underline{AA}}/p_{A^{i-1}}$$
$$e(x_i = aa \mid x_{i-1} = AA,\ \pi_{i-1} = \pi_i = 1) = p_{\underline{Aa}}/p_{A^{i-1}}$$
$$e(x_i = AA \mid x_{i-1} = aa,\ \pi_{i-1} = \pi_i = 1) = p_{\underline{aA}}/p_{a^{i-1}}$$
$$e(x_i = aa \mid x_{i-1} = aa,\ \pi_{i-1} = \pi_i = 1) = p_{\underline{aa}}/p_{a^{i-1}}.$$

We can then introduce a probability ε of genotyping error and κ of missing data. If we assume that, in presence of genotyping error, alleles are independent and observed according to their population frequencies, one obtains the following expressions for the conditional probabilities of the genotypes x_i given $\pi_{i-1} = \pi_i = 1$ (that can be used to define the appropriate $e(x_i \mid x_{i-1}, \pi_i = \pi_{i-1} = 1)$):

$$p(x_i = AA \mid x_{i-1} = AA) =$$
$$\frac{(1-\varepsilon)^2 p_{\underline{AA}} + \varepsilon(1-\varepsilon) p_{A^{i-1}} p_{A^i} \left(p_{A^{i-1}} + p_{A^i}\right) + \varepsilon^2 p^2_{A^{i-1}} p^2_{A^i}}{(1-\varepsilon) p_{A^{i-1}} + \varepsilon p^2_{A^{i-1}}} (1-\kappa)$$
$$p(x_i = Aa \mid x_{i-1} = AA) = 2\varepsilon p_{A^i} p_{a^i} (1-\kappa)$$
$$p(x_i = aa \mid x_{i-1} = AA) =$$
$$\frac{(1-\varepsilon)^2 p_{\underline{aa}} + \varepsilon(1-\varepsilon) p_{A^{i-1}} p_{a^i} \left(p_{A^{i-1}} + p_{a^i}\right) + \varepsilon^2 p^2_{A^{i-1}} p^2_{a^i}}{(1-\varepsilon) p_{A^{i-1}} + \varepsilon p^2_{A^{i-1}}} (1-\kappa)$$
$$p(x_i = AA \mid x_{i-1} = Aa) = \left((1-\varepsilon) p_{A^i} + \varepsilon^2 p^2_{A^i}\right)(1-\kappa)$$
$$p(x_i = Aa \mid x_{i-1} = Aa) = 2\varepsilon p_{A^i} p_{a^i} (1-\kappa)$$
$$p(x_i = aa \mid x_{i-1} = Aa) = \left((1-\varepsilon) p_{a^i} + \varepsilon^2 p^2_{a^i}\right)(1-\kappa)$$
$$p(x_i = - \mid x_{i-1}) = \kappa$$

Replacing A with a in $p(x_i \mid x_{i-1} = AA)$ will generate probablities $p(x_i \mid x_{i-1} = aa)$. Note that $p(x_i \mid x_{i-1} = Aa)$ is actually the marginal distribution of $p(x_i)$, which is a function of allele frequencies of the i-th marker. This will be the case no matter what the values of (π_{i-1}, π_i) are. When $x_{i-1} = -$, the genotype probability distribution of x_i is the marginal.

Let us now consider how the expression above would need to be modified for other values of π_{i-1}, π_i: in the case where $\pi_{i-1} = \pi_i = 0$, there are two haplotypes contributing to the genotype; when $\pi_{i-1} = 0$, $\pi_i = 1$ or $\pi_{i-1} = 1$, $\pi_i = 0$, only one chromosome contributes alleles at both markers. The probabilities $e(x_i \mid x_{i-1}, \pi_{i-1}, \pi_i)$ that can be derived taking into account these observations and the haplotype distribution (5) are given in the appendix.

Given our modified emission probabilities, our model no longer has precisely the structure of a HMM; however, it is still possible to evaluate the probability of the entire genotype sequence X using a recursive algorithm. For analogy with the HMM literature, we refer to forward and backward algorithm to indicate the two recursions that can be defined on our model.

Forward algorithm:

$$\alpha(\pi_i) = P(x_1, \ldots, x_i, \pi_i) = \sum_{\pi_{i-1}} \alpha(\pi_{i-1}) t(\pi_i \mid \pi_{i-1}) e(x_i \mid x_{i-1}, \pi_{i-1}, \pi_i) \tag{6}$$

with initial conditions:

$$\alpha(\pi_1 = 1) = P(x_1 \mid \pi_1 = 1) P(\pi_1 = 1) = e(x_1 \mid \pi_1 = 1)\delta$$
$$\alpha(\pi_1 = 0) = P(x_1 \mid \pi_1 = 0) P(\pi_1 = 1) = e(x_1 \mid \pi_1 = 0)(1-\delta)$$

Backward algorithm:

$$\beta(\pi_i) = P(x_{i+1}, \ldots, x_m \mid x_i, \pi_i) = \sum_{\pi_{i+1}} \beta(\pi_{i+1}) t(\pi_{i+1} \mid \pi_i) e(x_{i+1} \mid x_i, \pi_i, \pi_{i+1}) \tag{7}$$

with initial conditions:

$$\beta(\pi_m = 1) = \beta(\pi_m = 0) = 1$$

With $\alpha()$ and $\beta()$ defined as above, we also obtain other expressions that are typical of HMM:

$$P(X, \pi_i, \pi_{i+1}) = \alpha(\pi_i) \beta(\pi_{i+1}) t(\pi_{i+1} \mid \pi_i) e(x_{i+1} \mid x_i, \pi_i, \pi_{i+1})$$

and

$$P(X, \pi_i) = \alpha(\pi_i) \beta(\pi_i).$$

These similarities with the HMM formulation allow us to calculate efficiently the probability of a genotype sequence X and to reconstruct with a Viterbi style algorithm the unobserved values Π. The portion in the genomes where $\pi_i = 1$ correspond to 'interesting' homozygous segments: that is, when the data is evaluated taking into account allele frequency and linkage disequilibrium (even if in a quite minimal form) these regions appear as IBD or subject to genomic losses. As we have already pointed out, it is impossible to distinguish between these two phenomena on the basis of the genotypes of one individual alone; however, this is not our goal here. We are simply aiming at defining an algorithm that processes genotypes of one individual and highlights interesting homozygous segments. These can be due to IBD status, genomic loss, selection, or even linkage disequilibrium patterns that are not well described by a Markov model of order one, but rather imply much longer range dependence. To see how the described model can be used with this purpose, we need to clarify how we specify the parameter values.

We assume that the values of κ and ε are obtained by prior genotyping studies conducted using the same technology. The collection of two-marker haplotype population frequencies (5) is obtained from the genotype data using a standard EM algorithm. Once these parameters have been fixed, we estimate – for each individual – δ and η using a maximum likelihood approach. The numerical optimization is performed using a gradient algorithm described in the appendix.

Applying the algorithm to the dataset, we are able to identify the regions of non-random homozygosity (NRH)

Table 1. The 5-number summary of length of NRH segments (cM) and overall percentage of the chromosomes covered by NRH segments in the 12 populations

	NRH segment length (cM)				NRH chromosome proportion			
	1st Qu.	median	3rd Qu.	max.	1st Qu.	median	3rd Qu.	max.
azo	0.0000	0.1473	0.4665	23.63	0.0043	0.0129	0.0276	0.2565
fip	0.0000	0.2272	0.5752	39.36	0.0059	0.0148	0.0299	0.5522
ant	0.0563	0.2885	0.6546	24.24	0.0071	0.0162	0.0304	0.4694
ash	0.1057	0.2892	0.6324	7.937	0.0068	0.0142	0.0258	0.1086
erf	0.0327	0.3164	0.6565	20.90	0.0063	0.0153	0.0299	0.2268
nfl	0.1290	0.3164	0.6325	20.97	0.0061	0.0108	0.0222	0.2471
cau	0.1966	0.3621	0.6888	9.186	0.0064	0.0160	0.0270	0.0997
sar	0.1422	0.3661	0.8127	25.10	0.0077	0.0158	0.0319	0.3042
saf	0.1667	0.3687	0.7266	17.00	0.0062	0.0131	0.0231	0.3167
cr	0.1962	0.4219	0.7171	19.54	0.0057	0.0169	0.0326	0.2685
fic	0.2800	0.4845	0.8968	12.14	0.0095	0.0184	0.0334	0.1473
fik	0.2651	0.5168	1.0370	22.90	0.0083	0.0197	0.0431	0.2645

We do not display the minimum values of the NRH segments and overall percentages in the 12 populations in the table because they are all zeros.

in each of the individuals in the 12 populations. Table 1 illustrates some summary statistics across populations. We report the five number summaries of the NRH segments lengths, and the proportion of chromosome 22 on which they span in each individual. (A more accurate description of the distributions of these two quantities is available in the supplementary material). In terms of population comparisons, the Finnish isolate (FIK) and the Finnish cohort sample (which include individuals from isolate) (FIC) show the longest and most frequent NRH segments. This is consistent with the higher values of LD in FIK, as measured both here and in Service et al. [4]. The nature of this analysis is substantially exploratory and hence we do not claim that differences between the distributions illustrated in table 1 are statistically significant. This question can be addressed with appropriate confirmatory techniques; a first step in this direction is provided in the supplementary material.

Once the regions of NRH are reconstructed, it is of interest to try to interpret what genetic process contributed to their formation. The first hypothesis we considered is inbreeding. The Costa Rica population, for which we had careful genealogical records, offered an ideal test case. A total of 6 individuals were known to have positive – albeit small – inbreeding coefficient. Our algorithm estimated a positive δ for each of them, and a likelihood ratio test, whose significance was evaluated with simulations, rejected the hypothesis $\delta = 0$ for all cases. These results testify that our methodology can effectively detect inbreeding. However, the distribution of the estimated values of δ in CR, as well in the other populations, makes it unrealistic to attribute to inbreeding the bulk of the detected NRH segments: there are too many $\delta > 0$ and too large values of δ for these to be due only to previously undetected inbreeding. For example, figure 1 presents some individuals that are homozygous for as much as 30% of their genotypes. This distortion could be due to the fact that we are observing a small fraction of the entire genome of these individuals, but it is also important to explore other possible causes of NRH.

Chromosome 22 is known to harbor one of the most well-studied microdeletions: a deletion on 22q11.2 is known to generate susceptibility to DiGeorge [19] and velocardiofacial [20] syndromes and has been associated to schizophrenia [21]. It was then natural to investigate the possibility that some of the genotyped individuals harbored a deletion in this region. We did not find, however, any evidence of this. A few individuals from the Costa Rica showed a NRH region overlapping 22q11.2, but when we inspected the raw intensity values of the genotyping reaction, we saw no evidence of deletion. Moreover, it is worth emphasizing that the NRH segments we identify are generally quite long: shorter segments as the ones resulting from a microdeletion may not be judged as interesting by our methodology, as their length is entire-

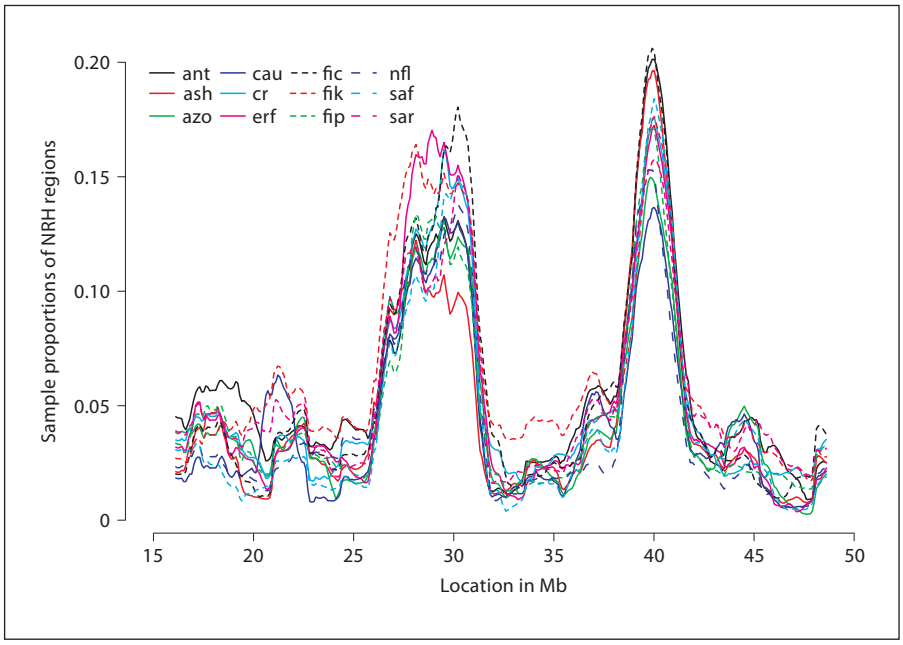

Fig. 6. Localization of NRH regions across populations. For each population, we evaluated the proportion of individuals that exhibit NRH at each SNP. The average of these proportions across a 1.7 Mb sliding window is plotted against the position of the window center, with the x axis representing the length of chromosome 22.

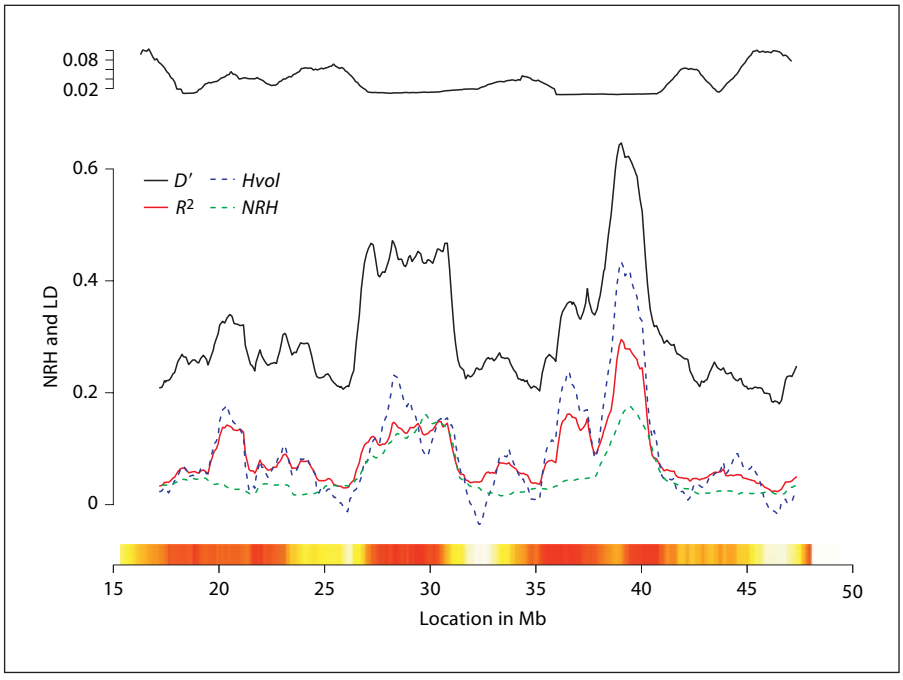

Fig. 7. Relation between NRH regions and patterns of linkage disequilibrium in Costa Rica. *D′*, R^2, *Hvol* and NRH proportions are represented, respectively, with a solid black, a solid red, a broken blue, and a broken green line (curves are obtained as in previous pictures). The heat-colored bar at the bottom displays the known gene density on chromosome 22, as obtained averaging information from the UCSC genome browser within 1.7 Mb sliding windows. On the top of the picture, a smoothed derivative of the genetic map of Chr. 22.

ly comparable to that of random homozygosity segments that are found all over the genome.

Segment of NRH could reflect the action of selection. In such case, we would expect NRH segments to occupy roughly the same genomic location in different individuals. To explore this possibility, we tallied, within each population, the proportion of individuals whose genotypes were interpreted as NRH at each genomic position. The results, presented in figure 6 revealed a consistent pattern across populations: while there are some specific population differences, these may very well be attributed to sample variation and are much less striking than the overall consistency. There are clearly two regions on chr 22 where NRH is prevalent: between 26 and 31 Mb and 38 and 41 Mb. These are comparatively gene rich regions, with, respectively 38 and 36 genes per Mb, versus the 20

genes per Mb of chr 22 as a whole. This is consistent with the hypothesis of selection playing a role in determining these NRH segments. However, there is one more possible reason for the aggregation of these NRH segments that we need to explore.

Figure 7 illustrates the relation between patterns of NRH and linkage disequilibrium. For clarity we have focused on the sample from one population only – Costa Rica – but the results are consistent across populations. As it can be seen, the regions of high density of NRH correspond very well with the regions of high LD. Recall that our methods for the identification of NRH takes into account LD in as much as it can be modeled with a Markov process of order one. Effects of higher order LD are not incorporated in the model and hence could lead to the identification of NRH segments: these would be signatures of long range linkage disequilibrium. Figure 7 clearly seems to suggest this to be the case. Linkage disequilibrium can indeed be caused by the effect of selection, however, it is interesting to note how this pattern correlates well also with variation in recombination fraction across chromosome 22, as captured by the genetic map and illustrated in figure 7. This suggests that the high LD due to low recombination may suffice to explain a large portion of the NRH regions. A similar finding was recently reported by Gibson et al. [22] in an analysis of hap-map data.

4 Conclusion

Homozygosity is a very natural genetic concept. We illustrate here how it can be constructively used to measure linkage disequilibrium, as well as a first pass detection tool for other genetic phenomena.

A number of studies have recently noted the presence of long stretches of homozygous markers in highly densely genotyped individuals. We provide a model and an algorithm to identify such segments characterized by non-random homozygosity. Our analysis of 12 population samples comprising 200 individuals each genotyped at 2,500 markers on chromosome 22 suggests that the vast majority of these NRH segments is due to linkage disequilibrium, that could result from selection effects, but that is actually well correlated with variation in recombination frequency. Caution has to be used, then, when homozygous haplotypes are used to estimate inbreeding or localize disease genes, that an accurate consideration is given to the specific linkage disequilibrium levels.

Acknowledgments

Chiara Sabatti and HuiWang were partially supported by NSF grant DMS0239427 and NIH/NIDOCD grant DC04224. Chiara Sabatti also acknowledges support from ASA/Ames grant NCC2-1364, USPHS grant GM53275, NIH grants R01NS037484 and R01MH049499. Nelson Freimer is partially supported by NIH grants R01NS037484 and R01MH049499. Yuguo Chen is partially supported by NSF grant DMS0503981. We thank two anonymous reviewers for their significant input.

Appendix: The IBD Reconstruction Model and Algorithm

Model with Differential Probabilities of Missing Genotype

In the main text, in the interest of readability, we have considered only one missing rate. However, it is quite reasonable to assume that the probability of a missing genotype depends on the hidden status Π, at least when this indicates genomic alterations. To accommodate this case we have considered a model with differential missing rate κ for genotypes corresponding to $\pi_i = 0$, and τ for genotypes corresponding to $\pi_i = 1$. In the detailed description of our model and algorithm that follows we will use these two differential missing rates. As in the main text, κ is assumed as given (estimated on the base of prior data), while τ is estimated with maximum likelihood.

Modified Emission Probabilities

Below we list under different hidden states, the emission probabilities of the genotypes.

When $\pi_{i-1} = 1$, $\pi_i = 1$, we have

$$p\left(x_i = AA \,|\, x_{i-1} = AA\right) = \frac{(1-\varepsilon)^2 p_{\underline{AA}} + \varepsilon(1-\varepsilon) p_{A^{i-1}}\left(p_{A^{i-1}} + p_{A^i}\right) + \varepsilon^2 p^2_{A^{i-1}} p^2_{A^i}}{(1-\varepsilon) p_{A^{i-1}} + \varepsilon p^2_{A^{i-1}}}(1-\tau)$$

$$p\left(x_i = Aa \,|\, x_{i-1} = AA\right) = 2\varepsilon p_{A^i} p_{a^i}(1-\tau)$$

$$p\left(x_i = aa \,|\, x_{i-1} = AA\right) = \frac{(1-\varepsilon)^2 p_{\underline{Aa}} + \varepsilon(1-\varepsilon) p_{A^{i-1}} p_{a^i}\left(p_{A^{i-1}} + p_{a^i}\right) + \varepsilon^2 p^2_{A^{i-1}} p^2_{a^i}}{(1-\varepsilon) p_{A^{i-1}} + \varepsilon p^2_{A^{i-1}}}(1-\tau)$$

$$p\left(x_i = AA \,|\, x_{i-1} = Aa\right) = \left((1-\varepsilon) p_{A^i} + \varepsilon^2 p^2_{A^i}\right)(1-\tau)$$

$$p\left(x_i = Aa \,|\, x_{i-1} = Aa\right) = 2\varepsilon p_{A^i} p_{a^i}(1-\tau)$$

$$p\left(x_i = aa \,|\, x_{i-1} = Aa\right) = \left((1-\varepsilon) p_{a^i} + \varepsilon^2 p^2_{a^i}\right)(1-\tau)$$

Similarly, when $\pi_{i-1} = 1$, $\pi_i = 0$,

$$p\left(x_i = AA \,|\, x_{i-1} = AA\right) = \frac{(1-\varepsilon)^2 p_{\underline{AA}} p_{A^i} + \varepsilon(1-\varepsilon) p_{A^{i-1}} p^2_{A^i} + \varepsilon p^2_{A^{i-1}} p^2_{A^i}}{(1-\varepsilon) p_{A^{i-1}} + \varepsilon p^2_{A^{i-1}}}(1-\kappa)$$

$$p\left(x_i = Aa \,|\, x_{i-1} = AA\right) =$$
$$\frac{(1-\varepsilon)^2\left(p_{\underline{Aa}}p_{A^i} + p_{\underline{AA}}p_{a^i}\right) + 2\varepsilon(1-\varepsilon)p_{A^{i-1}}p_{A^i}p_{a^i} + 2\varepsilon p^2_{A^{i-1}}p_{A^i}p_{a^i}}{(1-\varepsilon)p_{A^{i-1}} + \varepsilon p^2_{A^{i-1}}}(1-\kappa)$$
$$p\left(x_i = aa \,|\, x_{i-1} = AA\right) =$$
$$\frac{(1-\varepsilon)^2 p_{\underline{Aa}}p_{a^i} + \varepsilon(1-\varepsilon)p_{A^{i-1}}p^2_{a^i} + \varepsilon p^2_{A^{i-1}}p^2_{a^i}}{(1-\varepsilon)p_{A^{i-1}} + \varepsilon p^2_{A^{i-1}}}(1-\kappa)$$
$$p\left(x_i = AA \,|\, x_{i-1} = Aa\right) = p^2_{A^i}(1-\kappa)$$
$$p\left(x_i = Aa \,|\, x_{i-1} = Aa\right) = 2p_{A^i}p_{a^i}(1-\kappa)$$
$$p\left(x_i = aa \,|\, x_{i-1} = Aa\right) = p^2_{a^i}(1-\kappa)$$
$$p\left(x_i = - \,|\, x_{i-1}\right) = \kappa$$

When $\pi_{i-1} = 0,\ \pi_i = 1,$

$$p\left(x_i = AA \,|\, x_{i-1} = AA\right) = \left((1-\varepsilon)^2\frac{p_{\underline{\underline{AA}}}}{p_{A^{i-1}}} + \varepsilon(1-\varepsilon)p_{A^i} + \varepsilon p^2_{A^i}\right)(1-\tau)$$
$$p\left(x_i = Aa \,|\, x_{i-1} = AA\right) = 2\varepsilon p_{A^i}p_{a^i}(1-\tau)$$
$$p\left(x_i = aa \,|\, x_{i-1} = AA\right) = \left((1-\varepsilon)^2\frac{p_{\underline{\underline{Aa}}}}{p_{A^{i-1}}} + \varepsilon(1-\varepsilon)p_{a^i} + \varepsilon p^2_{a^i}\right)(1-\tau)$$
$$p\left(x_i = AA \,|\, x_{i-1} = Aa\right) =$$
$$\left((1-\varepsilon)^2\frac{\left(p_{\underline{AA}}p_{a^{i-1}} + p_{\underline{aA}}p_{A^{i-1}}\right)}{2p_{A^{i-1}}p_{a^{i-1}}} + 2\varepsilon(1-\varepsilon)p_{A^i} + 2\varepsilon p^2_{A^i}\right)(1-\tau)$$
$$p\left(x_i = Aa \,|\, x_{i-1} = Aa\right) = 2\varepsilon p_{A^i}p_{a^i}(1-\tau)$$
$$p\left(x_i = aa \,|\, x_{i-1} = Aa\right) =$$
$$\left((1-\varepsilon)^2\frac{\left(p_{\underline{Aa}}p_{a^{i-1}} + p_{\underline{aa}}p_{A^{i-1}}\right)}{2p_{A^{i-1}}p_{a^{i-1}}} + 2\varepsilon(1-\varepsilon)p_{a^i} + 2\varepsilon p^2_{a^i}\right)(1-\tau)$$
$$p\left(x_i = - \,|\, x_{i-1}\right) = \tau$$

When $\pi_{i-1} = 0,\ \pi_i = 0,$

$$p\left(x_i = AA \,|\, x_{i-1} = AA\right) = \left((1-\varepsilon)^2\frac{p^2_{\underline{\underline{AA}}}}{p^2_{A^{i-1}}} + \varepsilon(2-\varepsilon)p^2_{A^i}\right)(1-\kappa)$$
$$p\left(x_i = Aa \,|\, x_{i-1} = AA\right) = \left((1-\varepsilon)^2\frac{2p_{\underline{\underline{AA}}}p_{\underline{\underline{Aa}}}}{p^2_{A^{i-1}}} + 2\varepsilon(2-\varepsilon)p_{A^i}p_{a^i}\right)(1-\kappa)$$
$$p\left(x_i = aa \,|\, x_{i-1} = AA\right) = \left((1-\varepsilon)^2\frac{p^2_{\underline{\underline{Aa}}}}{p^2_{A^{i-1}}} + \varepsilon(2-\varepsilon)p^2_{a^i}\right)(1-\kappa)$$
$$p\left(x_i = AA \,|\, x_{i-1} = Aa\right) = \left((1-\varepsilon)^2\frac{p_{\underline{AA}}p_{\underline{aA}}}{p_{A^{i-1}}p_{a^{i-1}}} + \varepsilon(2-\varepsilon)p^2_{A^i}\right)(1-\kappa)$$
$$p\left(x_i = Aa \,|\, x_{i+1} = Aa\right) =$$
$$\left((1-\varepsilon)^2\frac{p_{\underline{AA}}p_{\underline{aa}} + p_{\underline{Aa}}p_{\underline{aA}}}{p_{A^{i-1}}p_{a^{i-1}}} + 2\varepsilon(2-\varepsilon)p_{A^i}p_{a^i}\right)(1-\kappa)$$
$$p\left(x_i = aa \,|\, x_{i-1} = Aa\right) = \left((1-\varepsilon)^2\frac{p_{\underline{Aa}}p_{\underline{aa}}}{p_{A^{i-1}}p_{a^{i-1}}} + \varepsilon(2-\varepsilon)p^2_{a^i}\right)(1-\kappa)$$
$$p\left(x_i = - \,|\, x_{i-1}\right) = \kappa$$

MLE and Derivation of the Gradient Algorithm

The dataset we analyze consists in the genotypes of multiple individuals. Depending on the interpretation that one is willing to give to δ and η it may be appropriate to estimate a separate value of these parameters for each individual or a single value for the entire population. In particular, if inbreeding is perceived as the most likely cause of extended segments of homozygosity, it is important to estimate a subject specific value of the two parameters describing the distribution of the unknown states, as every individual will have his/her own inbreeding coefficient. If, instead, the genotypes under analysis come from cancer cell lines, so that the most likely cause of extended homozygous segments is genomic loss, it may be more appropriate to estimate only one value of the parameters describing the loss process across all individuals. The gradient algorithm we describe can be easily adapted to both cases. As specified in the main text, to analyze our data we estimated individual specific values of δ and η. Generally speaking, when one decides to use the model with differential missing rate, the value of τ should be assumed constant across individuals. We now give the details of the gradient algorithm.

When the hidden state $\Pi = (\pi_i)$ is known, the likelihood is

$$P(X) = \sum_{\Pi} P\left(X, \Pi\right) =$$
$$\sum_{\Pi} t(\pi_1)\prod_{i=1}^{m-1} t\left(\pi_{i+1} \,|\, \pi_i\right) e\left(x_1 \,|\, \pi_1\right)\prod_{i=1}^{m-1} e\left(x_{i+1} \,|\, x_i, \pi_i, \pi_{i+1}\right)$$

Note that

$$\frac{\partial P\left(X, \Pi\right)}{\partial t\left(\pi_{i+1} \,|\, \pi_i\right)} = \frac{P\left(X, \Pi\right)}{t\left(\pi_{i+1} \,|\, \pi_i\right)}$$
$$\frac{\partial P\left(X, \Pi\right)}{\partial e\left(x_{i+1} \,|\, x_i, \pi_i, \pi_{i+1}\right)} = \frac{P\left(X, \Pi\right)}{e\left(x_{i+1} \,|\, x_i, \pi_i, \pi_{i+1}\right)}$$

The derivation of the first derivatives of the parameters are similar to the non-LD case:

$$\frac{\partial \log P(X)}{\partial \eta} = \frac{1}{P(X)}\sum_{\Pi}\frac{\partial P\left(X, \Pi\right)}{\partial \eta}$$
$$= \frac{1}{P(X)}\sum_{\Pi}\sum_{i=1}^{m-1}\frac{\partial P\left(X, \Pi\right)}{\partial t\left(\pi_{i+1} \,|\, \pi_i\right)}\frac{\partial t\left(\pi_{i+1} \,|\, \pi_i\right)}{\partial \eta}$$
$$= \frac{1}{P(X)}\sum_{i=1}^{m-1}\sum_{\Pi}\frac{P\left(X, \Pi\right)}{t\left(\pi_{i+1} \,|\, \pi_i\right)}\frac{\partial t\left(\pi_{i+1} \,|\, \pi_i\right)}{\partial \eta}$$

$$= \frac{1}{P(X)} \sum_{i=1}^{m-1} \sum_{\pi_i} \sum_{\pi_{i+1}} \frac{P(X, \pi_i, \pi_{i+1})}{t(\pi_{i+1}|\pi_i)} \frac{\partial t(\pi_{i+1}|\pi_i)}{\partial \eta}$$

$$= \frac{1}{P(X)} \sum_{i=1}^{m-1} \sum_{\pi_i} \sum_{\pi_{i+1}} \alpha(\pi_i)\beta(\pi_{i+1}) e(x_{i+1}|x_i, \pi_i, \pi_{i+1}) \frac{\partial t(\pi_{i+1}|\pi_i)}{\partial \eta}$$

$$\frac{\partial \log P(X)}{\partial \delta} = \frac{1}{P(X)} \sum_{\Pi} \frac{\partial P(X, \Pi)}{\partial \delta}$$

$$= \frac{1}{P(X)} \left(\sum_{\Pi} \frac{\partial P(X, \Pi)}{\partial t(\pi_1)} \frac{\partial t(\pi_1)}{\partial \delta} + \sum_{\Pi} \sum_{i=1}^{m-1} \frac{\partial P(X, \Pi)}{\partial t(\pi_{i+1}|\pi_i)} \frac{\partial t(\pi_{i+1}|\pi_i)}{\partial \delta} \right)$$

$$= \frac{1}{P(X)} \left(\sum_{\Pi} \frac{P(X, \Pi)}{t(\pi_1)} \frac{\partial t(\pi_1)}{\partial \delta} + \sum_{i=1}^{m-1} \sum_{\Pi} \frac{P(X, \Pi)}{t(\pi_{i+1}|\pi_i)} \frac{\partial t(\pi_{i+1}|\pi_i)}{\partial \delta} \right)$$

$$= \frac{1}{P(X)} \left(\sum_{\pi_1} \frac{P(X, \pi_1)}{t(\pi_1)} \frac{\partial t(\pi_1)}{\partial \delta} + \sum_{i=1}^{m-1} \sum_{\pi_i} \sum_{\pi_{i+1}} \frac{P(X, \pi_i, \pi_{i+1})}{t(\pi_{i+1}|\pi_i)} \frac{\partial t(\pi_{i+1}|\pi_i)}{\partial \delta} \right)$$

$$= \frac{1}{P(X)} \left(\sum_{\pi_1} \frac{\alpha(\pi_1)\beta(\pi_1)}{t(\pi_1)} \frac{\partial t(\pi_1)}{\partial \delta} + \sum_{i=1}^{m-1} \sum_{\pi_i} \sum_{\pi_{i+1}} \alpha(\pi_i)\beta(\pi_{i+1}) e(x_{i+1}|x_i, \pi_i, \pi_{i+1}) \frac{\partial t(\pi_{i+1}|\pi_i)}{\partial \delta} \right)$$

The gradient algorithm to maximize δ, η, τ is as following, t is the time, and λ is the step size.

$$\delta^{(t+1)} = \delta^{(t)} + \lambda_\delta \frac{\partial \log P(X)}{\partial \delta}_{|\delta=\delta^{(t)}}$$

$$\eta^{(t+1)} = \eta^{(t)} + \lambda_\eta \frac{\partial \log P(X)}{\partial \eta}_{|\eta=\eta^{(t)}}$$

$$\tau^{(t+1)} = \tau^{(t)} + \lambda_\tau \frac{\partial \log P(X)}{\partial \tau}_{|\tau=\tau^{(t)}}$$

These partial derivatives are additive, and the gradient algorithm can easily be extended to cases with multiple sequences where the δ and η are the same for the entire population. Suppose we have $j = 1, \ldots, n$ independent sequences, each with length m_j, and we denote each of them by X_j. Then the gradient algorithm becomes:

$$\frac{\partial \log P(X)}{\partial \eta} = \sum_j \frac{1}{P(X_j)} \left(\sum_{i=1}^{m_j-1} \sum_{\pi_{ji}} \sum_{\pi_{j,i+1}} \alpha(\pi_{ji})\beta(\pi_{j,i+1}) e(x_{j,i+1}|\pi_{j,i+1}, \pi_{ji}, x_{ji}) \frac{\partial t(\pi_{j,i+1}|\pi_{ji})}{\partial \eta} \right)$$

$$\frac{\partial \log P(X)}{\partial \delta} = \sum_j \frac{1}{P(X_j)} \left(\sum_{\pi_{j1}} \alpha(\pi_{j1})\beta(\pi_{j1}) \frac{\partial t(\pi_{j1})}{\partial \delta} + \sum_{i=1}^{m_j-1} \sum_{\pi_{ji}} \sum_{\pi_{j,i+1}} \alpha(\pi_{ji})\beta(\pi_{j,i+1}) e(x_{j,i+1}|x_{ji}, \pi_{ji}, \pi_{j,i+1}) \frac{\partial t(\pi_{j,i+1}|\pi_i)}{\partial \delta} \right)$$

$$\frac{\partial \log P(X)}{\partial \tau} = \sum_j \frac{1}{P(X_j)} \left(\sum_{\pi_{j1}} \alpha(\pi_{j1})\beta(\pi_{j1}) \frac{\partial e(x_{j1}|\pi_{j1})}{\partial \tau} + \sum_{i=1}^{m_j-1} \sum_{\pi_{ji}} \sum_{\pi_{j,i+1}} \alpha(\pi_{ji})\beta(\pi_{j,i+1}) t(\pi_{j,i+1}|\pi_{ji}) \frac{\partial e(x_{j,i+1}|x_{ji}, \pi_{ji}, \pi_{j,i+1})}{\partial \tau} \right)$$

References

1 The International HapMap Consortium: A haplotype map of the human genome. Nature 2005;437:1299–1320.

2 Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: Whole-genome patterns of common DNA variation in three human populations. Science 2005;307:1072–1079.

3 Thomas DC, Haile RW, Duggan D: Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 2005;77:337–345.

4 Service S, DeYoung J, Karayiorgou M, Louw Roos J, Pretorious H, Bedoya G, Ospina J, Ruiz-Linares A, Macedo A, Palha J, Heutink P, Aulchenko Y, Oostra B, van Duijn C, Jarvelin M, Varilo T, Peddle L, Rahman P, Piras G, Monne M, Murray S, Galver L, Peltonen L, Sabatti C, Collins A, Freimer N: Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nature Genetics 2006;38:556–560.

5 Sabatti C, Risch N: Homozygosity and linkage disequilibrium. Genetics 2002;160:1707–1719.

6 Sabatti C: Measuring dependence with volume tests. The American Statistician 2002;50:191–195.

7 Chen Y, Lin C, Sabatti C: Volume measures for linkage disequilibrium. Submitted, 2006.

8 Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES: Detecting recent positive selection in the human genome from haplotype structure. Nature 2002;419:832–837.

9 Leutenegger A, Prum B, Genin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson E: Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 2003;73:516–523.

10 Newton M, Gould M, Reznikoff C, Haag J: On the statistical analysis of allelic-loss data. Stat Med 2000;17:1425–1445.
11 Ohta T: Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. Genet Res 1980;36:181–197.
12 Hotelling H: Tubes and spheres in n-spaces, and a class of statistical problems. American Journal of Mathematics 1939;61:440–460.
13 Diaconis P, Efron B: Testing for independence in a two-way table: new interpretations of the Chi-square statistics. The Annals of Statistics 1985;13:845–874.
14 Woods CG, Cox J, Springell K, Hampshire DJ, Mohamed MD, McKibbin M, Stern R, Raymond FL, Sandford R, Sharif SM, Karbani G, Ahmed M, Bond J, Clayton D, Inglehearn CF: Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. Am J Hum Genet 2006;78:889–896.
15 Wang H, Lee Y, Nelson S, Sabatti C: Inferring genomic loss and location of tumor suppressor genes from high density genotypes. Journal of the French Statistical Society 2005; 146:153–171.
16 Iafrate A, Feuk L, Rivera M, Listewnik M, Donahoe P, Qi Y, Scherer SW, Lee C: Detection of large-scale variation in the human genome. Nat Genet 2004;36:949–951.
17 Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam T, Trask B, Patterson N, Zetterberg A, Wigler M: Large-scale copy number polymorphism in the human genome. Science 2004;305:525–528.
18 Newton M, Lee Y: Inferring the location and effect of tumor suppressor genes by instabilityselection modeling of allelic-loss data. Biometrics 2000;56:1088–1097.
19 de la Chapelle A, Herva R, Koivisto M, Aula P: A deletion in chromosome 22 can cause DiGeorge syndrome. Hum Genet 1981;57: 253–256.
20 Scambler PJ, Kelly D, Lindsay E, Williamson R, Goldberg R, Shprintzen R, Wilson DI, Goodship JA, Cross IE, Burn J: Velo-cardio-facial syndrome associated with chromosome 22 deletions encompassing the DiGeorge locus. Lancet 1992;339:1138–1139.
21 Karayiorgou M, Morris MA, Morrow B, Shprintzen RJ, Goldberg R, Borrow J, Gos A, Nestadt G, Wolyniec PS, Lasseter VK, Eisen H, Childs B, Kazazian HH, Kucherlapati R, Antonarakis SE, Pulver AE, Housman DE: Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. Proc Nat Acad Sci 1995;92:7612–7616.
22 Gibson J, Morton N, Collins A: Extended tracts of homozygosity in outbred human populations. Human Molecular Genetics 2006;15:789–795.

View publication stats