Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mathematical Biosciences 185 (2003) 205–216
Mathematical Biosciences
www.elsevier.com/locate/mbs

Some notes on the combinatorial properties of haplotype tagging

Carsten Wiuf [a,*], Zoë Laidlaw [b,c], Michael P.H. Stumpf [c]

[a] *Variagenics, 60 Hampshire Street, Cambridge, MA 02139, USA*
[b] *Department of History, University of Sheffield, Sheffield S10 1HR, UK*
[c] *Department of Biology, University College London, London WC1E 6BT, UK*

Received 15 July 2002; accepted 16 June 2003

Abstract

Tagging haplotypes with a small number of genetic markers is becoming an increasingly interesting and important problem. Surprisingly little work has been done to characterize the mathematical framework of this problem. In this paper we present a mathematical frame, based on Boolean algebras, that adequately describe the structure of a set of genetic bi-allelic markers and the corresponding set of haplotypes. We derive a number of results that relate the number of markers required to tag a set of haplotypes to the set of markers themselves.

Keywords: Haplotype tagging; SNPs; Block recombination

1. Introduction

Understanding of human genetic variation will be key for unraveling the genetic factors involved in many common complex diseases such as asthma, common cardiovascular diseases, cancer and most infectious diseases [1,2]. At the molecular level human genetic variation is determined by the forces of mutation and recombination, while on the level of the population random genetic drift, selection and demographic factors influence the distribution of genetic variants. Together they interact to determine the frequency of a genetic variant in a population.

* Corresponding author. Present address: Department of Computer Science, Bioinformatics Research Center (BiRC), University of Aarhus, Ny Munkegade, Bldg. 540, 8000 Aarhus C, Denmark. Fax: +45-89 423 077.
E-mail address: wiuf@daimi.au.dk (C. Wiuf).

0025-5564/$ - see front matter
doi:10.1016/S0025-5564(03)00089-0

The most common polymorphisms in the human genome are single nucleotide polymorphisms (SNP). These single base substitutions can either be causal for a disease phenotype (e.g., if it is in the coding region of a gene and results in an amino acid change) or silent. If a silent SNP is sufficiently closely linked to a causal variant then both polymorphisms will often be coinherited in cases; thereby giving rise to linkage disequilibrium (LD). Thus knowledge of neutral markers with increased frequencies in cases may be used to map the position of the causal variant. It has been estimated that in order to be certain that a marker is in significant LD with a disease causing polymorphism it will be necessary to have a known marker SNP every 3000 bases along the human genome. With a total genome size of 3.3 billion bases this would correspond to a map of 110 000 SNPs that would be required for whole genome association studies [1].

Recently it has become apparent that the central assumption, of uniform recombination rate along the whole genome, underlying the predictions of 50 000 SNPs being necessary is probably too pessimistic [3–5]. Several studies now demonstrate in some detail that large stretches of DNA tend to be coinherited without recombination or with only very little recombination. Common variation, i.e., the variation widely believed to underly common phenotypes, along those stretches, frequently referred to as blocks, has been shown to be most straightforwardly described in terms of haplotypes. If we denote the states of a SNP by 1 and 0, then a haplotype is the set of states taken by the SNPs along a stretch of DNA. There is often very little variation within these blocks and regularly 90% of chromosomes belong to the 3–5 most common haplotypes [4,5].

Shifting the focus from individual SNPs to haplotypes will have the advantage that a subset of SNPs can be found that captures all of the variation. These SNPs are generally referred to as haplotype tagging SNPs (htSNP) and several approaches have been or are currently being developed. Here we will be concerned with the problem of how many htSNPs are required to tag a given number of haplotypes. This question has been addressed by others, e.g. [6] show the problem is NP-complete and thus hard to solve, and some results are known if the SNPs conform to a tree. We will provide the mathematical frame in which this problem adequately is dealt with: Due to the mathematical structure of recombination it is a combinatorial problem and we will show that the quantities of interest are modeled straightforwardly using the apparatus of Boolean algebras.

In what follows we will first present basic properties of Boolean algebras before deriving various results about the required number of htSNPs. All of our results revolve around what we define to be the dimension dimension of a set of SNPs. This is the size of a set of SNPs that suffices to identify all observed haplotypes; it is thus just the number of htSNPs required.

2. Boolean algebras

A set of elements $\mathscr{B}$ with two binary operations, $\wedge$ and $\vee$, is a Boolean algebra if the following postulates hold:

(A) The operations $\wedge$ and $\vee$ are commutative
(B) Each operation is distributive over the other
(C) There exist in $\mathscr{B}$ distinct identity elements $\mathbf{0}$ and $\mathbf{1}$ relative to the operations $\wedge$ and $\vee$, respectively
$x \wedge \mathbf{1} = x \quad \text{and} \quad x \vee \mathbf{0} = x.$

(D) For every x in $\mathscr{B}$ there exists an element $\neg x$ (the inverse of x) such that

$$x \wedge \neg x = \mathbf{0} \quad \text{and} \quad x \vee \neg x = \mathbf{1}$$

(see [7] for an introduction).

Examples of Boolean algebras are many and they found their use in many different contexts, for example in propositional logic, set theory and various branches of computer science. Often the operations $\wedge$ and $\vee$ are related to either (i) the minimum and maximum of numbers or to (ii) the intersection and union of sets. The basic notation outlined above is adopted from propositional logic. Here we will be concerned with a Boolean algebra that arises from considering population genetic data, i.e., SNPs. Both analogues (i) and (ii) play a role here.

Consider a panel of SNPs obtained from n chromosomes sampled from a population. Assume there are at most two different alleles, 0 and 1, present in any given site and let $x = (x_1, \ldots, x_n)$ denote a vector of alleles for a given site in the n chromosomes. We allow x to be either all zeros or all ones (i.e., non-polymorphic), for reasons that will become clear later. Hereafter, x is referred to as a SNP. In general, there are 2^n different SNPs, but for any SNP, x, the SNP obtained from x by swapping 0 and 1 is practically identical to x, unless we for some reason impose a specific biological interpretation on 0 and 1. This might happen for example if the SNP is not selectively neutral or if an outgroup is used to decide which allele is the oldest.

Let two SNPs, x and y, be given. The SNPs x and y cluster the n chromosomes into four groups, those for which $(x_i, y_i) = (0,0)$, $(x_i, y_i) = (0,1)$, $(x_i, y_i) = (1,0)$, and $(x_i, y_i) = (1,1)$, respectively. Two binary and one unary operation can be defined on SNPs as follows:

$$x \wedge y = (\min(x_1, y_1), \ldots, \min(x_n, y_n)), \tag{1}$$

$$x \vee y = (\max(x_1, y_1), \ldots, \max(x_n, y_n)) \tag{2}$$

and

$$\neg x = (1 - x_1, \ldots, 1 - x_n). \tag{3}$$

If x and y are SNPs then $x \wedge y$ is the SNP that singles out the chromosomes for which $(x_i, y_i) = (1,1)$ and $x \vee y$ is the SNP that singles out all chromosomes for which $(x_i, y_i) = (0,0)$. The complementary operation $\neg x$ results in the SNP 'identical' to x with 0 and 1 reversed.

By direct inspection it can be seen that the set, $\mathscr{P}_n$, ($\mathscr{P}_n$ for polymorphism) of all SNPs on n chromosomes forms a Boolean algebra with binary operations $\wedge$, $\vee$ and unary operation $\neg$ as defined in Eqs. (1)–(3). Note that the two constant SNPs $\mathbf{0} = (0, \ldots, 0)$ and $\mathbf{1} = (1, \ldots, 1)$ can be obtained from any other SNP, x, by applying two of the three operations to x: $\mathbf{0} = x \wedge \neg x$ and $\mathbf{1} = x \vee \neg x$. Either of the operations $\wedge$ or $\vee$ can be expressed in terms the other and $\neg$; for example

$$x \vee y = \neg(\neg x \wedge \neg y) \tag{4}$$

and $\vee$ (or $\wedge$) is thus redundant. If a binary operation, $|$, similar to Sheffer's stroke (attributed to Sheffer, [8], but originally due to C.S. Peirce) in propositional logic is introduced all three operations, $\vee$, $\wedge$, and $\neg$, can be explained in terms of $|$: $x|y = 0$ if and only if $x = y = 1$, thus $x|x = \neg x$ and $(x|y)|(x|y) = x \wedge y$. We shall stick to $\wedge$ and $\neg$ here.

In order to establish our results we first need a few definitions.

Definition 1. For any given set, $\mathscr{S} \subseteq \mathscr{P}_n$, of SNPs, denote by $\mathrm{cl}(\mathscr{S})$ the smallest set containing $\mathscr{S} \cup \{\mathbf{0}\}$ that is closed under the operations $\wedge$ and $\neg$, i.e., if $x, y \in \mathrm{cl}(\mathscr{S})$ then $\neg x \in \mathrm{cl}(\mathscr{S})$ and $x \wedge y \in \mathrm{cl}(\mathscr{S})$.

Since $\vee$ can be expressed in terms of $\wedge$ and $\neg$, $\mathrm{cl}(\mathscr{S})$ is automatically closed under $\vee$. The constant SNPs are always included in $\mathrm{cl}(\mathscr{S})$ and as a consequence $\mathrm{cl}(\emptyset) = \{\mathbf{0}, \mathbf{1}\}$. Clearly, $\mathrm{cl}(\mathscr{S}) \subseteq \mathscr{P}_n$. In many cases strict inequality holds. For example if $\mathscr{S} = \{\mathbf{0}\}$ then $\mathrm{cl}(\mathscr{S}) = \{\mathbf{0}, \mathbf{1}\} \subset \mathscr{P}_n$, and if $\mathscr{S} = \{(0,0,1,1),(0,0,0,1)\}$ then $\mathrm{cl}(\mathscr{S}) = \{(0,0,1,1),(0,0,0,1),(0,0,1,0),(0,0,0,0),(1,1,0,0),(1,1,1,0),(1,1,0,1),(1,1,1,1)\} \subset \mathscr{P}_n$. The set $\mathrm{cl}(\mathscr{S})$ is also a Boolean algebra, a subalgebra of $\mathscr{P}_n$. Intuitively, $\mathrm{cl}(\mathscr{S})$ consists of all partitions, that can be formed given the information about the SNPs in $\mathscr{S}$, of n chromosomes into two sets.

We write haplotypes as stretches of zeros and ones without comma separation, e.g., 0 0 1 0. A haplotype is the allelic type of a chromosome.

Definition 2. An $\mathscr{S}$-haplotype is a haplotype defined by the SNPs in $\mathscr{S}$.

Example 1. If $\mathscr{S}_1 = \{(0,0,1,1),(0,0,0,1)\}$ then the $\mathscr{S}_1$-haplotypes are 0 0, 1 0, and 1 1. The $\mathrm{cl}(\mathscr{S}_1)$-haplotypes are 0 0 0 0 1 1 1 1, 1 0 1 0 0 1 0 1, and 1 1 0 0 0 0 1 1. If $\mathscr{S}_2 = \{(0,0,1,1),(0,0,0,1),(0,0,1,0)\}$ then the $\mathscr{S}_2$-haplotypes are 0 0 0, 1 0 1, and 1 1 0. Here, $\mathrm{cl}(\mathscr{S}_1) = \mathrm{cl}(\mathscr{S}_2)$.

Lemma 1. *Any SNP, x, in $\mathrm{cl}(\mathscr{S})$ defines a partition of chromosomes into two subsets such that chromosomes with the same $\mathscr{S}$-haplotype belong to the same subset. Oppositely, if x defines such a partition then x is in $\mathrm{cl}(\mathscr{S})$. If $\mathscr{S}_1 \subseteq \mathscr{S}_2$ then the $\mathscr{S}_1$-haplotypes group the $\mathscr{S}_2$-haplotypes into disjoint sets.*

Proof. List all SNPs in $\mathscr{S}$, $\mathscr{S} = \{x^{(1)}, \ldots, x^{(m)}\}$, and consider a particular chromosome, C. Define $z^{(i)}$ in the following way: $z^{(i)} = x^{(i)}$ if the entry for C in $x^{(i)}$ is 1, otherwise let $z^{(i)} = \neg x^{(i)}$. Then $z = z^{(1)} \wedge, \ldots, \wedge z^{(m)}$ has a 1 in the entry for C and in the entries for all other chromosomes with the same $\mathscr{S}$-haplotype as C and 0 otherwise; $z = (0, \ldots, 0, 1, \ldots, 1, 0 \ldots, 0)$ for some ordering of the chromosomes. Any partition fulfilling the requirement in Lemma 1 can now be formed using the $\vee$ operation. Note that using the operations $\wedge$, $\vee$, and $\neg$ one cannot split chromosomes with the same $\mathscr{S}$-haplotype into different subsets. To prove the second part, note that all $\mathscr{S}_2$-haplotypes with the same value of $z = z^{(1)} \wedge, \ldots, \wedge z^{(m_1)}$, with $\mathscr{S}_1 = \{x^{(1)}, \ldots, x^{(m_1)}\} \subseteq \mathscr{S}_2$, correspond to the same $\mathscr{S}_1$-haplotype. The proof is completed. □

Example 2. If $\mathscr{S}_1 \subset \mathscr{S}_2$ and $\mathrm{cl}(\mathscr{S}_1) = \mathrm{cl}(\mathscr{S}_2)$ then the $\mathscr{S}_1$-haplotypes and the $\mathscr{S}_2$-haplotypes define the same partitions of the chromosomes, but an $\mathscr{S}_2$-haplotype is defined from more SNPs than the $\mathscr{S}_1$-haplotype; compare Example 1. This is a direct consequence of Lemma 1.

A set $\mathscr{S}_1 \subseteq \mathscr{S}_2$ is said to span $\mathscr{S}_2$ if all SNPs in $\mathscr{S}_2$ can be formed from SNPs in $\mathscr{S}_1$ by repeated applications of the binary and unary operations, $\wedge$, $\vee$ and $\neg$. In the following we will discuss the minimum number of SNPs that are required to span $\mathrm{cl}(\mathscr{S})$ for a given set $\mathscr{S}$.

The concept of compatibility will play an important role. We define it here.

Definition 3. Two SNPs, x and y, are said to be (pairwise) compatible if at most three out of the four 2-locus haplotypes (0,0), (0,1), (1,0), and (1,1) are present in x and y. If x and y are not compatible they are said to be incompatible.

Ref. [9] shows that all SNPs are pairwise compatible if and only if the SNPs are compatible with a tree assuming no recurrent mutations.

3. Results

Let now $\mathscr{S} \subseteq \mathscr{P}_n$ be given. Define the dimension of $\mathscr{S}$ by

$$\dim(\mathscr{S}) = \min\{k | \mathrm{cl}(\mathscr{X}_k) = \mathrm{cl}(\mathscr{S}), \mathscr{X}_k = (x^{(1)}, \ldots, x^{(k)}) \subseteq \mathscr{S}\}. \tag{5}$$

Note that $\mathscr{X}_k$ is required to be a subset of $\mathscr{S}$ which implies that two sets, $\mathscr{S}_1$ and $\mathscr{S}_2$, with $\mathrm{cl}(\mathscr{S}_1) = \mathrm{cl}(\mathscr{S}_2)$ might have different dimensions. We call the set $\mathscr{X}_k$ a basis for $\mathscr{S}$ and say that $\mathscr{X}_k$ explains $\mathscr{S}$ and the set of $\mathscr{S}$-haplotypes, $\mathscr{H}$. The dimension of $\mathscr{S}$ is the smallest number of SNPs in $\mathscr{S}$ required to span all of $\mathscr{S}$ and as a consequence it also spans all of $\mathrm{cl}(\mathscr{S})$. The only exception is if $\emptyset \neq \mathscr{S} \subseteq \{\mathbf{0}, \mathbf{1}\}$. In that particular case, $\mathscr{X} = \emptyset$ is a basis and $\mathscr{X}$ does not span $\mathscr{S}$. Obviously, $\dim(\mathscr{S}) \leqslant \#\mathscr{S}$, the number of SNPs in $\mathscr{S}$, and removal of any x from a basis $\mathscr{X}_k$ reduces the number of explained haplotypes, thus $\dim(\mathscr{X}_k \setminus \{x\}) < \dim(\mathscr{S})$. We define the dimension of the closure, $\mathrm{cl}(\mathscr{S})$, of $\mathscr{S}$ by

$$\dim^*(\mathscr{S}) = \dim(\mathrm{cl}(\mathscr{S})). \tag{6}$$

To our knowledge the dimension of $\mathscr{S}$ is not a standard concept in the theory of Boolean algebras. However, there is a relation to combinatorial geometry which we will take up in Section 5.

It transpires in the proof of Lemma 1 that SNPs of the form

$$x = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0), \tag{7}$$

where x has entry 1 for all chromosomes with the same $\mathscr{S}$-haplotype, span $\mathrm{cl}(\mathscr{S})$. SNPs of the form in (7) are called index-SNPs. There is one for each haplotype. Index-SNPs, however, do not necessarily form a basis of $\mathrm{cl}(\mathscr{S})$.

Example 3. For example, if $\mathscr{S} = \{(0,0,0,1), (0,0,1,0), (0,1,0,0)\}$ then a basis of $\mathrm{cl}(\mathscr{S})$ is $\{(0,0,1,1), (0,1,0,1)\}$ with $\dim^*(\mathscr{S}) = 2$, but $\dim(\mathscr{S}) = 3$.

Corollary 1. *If* $\mathscr{S}_1 \subseteq \mathscr{S}_2$ *and* $\mathrm{cl}(\mathscr{S}_1) = \mathrm{cl}(\mathscr{S}_2)$ *then* $\dim(\mathscr{S}_1) \geqslant \dim(\mathscr{S}_2)$.

Proof. The proof follows directly from the definition of $\dim(\mathscr{S})$. □

Example 4. It is easy to give examples with strict inequality in Corollary 1: If $\mathscr{S}_2 = \{(0,0,0,1), (0,0,1,0), (0,1,0,0), (0,0,1,1), (0,1,0,1)\}$ and $\mathscr{S}_1 = \{(0,0,0,1), (0,0,1,0), (0,1,0,0)\}$, then $\dim(\mathscr{S}_2) = 2$, but $\dim(\mathscr{S}_1) = 3$.

Corollary 2. *If* $\mathrm{cl}(\mathscr{S}_1) \subset \mathrm{cl}(\mathscr{S}_2)$ *then* $\dim^*(\mathscr{S}_1) \leqslant \dim^*(\mathscr{S}_2)$. *If* $\mathscr{S}_1 \subseteq \mathscr{S}_2$ *and* $\dim^*(\mathscr{S}_1) < \dim^*(\mathscr{S}_2)$ *then* $\mathrm{cl}(\mathscr{S}_1) \subset \mathrm{cl}(\mathscr{S}_2)$.

Proof. The second part is trivial (a consequence of the definition of dimension). Regarding the first part of the corollary, Lemma 1 implies that all $\mathscr{S}_1$-haplotypes group one or more $\mathscr{S}_2$-haplotypes. For each $\mathscr{S}_1$-haplotype choose a representative for the corresponding $\mathscr{S}_2$-haplotypes. Let now $\mathscr{X} = \{x^{(1)}, \ldots, x^{(k)}\}$ be a basis for $\mathrm{cl}(\mathscr{S}_2)$ and define $\mathscr{Z} = \{z^{(1)}, \ldots, z^{(k)}\}$ in the following way: The entry of a particular chromosome is defined to be the value of the entry of the representative for the group the chromosome belongs to. Then $\mathscr{Z} \subseteq \mathrm{cl}(\mathscr{S}_1)$, $\mathrm{cl}(\mathscr{Z}) = \mathrm{cl}(\mathscr{S}_1)$ (again a consequence of Lemma 1 and its proof; compare Eq. (7) and the definition of index-SNPs) and $\dim^*(\mathscr{S}_1) \leqslant \#\mathscr{Z} \leqslant \dim^*(\mathscr{S}_2)$ as required. □

Note that the first part of Corollary 2 establishes a partial converse to Corollary 1. It is easy to come up with examples for which equality applies: $\mathscr{S}_1 = \{(1,0,0,0),(0,1,0,0)\}$ and $\mathscr{S}_2 = \{(0,0,1,1),(0,1,0,1)\}$. In both cases, $\dim^*(\mathscr{S}_i) = 2$, $i = 1,2$.

In general, it is not possible to relate the dimension of $\mathscr{S}_1$ to that of $\mathscr{S}_2$. The only cases where this appears to be possible are given in Corollaries 1 and 4 below. Simple examples show that even if $\mathscr{S}_1 \subseteq \mathscr{S}_2$ the dimension of $\mathscr{S}_1$ might be smaller or larger than the dimension of $\mathscr{S}_2$. As an example of the former case consider the following: $\mathscr{S}_1 = \{(0,0,1)\}$ and $\mathscr{S}_2 = \{(0,0,1),(0,1,0)\}$; and as an example of the latter case, consult Corollary 1 and Example 3. In contrast, the dimension of the closure has nice properties.

Corollary 3. *Let* $\mathscr{S}_1$ *and* $\mathscr{S}_2$ *be given. Then*

$$\max(\dim^*(\mathscr{S}_1), \dim^*(\mathscr{S}_2)) \leqslant \dim^*(\mathscr{S}_1 \cup \mathscr{S}_2) \tag{8}$$

and

$$\dim^*(\mathscr{S}_1 \cap \mathscr{S}_2) \leqslant \min(\dim^*(\mathscr{S}_1), \dim^*(\mathscr{S}_2)). \tag{9}$$

Proof. The first inequality follows from the first part of Corollary 2 because $\mathrm{cl}(\mathscr{S}_i) \subseteq \mathrm{cl}(\mathscr{S}_1 \cup \mathscr{S}_2)$, $i = 1,2$. The second inequality also follows from Corollary 2 in that $\mathrm{cl}(\mathscr{S}_1 \cap \mathscr{S}_2) \subseteq \mathrm{cl}(\mathscr{S}_i)$, $i = 1,2$. □

Corollary 4. *For arbitrary* $\mathscr{S}_1$ *and* $\mathscr{S}_2$,

$$\dim(\mathscr{S}_1 \cup \mathscr{S}_2) \leqslant \dim(\mathscr{S}_1) + \dim(\mathscr{S}_2). \tag{10}$$

Proof. Let $\mathscr{X}_i$, $i = 1,2$, be a basis for $\mathscr{S}_i$. It follows that

$$\dim(\mathscr{S}_1 \cup \mathscr{S}_2) \leqslant \dim(\mathscr{X}_1 \cup \mathscr{X}_2) \leqslant \#(\mathscr{X}_1 \cup \mathscr{X}_2) \leqslant \dim(\mathscr{S}_1) + \dim(\mathscr{S}_2), \tag{11}$$

where the first inequality follows from Corollary 1: $\mathscr{X}_1 \cup \mathscr{X}_2 \subseteq \mathscr{S}_1 \cup \mathscr{S}_2$ and $\mathrm{cl}(\mathscr{X}_1 \cup \mathscr{X}_2) = \mathrm{cl}(\mathscr{S}_1 \cup \mathscr{S}_2)$. This proves the corollary. □

Example 5. Let $\mathscr{S}_1 = \{(0,0,0,1),(0,0,1,1)\}$ and $\mathscr{S}_2 = \{(0,1,0,0),(1,0,0,0)\}$. Then $\dim(\mathscr{S}_1 \cup \mathscr{S}_2) = 3$ and $\dim(\mathscr{S}_1) + \dim(\mathscr{S}_2) = 2 + 2 \geqslant 3$. If $\mathscr{S}_2 = \{(0,1,0,0)\}$ then $\dim(\mathscr{S}_1 \cup \mathscr{S}_2) = 3$ and $\dim(\mathscr{S}_1) + \dim(\mathscr{S}_2) = 2 + 1 \geqslant 3$.

In particular Corollary 4 applies to $\mathscr{S}_1 := \mathrm{cl}(\mathscr{S}_1)$ and $\mathscr{S}_2 := \mathrm{cl}(\mathscr{S}_2)$ with dim replaced by $\dim^*$.

In the following we denote the number of SNPs in $\mathscr{S}$ by s and the number of different chromosomes, the haplotypes, by $h \leqslant n$. We identify x with $\neg x$, SNPs containing identical information about the haplotypes in the sample. In addition, we assume that all other SNPs are different from each other. The two quantities s and h are related through the two inequalities stated below.

Lemma 2. *In general, h and s are related through*

$$1 + \log_2(s+1) \leqslant h \leqslant 2^s, \quad \text{or} \quad \log_2(h) \leqslant s \leqslant 2^{h-1} - 1. \tag{12}$$

Assume all SNPs are pairwise compatible. Then

$$\frac{1}{2}(s+3) \leqslant h \leqslant s+1, \quad \text{or} \quad h-1 \leqslant s \leqslant 2h-3. \tag{13}$$

Proof. Inequality (13): Each new SNP can at most create one new haplotype, otherwise there would be an incompatible pair of SNPs. That proves $s+1 \geqslant h$. If all SNPs are pairwise compatible, then $\mathscr{S}$ can be represented by a tree. A tree has at most $h-3$ internal branches (a bifurcating tree has $h-3$; a multifurcating strictly less) and h external branches. In total a tree has less than $2h-3$ branches and in consequence $s \leqslant 2h-3$.

Inequality (12): $h \leqslant 2^s$ by simple combinatorics. Consider h different haplotypes. Each SNP is a partition of the h types into two sets. The number of different partitions of h types is

$$s_{\text{even}} = \binom{h}{1} + \binom{h}{2} + \cdots + \binom{h}{h/2-1} + \frac{1}{2}\binom{h}{h/2}$$

if h is even, identifying x with $\neg x$. If h is odd then

$$s_{\text{odd}} = \binom{h}{1} + \binom{h}{2} + \cdots + \binom{h}{(h-1)/2}.$$

Manipulating the terms gives

$$2s_{\text{even}} + 2 = 2s_{\text{odd}} + 2 = 2^h$$

or $s_{\text{even}} = s_{\text{odd}}$ and $s \leqslant 2^{h-1} - 1$. The proof is completed. □

Theorem 1. *For any $\mathscr{S}$, $\log_2(h) \leqslant \dim(\mathscr{S}) \leqslant h-1$, where $\log_2$ denotes the base-2 logarithm.*

Proof. If $\dim(\mathscr{S}) < \log_2(h)$ then there cannot be a set $\mathscr{S}' \subseteq \mathscr{S}$ that explains all h haplotypes (Lemma 2). The other inequality is proved by induction. If $h = 2$ the inequality is true. Assume now the inequality is true for $h < h_0$. Choose a SNP, x. The SNP divides the set of $\mathscr{S}$-haplotypes, $\mathscr{H}$, into two groups, $\mathscr{H}_1$ and $\mathscr{H}_2$. Choose among all SNPs in $\mathscr{S}$ two subsets, $\mathscr{S}_1$ and $\mathscr{S}_2$, such that $\mathscr{S}_1$ explains $\mathscr{H}_1$, $\mathscr{S}_2$ explains $\mathscr{H}_2$, $\dim(\mathscr{S}_1) \leqslant h_1 - 1$, and $\dim(\mathscr{S}_2) \leqslant h_2 - 1$, where $h_i = \#\mathscr{H}_i$, $i = 1, 2$. Let $\mathscr{S}_0 = \mathscr{S}_1 \cup \mathscr{S}_2 \cup \{x\}$. Surely $\mathscr{S}_0$ explains all haplotypes in $\mathscr{H}$ and $\dim(\mathscr{S}_0) \leqslant (h_1 - 1) + (h_2 - 1) + 1 = h_0 - 1$ (from Corollary 4). □

Theorem 1 has the obvious implication that if $\dim(\mathscr{S}) < h-1$ then there is at least one pair of incompatible SNPs in $\mathscr{S}$. However, there is no general relation between the incompatible pairs in

$\mathscr{S}$ and the dimension, as will be clear later. Myers [10] argues that the quantity $h - \dim(\mathscr{S}) - 1 \geqslant 0$ is a lower bound to the number of recombination events in a sample's history assuming an infinite-site mutation model [11]. Thus, $\dim(\mathscr{S})$ carries information about the sample's history. Note that $h - \dim(\mathscr{S}) - 1$ is the difference between the largest possible dimension that can be obtained for h haplotypes and the actual dimension of $\mathscr{S}$.

Theorem 2. *For any* $\mathscr{S}$, $\dim^*(\mathscr{S}) = \lfloor \log_2(h) \rfloor$, *where* $\lfloor a \rfloor$ *denotes the smallest integer larger than or equal to* a.

Proof. We only have to show that there is a set $\mathscr{X}$ of size $\lfloor \log_2(h) \rfloor$ that spans $\mathrm{cl}(\mathscr{S})$, because according to Lemma 2 the dimension cannot be smaller than $\lfloor \log_2(h) \rfloor$. If $h = 2^{s'}$ for some s' then a basis $\mathscr{X}_{s'}$ of $\mathrm{cl}(\mathscr{S})$ is (here shown for $h = 2^3$) $\mathscr{X}_{s'} = \{(0,0,0,0,1,1,1,1), (0,0,1,1,0,0,1,1), (0,1,0,1,0,1,0,1)\}$ (if $n > h$ some rows are duplicated enlarging the length of the vectors), and the theorem holds. If $2^{s'-1} < h \leqslant 2^{s'}$ then a basis for $\mathrm{cl}(\mathscr{S})$ can be obtained from $\mathscr{X}_{s'}$ by deletion of some rows in the vectors of $\mathscr{X}_{s'}$ (and duplication of others to obtain n in total). □

Theorem 3. *All SNPs are pairwise compatible if and only if* $\dim(\mathscr{S}) = h - 1$. *If all SNPs in* $\mathscr{S} = \mathscr{S}_1 \cup \mathscr{S}_2$ *are pairwise compatible, then*

$$\dim(\mathscr{S}_1 \cup \mathscr{S}_2) + \dim(\mathscr{S}_1 \cap \mathscr{S}_2) \leqslant \dim(\mathscr{S}_1) + \dim(\mathscr{S}_2). \tag{14}$$

Further if $\mathscr{S}_1 \cap \mathscr{S}_2 = \emptyset$ *Eq.* (14) *simplifies*:

(*I*) *Assume* $\mathrm{cl}(\mathscr{S}_1) \cap \mathrm{cl}(\mathscr{S}_2) = \{\mathbf{0}, \mathbf{1}\}$, *then*

$$\dim(\mathscr{S}_1 \cup \mathscr{S}_2) = \dim(\mathscr{S}_1) + \dim(\mathscr{S}_2); \tag{15}$$

(*II*) *Assume* $\mathrm{cl}(\mathscr{S}_1) \cap \mathrm{cl}(\mathscr{S}_2) = \{\mathbf{0}, \mathbf{1}, x, \neg x\}$, $x \neq \mathbf{0}, \mathbf{1}$, *then*

$$\dim(\mathscr{S}_1 \cup \mathscr{S}_2) + 1 = \dim(\mathscr{S}_1) + \dim(\mathscr{S}_2); \tag{16}$$

Proof. 'Only if': A direct consequence of Theorem 1 and Eq. (13). 'If': Assume two SNPs, x and y, are incompatible. They divide the haplotypes into four groups, $\mathscr{H}_i$, $i = 1, \ldots, 4$, according to the 2-locus haplotypes (0,0), (0,1), (1,0), and (1,1). Choose sets of SNPs, $\mathscr{X}_i$, that explain $\mathscr{H}_i$. Then

$$\dim(\mathscr{S}) \leqslant \#(\mathscr{X}_1 \cup \cdots \cup \mathscr{X}_4 \cup \{x, y\}) \leqslant \sum_{i=1}^{4} h_i - 1 + 2 = h - 2,$$

where $\#\mathscr{H}_i = h_i$. But this contradicts $\dim(\mathscr{S}) = h - 1$.

Eq. (14) follows from the following argument: Select a basis, $\mathscr{X}_{12}$, for $\mathscr{S} = \mathscr{S}_1 \cup \mathscr{S}_2$, extend this with SNPs, $\mathscr{Y}_1$, in $\mathscr{S}_1 \setminus \mathscr{S}_2$ to a basis, $\mathscr{X}_1$, for $\mathscr{S}_1$, and finally extend this with SNPs, $\mathscr{Y}_2$, in $\mathscr{S} \setminus \mathscr{S}_1 \subseteq \mathscr{S}_2$ to a basis, $\mathscr{X}$, for $\mathscr{S}$. It is possible according to Theorem 4 below. We have

$$\dim(\mathscr{S}) + \dim(\mathscr{S}_1 \cap \mathscr{S}_2) = 2\#\mathscr{X}_{12} + \#\mathscr{Y}_1 + \#\mathscr{Y}_2 \leqslant \dim(\mathscr{S}_1) + \dim(\mathscr{S}_2),$$

because $\#\mathscr{X}_{12} + \#\mathscr{Y}_1 = \dim(\mathscr{S}_1)$ and $\#\mathscr{X}_{12} + \#\mathscr{Y}_2 \leqslant \dim(\mathscr{S}_2)$. The latter follows from the fact that $\mathscr{X}_{12} \cup \mathscr{Y}_2 \subseteq \mathscr{S}_2$ and each $x \in \mathscr{X}_{12} \cup \mathscr{Y}_2$ explains a new $\mathscr{S}_2$-haplotype, thus there are less than $\dim(\mathscr{S}_2)$ SNPs in the set (see also Theorem 4).To prove the last part of the theorem proceed as follows: (I) Consider the index-SNPs of $\mathrm{cl}(\mathscr{S}_1 \cup \mathscr{S}_2)$. Exactly one of these is neither in $\mathrm{cl}(\mathscr{S}_1)$ nor in $\mathrm{cl}(\mathscr{S}_2)$. If all were in either $\mathrm{cl}(\mathscr{S}_1)$ or $\mathrm{cl}(\mathscr{S}_2)$ then $\mathrm{cl}(\mathscr{S}_1) \cap \mathrm{cl}(\mathscr{S}_2)$ could not be $\{\mathbf{0}, \mathbf{1}\}$ (because

then $z_1^{(1)} \vee, \ldots, \vee z_1^{(m_1)} = \neg(z_2^{(1)} \vee \cdots \vee z_2^{(m_2)})$, where $z_i^{(j)}$ is an index-SNP in $\mathscr{S}_i$), and if more than one were not in $\mathrm{cl}(\mathscr{S}_1)$ or $\mathrm{cl}(\mathscr{S}_2)$ then $\mathscr{S}_1 \cup \mathscr{S}_2$ could not span $\mathrm{cl}(\mathscr{S}_1 \cup \mathscr{S}_2)$. Further, $\dim(\mathscr{S}_i) = h_i - 1$ where h_i is the number of $\mathscr{S}_i$-haplotypes (first part of the theorem). It follows that h_i is one more than the number of index-SNPs in $\mathrm{cl}(\mathscr{S}_i)$ and thus $h_1 + h_2 = h + 1$ and $\dim(\mathscr{S}_1) + \dim(\mathscr{S}_2) = (h_1 - 1) + (h_2 - 1) = h - 1 = \dim(\mathscr{S}_1 \cup \mathscr{S}_2)$. (II) Can be proved similarly to (I). The proof is completed. □

Eq. (15) in Theorem 3 can be seen as an orthogonality relation. A similar relation does not hold for $\dim^*$. Eqs. (15) and (16) can also be proved graphically representing the SNPs with trees.

Example 6. Eq. (14) cannot be improved. Denote the left side of (14) by L, the right side by R. Let $\mathscr{S}_1 = \{(0,0,0,1),(0,0,1,1)\}$ and $\mathscr{S}_2 = \{(0,0,1,0),(0,0,1,1)\}$. Then $L = 2 + 1$ and $R = 2 + 2$, or $L < R$. Let $\mathscr{S}_1 = \{(0,1,0,0),(0,0,1,1)\}$ and $\mathscr{S}_2 = \{(0,0,1,0),(0,0,1,1)\}$. Then $L = 3 + 1$ and $R = 2 + 2$, or $L = R$.

Example 7. Eq. (14) is not true in general. Let $\mathscr{S}_1 = \{(0,0,0,1),(0,0,1,0),(0,1,0,1),(0,1,1,0)\}$ and $\mathscr{S}_2 = \{(0,0,0,1),(0,0,1,0),(0,1,0,1),(0,0,1,1)\}$, then $\mathscr{S}_1 \cap \mathscr{S}_2 = \{(0,0,0,1),(0,0,1,0),(0,1,0,1)\}$ and $\dim(\mathscr{S}_1 \cup \mathscr{S}_2) + \dim(\mathscr{S}_1 \cap \mathscr{S}_2) = 2 + 3 > 2 + 2 = \dim(\mathscr{S}_1) + \dim(\mathscr{S}_2)$.

Ref. [6] shows that to find $\dim(\mathscr{S})$ and a basis $\mathscr{X}$ for $\mathscr{S}$ are NP-complete problems. Thus, in general there cannot be a polynomial time algorithm that outputs $\dim(\mathscr{S})$ or $\mathscr{X}$ and some sort of exhaustive search is necessary. Potential bases, $\mathscr{X}$, must fulfill $\lfloor \log_2(h) \rfloor \leqslant \#\mathscr{X} \leqslant h - 1$ (Theorem 1), which can be used to decide whether a potential basis can be a basis at all. A dynamic algorithm that finds all bases can be constructed along the following lines.

(1) List all non-constant SNPs, $\mathscr{S} = \{x^{(1)}, \ldots, x^{(m)}\}$
(2) Put $\mathfrak{B}_1 = \{\mathscr{X}_1, \emptyset\}$ with $\mathscr{X}_1 = \{x^{(1)}\}$
(3) Define $\mathfrak{B}_k$ and $\mathfrak{B}'_k$ recursively in the following way
(3a) Let $\mathfrak{B}'_k = \{\mathscr{X}' | \mathscr{X}' = \mathscr{X} \cup \{x^{(k)}\}, \mathscr{X} \in \mathfrak{B}_{k-1}\}$, Remove from $\mathfrak{B}'_k$ all elements, $\mathscr{X}'$, that do not explain more haplotypes than its counterpart, $\mathscr{X}$, in $\mathfrak{B}_{k-1}$
(3b) Put $\mathfrak{B}_k = \mathfrak{B}'_k \cup \mathfrak{B}_{k-1}$
(4) Define two counters on $\mathfrak{B}_k$: $h(\mathscr{X}) = \#$haplotypes explained by $\mathscr{X}$, and $s(\mathscr{X}) = \#$SNPs in $\mathscr{X}$.

Note that the listing (step 1) can be done stepwise while building up $\mathfrak{B}_k$. After the final step, the elements in $\mathfrak{B}_m$ for which $s(\mathscr{X}) = \min\{s(\mathscr{X}') | \mathscr{X}' \in \mathfrak{B}_k\}$ and $h(\mathscr{X}) = \max\{h(\mathscr{X}') | \mathscr{X}' \in \mathfrak{B}_k\} = h$ are bases of $\mathscr{S}$. There might be efficient heuristic algorithms, in particular if $\dim(\mathscr{S})$ is small, because in that case only few SNPs are needed to span all of $\mathscr{S}$. As soon as a set is found that spans $\mathscr{S}$, further search can be restricted to sets with at most $\#\mathscr{S}$ SNPs.

If the widely discussed hypothesis of block structured recombination holds true then a basis for each block can be found in polynomial time: First apply the algorithm in [12] (see [13] for further explanation) to dissect the SNPs into apparently 'non-recombining' blocks and subsequently find a basis for each block. Alternatively, the algorithm in [10] can be used to find sets of SNPs in smaller regions that explain many haplotypes (corresponding to regions with high evidence of

historical recombination). His algorithm has the same time complexity as the algorithm in 1–4, but relatively fast and reliable heuristic approximations exist [10].

The next theorem is essentially due to [14]: it is easy to find a basis if all SNPs are compatible.

Theorem 4. *If all SNPs are pairwise compatible*, $\dim(\mathscr{S})$ *and a basis for* $\mathscr{S}$ *can be found in polynomial time using the following algorithm*;

(1) *List all non-constant SNPs*, $\mathscr{S} = \{x^{(1)}, \ldots, x^{(m)}\}$
(2) *Put* $\mathscr{X}_1 = \{x^{(1)}\}$
(3) *Add* $x^{(k)}$ *to* $\mathscr{X}_k$, $k = 1, 2, \ldots$, *if an extra* $\mathscr{S}$*-haplotype is explained by adding* $x^{(k)}$. *Let* $\mathscr{X}_{k+1}$ *be the new set of SNPs*, $\mathscr{X}_{k+1} = \mathscr{X}_k$ *or* $\mathscr{X}_{k+1} = \mathscr{X}_k \cup \{x^{(k)}\}$, *respectively*
(4) *Stop when* $\#\mathscr{X}_k = h - 1$. *Then* $\mathscr{X}_k$ *is a basis for* $\mathscr{S}$.

Proof. If $\#\mathscr{X}_k = h - 1$ then $\mathscr{X}_k$ must be a basis for $\mathscr{S}$. Assume $\#\mathscr{X}_k = h - 1$ is not reached, i.e. $\#\mathscr{X}_k < h - 1$. If $x^{(i)}$ is not added when proposed, say at step j, then also $\dim(\mathscr{X}_{j'} \cup \{x^{(i)}\}) < h - 1$, for $j' > j$. Thus, $\dim(\mathscr{S}) = \dim(\mathscr{X}_k \cup (\mathscr{S} \setminus \mathscr{X}_k)) < h - 1$, which contradicts that the dimension of $\mathscr{S}$ is $h - 1$. □

The algorithm in Theorem 4 can be implemented in time of $O(nm)$ [14]. Also here listing of the SNPs can be done stepwise. Since two incompatible SNPs give rise to four different haplotypes one might expect that a group of pairwise incompatible SNPs could be useful in building up a basis for $\mathscr{S}$. However, a SNP that is incompatible with every other SNP in a group does not necessarily create new haplotypes. As an example consider the following four chromosomes (or haplotypes) and three SNPs: $\mathscr{S} = \{(0,0,1,1),(0,1,0,1),(0,1,1,0)\}$. All three SNPs are pairwise incompatible but any two of them explain all four chromosomes, so the remaining SNP is always redundant. In general one can form large sets of pairwise incompatible SNPs. Below we give a lower bound on the size of the largest set of pairwise incompatible SNPs for a given number of haplotypes.

Theorem 5. *Assume* $h = 2^k$, *for some* $k \geqslant 2$. *Then there exists a set*, $\mathscr{I}_k$, *of SNPs of size* I_k *such that all SNPs in* $\mathscr{I}_k$ *are pairwise incompatible. The number* I_k *is recursively given by* $I_k = I_{k-1}^2 + 1$ *with* $I_2 = 3$, *and* I_k *fulfills* $2^{h-1} - 1 \geqslant I_k \geqslant 3^{\frac{1}{4}h}$.

Proof. If $k = 2$, $\mathscr{I}_2 = \{(0,0,1,1),(0,1,0,1),(0,1,1,0)\}$ is such a set. Assume now we have a set $\mathscr{I}_{k-1}$ of size I_{k-1} of pairwise incompatible SNPs. Define

$$\mathscr{I}_k = \{(0,\ldots,0,1,\ldots,1)\} \cup \{(x,y)|x,y \in \mathscr{I}_{k-1}\},$$

where the first SNP is of length 2^k with equal numbers of ones and zeros and (x,y) denotes concatenation of the vectors x and y. Then all pairs of SNPs in $\mathscr{I}_k$ are incompatible and $I_k = I_{k-1}^2 + 1$. The lower bound on I_k is obtained by induction and the upper bound is from Lemma 2. □

I_k is rapidly increasing even for small k, e.g., if $h = 2^4 = 16$ then $I_4 = 101$, and if $h = 2^5 = 32$ then $I_5 = 10202$.

4. Biogeographical Information

It is of interest to find SNPs that determine the biogeographical ancestry of a chromosome or an individual. The biogeographical information (BGI) can be coded in the form of binary vectors, just like SNPs, such that, e.g., African origin is coded as $(1,\ldots,1,0,\ldots,0)$ where 1 indicates that the chromosome has African ancestry, 0 that it has not. We assume BGI is defined from sources external to genetic information, e.g., using information about the place of birth of relatives of the individual that carries the chromosome.

Let BGI be coded in a set of vectors $\mathscr{G}$. Similar to the definition of the dimension of $\mathscr{S}$ we can ask for the minimum number of SNPs that uniquely place a chromosome in a biogeographical group,

$$\dim_{\mathscr{G}}(\mathscr{S}) = \min\{k|\mathrm{cl}(\mathscr{X}_k) \supseteq \mathscr{G}, \mathscr{X}_k = (x^{(1)},\ldots,x^{(k)}) \subseteq \mathscr{S}\}, \tag{17}$$

with the further natural requirements that $\mathscr{G} \subseteq \mathrm{cl}(\mathscr{S})$ and that $\mathscr{G}$ itself is closed under the operations $\wedge$ and $\neg$, i.e., $\mathrm{cl}(\mathscr{G}) = \mathscr{G}$. If $\mathscr{G} \not\subseteq \mathrm{cl}(\mathscr{S})$ then $\mathscr{S}$ cannot represent $\mathscr{G}$ faithfully. It follows that

$$\dim_{\mathscr{G}}(\mathrm{cl}(\mathscr{S})) = \dim(\mathscr{G}) = \dim^*(\mathscr{G}), \tag{18}$$

thus, $\dim_{\mathrm{cl}(\mathscr{S})}(\mathrm{cl}(\mathscr{S})) = \dim^*(\mathscr{S})$, and further that

$$\dim_{\mathscr{G}_1}(\mathrm{cl}(\mathscr{S})) \leqslant \dim_{\mathscr{G}_2}(\mathrm{cl}(\mathscr{S})), \tag{19}$$

if $\mathscr{G}_1 \subseteq \mathscr{G}_2$.

We can go on and derive results analogous to those derived in the previous section in either two ways: (1) For fixed $\mathscr{G}$ with dim replaced by $\dim_{\mathscr{G}}$, or (2) for $\mathscr{G}_1 \subseteq \mathscr{G}_2$. For example we find

$$\dim_{\mathscr{G}_1\cup\mathscr{G}_2}(\mathscr{S}) \leqslant \dim_{\mathscr{G}_1}(\mathscr{S}) + \dim_{\mathscr{G}_2}(\mathscr{S}). \tag{20}$$

We will refrain from deriving these results in general.

5. Discussion

In this exposition we have focused on the minimum set that spans a given set $\mathscr{S}$ of SNPs. The focus is naturally on describing $\mathscr{S}$ and $\mathscr{S}$'s relation to the set of different haplotypes, $\mathscr{H}$. In the special case $\mathscr{S} = \mathrm{cl}(\mathscr{S})$, $\mathrm{cl}(\mathscr{S})$ is generated by the set of index-SNPs, but these do not form a basis of $\mathrm{cl}(\mathscr{S})$.

Often combinatorial geometry is the most adequate description of the structure of finite sets. However, it seems that $\mathscr{H}$ (or $\mathrm{cl}(\mathscr{S})$) is the natural object, rather than $\mathscr{S}$, if we want to apply concepts of combinatorial geometry to the setting in this paper, e.g., $\mathscr{H}$ can be seen as a combinatorial geometry generated by the set of index-SNPs and the dimension of $\mathscr{H}$ would be defined in terms of the number of index-SNPs (see, e.g., [15]). Unfortunately, the index-SNPs are not in general part of $\mathscr{S}$ and the combinatorial geometry structure of $\mathscr{H}$ does not transfer to $\mathscr{S}$. Only if all SNPs in $\mathscr{S}$ are pairwise compatible do we have a correspondence. The set of haplotypes $\mathscr{H}$ stands in relation to the set of splits in a tree representing $\mathscr{H}$, and the splits are in one-to-one relation with the SNPs. This correspondence is also clear from the equality $\dim(\mathscr{S}) = h - 1$, that directly relates the number of haplotypes to $\mathscr{S}$. Eq. (14) is a general property of combinatorial geometries.

The algorithm to determine a basis of $\mathscr{S}$ (or $\mathscr{H}$) is inefficient by nature of the problem. However, efficient heurestic algorithms might be applied to find sets, $\mathscr{X}$, that span $\mathscr{S}$ and such that $\#\mathscr{X}$ is near the optimal possible, $\dim(\mathscr{S})$. One strategy could be to identify a basis for the haplotypes, $\mathscr{H}_0$, defined by the most common SNPs, e.g., those with minor allele frequency above 10%. Often the haplotypes in $\mathscr{H}_0$ are few in number (e.g., [16]) and can easily be tagged with the algorithm presented in the paper. In a second step, for each h in $\mathscr{H}_0$ the group of halotypes in $\mathscr{H}$ defining h could be tagged, and so forth. This heurestic algorithm is easy to implement and fast to run.

Acknowledgements

M.H.P.S. is a Wellcome Trust Research Fellow.

References

[1] L. Kruglyak, Prospects for whole-genome disequilibrium mapping of common disease genes, Nat. Genet. 22 (1999) 139.
[2] J.K. Pritchard, M. Przeworski, Linkage disequilibrium in humans: Models and data, Am. J. Hum. Genet. 69 (2001) 1.
[3] A.J. Jeffreys, L. Kauppi, R. Neumann, Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex, Nat. Genet. 29 (2001) 217.
[4] M.J. Daly, J.D. Rioux, S.E. Schaffner, T.J. Hudson, E.S. Lander, High-resolution haplotype structure in the human genome, Nat. Genet. 29 (2001) 229.
[5] N. Patil, A.J. Berno, W.A. Barrett, J.M. Doshi, C.P. Hacker, et al., Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, Science 294 (2001) 1719.
[6] M.R. Garey, D.S. Johnson, Computers and Intractability, Freeman, New York, 1979.
[7] J.E. Whitesitt, Boolean Algebras and their Applications, Addison-Wesley, Reading, MA, 1995.
[8] H.M. Sheffer, A set of five independent postulates for Boolean algebras with applications to logical constants, Trans. Am. Math. Soc. 14 (1913) 481.
[9] G.C. Estabrook, C. Johnson, F. McMorris, An idealized concept of the true cladistic character, Math. Biosci. 23 (1975) 263.
[10] S. Myers, Bounds on the minimum number of recombination events in a sample history, Genetics 163 (2003) 375.
[11] G. Watterson, On the number of segregating sites in genetic models without recombination, Theor. Pop. Biol. 52 (1975) 43.
[12] R.R. Hudson, N. Kaplan, Statistical properties of the number of recombination events in the history of a sample of DNA sequences, Genetics 111 (1985) 147.
[13] C. Wiuf, On the minimum number of topologies explaining a sample of DNA sequences, Theor. Pop. Biol. 62 (2002) 357.
[14] D. Gusfield, Efficient algorithms for inferring evolutionary trees, Networks 21 (1991) 19.
[15] J.H. van Lint, R.M. Wilson, A Course in Combinatorics, Cambridge University, Cambridge, 1992.
[16] G.C.L. Johnson, L. Esposito, B.J. Barratt, A.N. Smith, J. Heward et al., Haplotype tagging for the identification of common disease genes, Nat. Genet. 29 (2001) 233.