BIOINFORMATICS

Vol. 20 no. 0 2004, pages 1–6
doi:10.1093/bioinformatics/bth148

Haplotypes histories as pathways of recombinations

Nadia El-Mabrouk[1,*] and Damian Labuda[2]

[1]Département d'informatique et de recherche opérationnelle, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, QC, Canada H3C 3J7 and [2]Centre de recherche, Hopital Sainte-Justine, Département de Pédiatrie, Université de Montréal, 3175 Côte-Sainte-Catherine, Montréal, QC, Canada H3T 1C5

Received on April 21, 2003; revised on September 8, 2003; accepted on September 9, 2003

ABSTRACT

Motivation: The diversity of a haplotype, represented as a string of polymorphic sites along a DNA sequence, increases exponentially with the number of sites if recombinations are taking place. Reconstructing the history of recombinations compared with that of the polymorphic sites is thus extremely difficult. However, in the human genome, because of the relatively simple pattern of haplotype diversity dominated by a few ancestral haplotypes, the complexity of the recombinational network can be reduced, thus making its reconstruction feasible. We focus on the problem of inferring the recombination pathways starting with putative ancestral haplotypes and leading to new rare recombinant haplotypes.

Results: We describe classes of recombinations that represent the whole set of minimal recombination pathways leading to a new haplotype. We present an $O(n^2)$ algorithm that outputs such representative recombination pathways. We apply it to haplotypes of the 8 kb dystrophin gene segment *dys44*.

Availability: A software implementing the algorithm and some other extentions has been developed on a Java platform (JDK 1.3.1). It is freely available at http://www.iro.umontreal.ca/~mabrouk/

Contact: mabrouk@iro.umontreal.ca

1 INTRODUCTION

The diversity of the human genome, seen across individuals in populations, is quantal rather than continuous. Mutations lead to discrete new alleles, whereas recombinations, redistributing these alleles among homologous segments, create discrete haplotypes. Haplotypes can thus be represented as strings of alleles at polymorphic sites along DNA segments. Historically, the focus was on the mutation process, but recombinations, influencing allelic relations between adjacent variant sites, get increasing attention. The effect of the latter can be described quantitatively in term of linkage disequilibrium or qualitatively by cataloguing the underlying haplotypes.

The evolutionary history of human species can be represented by a long stationary phase characterized by an effective population size of about 10 000 followed by rapid, almost starlike, population expansion (Heyer *et al*., 2001), well reflected in mitochondrial DNA (Di Rienzo and Wilson, 1991; Rogers and Harpending, 1992) and microsatellite data (Gonser *et al*., 2000). This is also consistent with a number of studies examining nuclear loci where a relatively simple pattern of diversity is observed: a few frequent haplotype variants dominate over a flat distribution of rare haplotypes due to recent recombinations or new mutations (Fullerton *et al*., 2000; Harding *et al*., 1997; Jaruzelska *et al*., 1999; Kaessmann *et al*., 1999; Labuda *et al*., 2000). The dominating common haplotypes are likely to represent the founder ancestral haplotypes issued either from the early stationary phase or from subsequent founder effects following expansion. New haplotypes derived through crossovers or by mutation from common haplotypes are usually rare and have a limited geographic distribution.

Prior work on recombination has largely focused on statistical tests estimating recombination events (Hudson and Kaplan, 1985; Myers and Griffiths, 2002), or on reconstructing the coalescent in the presence of recombination (Griffiths and Marjoram, 1996; Wiuf and Hein, 1999a,b). Hein (1990, 1993) was the first to consider designing algorithms to reconstruct the history of a genomic segment with recombination. Recently, there is a renewed interest in the analysis of haplotype diversity and the underlying recombinations, based on different theoretical evolutionary models (Kececioglu and Gusfield, 1998; Ukkonen, 2002; Wang *et al*., 2001; Wu and Gu, 2001).

In this paper, we focus on the problem of reconstructing the most parsimonious recombination pathways from the putative ancestral frequent to new rare haplotypes, thus reconstructing their plausible genealogies. It is the first step in the attempt to establish a recombination network that explains the observed haplotype diversity. In Section 2, we formalize the problem and describe various, well-defined classes of recombinations that are representative of the whole set of minimal recombination pathways. In Section 3, we present

*To whom correspondence should be addressed.

an $O(n^2)$ algorithm that outputs such a set of representative recombination pathways. It has been implemented as a tool with a graphical interface. In Section 4, we apply our tool to haplotypes comprising 35 segregating sites from the 8 kb dystrophin gene segment *dys44* described in Labuda *et al*. (2000) and Zietkiewicz *et al*. (2003).

The essential proofs can be consulted at http://www.iro.umontreal.ca/~mabrouk/

2 METHODS

2.1 Formalizing the problem

A haplotype of n sites is a string of nucleotides of size n. It models a chromosomal segment with n polymorphisms due to nucleotide substitutions, termed as single nucleotide polymorphisms (SNPs). SNPs are usually bi-allelic such that in a population only two nucleotides are observed at each site: the ancestral and the new (derived) allele.

A recombination between two haplotypes X and Y can be modeled as an operation that breaks and exchanges the opposite parts of X and Y. That is, it is an operation of the form: $X^5X^3, Y^5Y^3 \longrightarrow X^5Y^3, Y^5X^3$ where $X = X^5X^3$, $Y = Y^5Y^3$ (5 and 3 denote the $5'$- and $3'$-terminal segments of X and Y) and X^5, Y^5, as well as X^3, Y^3, have the same length.

In the human model considered here, only one of the resulting haplotypes is transmitted. Therefore, a recombination can be represented as $X, Y \longrightarrow Z$, where Z is a recombinant.

Let $\mathcal{C} = \{C_1, \ldots, C_h\}$ be a set of common haplotypes, and R be a new recombinant one. The problem is to find a minimal recombination pathway (minimal series of recombinations) generating R from a subset $\mathcal{C}_R$ of $\mathcal{C}$. A recombination $X, Y \longrightarrow Z$ in such a pathway is such that: X and/or Y is in $\mathcal{C}_R$, or is generated from previous recombinations of elements of $\mathcal{C}_R$.

We say that R is allelic with $\mathcal{C}$ if and only if, for any position i in R, there exists a haplotype C_k of $\mathcal{C}$ such that $R[i] = C_k[i]$, where, for any haplotype X, $X[i]$ denotes the i-th element of X. For example, the haplotype:

$$R = \texttt{C A C T T G A A C G}$$

allelic with $\mathcal{C} = \{C_1, C_2, C_3, C_4, C_5\}$:

$$C_1 = \texttt{A C G T C T G A T T}$$
$$C_2 = \texttt{C A G A T G G A C G}$$
$$C_3 = \texttt{C C G A T G G C C G}$$
$$C_4 = \texttt{A A C T T T G A C T}$$
$$C_5 = \texttt{A C C T C G A A T G}$$

If R is not allelic with $\mathcal{C}$, then R cannot be generated from $\mathcal{C}$ by recombination. Non-allelic sites require new mutations. Conversely, it is evident that if R is allelic with $\mathcal{C}$, then R can be generated from $\mathcal{C}$ by a series of recombinations.

Table 1. The set HAP corresponding to the haplotypes $\mathcal{C}$ of the last section

H_k	pos 1	2	3	4	5	6	7	8	9	10
HAP_1	0	0	0	1	0	0	0	1	0	0
HAP_2	1	1	0	0	1	1	0	1	1	1
HAP_3	1	0	0	0	1	1	0	0	1	1
HAP_4	0	1	1	1	1	0	0	1	1	0
HAP_5	0	0	1	1	0	1	1	1	0	1

2.2 Haplotypes recoding

To simplify the ensuing algorithmic developments, we model haplotypes as binary strings of 0s and 1s, and reformulate the problem as one of generating the unitary haplotype, i.e. the haplotype H such that $H[i] = 1$ for any i. To do so, we recode the haplotypes by computing the set HAP $= \{\text{HAP}_1, \ldots, \text{HAP}_h\}$ such that for any k and any position pos, $\text{HAP}_k[\text{pos}] = 1$ if and only if $R[\text{pos}] = C_k[\text{pos}]$ (Table 1).

There is a one-to-one correspondence between the recombination pathways generating R from $\mathcal{C}$ and those generating the unitary haplotype H from HAP.

2.3 Canonical pathways

A canonical pathway generating H from HAP is a sequence of recombinations that does not contain any event on two new (i.e. R) haplotypes. Namely, it is a pathway of the form:

$$\begin{array}{l} H_1,\ H_2 \longrightarrow^{r_1} R_2 \\ R_2,\ H_3 \longrightarrow^{r_2} R_3 \\ \quad\vdots \qquad\quad \vdots \\ R_p,\ H_{p+1} \longrightarrow^{r_p} H \end{array}$$

where $\{H_1, \ldots, H_{p+1}\} \subset$ HAP and $R_2, \ldots, R_p$ are new or extinct (i.e. not seen in the population sample) haplotypes. The ordered series $(H_1, \ldots, H_p, H_{p+1})$ of all haplotypes of HAP appearing in the pathway is called its associated haplotype table.

The reason behind considering canonical recombination pathways is that a recombination between two common or between a common and a new rare haplotype is much more likely than a recombination of two new rare haplotypes. Moreover, any pathway generating R from $\mathcal{C}$ can be 'reduced' to a canonical pathway (Lemma 1) such that we can limit our considerations to canonical pathways only, as illustrated in Figure 1.

By reordering a pathway of haplotype table T we mean creating a new pathway with a haplotype table being a permutation of T.

LEMMA 1. *Any minimal pathway generating R from $\mathcal{C}$ can be reordered into a canonical pathway.*

$$\begin{array}{rcl} \mathrm{HAP}_2 = \mathbf{1\,1} \mid 0\,0\,1\,1\,0\,1\,1\,1 & , & \mathrm{HAP}_4 = 0\,1 \mid \mathbf{1\,1\,1}\,0\,0\,1\,1\,0 \longrightarrow^{r_1} R'_2 = \mathbf{1\,1\,1\,1\,1}\,0\,0\,1\,1\,0 \\ R'_2 = \mathbf{1\,1\,1\,1\,1} \mid 0\,0\,1\,1\,0 & , & \mathrm{HAP}_5 = 0\,0\,1\,1\,0 \mid \mathbf{1\,1\,1}\,0\,1 \longrightarrow^{r_2} R'_3 = \mathbf{1\,1\,1\,1\,1\,1\,1\,1}\,0\,1 \\ R'_3 = \mathbf{1\,1\,1\,1\,1\,1\,1\,1} \mid 0\,1 & , & \mathrm{HAP}_2 = 1\,1\;\;0\,0\,1\,1\,0\,1 \mid \mathbf{1\,1} \longrightarrow^{r_3} H = \mathbf{1\,1\,1\,1\,1\,1\,1\,1\,1\,1} \end{array}$$

Fig. 1. A canonical recombination pathway generating H from the set HAP of Table 1.

2.4 Greedy pathways

We call a strip the maximal substring of 1s in a haplotype, and the strip prefix of a haplotype its longest prefix of 1s. A recombination $r_i : R_i, H_{i+1} \longrightarrow R_{i+1}$ is said to be greedy iff R_{i+1} has a prefix inherited from R_i, and the strip prefix of R_{i+1} is longer than the strip prefix of R_i. In other words, a greedy recombination creates a new haplotype with the largest possible strip prefix. Each recombination shown in Figure 1 is greedy whereas the following recombination is not greedy:

$$0\,1 \mid 1\,1\,1\,0\,0\,1\,1\,0,\;\; 1\,1 \mid 0\,0\,1\,1\,0\,1\,1\,1$$
$$\longrightarrow 0\,1\,0\,0\,1\,1\,0\,1\,1\,1.$$

The problem is reduced to greedy pathways as follows.

LEMMA 2. *A minimal canonical pathway can be reordered into a greedy pathway.*

From Lemmas 1 and 2, if one is interested in all the haplotype sets leading to H, rather than in the precise order of haplotypes, then it is sufficient to generate the set of all minimal canonical pathways of perfectly greedy recombinations. This is reasonable as there are no criteria that allow one particular table of haplotypes, among all those corresponding to the same set, to be favored.

2.5 Perfectly greedy recombinations

Two haplotypes can recombine at different positions yet leading to the same resulting haplotype. Two recombinations are said to be equivalent if they involve the same parental haplotypes and give rise to the same daughter haplotype. For example, the following recombination r'_1 is equivalent to the first recombination r'_1 of Figure 1:

$$\mathrm{HAP}_2 = \mathbf{1}|\;1\,0\,0\,1\,1\,0\,1\,1\,1,$$
$$\mathrm{HAP}_4 = 0\,|\mathbf{1\,1\,1\,1}\,0\,0\,1\,1\,0$$
$$\longrightarrow^{r'_1}\;\; R'_2 = \mathbf{1\,1\,1\,1\,1}\,0\,0\,1\,1\,0.$$

Among all possible equivalent recombinations, we only consider the perfectly greedy ones that cut the strips at their last position. For example, among r_1 and r'_1, r_1 is the one chosen. This simplifies the presentation. However, all equivalent recombinations can be generated as well, as offered by the software provided: one option outputs only perfectly greedy recombinations, the other all greedy recombinations.

3 ALGORITHM

In the following, a solution is a haplotype table associated to a minimal pathway of min perfectly greedy recombinations.

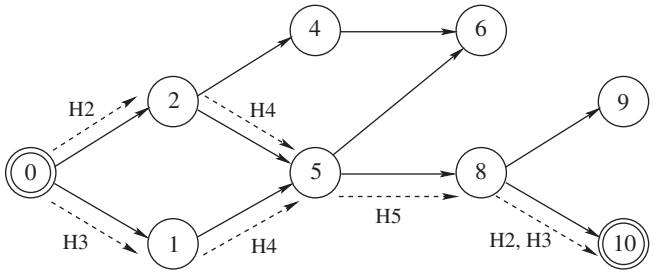

Fig. 2. The graph obtained by Construct-graph for the set of haplotypes listed in Table 1. There are two paths: $P_1 = (0, 1, 5, 8, 10)$ and $P_2 = (0, 2, 5, 8, 10)$. Each edge of these paths is labeled by the set of associated haplotypes. The two solutions that are associated to P_1 are (H_3, H_4, H_5, H_2) and (H_3, H_4, H_5, H_3), and the two associated to P_2 are (H_2, H_4, H_5, H_2) and (H_2, H_4, H_5, H_3).

The algorithm is based on an oriented graph $G(V, E)$ containing an initial vertex and a final vertex. Each vertex is labeled with a position (the site number, from 0 to n). The initial vertex has the default label 0, and the final one is labeled n. As two different vertices are labeled differently, we do not distinguish between a vertex and its label. A path is a sequence of vertices beginning with vertex 0, ending with vertex n and connected by edges of the graph. Roughly speaking, an edge represents a set of perfectly greedy recombinations, and a path a set of solutions. More precisely, the edge $(0, v_1)$ indicates that the first haplotype to be considered should have a strip beginning at position 1 and ending at v_1. Each following edge (v_i, v_{i+1}) indicates that the next recombination should be performed on a common haplotype with a strip beginning at or before position v_i, and ending at position v_{i+1}. We label each edge (v_i, v_{i+1}) by such associated haplotypes.

The graph is constructed in a breadth-first manner. We begin by constructing the initial vertex pos $= 0$. Then, for each position corresponding to the end of a strip begining at or before position pos $+ 1$, if it does not yet correspond to a vertex of the graph, we construct it and we link it to pos by a new edge directed from pos to that vertex. These new vertices are said to be of level one, because they are reachable from 0 by a path of one edge. We repeat the procedure with each new vertex of level one, and construct vertices of level 2. We stop the procedure as soon as the terminal vertex is added to the graph, and all the vertices of the preceding level are treated (Fig. 2).

Please check the citation to Figure 2.

PROPOSITION 1. *A haplotype table is a solution if and only if it is associated to a path of $G(V, E)$.*

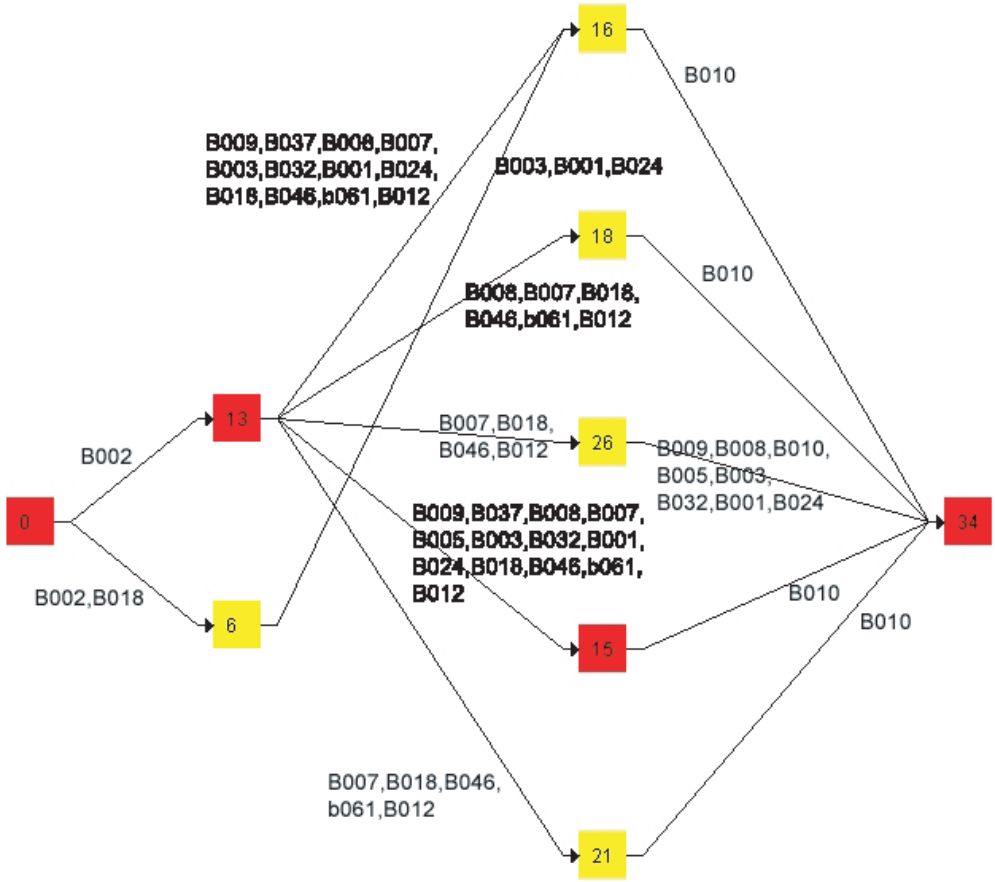

Fig. 3. The software output for the haplotype B022. A recombination path is obtained by choosing a path from source to sink (here from 0 to 34), and one haplotype from the haplotype set labeling each edge. In this output, the best path (red labels) has been chosen on the basis of the crossover probability at each site.

Complexity Let h be the number of haplotypes of HAP and n be the size of each haplotype. The preprocessing required to find the strip positions of all haplotypes can be done by traversing, one after the other, each column of the table representing HAP. Therefore, this step has a time complexity $O(nh)$. The graph construction can then be done in a time proportional to the total size of the graph, i.e. $O(n)$. Therefore, the time complexity of the algorithm is $O(nh)$.

4 AN APPLICATION

Labuda *et al*. (2000) analyzed haplotypes comprising 34 polymorphisms from the *dys44* segment of the dystrophin gene (see also Zietkiewicz *et al*., 2003). In the genealogical reconstructions involving recombinations of the most common haplotypes, they were able to derive rare and presumably young non-African haplotypes through at most two recombination events. In contrast, the haplotypes found only in the sub-Saharan Africans could not be simply related to the set of common frequent haplotypes (e.g. `B022, B047, b059`). Here, we show plausible pathways for B022 starting from a parental sample of the 15 most frequent haplotypes that occur in the African sample (Fig. 3). Figure 4 shows one particular path of this network.

However, not all of these pathways are equally probable given the genetic distances between sites (i.e. recombination probabilities) and/or the population frequencies of the parental haplotypes involved. Various approaches can be considered to choose the most likely pathway among the ones found:

(1) Minimize the number of different haplotypes in an optimal recombination pathway.

(2) Weight each recombination by its crossover probability.

   (a) An easy way to compute such a probability is to consider the nucleotide distance between the two sites flanking the crossover. For example, 417 nt separate sites 13 and 14, while only 258 nt separate sites 6 and 7. Therefore, a recombination within a segment following position 13 is 417/258 more probable than that within a segment following position 6, provided the same recombination rate per nucleotide. Thus, in Figure 3, the path

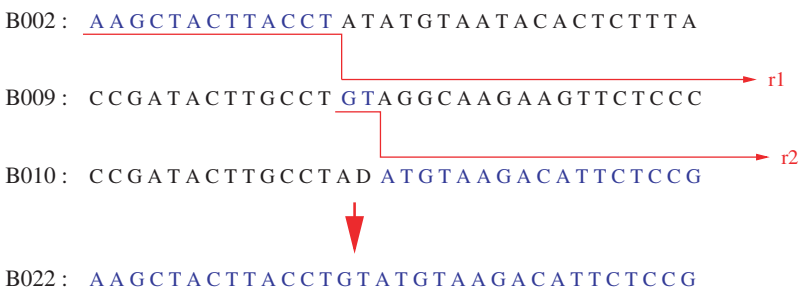

Fig. 4. The recombination pathway (r_1, r_2) corresponding to the red path of Figure 3 and its haplotype table (B002, B009, B010)

(0, 13, 15, 34) (red labels) is more likely than (0, 6, 15, 34). More appropriate would be to use genetic rather than physical distances, once this information is available.

(b) Weight each common haplotype by its frequency in the population. Choose a recombination pathway through a path of maximal length, where the length of a path is the sum of weights of the common haplotypes of the pathway.

The software allows to choose one of the above criteria to select the most appropriate pathway among all possible ones. However, other statistical considerations can be used to weight the edges of the graph, without any modification to the algorithm.

5 DISCUSSION

Our algorithm represents the first attempt to establish a network of recombinations reconnecting observed haplotypes of a locus to explain their diversity. The output is a number of pathways showing a representative set of all the most parsimonious ways to derive a new haplotype from a set of common ones. Here, a recombination pathway is characterized only by the set of known haplotypes that participate in these recombinations. If the likelihood of a recombination pathway depends only on the frequency and/or geographical distribution of these observed haplotypes, this restriction is made without any loss of generality. However, if the unequal probability of crossover-sites along the haplotype plays a major role in the evaluation of different solutions, then our algorithm does not guarantee finding of the most likely recombination pathway. In that case, we have to output all possible recombination pathways, not only the perfectly greedy and canonical ones. Removing the 'perfectly greedy' condition is straightforward and has been implemented in the accompanying software. However, removing the 'canonical' and 'greedy' conditions gives rise to a much larger set of possible solutions, where all possible crossover-sites are considered.

Intermediate haplotypes involved in solutions produced by the algorithm are ignored. Such haplotypes can be considered as rare ones that are missing in the data set, and could be eventually added to the set of ancestral haplotypes for subsequent analysis. Finally, it should be emphasized that in this model only recombinations are considered. However, over small distances, double recombinants can be produced by gene conversion as well (Andolfatto and Nordborg, 1998; Przeworski and Wall, 2001). We envisage generalizing the algorithm to encompass evolutionary scenarios that involve both simple recombination and gene conversion.

ACKNOWLEDGEMENTS

We are grateful to Machavaram Sreeram for the implementation of the software tool. N.E.M. is supported by the Natural Sciences and Engineering Research Council of Canada, and the 'Fonds québécois de recherche sur la nature et les technologies'. D.L. is supported by the Canadian Institute of Health Research (grant MOP-12782) and the Quebec Network of Applied Genetics of the 'Fonds de Recherche en Santé du Québec'.

REFERENCES

Andolfatto,P. and Nordborg,M. (1998) The effect of gene conversion on intralocus associations. *Genetics*, **148**, 1397–1399.

Di Rienzo,A. and Wilson,A. (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl Acad. Sci., USA*, **88**, 1597–601.

Fullerton,S., Clark,A., Weiss,K., Nickerson,D., Taylor,S., Stengard,J., Salomaa,V., Vartiainen,E., Perola,M., Boerwinkle,E. and Sing,C. (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Ann. J. Hum. Genet.*, **67**, 881–900.

Gonser,R., Donnelly,P., Nicholson,G. and Di Rienzo,A. (2000) Microsatellite mutations and inferences about human demography. *Genetics*, **154**, 1793–1807.

Griffiths,R. and Marjoram,P. (1996) Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, **3**, 479–502.

Harding,R., Fullerton,S., Griffiths,R., Bond,J., Cox,M., Schneider,J., Moulin,D. and Clegg,J., (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Ann. J. Hum. Genet.*, **60**, 772–789.

Hein,J. (1990) Reconstructing the evolution of sequences subject to recombination using parsimony. *Math. Biosci.*, **98**, 185–200.

Hein,J. (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, **36**, 396–405.

Heyer,E., Zietkiewicz,E., Rochowski,A., Yotova,V., Puymirat,J. and Labuda,D. (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am. J. Hum. Genet.*, **69**, 1113–1126.

Hudson,R. and Kaplan,N. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **111**, 147–164.

Jaruzelska,J., Zietkiewicz,E., Batzer,M., Cole,D., Moisan,J., Scozzari,R., Tavaré,S. and Labuda,D. (1999) Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. *Genetics*, **152**, 1091–1101.

Kaessmann,H., Heissig,F., Von Haeseler,A. and Paabo,S. (1999) DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.*, **22**, 78–81.

Kececioglu,J. and Gusfield,D. (1998) Reconstructing a history of recombinations from a set of sequences. *Discrete Appl. Math.*, **88**, 239–260.

Labuda,D., Zietkiewicz,E. and Yotova,V. (2000) Archaic lineages in the history of modern humans. *Genetics*, **156**, 799–808.

Myers,S. and Griffiths,R. (2002) Bounds on the minimum number of recombination in a sample history. *Genetics*, **163**, 375–394.

Przeworski,M. and Wall,J. (2001) Why is there so little intragenic linkage disequilibrium in humans? *Genet. Res.*, **77**, 143–151.

Rogers,A. and Harpending,H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.*, **9**, 552–569.

Ukkonen,E. (2002) Finding founder sequences from a set of recombinants. *2nd International Workshop, Algorithms in Bioinformatics* (*WABI*), LNCS, Vol. 2452. Springer, pp. 277–286.

Wang,L., Zhang,K. and Zhang,L. (2001) Perfect phylogenetic networks with recombination. *J. Comput. Biol.*, **8**, 69–78.

Wiuf,C. and Hein,J. (1999a) The ancestry of a sample of sequences subject to recombination. *Genetics*, **151**, 1217–1228.

Wiuf,C. and Hein,J. (1999b) Recombination as a point process along sequences. *Theor. Popul. Biol.*, **55**, 248–259.

Wu,S. and Gu,X. (2001) A greedy algorithm for optimal recombination. In *COCOON*, LNCS, Vol. 2001, pp. 87–90.

Zietkiewicz,E., Yotova,V., Gehl,D., Wambach,T., Arrieta,I., Batzer,M., Cole,D., Hechtman,P., Kaplan,F., Modiano,D., Moisan,J., Michalski,R. and Labuda,D. (2003) Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human's diversity. *Ann. J. Hum. Genet.* (in press).

Please provide publisher details

Please update