An Integer Programming Formulation of the Parsimonious Loss of Heterozygosity Problem

Daniele Catanzaro, Martine Labbé, and Bjarni V. Halldórsson

Abstract—A *Loss of Heterozygosity* (LOH) event occurs when, by the laws of Mendelian inheritance, an individual should be heterozygote at a given site but, due to a deletion polymorphism, is not. Deletions play an important role in human disease and their detection could provide fundamental insights for the development of new diagnostics and treatments. In this article we investigate the *Parsimonious Loss of Heterozygosity Problem (PLOHP)*, i.e., the problem of partitioning suspected polymorphisms from a set of individuals into a minimum number of deletion areas. Specifically, we generalize Halldórsson *et al.*' work by providing a more general formulation of the PLOHP and by showing how one can incorporate different recombination rates and prior knowledge about the locations of deletions. Moreover, we show that the PLOHP can be formulated as a specific version of the clique partition problem in a particular class of graphs called *undirected catch-point interval graphs* and we prove its general $\mathcal{NP}$-hardness. Finally, we provide a state-of-the-art integer programming formulation and strengthening valid inequalities to exactly solve real instances of the PLOHP containing up to 9000 individuals and 3000 SNPs. Our results give perspectives on the mathematics of the PLOHP and suggest new directions on the development of future efficient exact solution approaches.

Index Terms—clique partitioning, submodular functions, polymatroid rank functions, undirected catch-point interval graph, combinatorial optimization, mixed integer programming, computational biology, loss of heterozygosity, genome-wide association studies, single nucleotide polymorphism.

1 Introduction

The recent completion of the Hap Map project [1] has shown that any two copies of the human genome differ from one another by approximately 0.1% of nucleotide sites, i.e., one variant per 1000 nucleotides on average [2], [3], [4], [5]. The most common variants, called *Single Nucleotide Polymorphisms* (SNPs, see Figure 1), together with the recombination process, constitute the predominant form of human variation [6], [7], [8]. A large number of other types of variations exist in nature, including insertions, inversions, translocations. One type of variation being *deletions*, which occur when a subsequence of the human genome is present in a reference genome but is not in the genome of an individual being analyzed.

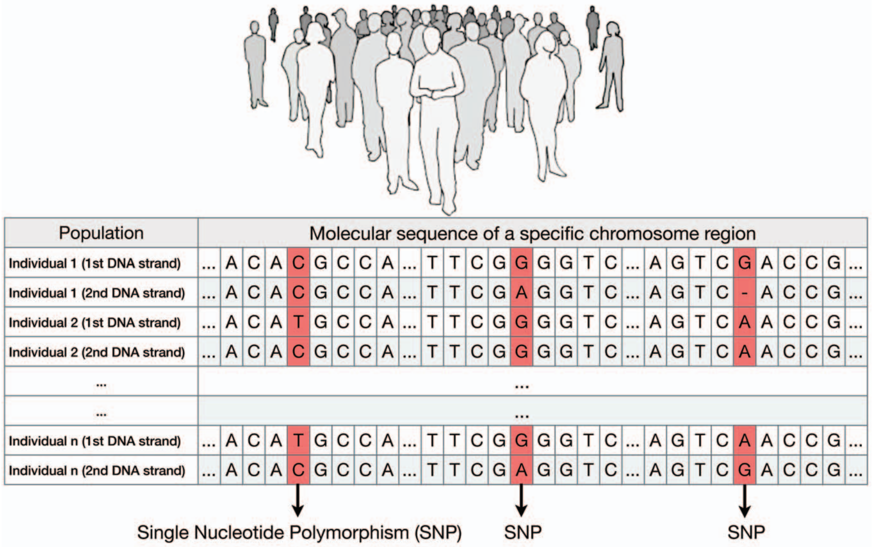
When the genotypes of a child and its two parents are known a deletion polymorphism may be observed as a *Loss of Heterozygosity* (LOH) event on the child chromosome. Specifically, the laws of Mendelian inheritance dictate that each individual inherits one copy of a chromosome from the father and one from the mother. Hence, for a given SNP, an individual can be either *homozygous*, i.e., the nucleotides of the parental DNA strands are equal, or *heterozygous*, i.e., the nucleotides of the parental DNA strands are different. For example, the first individual in Figure 1 is homozygous at the first SNP and heterozygous at the second SNP. When a deletion polymorphism occurs, an individual carries only a single copy of the chromosomal segment while the other is missing. As an example, the first individual in Figure 1 carries a deletion at the third SNP of the considered chromosome region (denoted by the symbol '-'). If the deletion is *de novo*, the lack of information concerns only the individual and not the respective parents. Otherwise, the deletion is said to be *inherited* i.e., passed from one of the two parents to the child. If the deletion event modifies the heterozygosity of an individual at a given site of a chromosomal region then we say that a LOH event occurred at that site.

Deletions may have a negative impact on the health of an individual and may give rise to several human diseases. For example, recent studies showed that schizophrenia [9], multiple sclerosis [10], Alzheimer [11], type I diabetes [12], obesity [13], and some cardiovascular diseases [14], [15], [16] are associated with large recurrent deletion events occurred across the genomes of the affected individuals [17]. A shared hope in the scientific community is that detecting deletions across the genome of individuals could be of fundamental assistance for the diagnosis and the treatment of certain human diseases, hence considerable research efforts have been dedicated to this task in recent years [1].

A natural approach to perform the task of finding deletions consists of comparing the genomes of a given pop-

- D. Catanzaro, and Martine Labbé belong to the Graphs and Mathematical Optimization Unit, Computer Science Department, Université Libre de Bruxelles (ULB), Boulevard du Triomphe, CP 210/01, B-1050, Brussels, Belgium. Phone: +32 2 650 5628. Fax: +32 2 650 5970.
- B. V. Halldórsson belongs to the Department of Biomedical Engineering, School of Science and Engineering, Reykjavk University, Menntavegur 1, 101 Reykjavik, Iceland. Phone: +354 599 6247.
Correspondence should be addressed to: bjarnivh@ru.is.

Digital Object Indentifier 10.1109/TCBB.2012.138



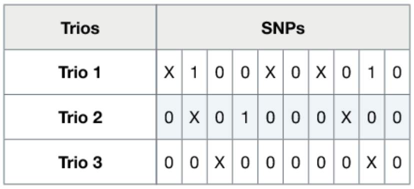
Population	Molecular sequence of a specific chromosome region
Individual 1 (1st DNA strand)	... A C A C G C C A ... T T C G G G G T C ... A G T C G A C C G ...
Individual 1 (2nd DNA strand)	... A C A C G C C A ... T T C G A G G T C ... A G T C - A C C G ...
Individual 2 (1st DNA strand)	... A C A T G C C A ... T T C G G G G T C ... A G T C A A C C G ...
Individual 2 (2nd DNA strand)	... A C A C G C C A ... T T C G G G G T C ... A G T C A A C C G ...
...	...
...	...
Individual n (1st DNA strand)	... A C A T G C C A ... T T C G G G G T C ... A G T C A A C C G ...
Individual n (2nd DNA strand)	... A C A C G C C A ... T T C G A G G T C ... A G T C G A C C G ...

Fig. 1. Any two copies of the human genome differ from one another by approximatively 0.1% of nucleotide sites. In this example, most of the DNA sequence is identical in a given chromosome region from a set of individuals, apart from three variant sites. These sites are called *Single Nucleotide Polymorphisms* (SNPs). The symbol '-' represents a deletion, i.e., a lack of a nucleotide.

ulation of affected individuals with the genomes from a population of unaffected ones. However, the genomes of the individuals are generally not readily available and even if they were, the comparison process would be laborious, time consuming and cost-prohibitive due to the large amount of data to analyze. Hence, the use of predictive models is usually considered as an alternative to the experimental approach [18]. In this context, a number of methods for the detection of deletions have been suggested in the literature, including tiling arrays [19] and high throughput sequencing [20], [21].

In this paper we focus on detecting germline deletions from genotype data of an offspring and his parents. These data may be derived from SNP arrays, which have been used for genome-wide association studies at a number of laboratories, see e.g., [10], [22], [23] and [24]. Somatic mutations might also be detected in a similar framework to the one presented here, given genotypes from multiple tissues of the individual being studied.

It is worth noting that detecting deletions from genotype data may not be straightforward due to the limit of current genotyping technology and the presence of uncertainty in the genotyping process. In fact, current SNP genotyping technology is not able to discern easily the difference between a homozygous site and a deletion, hence the output will always be a homozygous SNP even if the true genotype of the individual may carry only a single copy of the genotype. Moreover, even if a deletion polymorphism were observed in molecular data, such event could be due either to the presence of real deletions or to *genotyping errors*, i.e., misreadings caused by the genotyping technology [25]. In this article we address these major limitations. Specifically, our work is an extension of one of the problems presented in Halldórsson *et al.* [10], which dealt with the problem of detecting deletions as well as the problem of determining haplotypes from genotypes. Here, we extend their work on deletions and present a more general predictive model able to incorporate prior knowledge about the locations of deletions in the human genome and the probability of genotyping error. We show that the problem of detecting deletions from genotype data can be formulated as a specific version of the clique partition problem in a particular class of graphs called *undirected catch-point interval graphs*. We prove that this problem is $\mathcal{NP}$-hard in general and we provide a methodology to solve it based on Integer Programming (IP). Specifically, we present an IP formulation and strengthening valid inequalities to reduce the solution space. We then demonstrate through a series of empirical tests on real and artificial data that this formulation is often characterized by a small gap between the optimal solution and its non-integral linear programming bound relative to the prior art as well as often substantially faster processing of very large instances of the problem containing up to 9000 individuals and 3000 SNPs. The work thus gives perspective on the mathematics of detecting deletions from genotype data, provides methodology suitable for provably optimal solution of hard real instances that resist all prior approaches, and suggests new directions



Trios	SNPs									
Trio 1	X	1	0	0	X	0	X	0	1	0
Trio 2	0	X	0	1	0	0	0	X	0	0
Trio 3	0	0	X	0	0	0	0	0	X	0

Fig. 3. An example of three trios having 10 SNPs each.

on the development of future efficient exact solution approaches.

2 Notation and problem formulation

In this section, we state the problem of detecting deletions from genotype data in terms of an optimization problem. To this end, consider a *trio* t, i.e., a set of two parents and an offspring, and let s denote a SNP genotyped in t. Then, one of the following three situations may occur:

1) The SNP s can be *Inconsistent with a Loss of Heterozygosity* (ILOH), a situation that occurs when the child is heterozygous. In this case different alleles must have been inherited from each parent. For example, this is the case in the first highlighted column of the sequences of the trio shown in Figure 2.
2) The SNP s can be *Consistent with a Loss of Heterozygosity* (CLOH), a situation that occurs when a deletion may (but needs not) be introduced to explain the trio's inheritance pattern. For example, by referring to the second highlighted column of the sequences of the trio shown in Figure 2, the SNP of the child could be explained by means of a deletion of the paternal pattern.
3) The SNP s can show *Evidence of a Loss of Heterozygosity* (ELOH), a situation that occurs when a deletion or a genotyping error are the only possible explanation for the trio inheritance pattern. For example, by referring to the third highlighted column of the sequences of the trio shown in Figure 2, the SNP of the child can be explained only by means of a deletion of the maternal pattern.

Using the definitions described above, a trio genotyped at m SNPs can be encoded as a string of length m over an alphabet $\Sigma = \{1, 0, X\}$, where '1' codes for a SNP inconsistent with having a loss of heterozygosity; '0' codes for a SNP consistent with a loss of heterozygosity; and 'X' codes for a SNP showing evidence of loss of heterozygosity [10]. For example, the string $t = \langle X100X0X010\rangle$ in Figure 3 represents a trio genotyped at 10 SNPs, thereof 5 consistent with having a loss of heterozygosity, 3 showing evidence of a loss of heterozygosity, and 2 inconsistent with having a loss of heterozygosity.

Let $\mathcal{SNP}$ denote a set of m SNPs, and $\mathcal{T} = \{t^p\}$ as a set of n trios genotyped at the m SNPs in $\mathcal{SNP}$. Given a trio $t^p \in \mathcal{T}$, let t^p_s denote both the value and the position of s-th SNP in t^p. Further, let $\mathcal{T}_X = \{t^p_s \in \mathcal{T} \times \mathcal{SNP} : t^p_s = \text{'X'},\ t^p \in \mathcal{T}, s \in \mathcal{SNP}\}$.

Given a trio/SNP pair $t^p_s \in \mathcal{T}_X$, we denote ${}^l t^p_s$ and ${}^r t^p_s$ as the positions of the closest ILOH SNPs on t^p on the left and on the right of t^p_s, respectively, and we set $l^p_s = {}^l t^p_s + 1$ and $r^p_s = {}^r t^p_s - 1$. Let l^p_s and r^p_s be the *left* and *right margins* of t^p_s, respectively. We set $l^p_s = 1$ if there is no ILOH SNP on the left of t^p_s and $r^p_s = m$ if there is no ILOH SNP on the right of t^p_s, respectively. For example, by considering the SNP t^1_5 in Figure 3 we have that $l^1_5 = 3$ and $r^1_5 = 8$. Similarly, by considering the SNP t^3_3 in Figure 3 we have that $l^3_3 = 1$ and $r^3_3 = 10$. Finally, $l^1_1 = r^1_1 = 1$.

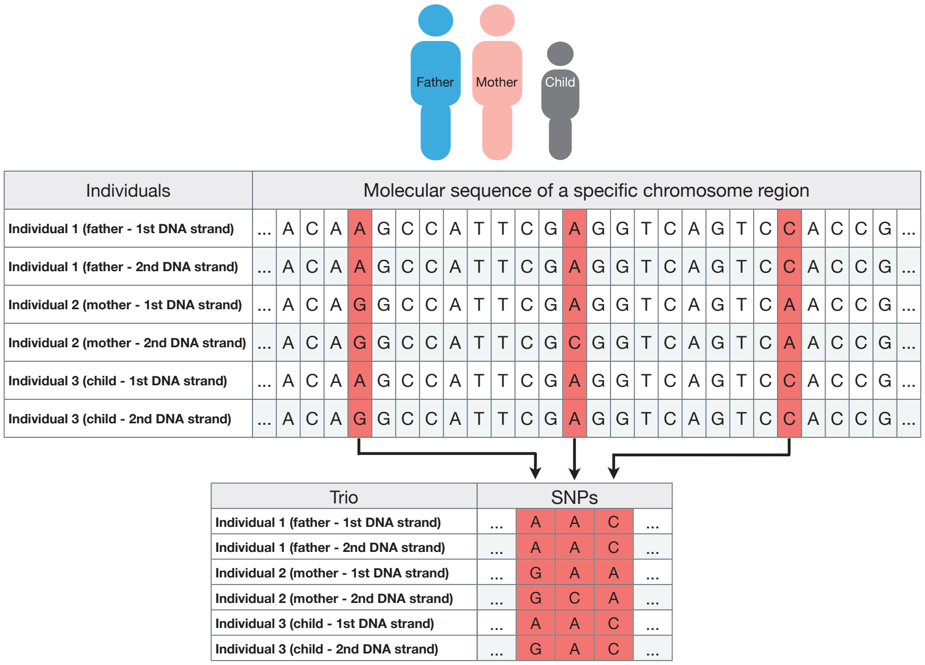
Given a trio $t^p \in \mathcal{T}$, we define an *interval* to be any contiguous subset of SNPs in t^p that does not contain ILOH SNPs. Consider two distinct trio/SNP pairs in $\mathcal{T}_X$, say $t^p_{s_1}$ and $t^q_{s_2}$. We say that $t^p_{s_1}$ and $t^q_{s_2}$ are *mutually compatible with a deletion* if both the position of $t^q_{s_2}$ falls inside the interval $[l^p_{s_1}, r^p_{s_1}]$ and the position of $t^p_{s_1}$ falls inside the interval $[l^q_{s_2}, r^q_{s_2}]$. In which case we also say that the corresponding intervals $[l^p_{s_1}, r^p_{s_1}]$ and $[l^q_{s_2}, r^q_{s_2}]$ are *mutually compatible*. For example, t^3_3 and t^1_5 are mutually compatible with a deletion, in fact the position of the SNP t^3_3 falls inside the substring delimited by the SNPs at positions $l^1_5 = 3$ and $r^1_5 = 8$ and vice-versa the position of the SNP t^5_1 falls inside the substring delimited by the SNPs at positions $l^3_3 = 1$ and $r^3_3 = 10$. Finally, it is easy to realize that the SNP t^1_1 is not compatible with any other as it does not fall inside any other interval induced by a trio/SNP pair in $\mathcal{T}_X$.

Consider a subset Q of trio/SNP pairs in $\mathcal{T}_X$ and their corresponding intervals. We define a *Region Compatible with a Deletion* (RCD) as any subset $S \subseteq Q$ of mutually compatible intervals $[l^p_s, r^p_s]$, for all $t^p_s \in Q$. As an example, the SNPs t^1_5, t^1_7 and t^2_8 in Figure 3 form a RCD, as their corresponding intervals are mutually compatible.

RCDs play a central role in detecting deletions in the human genome. In fact, as genotyping errors are usually sporadic during the genotyping phase, high concentrations of CLOH or ELOH SNPs located in specific areas of $\mathcal{T}$ are likely to indicate the presence of true underlying deletions. As deletion events are rare in nature, the number of such areas can be expected to be small. Halldórsson *et al.* [10] proposed to exploit these insights to detect deletions in $\mathcal{T}$. Specifically, denoted $\mathcal{R}$ and $\mathcal{E}$ as the set of RCDs and the set of genotyping errors in $\mathcal{T}$, respectively, and $h(\rho)$ and $g(\eta)$ as the costs of detecting a RCD $\rho \in \mathcal{R}$ and a genotyping error $\eta \in \mathcal{E}$, respectively, the authors proposed to solve the following optimization problem to accomplish the task:

Problem. The Parsimonious Loss of Heterozygosity Problem (PLOHP). *Given a set $\mathcal{T}$ of n trios having m SNP each, minimize the overall cost*

$$\chi = \sum_{\rho \in \mathcal{R}} h(\rho) + \sum_{\eta \in \mathcal{E}} g(\eta)$$



Individuals	Molecular sequence of a specific chromosome region																										
Individual 1 (father - 1st DNA strand)	...	A	C	A	A	G	C	C	A	T	T	C	G	A	G	G	T	C	A	G	T	C	C	A	C	C	G	...
Individual 1 (father - 2nd DNA strand)	...	A	C	A	A	G	C	C	A	T	T	C	G	A	G	G	T	C	A	G	T	C	C	A	C	C	G	...
Individual 2 (mother - 1st DNA strand)	...	A	C	A	G	G	C	C	A	T	T	C	G	A	G	G	T	C	A	G	T	C	A	A	C	C	G	...
Individual 2 (mother - 2nd DNA strand)	...	A	C	A	G	G	C	C	A	T	T	C	G	C	G	G	T	C	A	G	T	C	A	A	C	C	G	...
Individual 3 (child - 1st DNA strand)	...	A	C	A	A	G	C	C	A	T	T	C	G	A	G	G	T	C	A	G	T	C	C	A	C	C	G	...
Individual 3 (child - 2nd DNA strand)	...	A	C	A	G	G	C	C	A	T	T	C	G	A	G	G	T	C	A	G	T	C	C	A	C	C	G	...

Trio	SNPs				
Individual 1 (father - 1st DNA strand)	...	A	A	C	...
Individual 1 (father - 2nd DNA strand)	...	A	A	C	...
Individual 2 (mother - 1st DNA strand)	...	G	A	A	...
Individual 2 (mother - 2nd DNA strand)	...	G	C	A	...
Individual 3 (child - 1st DNA strand)	...	A	A	C	...
Individual 3 (child - 2nd DNA strand)	...	G	A	C	...

Fig. 2. An example of a *trio* and their genotypes; A set of SNPs in the genomes of two parents and their offspring. The first highlighted column in the molecular sequence of the trio represents a SNP inconsistent with a loss of heterozygosity; the second highlighted column represents a SNP consistent with a loss of heterozygosity; the third highlighted column represents a SNP showing an evidence of a loss of heterozygosity.

such that each entry in $\mathcal{T}_X$ is either compatible with a deletion or is classified as a genotyping error.

Halldórsson *et al.* [10] assumed that functions $h(\rho)$ and $g(\eta)$ always assign the same cost to each $\rho \in \mathcal{R}$ and $\eta \in \mathcal{E}$, respectively. We relax this aspect and generalize the PLOHP to the case where we can have different costs depending on the SNP and deletion being considered. In fact, genotyping technologies are usually characterized by a high variability in the quality of the SNP genotypes produced [26]. A common method for dealing with these is to remove from analysis markers that show many ELOH events [22], this method however may remove most of the signal from the data in the preprocessing step. Similarly, different regions in the genome may have different propensity for carrying deletions [20]. This fact justifies the need to weigh the different SNPs based on their probability of being a genotyping error. Hence, in what follows we shall assume that functions h and g are generic functions.

The PLOHP is based on the *parsimony principle* [27]. This fact implies that the optimal solutions to the problem provide estimations of deletion events that, in the worst case, are lower bounds on the overall number of true deletion events occurred in the set of trios being considered [28], [29]. Halldórsson *et al.* [10] conjectured the general $\mathcal{NP}$-hardness of the PLOHP but did not investigate the issue any further. In the next sections we shall address this major issue and provide an algorithm able to exactly solve practical-use instances of the problem.

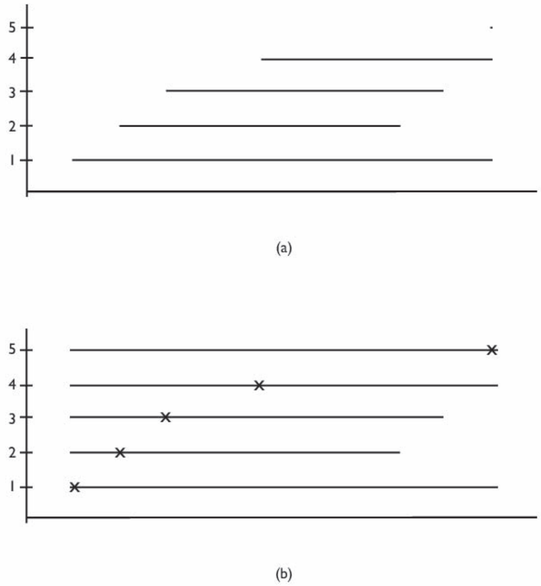


Fig. 4. An interval graph can be transformed into a LOHG by converting each interval $[l_k, r_k]$ in Figure (a) into the pointed interval $([1, r_k], l_k)$ as shown in Figure (b). The symbol 'x' represents the position of the k-th l_k point in the corresponding interval.

2.1 The LOH graph

We revisit the graphs defined in [10], which we shall term *LOH Graphs* and turn out useful in transforming the PLOHP into a particular version of the Minimum Clique Partition Problem (MCPP) [30].

In order to characterize such a class of graphs, consider a set of trios $\mathcal{T}$ and denote $I_k = ([l_k, r_k], x_k)$ as the interval induced by the k-th trio/SNP pair in $\mathcal{T}_X$, where x_k denote the position of the SNP having value 'X' in the interval and l_k and r_k denote the left and right margins of x_k, respectively. Moreover, denote $\mathcal{I}$ as the set of intervals induced by $\mathcal{T}$ and set $\nu = |\mathcal{I}|$. Consider a graph G_π having a vertex for each interval I_k, $k = 1, \ldots, |\mathcal{T}_X|$, and an edge between two vertices if a given intersection rule π is satisfied. If π concerns just the presence/absence of an intersection between two distinct intervals then G_π is a classical *interval graph* (see [31]). If the intersection rule π also involves the position x_k of the SNP having value 'X' in the interval then G_π is a *catch-point interval graph* (see [32]). If π concerns the presence/absence of mutual compatibility between two distinct intervals then G_π then becomes the symmetric restriction of a catch-point interval graph which is called *undirected catch-point interval graphs* or, more simply, a *LOH Graph* (LOHG).

The class of the LOHGs can be seen as a generalization of the class of the interval graphs. In fact, the following proposition holds:

Proposition 1. *The class of the LOHGs strictly contains the class of the interval graphs.*

Proof: A generic interval $[l_k, r_k]$ of an interval graph is completely characterized by the left and right margins l_k and r_k, respectively. In contrast, a generic interval of a LOHG G is completely characterized by the pair $([l_k, r_k], x_k)$, i.e., by an interval and a point belonging to that interval. Now, denote $\hat{l} = \min_{k \in \{1,2,\ldots,\nu\}} l_k$ and observe that we can transform any interval graph G_I into a LOHG G by mapping each interval $[l_k, r_k]$ of G_I into the interval $([\hat{l}, r_k], l_k)$ of G (see e.g., Figure 4). Hence, any interval graph is also a LOHG. To complete the proof, it is sufficient to show that the converse is not true. In fact, the class of interval graphs belongs to the class of the *perfect graphs* [33] which in turn implies that interval graphs do not contain odd cycles of length greater than or equal to 5 [34]. In contrast, the class of the LOHGs may also include such cycles, as shown e.g., in Figure 5. □

Given an instance of the PLOHP and its corresponding LOHG, G, we observe that by definition, the RCDs in $\mathcal{T}$ correspond to *cliques* in G, i.e., to complete subgraphs of G. We also recall that a *maximal clique* of a graph is a clique that is not a subset of a larger clique and a *maximum clique* is a clique of maximum size. Then, the following result holds for LOHGs:

Proposition 2. *Let G be a LOHG. Then G contains at most $\nu(\nu-1)/2$ maximal cliques.*

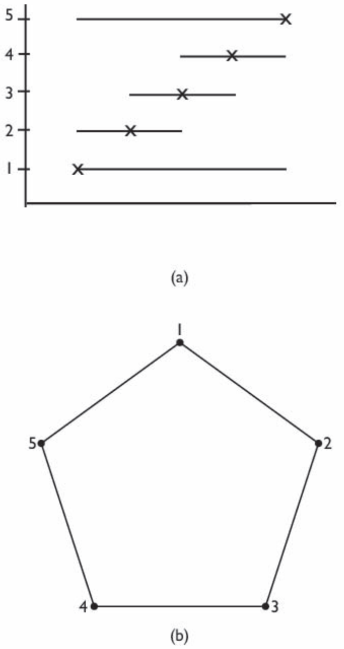


Fig. 5. A counter example showing that in general a LOHG is not a perfect graph. In fact, the set of pointed intervals shown in the subfigure (a) induces the odd cycle of length 5 shown in the subfigure (b).

Proof: Take any set of vertices that forms a maximal clique C in G and consider the corresponding set of pointed intervals in $\mathcal{I}$. Let v_l and v_r be the nodes whose corresponding 'x' values x_{v_l} and x_{v_r} are the furthest to the left and to the right, respectively, in the set of pointed intervals induced by C. Each vertex v in the clique is connected to these two vertices and its corresponding pointed interval is such that $x_v \in [x_{v_l}, x_{v_r}] \subseteq [l_v, r_v]$. Hence, a clique can be defined by the leftmost and rightmost vertices, respectively, and as consequence G contains at most $\nu(\nu-1)/2$ maximal cliques. □

We note that this implies that the maximum clique problem can be solved in polynomial time in a LOH Graph. Enumerating all cliques can be performed in polynomial time by choosing all possible distinct pairs of values x_{v_l} and x_{v_r} in $\mathcal{I}$ and, for each of them, by listing the pointed intervals such that $x_v \in [x_{v_l}, x_{v_r}] \subseteq [l_v, r_v]$.

In the Section 4 we shall exploit Proposition 2 to develop an exact approach to solution of the PLOHP based on integer programming.

3 The complexity of the PLOHP

Given an instance of the PLOHP and its corresponding LOHG, G, we note that in any optimal solution to the instance, a genotyping error will always be a SNP that does not belong to any RCD selected. Hence, in any optimal solution to the instance, a genotyping error will correspond to a clique of G having cardinality 1. This insight allows us to consider the PLOHP as an instance of the MCPP in a particular class of graphs [30]. The MCPP is known to be $\mathcal{NP}$-hard in general [30]; in this section we will show that the MCPP (i) remains hard even when restricted to the class of the LOHGs and (ii) can be solved in polynomial time if the LOHG and the cost functions h and g satisfy some specific properties. Before proceeding, we shall introduce some notation that will prove useful throughout the section.

We say that a set function f is *zero-cardinal* if $f(\emptyset) = 0$; *non-negative* if f assumes only non-negative values; and *non-decreasing* if $f(T) \leq f(S)$ for any $T \subseteq S \subseteq V$. We say that f is *submodular* if it satisfies the following property [35]:

$$f(S \cup \{u\}) + f(T) \leq f(S) + f(T \cup \{u\}) \\ \forall\ T \subseteq S \subseteq V,\ u \in V \setminus S.$$

We say that f is a *polymatroid rank function* if it is zero-cardinal, non-decreasing, non-negative, and submodular. Moreover, similarly to [36], we define a *value-polymatroid* set function f as a zero-cardinal, non-decreasing, non-negative set function that satisfies the following property:

$$f(S \cup \{u\}) + f(T) \leq f(S) + f(T \cup \{u\}) \\ \forall\ S, T \subseteq V : f(S) \geq f(T),\ u \in V \setminus (S \cup T).$$

Note that a value-polymatroid set function is also polymatroidal, but the converse is generally not true [36]. Finally, a set function f is *size-defined submodular* if there exists a function $\psi : [0 \ldots |V|] \to \mathbb{R}_0^+$ such that $f(S) = \psi(|S|)$, for any $S \subseteq V$. As shown in [36], a size-defined submodular set function f is both value-polymatroidal and polymatroidal.

Proposition 1 together with the previous definitions turn out to be useful to investigate the complexity of the PLOHP. Specifically, denote $\mathcal{C}(G)$ the set of cliques of G and set

$$f(C) = \begin{cases} 0 & \text{if } C = \emptyset \\ g(C) & \text{if } |C| = 1 \\ h(C) & \text{if } |C| \geq 2 \end{cases} \qquad \forall\ C \in \mathcal{C}(G).$$

Then, the following proposition holds:

Proposition 3. *The decision version of the PLOHP is $\mathcal{NP}$-complete even when the cost function f is restricted to polymatroidal set functions.*

Proof: The statement follows by observing that the class of the LOHGs strictly contains the class of interval graphs and that the minimum clique partition problem on an interval graph is $\mathcal{NP}$-complete when the cost function is polymatroidal [36]. □

In general, it is easy to realize that the decision version of the PLOHP is $\mathcal{NP}$-complete for any cost function $f(C) = \psi(|C|)\sigma(C)$ such that $\psi : [0 \ldots |V|] \to \mathbb{R}_0^+$ and $\sigma(C)$ is a generic function on C. In fact, such a case also includes the *rooted-TSP cost function on a tree* (see [36]) which is trivially polymatroidal.

Although Proposition 3 states that the decision version of the PLOHP is in general $\mathcal{NP}$-complete, it is worth noting that in some special cases the problem can be still solved in polynomial time. For example, the following proposition holds:

Proposition 4. *Let $G = (V, E)$ be a LOHG and $f : \mathcal{C}(G) \to \mathbb{R}_0^+$ a value-polymatroidal cost function. If G is also an interval graph then it is possible to compute a minimum cost partition into cliques of G in polynomial time.*

Proof: The statement follows from Proposition 1 and from the fact that the minimum clique partition problem on an interval graph can be solved in polynomial time when the cost function is a value-polymatroidal set function [36]. □

Proposition 4 turns out useful to show that if G is an interval graph the PLOHP can be solved in polynomial-time when the following objective function, introduced in [10], is used:

$$f_\alpha(C) = \begin{cases} 0 & \text{if } C = \emptyset \\ c_1 & \text{if } |C| = 1 \\ c_2 & \text{if } |C| \geq 2, \end{cases} \qquad \forall\ C \in \mathcal{C}(G) \qquad (1)$$

where c_1 and c_2 are two constants such that $0 < c_1 \leq \alpha c_1 < c_2 \leq (\alpha + 1)c_1$, and α is a positive integer such that $2 \leq \alpha \leq |V| - 1$. In fact, in such a case it is easy to see that the set function $f_\alpha(C)$ is size-defined submodular, hence value-polymatroidal; thus, if G is an interval graph, by Proposition 4 the PLOHP can be solved in polynomial time. Moreover, it is worth noting that the optimal solution to the problem can be characterized when considering function (1). In fact, the following proposition holds:

Proposition 5. *Consider a graph $G = (V, E)$ and a cost function f_α defined as in* (1). *Then, there exists a minimum cost partition into cliques of G, say $P^\star$, such that none of the cliques in $P^\star$ has cardinality greater than 1 and smaller than $\alpha + 1$. Moreover, if $P^\star$ contains cliques of cardinality greater than or equal to $\alpha + 1$ then at least one of them is a maximal clique of G.*

Proof: By contradiction, suppose there exists a clique $C \in P^\star$ such that $2 \leq |C| \leq \alpha$. Then, due to the nature of f_α, it is possible to obtain a lower cost partition into cliques of G by just breaking C into $|C|$ cliques of cardinality 1. In fact, in such a case we would have that $\sum_{v \in C} f_\alpha(\{v\}) = |C|c_1 < c_2 = f_\alpha(C) \leq (\alpha + 1)c_1$. However, this contradicts the hypothesis that $P^\star$ has minimum cost. Hence, $P^\star$ does not contain cliques having cardinality between 2 and α. Now, assume that $P^\star$ contains a clique $C \in P^\star$ such that $|C| \geq \alpha + 1$. Since f_α is non-decreasing we have that $f_\alpha(C) \geq f_\alpha(T)$, for any $T \in P^\star$. If C is not a maximal clique of G then there exists some $t \in V \setminus C$ such that $C \cup \{t\}$ is a clique in G. Note that t belongs to some $T \in P^\star \setminus C$ and that since f_α is non-decreasing, it holds that $f_\alpha(C) \geq f_\alpha(T) \geq f_\alpha(T \setminus \{t\})$.

Observe also that since f_α is α-value-polymatroidal, it holds that $f_\alpha(C \cup \{t\}) + f_\alpha(T \setminus \{t\}) \leq f_\alpha(C) + f_\alpha(T)$. Hence, it is possible to enlarge C until it becomes a maximal clique of G without getting worse the cost of $P^\star$. □

4 An integer programming model for the PLOHP

The $\mathcal{NP}$-hardness of the PLOHP justifies the development of exact and approximate solution approaches for the problem. In this section we present an integer programming model for the PLOHP. The algorithm is guaranteed to return an optimal solution and its time performance is significantly faster than the current state-of-the-art exact algorithm described in [10].

To this end, given a vertex $v \in V$, we denote $\mathcal{C}_v = \{C \in \mathcal{C}(G) : v \in C\}$. Moreover, we denote y_C as a decision variable equal to 1 if the clique $C \in \mathcal{C}(G)$ is selected in the optimal solution to the problem and 0 otherwise. Then, Formulation 1 is a valid formulation for the PLOHP.

Formulation 1.

$$\min \sum_{C \in \mathcal{C}(G)} f(C) y_C \tag{2a}$$

$$s.t. \sum_{C \in \mathcal{C}_v} y_C = 1 \qquad \forall\, v \in V \tag{2b}$$

$$y_C \in \{0,1\} \qquad \forall\, C \in \mathcal{C}(G). \tag{2c}$$

Constraints (2b) impose that each vertex $v \in V$ belongs to the clique $C \in \mathcal{C}(G)$ and constraints (2c) impose the integrality on variables y_C.

Formulation 1 is characterized by an exponential number of variables and constraints and its linear relaxation can be exactly solved by using column generation techniques. Specifically, observe that a variable with negative reduced cost in the linear relaxation of Formulation 1 corresponds to a dual constraint violated by the current dual solution. Denoted μ_v as the dual variables associated with constraints (2b), the dual of the linear relaxation of Formulation 1, denoted by LP1, is characterized by the following constraints:

$$\sum_{v \in V : v \in C} \mu_v \leq f(C) \qquad \forall\, C \in \mathcal{C}(G). \tag{3}$$

Constraints (3) are violated if there exists a clique $\hat{C} \in \mathcal{C}(G)$ such that $\sum_{v \in V : v \in C} \mu_v > f(C)$. The existence of such a clique can be checked in polynomial time by using Proposition 2 and this in turn implies that the linear relaxation of LP1 can be solved in polynomial time.

Interestingly, if the cost function f is defined as in (1) then Formulation 1 can be rewritten as follows. Denote x_v as a decision variable equal to 1 if vertex $v \in V$ forms a clique of cardinality 1 in the optimal solution to the problem and 0 otherwise. Moreover, denote $\hat{\mathcal{C}}(G)$ as the set of all maximal cliques in G having cardinality greater or equal to 2 and $\hat{\mathcal{C}}_v(G) = \{C \in \mathcal{C}(G) : v \in C, |C| \geq 2\}$. Then Formulation 2 is a valid formulation for the PLOHP.

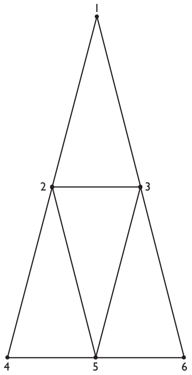


Fig. 6. A counter example showing that in general the linear relaxation of Formulation 2 is not integral.

Formulation 2.

$$\min \sum_{C \in \hat{\mathcal{C}}(G)} c_2 y_C + \sum_{v \in V} c_1 x_v \tag{4a}$$

$$s.t. \sum_{C \in \hat{\mathcal{C}}_v(G)} y_C + x_v \geq 1 \qquad \forall\, v \in V \tag{4b}$$

$$x_v \in \{0,1\} \qquad \forall\, v \in V \tag{4c}$$

$$y_C \in \{0,1\} \qquad \forall\, C \in \hat{\mathcal{C}}(G). \tag{4d}$$

Formulation 2 has the benefit of being polynomial-sized, due to Proposition 2, hence in principle its relaxation does not require the use of column generation techniques to be solved.

Both Formulations 1 and 2 can be strengthened by adapting appropriately the inequalities described in [37], [38], [39]. Moreover, additional valid inequalities can be considered. Specifically, given a pair of distinct vertices $v, w \in V$, we say that v *dominates* w if (i) $N(w) \subset N(v)$ and (ii) for any $C_v \in \hat{\mathcal{C}}_v(G)$ and $C_w \in \hat{\mathcal{C}}_w(G)$ it holds that $|C_v| > |C_w|$. In such a case, we also say that v is *dominating* and w is *dominated*. Dominated vertices are irrelevant to the clique partitioning of G, since in any optimal solution to the problem they will always be identified as cliques of cardinality one. Hence, both formulations can be strengthened by adding the following valid inequalities, whose proof trivially follows from Proposition 4 and the definition of domination:

Proposition 6. *Let v be a dominating vertex in V. Then, the inequality*

$$\sum_{C \in \hat{\mathcal{C}}_v(G)} y_C \geq 1 \tag{5}$$

is valid for both Formulations 1 and 2.

4.1 LOHP polyhedra

Remarkably, in our initial set of computational experiments we carried out on a set of real and random instances of the problem (see Section 5) we observed that when using function (1) the linear relaxation of Formulation 2 is always integral. This result could lead one to suspect that, due to the particular nature of function (1) and the structural properties of the class of the LOHGs, the polyhedron obtained by relaxing the integrality constraints (4c)-(4d) in Formulation 2 could coincide with the convex hull of all feasible solutions to the PLOHP. However, it is possible to provide at least two counter examples to this conjecture. Specifically, consider the LOHG shown in Figure 6 and assume, for ease of exposition, that $c_1 = 1$ and $c_2 = 2$. Then, it is easy to see that the value of the optimal integral solution for such a graph is 5 (obtained by taking any triangle and leaving the other three vertices isolated) while the value of the linear relaxation of Formulation 2 is 9/2 (obtained by taking 1/2 of each triangle and 1/2 of each vertex having degree 2).

Using a similar approach, a second case in which a fractional linear relaxation of Formulation 2 occurs is when considering the LOHG on 9 vertices obtained by merging the following four cliques $C_1 = \{1, 2, 3, 4\}$, $C_2 = \{2, 5, 6, 7\}$, $C_3 = \{1, 7, 8, 9\}$, $C_4 = \{3, 4, 5, 8\}$. In fact, by assuming again $c_1 = 1$ and $c_2 = 2$, the value of the optimal integral solution is 6 while the value of the linear relaxation is 5, leading to larger a integrality gap than the previous case. Although these two cases did not occur in our instances, there is no biological insight to exclude a-priori their existence in real instances of the problem. Hence, it may turn out useful to investigate possible valid inequalities to prevent the occurrence of at least both cases. To this end, it is worth noting that the following proposition holds:

Proposition 7. *Let G be the LOHG shown in Figure 6 and denote $\overline{C}$ as the clique formed by vertices 2, 3 and 5. Let v be a vertex in V such that $\hat{\mathcal{C}}_v(G) \cap \overline{C} \neq \emptyset$. Then, the inequality*

$$x_v \le y_{\overline{C}} + \sum_{u \in \hat{\mathcal{C}}_v(G) \cap \overline{C}} x_u \tag{6}$$

is valid for both Formulations 1 and 2.

Proof: In any feasible solution to the problem variables y and x can assume value either 0 or 1. If at least one variable in the right-hand-side of (6) is equal to 1 then the inequality is trivially valid. If all variables in the right-hand-side of (6) are equal to 0 then the inequality is still valid. In fact, denote C' as the unique maximal clique containing vertex v, due to constraints (4b) we have that $y_{C'} = 1$, which in turn implies that $x_v = 0$. □

It is easy to realize that Proposition 7 also holds for the second case above considered.

5 Experiments

In this section we analyze the performance of our model in solving instances of the PLOHP. Our experiments were motivated by three main goals: to measure the runtime performances of our model in tackling real instances of the PLOHP, compare the performances of our model versus Halldórsson *et al.*' exact algorithm, and to allow the exact analysis of datasets potentially larger than to the ones analyzed in [10]. We emphasize that our experiments aim simply to evaluate the performances of our model; we neither attempt to study the efficiency of our model for LOH estimations nor compare the accuracy of our model to LOH estimation solvers that use an objective function that is different from the one used in [10]. The reader interested in these topics is referred to [22], [23], [24].

Parameter sets	Site error probability	Interval length	Probability of an ELOH
Set 1	0.0001	5	0.75
Set 2	0.0001	2	1
Set 3	0.0001	9	0.75
Set 4	0.0001	7	0.50
Set 5	0.0033	9	0.75

TABLE 1
Parameter sets used to generate the first set of random instances of the PLOHP.

We tested our model on three datasets, namely: a set of 5 real instances the PLOHP from [10], characterized by 3000 trios having 3575 SNP each, and two sets of simulated instances of the PLOHP. The first set of simulated instances were generated using the same procedure described in [10], with the following list of tunable parameters: the *site error probability* i.e., the Mendelian error added to a set of considered trios according to a given probability uniformly distributed in $[0, 1]$ and assumed independent for each site; the *interval length* i.e., the exact length of the generated deletion; and the *probability of an ELOH* event within the generated deletion interval, uniformly distributed in $[0, 1]$. In the first set of simulated data we used 5 parameters sets (showed in Table 1). These instances were generated so as to have a similar structure as real datasets, but with a higher rate of ELOH sites, as higher rates of ELOH sites increase the time complexity of the algorithm. For each possible combination of parameter set, trios number, and SNP number we created 20 random instances of the problem by using the Mersenne Twister library [40] as pseudorandom number generator, for an overall number of 600 instances of the PLOHP per simulated set.

The second set of simulated instances was constructed with the sole purpose of determining which problem characteristics would cause the biggest difficulties for the algorithm presented. Here we varied the number of SNPs as 100 and 200 and also varied the number of trios as 100 and 200. We then fixed the number of sites consistent with LOH sites to 50% and varied the number of evidence of LOH sites in the set $0.1, 1, 5, 25, 45, 49, 49.9\%$,

ELOH	Time (sec.)	
Probability (%)	100x100	100x200
0.1	0.004±0.001	0.006±0.001
1	0.007±0.001	0.016±0.002
5	0.019±0.013	0.039±0.003
25	0.134±0.018	0.337±0.035
45	2.275±0.213	15.356±4.061
49	14.01±0.771	850.369±247.686
49.9	11.51±0.992	588.161±162.456
ELOH	Time (sec.)	
Probability (%)	200x100	200x200
0.1	0.005±0.001	0.01±0.002
1	0.013±0.001	0.04±0.005
5	0.036±0.001	0.116±0.009
25	0.279±0.015	0.852±0.104
45	8.444±2.056	70.653±18.177
49	742.231±148.937	2169.624±901.948
49.9	469.526±63.149	1092.288±182.949

TABLE 2
Average computing time required to solve the second set of random instances of the PLOHP.

letting all other sites be sites being inconsistent with LOH.

In order to compare the performances of our model versus Halldórsson *et al.*' exact algorithm, we used function (1) to deal with the random instances of the PLOHP and used the same coefficients described in [10] to set the constants c_1 and c_2. Moreover, in order to simulate the high variability in the quality of the SNP genotypes produced by genotyping technologies and the different propensity of the regions in the genome to carry deletions, we also considered an alternative objective function to analyze Halldórsson *et al.*' instances. Specifically, we used the first 3500 recombination rates between sexes, populations, and individuals [41] (appropriately rescaled in the interval $[0, 1]$) relative to chromosome 1 in order to associate a weight to each SNP of the considered real instances. Then, we computed the objective function as

$$f(C) = \begin{cases} 0 & \text{if } C = \emptyset \\ b * \alpha_r & \text{if } |C| = 1 \\ |C|\gamma_C & \text{if } |C| \geq 2, \end{cases} \qquad \forall\ C \in \mathcal{C}(G) \tag{7}$$

where b is a random number uniformly distributed in $[0, 1]$, α_r equal to the average rate in the considered chromosomic region, and γ_C is the average of the recombination rates associated to the SNPs involved in the clique C. Codes and datasets can be downloaded upon request.

5.1 Numerical results

Figure 7 shows the average runtime performances of the model when tackling the random instances of the PLOHP under different parameter settings. Similarly, Figure 8 shows the average number of nodes by the integer program when tackling the random instances of the PLOHP under different parameter sets. We observe that in a very large fraction of the integer programs no nodes are expanded, i.e. the problem is a solved at the root node, implying that a linear relaxation of the integer program provides optimal integer solutions.

As a general trend, we observed that our model constitutes a very tight formulation for the PLOHP, being characterized by gaps that are often very close to 0% and by an average number of branches that is largely very close to 1. This fact has a positive impact on the solution times and can be noted both in Figure 7 and in Figures 8. Specifically, Figure 7, 8 shows that the random instances of the problem can be solved within 1 hour computing time, and in some cases even within a minute. In contrast, in no case was the exact algorithm described in [10] able to tackle such instances within the considered limit runtime.

In Table 2 we vary the number of ELOH sites in our dataset. We observe that the solution time of the instances increases as the number of sites showing evidence of LOH increases (and the number of sites inconsistent with LOH decreases), up to the point where there are very few sites being inconsistent with LOH, at which point the solution time decreases again.

A closer look to the runtime shows that, independent of the parameter set considered, the average solution times appear to grow exponentially with the size of the instance. The parameter sets influence the average solution times by determining their scale factor. In general, the time required to solve the instances increases when the length of the deletion is increased and when the probability of an ELOH decreases. In more detail, for a fixed generic instance of the PLOHP, we observed that the vast majority of compute time is taken by the function that lists all possible maximal cliques in $\hat{\mathcal{C}}(G)$, while the solution time of the model tout court is usually comparatively very short. We observed that when the instances had a very large number of ELOH sites then the compute time became large. A large part of this compute time was spent on generating the problem instances as the number of cliques became quite large. Possibly, the use of column generation techniques, although not strictly necessary when dealing e.g., with Formulation 2, together with the use of graph decomposition methods and divide and conquer techniques could improve the solution time of the model and make it less computationally demanding.

A second interesting observation from the experiments is the low average number of branches performed by the model. Specifically, Figures 8 show that the number of branches is often quite close to 1; nevertheless exceptions do arise (see e.g., Figure 8 Parameter Sets 1 and 4: some instances are characterized by larger numbers of expanded nodes). This fact could appear in contrast with the trend of the gap, usually equal to 0%. However, it is

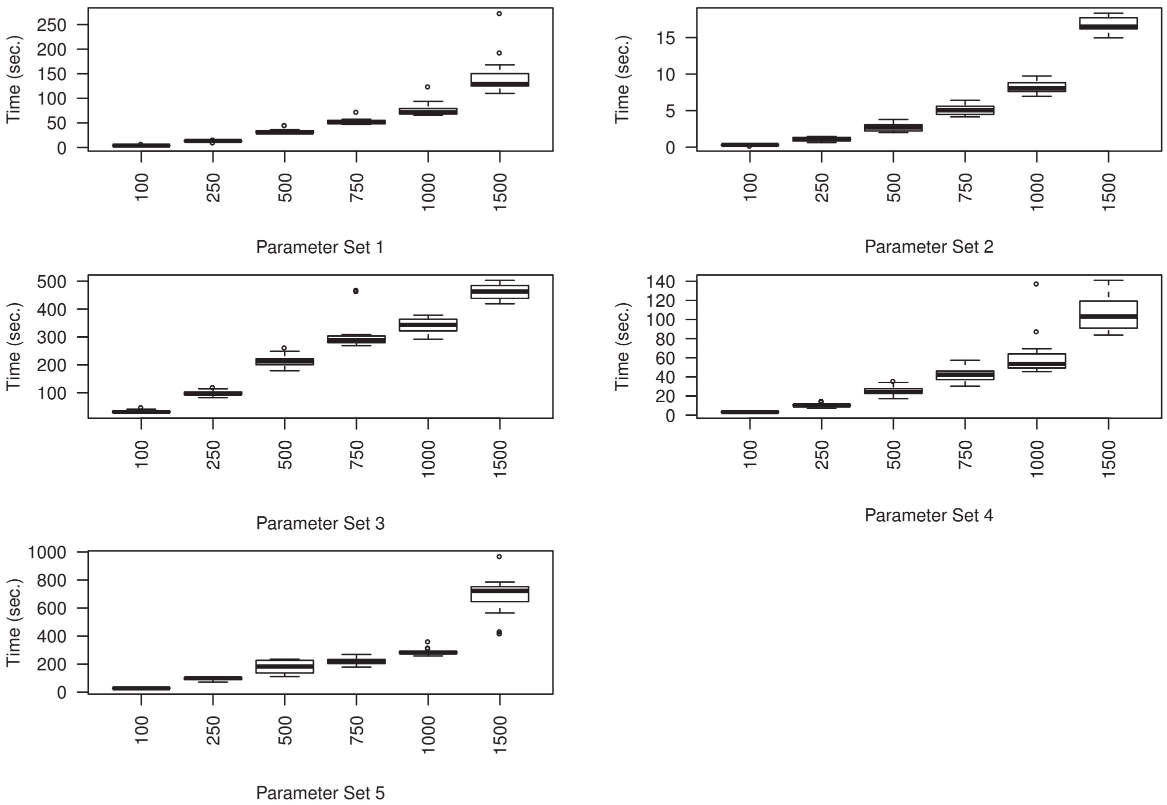


Fig. 7. The solution times on the first set of random instances under parameter sets 1, 2, 3, 4, and 5.

worth noting that the number of branches performed by the solver does not include only the branches in the search tree but also the branches performed by the heuristic strategy to find a primal bound to the instance. While the former is usually zero (i.e., the instance is solved at the root node) the latter sometimes may not, causing therefore some larger numbers of expanded nodes in some instances.

In Tables 3, 4 and 5 we compare the runtimes in the Halldórsson *et al.* under different parameter settings. In all of the instance in Tables 3, 4 and 5 the optimal solution was found at the root node without any branching.

In Table 3 we observe that the run time does not appear to be affected by the choice of the parameters $c1$ and $c2$. We observed longer runtimes when considering the objective function (7). Specifically, in no case was the model able to tackle Halldórsson *et al.*' instances within 12 hours. We observed that in this case the longer runtimes were caused by longer solution times (i.e., longer Simplex iterations and heuristic strategy runtimes) rather than longer model generation times. This phenomenon is a further confirmation of the hardness of the PLOHP instances when rates are considered. In order to provide a better evidence of this phenomenon, we analyzed the leading principal submatrices 1000x1500 and 2000x2000 of each instance in Halldórsson *et al.*' dataset, both in absence and in presence of rates (i.e., both when considering the objective functions (1) and (7), respectively; see Tables 4 and 5). The results showed that, when considering the rates, the solution times increased from 105% up to 317%, showing a similar trend in terms of gaps and nodes. It is possible that the use of column generation techniques and divide and conquer strategies (e.g., graph decomposition methods) could improve the solution time of the model and allow for quicker solution time of these instances. This however, is outside of the scope of the present article.

Instance	c1	c2	Time (sec.)
gens1	1	2	14741.9
gens2	1	2	17479.6
gens3	1	2	17205.7
gens4	1	2	13914.5
gens5	1	2	8318.1
gens1	2	11	15049.5
gens2	2	11	15975.8
gens3	2	11	15234.1
gens4	2	11	13490.8
gens5	2	11	15337.4

TABLE 3
Performance on Halldórsson *et al.*' instances.

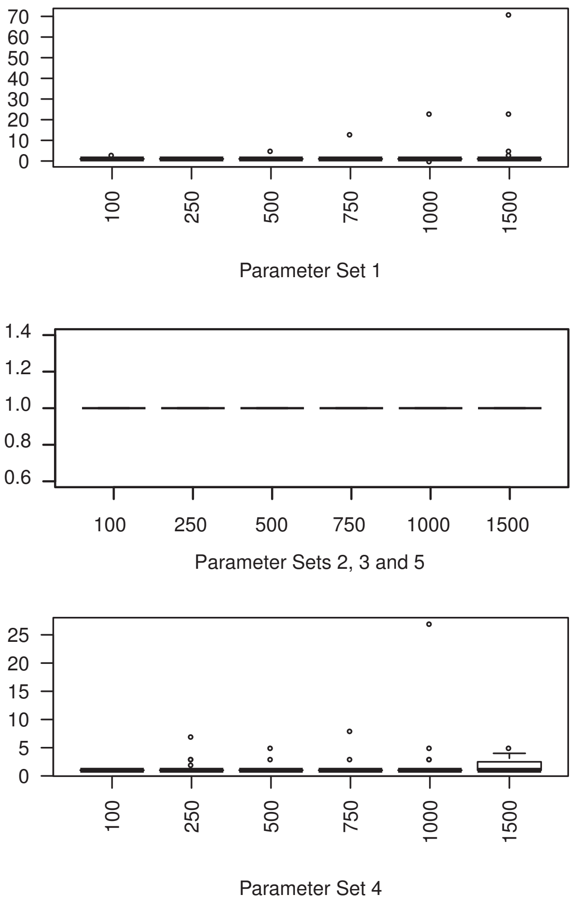


Fig. 8. The number of nodes expanded on the first set of random instances under parameter sets 1, 2, 3, 4, and 5.

Instance	# SNPs	# Trios	Time (sec.)
gens1	1500	1000	404.76
gens2	1500	1000	401.23
gens3	1500	1000	442.85
gens4	1500	1000	399.51
gens5	1500	1000	444.49
gens1	2000	2000	982.49
gens2	2000	2000	900.74
gens3	2000	2000	886.69
gens4	2000	2000	920.76
gens5	2000	2000	985.95

TABLE 4
Performance on Halldórsson *et al.*' instances when assuming objective function (1) and the constants $c_1 = 2$ and $c_2 = 11$.

Instance	# Trios	# SNPs	Time (sec.)
gens1	1000	1500	458.36
gens2	1000	1500	457.89
gens3	1000	1500	442.85
gens4	1000	1500	352.61
gens5	1000	1500	999.07
gens1	2000	2000	2823.77
gens2	2000	2000	2883.34
gens3	2000	2000	2852.16
gens4	2000	2000	2926.84
gens5	2000	2000	2878.87

TABLE 5
Performance on subinstances of Halldórsson *et al.*' instances when assuming objective function (7).

6 Conclusion

In this article we investigated the *Parsimonious Loss of Heterozygosity Problem (PLOHP)*, i.e., the problem of partitioning suspected polymorphisms of a set of individuals into the minimum number of deletion areas. Specifically, we showed that the PLOHP can be formulated as particular instance of the clique partition problem in a *rule-constrained interval graph* G_π, i.e., an interval graph having an edge between two vertices when a secondary intersection rule π is satisfied, and we proved the general $\mathcal{NP}$-hardness of the problem. Moreover, we extended the results described in Halldórsson *et al.* [10] by providing a state-of-the-art integer programming formulation and a possible strengthening valid inequalities able to exactly solve real instances of the PLOHP containing up to 9.000 individuals and 3000 SNPs within 12 hour computing time. Our results give perspective on the development of future approaches to solution of the problem that may turn out to be useful in practical applications.

Acknowledgments

The first author acknowledges support from the Belgian National Fund for Scientific Research (F.N.R.S.), of which he is "Chargé de Recherches". Both the first and the second authors also acknowledge support from Communauté Française de Belgique — Actions de Recherche Concertées (ARC) and "Ministerio de Ciencia e Innovación" through the research project MTM2009-14039-C06. Part of this work has been developed when Dr. Catanzaro was visiting Reykjavik University.

References

[1] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, no. 18, pp. 851–861, 2007.

[2] W. H. Li and L. A. Sadler, "Low nucleotide diversity in man," *Genetics*, vol. 129, pp. 513–523, 1991.

[3] D. G. Wang, J.-B. Fan, C.-J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander, "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome," *Science*, vol. 280, no. 5366, pp. 1077–1082, 1998.

[4] M. Cargill and *et al.*, "Characterization of single-nucleotide polymorphisms in coding regions of human genes," *Nature Genetics*, vol. 22, pp. 231–238, 1999.

[5] M. Halushka, J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti, "Patterns of single nucleotide polymorphisms in candidate genes of blood pressure homeostasis," *Nature Genetics*, vol. 22, pp. 239–247, 1999.

[6] J. Terwilliger and K. Weiss, "Linkage disquilibrium mapping of complex disease: Fantasy and reality?" *Current Opinions in Biotechnology*, vol. 9, pp. 579–594, 1998.

[7] M. Hoehe, K. Kopke, B. Wendel, K. Rohde, C. Flachmeier, K. Kidd, W. Berrettini, and G. Church, "Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence," *Human Molecular Genetics*, vol. 9, pp. 2895–2908, 2000.

[8] D. Catanzaro, M. Andrien, M. Labbé, and M. Toungouz-Nevessignsky, "Computer-aided human leukocyte antigen association studies: A case study for psoriasis and severe alopecia areata," *Human Immunology*, vol. 71, no. 8, pp. 783–788, 2010.

[9] H. Stefansson, D. Rujescu, S. Cichon, O. P. H. Pietiläinen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, T. Hansen, K. D. Jakobsen, P. Muglia, C. Francks, P. M. Matthews, A. Gylfason, B. V. Halldorsson, D. Gudbjartsson, T. E. Thorgeirsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, A. Bjornsson, S. Mattiasdottir, T. Blondal, M. Haraldsson, B. B. Magnusdottir, I. Giegling, H. J. Moeller, A. Hartmann, K. V. Shianna, D. Ge, A. C. Need, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, T. Paunio, T. Toulopoulou, E. Bramon, M. D. Forti, R. Murray, M. Ruggeri, E. Vassos, S. Tosato, M. Walshe, T. Li, C. Vasilescu, T. W. Moehleisen, A. G. Wang, H. Ullum, S. Djurovic, I. Melle, J. Olesen, L. A. Kiemeney, B. Franke, C. Sabatti, N. B. Freimer, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, O. A. Andreassen, R. A. Ophoff, A. Georgi, M. Rietschel, T. Werge, H. Petursson, D. B. Goldstein, M. M. Nöthen, L. Peltonen, D. A. Collier, D. S. Clair, and K. Stefansson, "Large recurrent microdeletions associated with schizophrenia," *Nature*, vol. 455, pp. 232–236, 2008.

[10] B. Halldórsson, D. Aguiar, R. Tarpine, and S. Istrail, "The Clark phase-able sample size problem: Long-range phasing and loss of heterozygosity in GWAS," *Journal of Computational Biology*, vol. 18, no. 3, pp. 323–333, 2011.

[11] M. Goedert and M. G. Spillantini, "A century of alzheimer's disease," *Science*, vol. 314, no. 5800, pp. 777–781, 2006.

[12] D. A. Elder, K. Kaiser-Rogers, A. S. Aylsworth, and A. S. Calikoglu, "Type i diabetes mellitus in a patient with chromosome 22q11.2 deletion syndrome," *American Journal of Medical Genetics*, vol. 101, no. 1, pp. 17–19, 2001.

[13] M. Shinawi, T. Sahoo, B. Maranda, S. A. Skinner, C. Skinner, C. Chinault, R. Zascavage, S. U. Peters, A. Patel, R. E. Stevenson, and A. L. Beaudet, "11p14.1 microdeletions associated with ADHD, autism, developmental delay, and obesity," *American Journal of Medical Genetics*, vol. 155, no. 6, pp. 1272–1280, 2011.

[14] K. Momma, R. Matsuoka, and A. Takao, "Aortic arch anomalies associated with chromosome 22q11 deletion," *Pediatric Cardiology*, vol. 20, no. 2, pp. 97–102, 1999.

[15] C. M. Ogilvie, J. W. Ahn, K. Mann, R. G. Roberts, and F. Flinter, "A novel deletion in proximal 22q associated with cardiac septal defects and microcephaly: A case report," *Molecular Cytogenetics*, vol. 2, no. 9, pp. 1–5, 2009.

[16] S. Puvabanditsin, M. S. Nagar, M. Joshi, G. Lambert, E. Garrow, and E. Brandsma, "Microdeletion of 16p11.2 associated with endocardial fibroelastosis," *American Journal of Medical Genetics*, vol. 152, no. 9, pp. 2382–2386, 2010.

[17] J. McClellan and M. C. King, "Genetic heterogeneity in human disease," *Cell*, vol. 141, pp. 210–217, 2010.

[18] D. Catanzaro and M. Labbé, "The pure parsimony haplotyping problem: Overview and computational advances," *Intenational Transactions in Operations Research*, vol. 16, no. 5, pp. 561–584, 2009.

[19] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, T. W. T. C. C. Consortium", C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles, "Origins and functional impact of copy number variation in the human genome," *Nature*, no. 464, pp. 704–712, 2009.

[20] J. A. Corbel, A. Abyzov, X. Mu, N. Carreiro, P. Cayting, Z. Zhang, M. Snyder, and M. Gerstein, "PEMer: a computational framework with simulation-based error models: for inferring genome structural variants from massive paired-end sequencing data," *Genome Biology*, vol. 38, no. 10, p. R23, 2009.

[21] K. Chen, J. Wallis, M. McLellan, D. Larson, J. Kallick, C. Pohl, S. McGrath, M. Wendl, Q. Zhang, D. Locke, X. Shi, R. Fulton, T. Ley, R. Wilson, L. Ding, and E. Mardis, "BreakDancer: an algorithm for high resolution mapping of genomic structural variation," *Nature Methods*, vol. 6, pp. 677–81, 2009.

[22] D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles, and J. K. Pritchard, "A high-resolution survey of deletion polymorphism in the human genome," *Nature Genetics*, vol. 38, pp. 75–81, 2006.

[23] E. Corona, B. Raphael, and E. Eskin, "Identification of deletion polymorphisms from haplotypes," in *RECOMB 2007 - Proceedings of the 11th annual international conference on Research in computational molecular biology*, ser. Lecture Note in Computer Science, S. Istrail, P. Pevzner, and M. Waterman, Eds. Springer, NY, 2007, pp. 354–365.

[24] S. A. McCarroll, F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. H. Shapero, P. I. W. de Bakker, J. B. Maller, A. Kirby, A. L. Elliott, M. Parkin, E. Hubbell, T. Webster, R. Mei, J. Veitch, P. J. Collins, R. Handsaker, S. Lincoln, M. Nizzari, J. Blume, K. W. Jones, R. Rava, M. J. Daly, S. B. Gabriel, and D. Altshuler, "Integrated detection and population-genetic analysis of snps and copy number variation," *Nature Genetics*, vol. 40, pp. 1166–1174, 2008.

[25] D. Catanzaro, M. Labbé, and L. Porretta, "A class representative model for pure parsimony haplotyping under uncertain data," *PLoS one*, vol. 6, no. 3, p. e17937, 2011.

[26] T. I. H. Consortium, "The international hapmap project," *Nature*, vol. 426, no. 18, pp. 789–796, 2003.

[27] V. A. Albert, *Parsimony, phylogeny, and genomics*. Oxford Bioscience, UK, 2005.

[28] D. Catanzaro, "The minimum evolution problem: Overview and classification," *Networks*, vol. 53, no. 2, pp. 112–125, 2009.

[29] ——, "Estimating phylogenies from molecular data," in *Mathematical approaches to polymer sequence analysis and related problems*, R. Bruni, Ed. Springer, NY, 2011.

[30] M. R. Garey and D. S. Johnson, *Computers and Intractability: A guide to the theory of NP-Completeness*. Freeman, NY, 2003.

[31] P. C. Fishburn, *Interval orders and interval graphs: Study of partially ordered sets*. John Wiley and Sons Inc., NY, 1985.

[32] E. Prisner, "A characterization of interval catch digraphs," *Discrete Mathematics*, vol. 73, pp. 285–289, 1989.

[33] L. Lovász, "Perfect graphs," in *Selected topics in graph theory*, L. W. Beineke and R. J. Wilson, Eds. Academic Press, NY, 1983, vol. 2, pp. 55–87.

[34] M. Chudnovsky, N. Robertson, P. Seymour, and R. Thomas, "The strong perfect graph theorem," *Annals of Mathematics*, vol. 164, no. 1, pp. 51–229, 2006.

[35] A. Schrijver, *Combinatorial optimization: Polyhedra and efficiency*. Springer, NY, 2003.

[36] G. Dion, V. Jost, and M. Queyranne, "Clique partitioning of interval graphs with submodular costs on the cliques," *RAIRO Operations Research*, vol. 41, pp. 275–287, 2007.

[37] M. Grötschel and Y. Wakabayashi, "Facets of the clique partitioning polytope," *Mathematical Programming*, vol. 47, pp. 367–387, 1990.

[38] H. J. Bandelt, M. Oosten, J. H. G. C. Rutten, and F. C. R. Spieksma, "Lifting theorems and facet characterization for a class of clique partitioning inequalities," *Operations Research Letters*, vol. 24, no. 5, pp. 235–243, 1999.

[39] M. Oosten, J. H. G. C. Rutten, and F. C. R. Spieksma, "The clique partitioning problem: Facets and patching facets," *Networks*, vol. 38, no. 4, pp. 209–226, 2001.

[40] M. Matsumoto and T. Nishimura, "Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM Transactions on Modeling and Computer Simulation*, vol. 8, no. 1, pp. 3–30, 1998.

[41] A. Kong, G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, K. T. Kristinsson, S. A. Gudjonsson, M. L. Frigge, A. Helgason, and U. T. K. Stefansson, "Fine-scale recombination rate differences between sexes, populations and individuals," *Nature*, vol. 467, pp. 1099–1103, 2010.

Daniele Catanzaro received the BS-Eng. from the Department of Electrical Engineering and Computer Science of the University of Palermo, Italy, and the PhD from the Computer Science Department of the Free University of Brussels, Belgium, in 2008. Currently, he is Chargé de Recherches at the Belgian National Fund for Scientific Research (FNRS) and affiliated to the Graphs and Mathematical Optimization Unit of the Free University of Brussels.

Martine Labbé received the BS-Math. and PhD degrees from the Department of Mathematics of the Free University of Brussels, Belgium, the last in 1985. After being visiting professor at Université Louis Pasteur, France, and assistant professor at the Econometrisch Instituut of the Erasmus Universiteit, the Netherlands, she joined the Department of Computer Science of the Free University of Brussels, where currently she is full professor in operations research.

Bjarni V. Halldrsson received a BS-Math. degree from the University of Iceland and a PhD degree in Algorithms, Combinatorics and Optimization from Carnegie Mellon University, Pittsburgh, PA, USA, 2001. He was a computer scientist at Celera Genomics 2001-2004, a statistician at deCODE genetics 2004-2010. Bjarni joined Reykjavk University in 2006, where he is currently an Associate Professor.