PROBABILISTIC RECONSTRUCTION OF GENEALOGIES FOR POLYPLOID PLANT SPECIES

PROÏA FRÉDÉRIC, PANLOUP FABIEN, TRABELSI CHIRAZ, AND CLOTAULT JÉRÉMY

Abstract. A probabilistic reconstruction of genealogies in a polyploid population (from 2x to 4x) is investigated, by considering genetic data analyzed as the probability of allele presence in a given genotype. Based on the likelihood of all possible crossbreeding patterns, our model enables to infer and to quantify the whole potential genealogies in the population. We explain in particular how to deal with the uncertain allelic multiplicity that may occur with polyploids. Then we build an *ad hoc* penalized likelihood to compare genealogies and to decide whether a particular individual brings a sufficient information to be included in the taken genealogy. This decision criterion enables us in a next part to suggest a greedy algorithm in order to explore missing links and to rebuild some connections in the genealogies, retrospectively. As a by-product, we also give a way to infer the individuals that may have been favored by breeders over the years. In the last part we highlight the results given by our model and our algorithm, firstly on a simulated population and then on a real population of rose bushes. Most of the methodology relies on the maximum likelihood principle and on the graph theory.

1. Introduction

1.1. **Motivations.** Pedigrees depict the genealogical relationships between individuals of a given population. They can be built thanks to mating knowledge or they can be inferred from molecular markers. The identification of pedigrees allows a broad variety of applications: genealogy identification, like in grapevine [12], improvement of conservation programs for endangered species [14], inference of statistics used in quantitative and population genetics like heritability or population effective size [1, 10], etc. Like for most population genetics analyses, pedigree reconstruction methods and their implementation were firstly developed for diploid species (but see [21]). Polyploids, *i.e.* species with more than two alleles for a given locus, represent approximately 25% of plant species [2], and among them a large number of cultivated species. Polyploidy in animals is more rare but some examples were described in insects, fishes, amphibians and reptiles [18, 15].

Several strategies were used to reconstruct the genealogical relationships from molecular markers (reviewed in [9]). Exclusion methods eliminate potential parents which do not show at least one allele per locus shared with a putative offspring. If more than two parents are possible, categorical allocation methods allow to identify the most likely parents according to their probability to transmit alleles shared with the potential progeny. Parental reconstruction methods use full- or half-siblings in order to identify the most likely parents. By comparison, sibling reconstruction methods add a preliminary step of inference of siblings when they are unknown. In this paper, the objective is to adapt and to extend the approach

Key words and phrases. Allelic multiplicity, Crossbreeding patterns, Genealogies of plant species, Graph theory, Maximum likelihood principle, Missing links, Pedigree reconstruction, Polyploid population.

of [4], namely to determine for each individual the most likely couple of parents amongst all older individuals, so as to build some family trees in polyploid plant species. Our study certainly intends to be applied on real genetic datasets, in particular the main practical motivation is to find some retrospective links in a population of rose bushes that will now be described.

The empirical dataset used in the last section of this article was obtained on cultivated roses bred mainly during the nineteenth century (*Rosa* sp.). Rose breeding activities were particularly abundant during the nineteenth centuries and were very documented. As an example, breeding year is known for a majority of roses from this period. However, the genealogical relationships described in archives are highly hypothetical, due to the lack of control of artificial hybridization until the end of the nineteenth century. Among the approximately 200 species of the genus *Rosa*, ploidy level varies between 2x and 10x [8]. Rose breeding activities from the nineteenth century involved interspecific crossings between diploid species and tetraploid species, with a small contribution of genotypes with higher ploidy like species from the *Caninae* section (4x, 5x and 6x) [17, 13]. Cultivated roses bred during the nineteenth century can exhibit all ploidy values between 2x and 6x, even if 5x and 6x are rare [13]. The mode of inheritance in these rose cultivars remains highly unknown. It is generally considered that modern tetraploid cultivated roses exhibit a tetrasomic inheritance (no preferred pairing among the set of four homologous chromosomes and creation of tetravalents during meiosis) [11]. But a mixture of disomic (preferred pairing of two bivalent pairs during meiosis according to their genomic similarity) and tetrasomic inheritance could be observed according to chromosomes and according to genotypes [3]. Triploid roses have played a major role in rose hybridizations. Like in other species, triploid roses exhibit a low fertility rate, due to irregular meiosis leading to aneuploidy [16]. However, even if the production of fertile gametes from triploids remains rare, these events were selected by breeders, especially as bridges between different ploidy levels. For example, *Bourbon*, *Hybrid China* and *Hybrid Tea* rose groups were both obtained by a cross between a Chinese diploid cultivar and a European tetraploid cultivar. First cultivars from these groups were triploid [7]. Triploids form both haploid and diploid gametes [20]. Following the obtention of a variety by hybridization, it was then propagated vegetatively by cutting or grafting and often conserved in rose gardens. Therefore rose varieties can be considered as immortal and they could have been involved at different periods in rose pedigrees. As most of plants, roses are hermaphrodites and can therefore have been used as female or male on different hydridization events. Selfing rate in roses is very low mainly because of self-incompatibility ([19] and J. Mouchotte, pers. comm.). These specific breeding behaviors are the cornerstone of our probabilistic model.

In a general way, the polyploidy of the population may give rise to complications in terms of multiplicity of the alleles, being only aware of their presence or absence: that will be one of our strategic challenges to deal with this lack of information, widely discussed throughout the manuscript. Whereas for diploids the presence or absence of alleles is sufficient – for $\{a\}$ and $\{a, b\}$ undoubtedly correspond to $\{a, a\}$ and $\{a, b\}$ – the observation of $\{a, b\}$ for a tetraploid can correspond to $\{a, a, a, b\}$, $\{a, a, b, b\}$ or $\{a, b, b, b\}$. Reading the presence or absence of alleles on electrophoregrams and interpreting theoretical ratios between peak intensities is an option to determine the number of copies of each allele [6]. Unfortunately, we will explain in good time the reasons why this strategy is not reliable in our context and

we will introduce a way to deal with this allelic multiplicity through the intermediary of probabilities related to each configuration. Before getting to the heart of the matter, let us point out that the objective of this work is not to introduce a biological issue, but rather to build and justify the more realistic mathematical framework regarding the biological model of roses bred during the nineteenth century. This work is above all a methodological one.

The paper is organized as follows. In Section 2, we present a probabilistic method in order to reconstruct genealogies for species with several ploidy levels, from 2x to 4x, by considering genetic data analyzed as the probability of allele presence in a given genotype. In particular, we compute the likelihood associated with all crossbreeding patterns and we explain how to build and quantify the whole possible genealogies of the population and how to treat the unknown allelic multiplicity. As a by-product we also give a way to find the individuals favored by breeders, retrospectively. Section 3 treats the isolated individuals, more precisely, the missing links. Under some criteria, we suggest an algorithm computing virtual individuals to improve the genealogy. Whereas Sections 2 and 3 are mainly theoretical, all our results will be tested in Section 4, both on a simulated population and on a rose bushes population. We conclude by highlighting some weaknesses of our methodology and by giving, in accordance, some trails for future studies.

1.2. **Preliminary considerations and notations.** In the whole paper, $\mathcal{P}$ stands for the population of size $n = \mathrm{Card}(\mathcal{P})$ and m is the number of genes involved in the reconstruction process. Technically, m corresponds to the number of signals on which we read the *peaks*, expressing the set of alleles detected on each gene. We make the crucial hypothesis that signals are *mutually independent*, which can be argued on a genetic as well as statistical point of view (genes are chosen for their absence of known interaction and a prior statistical treatment tends to decorrelate them). For an individual $e \in \mathcal{P}$, we denote by $g_s(e)$ the *genotype* of gene s, that is, the *set of alleles* present for this gene, shortened in $g(e)$ when we deal with an unspecified gene (to be precise, we should in fact speak of *multiset* since we may have multiple instances of the same allele in the genotype, however we shall not make this kind of distinctions). We also denote by $x(e) = \mathrm{Card}(g(e)) \in \{2, 3, 4\}$ the *ploidy* of e, the number of sets of chromosomes in a cell. In addition, we assume that the birth dates are known and that no death occurs, which is consistent with the fact that the work is related to plant cultivars. We also assume that gametes are produced according to strict polysomic inheritance and we neglect double reduction.

2. Likelihood of a genealogy

This section is the heart of the paper. Firstly we will describe the genetic patterns that we retain to cross the polyploid individuals, and we will discuss on the probabilistic treatment of the allelic multiplicity that may appear for triploids and tetraploids. Thereafter, we will be in the position to estimate some retrospective links and to compute an *ad hoc* penalized likelihood for the genealogy.

2.1. **Crossbreeding patterns.** To simplify the combinatorial analysis, we use the following natural models. Diploids produce haploid gametes, genotype $\{a, b\}$ leads to gametes $\{a\}$ and $\{b\}$ with probability 1. Triploids produce haploid and diploid gametes, genotype $\{a, b, c\}$ leads to gametes $\{a\}$ and $\{b, c\}$ with probability $\frac{1}{3}$, gametes $\{b\}$ and $\{a, c\}$ with probability $\frac{1}{3}$ and gametes $\{c\}$ and $\{a, b\}$ with probability $\frac{1}{3}$. Tetraploids produce diploid

gametes, genotype $\{a, b, c, d\}$ leads to gametes $\{a, b\}$ and $\{c, d\}$ with probability $\frac{1}{3}$, gametes $\{a, c\}$ and $\{b, d\}$ with probability $\frac{1}{3}$ and gametes $\{a, d\}$ and $\{b, c\}$ with probability $\frac{1}{3}$. In addition, each individual can either be male or female, the set of gametes is treated as an *urn problem*. Crossing is made by choosing at random two gametes among all these possibilities, bringing them together to obtain the offspring's genotype. Figures 1–2–3 are schematic representations of the gametes production, indicated by arrows, of a parent cell.

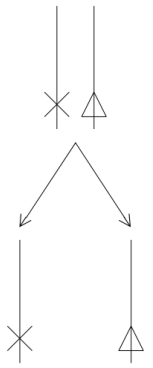

FIGURE 1. Schematic representation of the gametes production (in the bottom) for a diploid cell (in the top). Symbols represent the alleles of a given gene on its chromosome (line).

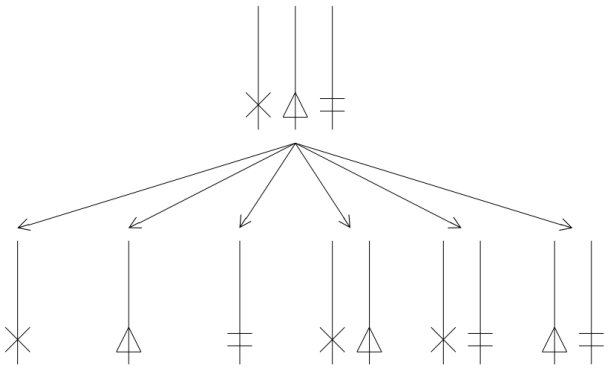

FIGURE 2. Schematic representation of the gametes production (in the bottom) for a triploid cell (in the top). Symbols represent the alleles of a given gene on its chromosome (line).

Let p_1 and p_2 be two individuals having ploidies $x(p_1)$ and $x(p_2)$ with genotypes $g(p_1) = \{a_1, \ldots, a_{x(p_1)}\}$ and $g(p_2) = \{b_1, \ldots, b_{x(p_2)}\}$, respectively. In the sequel, p_1 and p_2 are the parents of the offspring e. The different ploidy levels lead to six patterns that one is now going to describe in details.

(P_1) $x(p_1) = x(p_2) = 2$. Let $g(p_1) = \{a_1, a_2\}$ and $g(p_2) = \{b_1, b_2\}$. Then, e has 4 potential diploid genotypes $g(e) = \{a_i, b_k\}$. Each one has probability $\frac{1}{4}$.

(P_2) $x(p_1) = 2$ and $x(p_2) = 3$. Let $g(p_1) = \{a_1, a_2\}$ and $g(p_2) = \{b_1, b_2, b_3\}$. Then, e has 6 potential diploid genotypes $g(e) = \{a_i, b_k\}$, and 6 potential triploid genotypes $g(e) = \{a_i, b_k, b_\ell\}$. Each one has probability $\frac{1}{12}$.

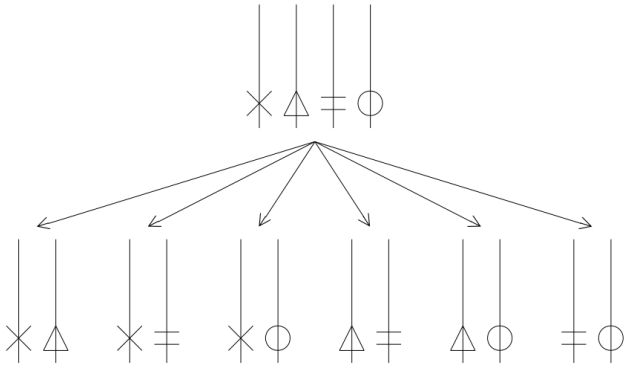

FIGURE 3. Schematic representation of the gametes production (in the bottom) for a tetraploid cell (in the top). Symbols represent the alleles of a given gene on its chromosome (line).

(P_3) $x(p_1) = 2$ and $x(p_2) = 4$. Let $g(p_1) = \{a_1, a_2\}$ and $g(p_2) = \{b_1, b_2, b_3, b_4\}$. Then, e has 12 potential triploid genotypes $g(e) = \{a_i, b_k, b_\ell\}$. Each one has probability $\frac{1}{12}$.

(P_4) $x(p_1) = x(p_2) = 3$. Let $g(p_1) = \{a_1, a_2, a_3\}$ and $g(p_2) = \{b_1, b_2, b_3\}$. Then, e has 9 potential diploid genotypes $g(e) = \{a_i, b_k\}$, 18 potential triploid genotypes $g(e) = \{a_i, b_k, b_\ell\}$ or $g(e) = \{a_i, a_j, b_k\}$, and 9 potential tetraploid genotypes $g(e) = \{a_i, a_j, b_k, b_\ell\}$. Each one has probability $\frac{1}{36}$.

(P_5) $x(p_1) = 3$ and $x(p_2) = 4$. Let $g(p_1) = \{a_1, a_2, a_3\}$ and $g(p_2) = \{b_1, b_2, b_3, b_4\}$. Then, e has 18 potential triploid genotypes $g(e) = \{a_i, b_k, b_\ell\}$, and 18 potential tetraploid genotypes $g(e) = \{a_i, a_j, b_k, b_\ell\}$. Each one has probability $\frac{1}{36}$.

(P_6) $x(p_1) = x(p_2) = 4$. Let $g(p_1) = \{a_1, a_2, a_3, a_4\}$ and $g(p_2) = \{b_1, b_2, b_3, b_4\}$. Then, e has 36 potential tetraploid genotypes $g(e) = \{a_i, a_j, b_k, b_\ell\}$. Each one has probability $\frac{1}{36}$.

To sum up, all diploid offsprings may come from patterns (P_1)–(P_2)–(P_4), all triploid offsprings from patterns (P_2)–(P_3)–(P_4)–(P_5) and all tetraploid offsprings from patterns (P_4)–(P_5)–(P_6). One can remark that the trickiest case is probably (P_4) since three different ploidies can be generated by crossing triploids, Figure 4 gives a streamlined representation of it. Now let $\{(p_1, p_2) \mapsto e\}$ be the event through which the pair (p_1, p_2) conceives e, let u denote the number of genotypes generated by the pattern ($u = 4$, $u = 12$ or $u = 36$) corresponding to the ploidy of p_1 and p_2, and let $e_1, \ldots, e_u$ name the potential offsprings of the cross. Our hypotheses show that, conditionally to the knowledge of the genotypes of the parents, each offspring is drawn through a uniform distribution. So, we set

$$\mathbb{P}(\{(p_1, p_2) \mapsto e\} \mid \{g(p_1), g(p_2), g(e)\}) = \frac{1}{u} \sum_{r=1}^{u} \mathbb{1}_{\{e_r = e\}} \tag{2.1}$$

where the genetic equality $e_r = e$ means that $g(e_r)$ and $g(e)$ coincide in a sense that we have to define. Specifically, we consider that $e_r = e$ once

$$g(e_r) = g(e) \quad \text{and hence} \quad x(e_r) = x(e) \tag{2.2}$$

which in this case amounts to say that e_r and e have the same ploidy and the same set of alleles (we remind that $x = \mathrm{Card}(g)$). However, it is important to highlight that (2.2) is only relevant from theoretical perspectives or on simulated data. We will see in Section 4.2

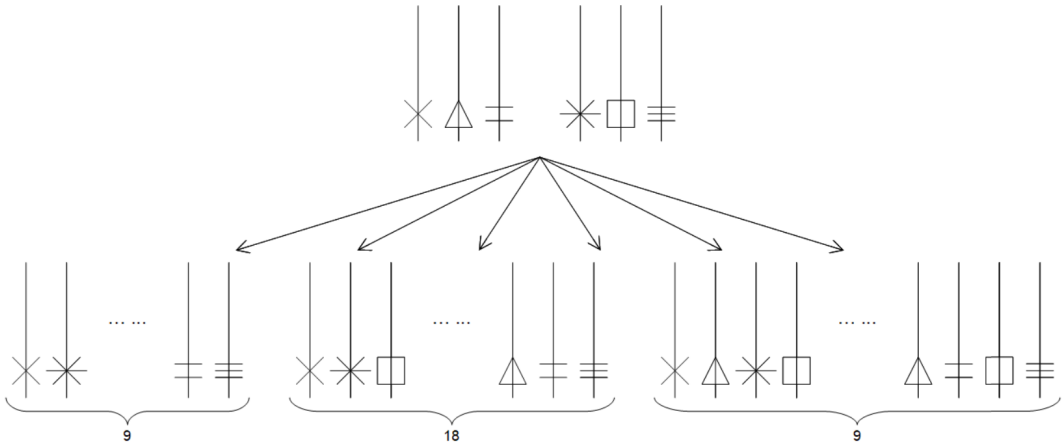

FIGURE 4. Schematic representation of pattern (P_4) leading to $u = 36$ potential offsprings including 9 diploids, 18 triploids and 9 tetraploids. Symbols represent the alleles of a given gene.

that real genotypes result from a calibration of the equipment and some rounded values to be interpreted as *base pairs*. Therefore,

$$x(e_r) = x(e) \quad \text{and} \quad \|g^*(e_r) - g^*(e)\|_\infty \leqslant 1 \tag{2.3}$$

where g^* stands for an ascending sorted vector containing the elements of g, should be an appropriate comparison on such data. Indeed, this criterion allows an offset of ± 1 base pairs for two corresponding alleles.

Examples. To illustrate this calculation method, let us consider $g(p_1) = \{a, a\}$ and $g(p_2) = \{a, b\}$. Then $u = 4$, the potential offsprings have genotypes $g(e_1) = g(e_3) = \{a, a\}$ and $g(e_2) = g(e_4) = \{a, b\}$. For $g(e) = \{a, a\}$ or $g(e) = \{a, b\}$, formula (2.1) gives probability $\frac{1}{2}$. It also gives probability 0 for all other genotypes. In the more intricate case where $g(p_1) = \{a, a, b, c\}$ and $g(p_2) = \{a, c, c\}$, then $u = 36$ and among the potential offsprings, 5 will have genotype $g(e) = \{a, b, c\}$. Formula (2.1) gives probability $\frac{5}{36}$ for such a triploid offspring.

2.2. **Allelic multiplicity.** For an individual $e \in \mathcal{P}$, the set $g(e)$ is the *true* genotype. However in our experimental studies, we only observe a *partial* genotype $\widehat{g}(e) \subset g(e)$ containing the distinct alleles – a set of peaks on the signal. Taking advantage of the ploidy $x(e)$, one is able to infer all possible $g(e)$ from $\widehat{g}(e)$. Explicitly, we use the following connections, where π names a probability of multiplicity in a generic way.

(C_1) $\widehat{g}(e) = \{a\}$ and $x(e) = 2$ lead to $g(e) = \{a, a\}$ with probability 1.
(C_2) $\widehat{g}(e) = \{a, b\}$ and $x(e) = 2$ lead to $g(e) = \{a, b\}$ with probability 1.
(C_3) $\widehat{g}(e) = \{a\}$ and $x(e) = 3$ lead to $g(e) = \{a, a, a\}$ with probability 1.
(C_4) $\widehat{g}(e) = \{a, b\}$ and $x(e) = 3$ lead to $g(e) = \{a, a, b\}$ with probability π_{21} and to $g(e) = \{a, b, b\}$ with probability π_{12}. We set $\pi_{21} + \pi_{12} = 1$.
(C_5) $\widehat{g}(e) = \{a, b, c\}$ and $x(e) = 3$ lead to $g(e) = \{a, b, c\}$ with probability 1.
(C_6) $\widehat{g}(e) = \{a\}$ and $x(e) = 4$ lead to $g(e) = \{a, a, a, a\}$ with probability 1.

(C$_7$) $\widehat{g}(e) = \{a, b\}$ and $x(e) = 4$ lead to $g(e) = \{a, a, a, b\}$ with probability π_{31}, $g(e) = \{a, a, b, b\}$ with probability π_{22} and $g(e) = \{a, b, b, b\}$ with probability π_{13}. We set $\pi_{31} + \pi_{22} + \pi_{13} = 1$.

(C$_8$) $\widehat{g}(e) = \{a, b, c\}$ and $x(e) = 4$ lead to $g(e) = \{a, a, b, c\}$ with probability π_{211}, $g(e) = \{a, b, b, c\}$ with probability π_{121} and $g(e) = \{a, b, c, c\}$ with probability π_{112}. We set $\pi_{211} + \pi_{121} + \pi_{112} = 1$.

(C$_9$) $\widehat{g}(e) = \{a, b, c, d\}$ and $x(e) = 4$ lead to $g(e) = \{a, b, c, d\}$ with probability 1.

Instead of selecting a genotype for e when several are conceivable, that is, for combinations (C$_4$)–(C$_7$)–(C$_8$), the model that we introduce in the next section takes account of all possibilities weighted by their related probabilities. In fact, our model enables to choose if necessary $\pi = \pi^{(s)}$ gene by gene or, equivalently, signal by signal, to consider the different interpretations of the relative amplitude of the peaks on each signal, for material reasons. We will describe it in more details in the beginning of Section 4.2.

2.3. **Probability of a genealogical link.** For any individual $e \in \mathcal{P}$, let $\mathcal{S}(e) \subset \mathcal{P}^2$ be the *compatible* subpopulation, that is, the set of non-ordered pairs (p_1, p_2) with $p_1 \neq p_2$ (excluding selfing) genetically and chronologically candidates to the genealogy of e. It is worth noting that the only chronological constraint is obviously to consider that birth dates of descendants cannot be prior to the ones of their parents. In particular, the probabilities of ancestry are considered as *time-invariant*: any individual has the same probability of being a parent, regardless of its birth date, excluding *de facto* any generational model like Galton-Watson trees. This point of view is specific to plant species, and would clearly be irrelevant for animal populations. Our objective is to build a probability measure on $\mathcal{S}(e) \cup \{\varnothing\}$ quantifying the whole possible genealogical links of e, the element $\varnothing$ being added to cover the case where no parents can be found in the population. The hypothesis of mutual independence of the signals allows us to work on each signal and to multiply the results. Let

$$\delta(e, p_1, p_2) = \prod_{s=1}^{m} \sum_{G \in \mathcal{G}_s} \mathbb{P}(\{(p_1, p_2) \mapsto e\} \,|\, G)\, \mathbb{P}(G) \tag{2.4}$$

where $\mathcal{G}_s$ is the set of all possible genotypes on signal s for the triplet (p_1, p_2, e). In the best-case scenario, $\mathrm{Card}(\mathcal{G}_s) = 1$ which means that $\widehat{g}_s(p_1)$, $\widehat{g}_s(p_2)$ and $\widehat{g}_s(e)$ lead to no uncertain allelic multiplicity, and thus $\mathbb{P}(G) = 1$. At worst, $\mathrm{Card}(\mathcal{G}_s) = 27$ meaning that $\widehat{g}_s(p_1)$, $\widehat{g}_s(p_2)$ and $\widehat{g}_s(e)$ are in the situation (C$_7$) or (C$_8$), and $\mathbb{P}(G)$ is the product of the related probabilities.

Example. Suppose that $x(p_1) = 3$, $x(p_2) = 4$, $x(e) = 4$ and that, on a particular signal s, we observe $\widehat{g}_s(p_1) = \{a, b\}$, $\widehat{g}_s(p_2) = \{a, c, d\}$ and $\widehat{g}_s(e) = \{a, d\}$. Then, $\mathrm{Card}(\mathcal{G}_s) = 18$. Indeed, we build $\mathcal{G}_s$ by combining $\{a, a, b\}$ and $\{a, b, b\}$ for p_1, $\{a, a, c, d\}$, $\{a, c, c, d\}$ and $\{a, c, d, d\}$ for p_2, and $\{a, a, a, d\}$, $\{a, a, d, d\}$ and $\{a, d, d, d\}$ for e. For the first combination we have $\mathbb{P}(G) = \pi_{21}^{(s)}\, \pi_{211}^{(s)}\, \pi_{31}^{(s)}$, for the second one $\mathbb{P}(G) = \pi_{21}^{(s)}\, \pi_{211}^{(s)}\, \pi_{22}^{(s)}$, and so on.

It only remains to renormalize. Explicitely, with

$$\Delta(e) = \sum_{(p_1, p_2) \in \mathcal{S}(e)} \delta(e, p_1, p_2) \tag{2.5}$$

where $\delta(e, p_1, p_2)$ is given in (2.4), let

$$\forall\,(p_1,p_2)\in\mathcal{S}(e),\quad \nu_e((p_1,p_2))=\begin{cases}\frac{\delta(e,p_1,p_2)}{\Delta(e)} & \text{if } \Delta(e)>0\\ 0 & \text{otherwise}\end{cases} \tag{2.6}$$

and fix $\nu_e(\varnothing) = 1$ as soon as $\Delta(e) = 0$, and $\nu_e(\varnothing) = 0$ otherwise. Then clearly, $\nu_e : \mathcal{S}(e) \cup \{\varnothing\} \to [0\,;1]$ is a probability measure that can be applied to look for the whole genealogy of $e \in \mathcal{P}$. To build the *most likely genealogy*, we must pick

$$c^*(e) = \underset{c\,\in\,\mathcal{S}(e)\,\cup\,\{\varnothing\}}{\arg\max}\ \nu_e(c). \tag{2.7}$$

To be precise, $c^*(e)$ defined as above is not necessarily unique, in such case we arbitrarily pick one optimum at random. We will see in the sequel that choosing a genealogical link amongst others is not necessarily relevant, hence we also consider

$$G(e) = \{c \in \mathcal{S}(e) \cup \{\varnothing\} \mid \nu_e(c) > 0\} \tag{2.8}$$

which represents the whole potential genealogical links of e in our population $\mathcal{P}$.

2.4. **A retrospective family tree.** Now the objective is to compute $G(e)$ – and thus $c^*(e)$ – for all $e \in \mathcal{P}$. In the framework of this study, a *family tree* of the population $\mathcal{P}$ is a set of triplets (p_1, p_2, e) having probabilities $\nu_e((p_1, p_2)) > 0$, on the basis of m genes, such that there is at most one triplet (p_1, p_2, e) for any individual e, interpretable as the realization of the event $\{(p_1, p_2) \mapsto e\}$, taking up the notation of the previous sections. In an equivalent way, we build a *graph* in which each individual is a vertex and each genealogical link is a couple of arcs (from the parents to the offspring). We also require that a triplet (p_1, p_2, e) is assigned to the node e of the family tree as soon as $c^*(e) \neq \varnothing$, that is as soon as there exists at least one potential genealogical link for e. Note that the chronological constraint applied on $\mathcal{S}(e)$ is sufficient to ensure that no cycle is present in the graph. The methods and algorithms that follow will be tested and applied in the Section 4.

2.4.1. *Most likely trees.* Combining all options of $G(e)$ for each $e \in \mathcal{P}$ gives an exhaustive set of trees, the whole potential genealogies of the population that we will denote as $\mathbb{G}(\mathcal{P})$ in (2.11). However, on large datasets, this can be difficult due to the exponential growth of the combinations. Thus we look for criteria of selection, and first we define the log-likelihood of a family tree $\mathcal{T}$ as follows,

$$\ell(\mathcal{T}) = \sum_{(p_1,p_2,e)\,\in\,\mathcal{T}} \ln \nu_e((p_1,p_2)). \tag{2.9}$$

Clearly $\mathcal{P}$ can be divided into $\mathcal{L} = \{e \in \mathcal{P} \,|\, c^*(e) \neq \varnothing\}$ and $\mathcal{I} = \{e \in \mathcal{P} \,|\, c^*(e) = \varnothing\}$, respectively the individuals having potential ancestors in the population, present as nodes in all family trees built according to our constraints, and the ones for which we have not been able to find any genealogical link, that we will describe as *isolated*. Our model guarantees that maximizing $\ell(\mathcal{T})$ amounts to locally maximizing the log-probability of each link. To sum up,

$$\max_{\mathcal{T}}\ \ell(\mathcal{T}) = \sum_{e\,\in\,\mathcal{L}} \ln \nu_e(c^*(e)) \tag{2.10}$$

and this upper bound is reached by the tree $\mathcal{T}^*$ built on all $e \in \mathcal{L}$ associated with the pairs $c^*(e)$. We shall note that formula (2.10) does not necessarily highlight a unique family

tree, for some pairs (p_1, p_2) may have the same probability of producing e. In this case, the maximization problem has more than one solution.

2.4.2. *Number of offsprings.* Suppose now that the population is small enough to be able to compute

$$\mathbb{G}(\mathcal{P}) = \prod_{e \in \mathcal{P}} G(e) \tag{2.11}$$

where $G(e)$ is given in (2.8). Namely, $\mathbb{G}(\mathcal{P})$ contains the exhaustive set of potential genealogies of the population. Due to the combination of the options of all $G(e)$, $\mathrm{Card}(\mathbb{G}(\mathcal{P}))$ may be very large. In fact such a Cartesian product is only conceptual, but quickly intractable for practical purposes leading to combinatorial explosions. Therefore, a *threshold probability* must be used to select the genealogies of $\mathbb{G}(\mathcal{P})$. Concretely, we can replace the definition of $G(e)$ in (2.8) by the more stringent

$$G(e) = \{c \in \mathcal{S}(e) \cup \{\varnothing\} \mid \nu_e(c) > \pi_{\min}\} \tag{2.12}$$

for a given choice of $0 \leqslant \pi_{\min} < 1$, and the construction of $\mathbb{G}(\mathcal{P})$ accordingly. If we define $N(i)$ as a random variable counting the offsprings of $i \in \mathcal{P}$, then it could be interesting to give an estimation of its probability distribution so as to infer, retrospectively, the individuals favored by breeders. Our model directly suggests to use

$$\forall\, k \in \mathbb{N}, \quad \widehat{\mathbb{P}}(N(i) = k) = \sum_{g \in \mathbb{G}(\mathcal{P})} w_g \, \mathbb{1}_{\{n_g(i) = k\}} \tag{2.13}$$

where $n_g(i)$ is the number of offsprings of i in the genealogy g and w_g is a *weighting* of the genealogy that can naturally be defined as the ratio between its likelihood and the sum of the whole likelihoods. It follows that

$$\widehat{\mathbb{E}}[N(i)] = \sum_{g \in \mathbb{G}(\mathcal{P})} w_g \, n_g(i) \tag{2.14}$$

may be a useful tool to decide whether i has been favored by breeders, by comparison with the global mean value and a classical *outlier threshold*. This approach will be illustrated on the rose bushes population of Section 4.2.

Example. Consider a set of 4 genealogies of likelihood 0.8, 0.6, 0.1 and 0.02, among which an individual i has 0, 1, 1 and 2 offsprings, respectively. Then we propose estimating $\widehat{\mathbb{P}}(N(i) = 0) \approx 0.526$, $\widehat{\mathbb{P}}(N(i) = 1) \approx 0.461$, $\widehat{\mathbb{P}}(N(i) = 2) \approx 0.013$ and $\widehat{\mathbb{P}}(N(i) > 2) = 0$. For this individual, $\widehat{\mathbb{E}}[N(i)] \approx 0.487$.

2.4.3. *Comparison of trees.* For a fixed population of size n, since each tree contains the same number of links, maximizing the likelihood *via* (2.9) seems a suitable criterion. However, it cannot be trusted to compare trees with a different number of links. To understand this, let $\mathcal{P}_i = \mathcal{P} \cup \{i\}$ be the same population enhanced with a new individual, from the last generation, such that $\delta(i, p_1, p_2) > 0$ for at least two pairs $(p_1, p_2) \in \mathcal{S}(i)$. Then, for these pairs we get $\ln \nu_i((p_1, p_2)) < 0$, implying that $\ell(\mathcal{T}) > \ell(\mathcal{T}_i)$, where $\mathcal{T}$ and $\mathcal{T}_i$ are the family trees maximizing the likelihood on $\mathcal{P}$ and $\mathcal{P}_i$, respectively. In other words, this criterion favors $\mathcal{T}$ rather than $\mathcal{T}_i$ whereas there exists a link between some individuals of $\mathcal{P}$ and i. In

order to overcome this negative impact, as soon as we have to compare family trees on two populations $\mathcal{P}$ and $\mathcal{P}_i$ such that $\mathcal{P}_i = \mathcal{P} \cup \{i\}$, we suggest to consider a trade-off like

$$\ell^*(\mathcal{T}_i) = \ell(\mathcal{T}_i) + \Psi(i) \tag{2.15}$$

where $\ell(\mathcal{T}_i)$ is the log-likelihood given by (2.9) of any genealogical tree $\mathcal{T}_i$ on $\mathcal{P}_i$ containing i, and $\Psi(i)$ is a measure of the *interaction ability* of the new individual i with $\mathcal{P}$, and reciprocally. Whence, to decide whether i has to be added into the genealogy, it will be possible to compare $\ell^*(\mathcal{T}_i)$ and $\ell(\mathcal{T})$ for the most likely tree $\mathcal{T}$ built on $\mathcal{P}$, provided a suitable adjustment of $\Psi(i)$. In this way, we intend to compensate the mechanical decrease of the log-likelihood due to the accumulation of potential links including i. This penalization of the log-likelihood is a strategy similar to the well-known AIC and BIC criteria. In the next section, when looking for missing individuals that could improve the family tree, we will see how to give a suitable explicit form to Ψ according to our purposes.

3. Missing links

Remind that our model assumes that no death occurs, which, as we have seen, is consistent with the fact that the work is related to perennial plant cultivars with asexual multiplication. However, individuals are obviously missing in the population – because they represent intermediate individuals never recorded as a cultivar and never distributed by the breeder, because the cultivar disappeared from rose gardens deliberately or accidentally, or because it was not sampled in the study. In this section, our objective is to look for some *missing links*. Since we do not know exactly how many individuals are missing, our strategy is to launch a *greedy algorithm* that explores the population and tries to detect an excess of information that might improve substantially the genealogy. The combinatorial complexity leads us to focus on some particular areas for the algorithm. More precisely, it seems that the isolated individuals are suitable starting points, for which we recall that $\mathcal{I} = \{e \in \mathcal{P} \,|\, c^*(e) = \varnothing\}$ is the set of individuals having no parents in the most likely genealogy. For all $e \in \mathcal{I}$, let $\mathcal{R}(e) \subset \mathcal{P}$ be the individuals in the population chronologically candidates to the genealogy of e and able to produce a gamete compatible with e. In addition, for each $p \in \mathcal{R}(e)$, consider

$$i^*(e,p) = \arg\max_i \, \delta(e,p,i) \tag{3.1}$$

as it is defined in (2.4). Namely, $i^*(e,p)$ is a virtual individual having a genotype which maximizes the probability of the event $\{(p,i) \mapsto e\}$, it can be seen as the perfect partner of p to produce e. Given $i = i^*(e,p)$, we now have to decide whether i significantly improves the genealogy. Let us carry on with the criterion introduced in (2.15), with an enhanced population $\mathcal{P}_i = \mathcal{P} \cup \{i\}$. To match with our study, the penalization $\Psi(i)$ must favor individuals i providing the maximum of interactions with $\mathcal{P}$. As we have seen in the last section, few interactions leave the likelihood almost unchanged whereas too much interactions tend to depreciate it, this was our motivation to look for a trade-off. We also want to give priority to any individual i reducing the number of *connected components* in the genealogy – that is, the number of subgraphs in which all nodes are connected. Define $\mathcal{T}$ and $\mathcal{T}_i$ as the maximum likelihood trees on $\mathcal{P}$ and $\mathcal{P}_i$, respectively, and suppose that i is contained in $\mathcal{T}_i$. Combining these requirements, we can write the penalization in the form

$$\Psi(i) = \lambda_i \, \frac{r(i)}{n} - \mu_i \, \Delta C(i) \tag{3.2}$$

where $r(i)$ is the number of individuals of $\mathcal{P}$ potentially interacting with i, $\Delta C(i)$ is the difference between the number of connected components in $\mathcal{T}$ and $\mathcal{T}_i$, $\lambda_i \geqslant 0$ and $\mu_i \geqslant 0$ are regularization parameters. Our decision rule consists in keeping an individual i which satisfies $\ell^*(\mathcal{T}_i) > \ell(\mathcal{T})$. We can formalize $r(i)$ like

$$r(i) = \sum_{p \in \mathcal{P}} \eta(i,p)$$

where $\eta(i,p) = 1$ if one can find $a \in \mathcal{P}$ such that $\delta(i,a,p) > 0$, $\delta(a,i,p) > 0$ or $\delta(p,a,i) > 0$, that is, if there is a nonzero probability for at least a link involving p and i, and $\eta(i,p) = 0$ otherwise. Note that a may be an offspring of i as well as a parent or a partner of i to be considered as an interaction involving i. To adapt our criterion, we can choose

$$\lambda_i = \frac{n}{2} \left| \ell(\mathcal{T}) - \ell(\mathcal{T}_i) \right| \tag{3.3}$$

since this guarantees that $\ell^*(\mathcal{T}_i) = \ell(\mathcal{T})$ when the new individual does not bring any connection except the one for which it has been created, not gathering connected components ($r(i) = 2$ and $\Delta C(i) = 0$), and thus when i should be rejected. A similar strategy enables to fix μ_i, for $r(i) = 2$ must at least coincide with $\Delta C(i) = -1$ to make an interesting link. This is the case when i has been created to fulfill the event $\{(p,i) \mapsto e\}$, and when p and e belong to different connected components. Of course that situation must be favored, and to simplify one can choose

$$\mu_i = \lambda_i + 1 \tag{3.4}$$

which amounts to say that $\ell^*(\mathcal{T}_i) > \ell(\mathcal{T})$ whenever $\Delta C(i) < 0$. To enhance the population, we suggest the following algorithm.

(0) Fix $n_v > 0$, the maximum number of virtual individuals allowed to be inserted in the population.
(1) Build $\mathcal{R}(e)$ for all $e \in \mathcal{I}$.
(2) For all $p \in \mathcal{R}(e)$, compute the maximum likelihood partner i such that $\{(p,i) \mapsto e\}$ is achieved.
(3) Add in $\mathcal{P}$ the individual maximizing $\ell^*(\mathcal{T}_i)$ provided

$$\max_i \ \ell^*(\mathcal{T}_i) > \ell(\mathcal{T}).$$

Set $t_e - 1$ as birth date of the new individual, where t_e is the one of e.
(4) Recalculate the most likely tree $\mathcal{T}$ and the set $\mathcal{I}$ according to the new population.
(5) Repeat steps (1)–(4) as long as the criterion increases and $\mathrm{Card}(\mathcal{P}) < n + n_v$.

4. An empirical study

The numerical processings were carried out through the `R` programming language and its software environment. In particular, we used the package *igraph*[1] to display the graphs. In the whole figures of this section, the geometric shapes that we use are circles to represent diploids, triangles for triploids and squares for tetraploids, gray individuals are real whereas white individuals are virtual. Similarly, we use solid lines for true links as well as dotted lines for the wrong links given by the model. The computations are conducted *via* the uniform probabilities $\pi_{21} = \pi_{12} = \frac{1}{2}$ and $\pi_{31} = \pi_{22} = \pi_{13} = \pi_{211} = \pi_{121} = \pi_{112} = \frac{1}{3}$.

[1] `https://cran.r-project.org/web/packages/igraph/igraph.pdf`

4.1. **On a simulated population.** Consider the simulated population $\mathcal{P}$ whose detailed description is provided in the Appendix. To sum up, there are $n = 54$ individuals among which 17 diploids, 17 triploids and 20 tetraploids have interacted throughout 8 generations. The simulation relies on $m = 4$ genes, dates of birth are known (*via* the generations) so as ploidies and observed genotypes. The goal is to apply our model on this population and to put the results into perspective, compared with the true genealogy $\mathcal{T}^0$ which is represented on the left of Figure 5.

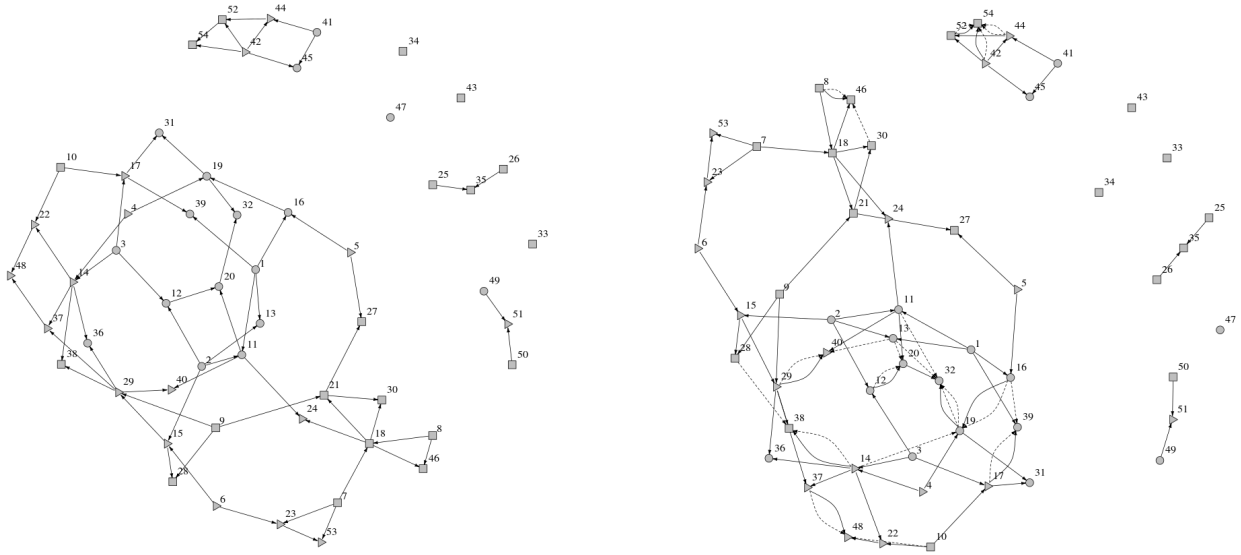

FIGURE 5. True genealogy $\mathcal{T}^0$ of the simulated population, on the left. Superposition of all genealogies of the simulated population found by the model, on the right.

4.1.1. *Family trees and most likely genealogy.* All genealogies found by the model have been superposed on the right of Figure 5, that is, the full content of $G(e)$ given in (2.8) for each $e \in \mathcal{P}$. We can first verify that the ancestors (individuals from 1 to 10) are only parents. On the one hand, we observe that the true genealogy is included in the graph, illustrating thereby the effectiveness of the exploratory algorithm. One can also notice, on the other hand, that some wrong links have been detected. We should however indicate that a wrong link is not an *impossible* link, for the reader can check that dotted arcs correspond to compatible crosses. Consider as an example the link $\{(14, 28) \mapsto 38\}$ appearing in Figure 5 but absent from the true genealogy. We have $x(14) = 3$, so $\widehat{g}_2(14) = \{160, 170, 180\} = g_2(14)$. Similarly, with $x(28) = 4$ and $x(38) = 4$, $\widehat{g}_2(28) = \{210, 290\}$ can correspond to $g_2(28) = \{210, 290, 290, 290\}$ and $\widehat{g}_2(38) = \{160, 170, 290\}$ to $g_2(38) = \{160, 170, 290, 290\}$. Through pattern (P_5), a genealogical link is possible on the signal 2 and we easily check that the same conclusion holds on each signal. This is an illustration of the fact that, from a practical point of view – namely, with an *unknown* true genealogy – it is preferable to produce a set of possible genealogies instead of a single one. Afterwards, the accumulation of genes enables to prune the trees, step by step, and to reinforce the remaining branches. To support this argument, Figure 6 shows on its left the family tree $\mathcal{T}^*$ maximizing the log-likelihood (2.9) in which

we observe that the true genealogy was *not* the most likely one, retrospectively. Let us have a look at the differences. The first one is the selection of $\{(14, 28) \mapsto 38\}$ instead of $\{(14, 29) \mapsto 38\}$. Knowing that 28 and 29 both have parents $(9, 15)$, we easily understand their genetic likeness. The second one is interpreted in the same way since $\{(8, 30) \mapsto 46\}$ stands in for $\{(8, 18) \mapsto 46\}$, and since 18 is a parent of 30. For the last two ones, 13 takes the place of 11 in the true connections $\{(11, 12) \mapsto 20\}$ and $\{(11, 29) \mapsto 40\}$, 11 and 13 having the same parents. To be precise, in the latter example each link leads to the same probability and the maximum of likelihood is not unique (in which case the algorithm chooses one solution at random). On this dataset, we get

$$\ell(\mathcal{T}^*) \approx -3.052 \ > \ -7.616 \approx \ell(\mathcal{T}^0).$$

Even so, wrong links maximizing the log-likelihood are usually relevant. In this example, the wrong parents detected are in fact close relatives of true parents.

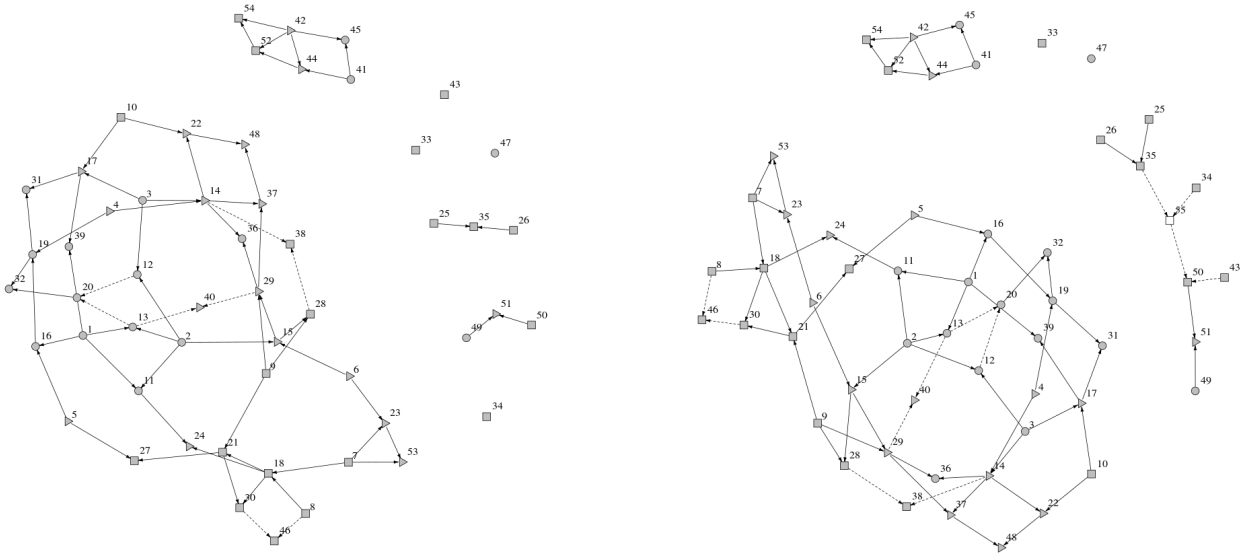

FIGURE 6. Genealogy $\mathcal{T}^*$ maximizing the log-likelihood of the simulated population found by the model, on the left. There are 8 connected components. Genealogy $\mathcal{T}_1$ maximizing the log-likelihood of the simulated population enhanced with one individual (55) found by the model, on the right. There are 5 connected components.

4.1.2. *Missing links.* We now look for missing links, following the algorithm described at the end of Section 3 with $n_v = 3$. Compared with the most likely tree $\mathcal{T}^*$ on the population $\mathcal{P}$, the largest increase of our penalized criterion ℓ^* given by (2.15) is reached by adding the tetraploid $g_1(55) = \{200, 200, 200, 200\}$, $g_2(55) = \{270, 270, 270, 270\}$, $g_3(55) = \{370, 370, 370, 370\}$ and $g_4(55) = \{410, 410, 520, 520\}$, respectively for the 4 genes, as a member of generation 5. We obtain the genealogy on the right of Figure 6. From 8 connected components in $\mathcal{T}^*$, only 5 remain in the maximum likelihood tree $\mathcal{T}_1$ on the population enhanced with the individual 55 having this precise genotype. Thus its role as a missing link is clearly highlighted and that explains the reason why it has been privileged, even if $\ell(\mathcal{T}_1) \approx -3.106$ has decreased compared to $\ell(\mathcal{T}^*) \approx -3.052$. A second

application of the algorithm generates the tetraploid having $g_1(56) = \{10, 10, 200, 200\}$, $g_2(56) = \{130, 130, 380, 380\}$, $g_3(56) = \{210, 210, 370, 370\}$ and $g_4(56) = \{430, 520, 520, 520\}$ on its 4 genes, in generation 3. Only 4 connected components remain, but the log-likelihood is now $\ell(\mathcal{T}_2) \approx -3.482$. The last application of the algorithm gives a diploid $g_1(57) = \{90, 90\}$, $g_2(57) = \{220, 220\}$, $g_3(57) = \{310, 310\}$ and $g_4(57) = \{510, 510\}$ in generation 5. Only 3 connected components remain while, for this last addition, the log-likelihood is unchanged. Figure 7 depicts $\mathcal{T}_2$ and $\mathcal{T}_3$, respectively on the left and on the right. This simulated example seems to clearly illustrate the operation of the exploratory algorithm, focusing on connected components to build missing links, retrospectively.

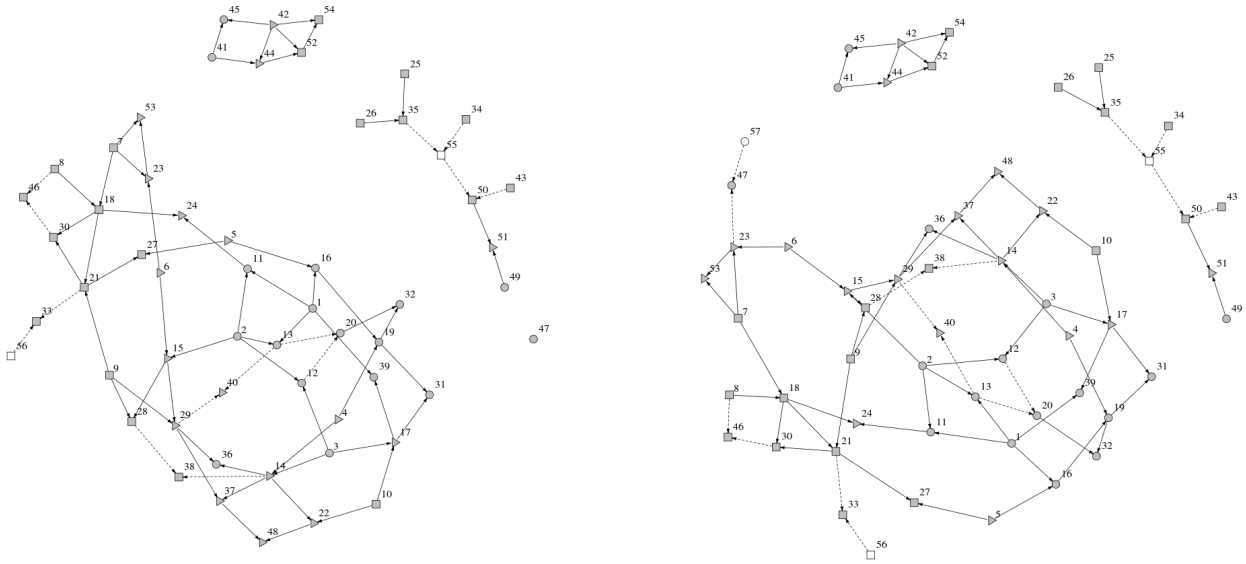

FIGURE 7. Genealogy $\mathcal{T}_2$ maximizing the log-likelihood of the simulated population enhanced with two individuals (55 and 56) found by the model, on the left. There are 4 connected components. Genealogy $\mathcal{T}_3$ maximizing the log-likelihood of the simulated population enhanced with three individuals (55, 56 and 57) found by the model, on the right. There are 3 connected components.

4.2. **On a rose bushes population.** To conclude the study, one is now going to launch our model on a subpopulation of rose bushes collected on the basis of $m = 4$ genes. We start by giving some explanations about the experimental gathering of the data. Among molecular markers, microsatellite markers are still a reference for pedigree reconstruction because they are highly multiallelic codominant markers [9]. After Polymerase Chain Reaction (PCR), amplified fragments are generally separated by capillary electrophoresis. According to their size, amplified fragments are detected at a given time of the electrophoresis and are depicted as a peak in the electrophoregram, whose area varies according to the intensity of the signal. Thus, a statistical treatment of the four signals of the individual i gives the observed genotypes $\widehat{g}(i)$. To deal with allelic multiplicity, theoretical ratios between peak intensities could be used to determine the relative number of copies of each allele in polyploids [6]. Unfortunately this strategy is very difficult to apply, especially because signal intensity is also dependent of amplification competition between alleles during PCR. Therefore, in most cases electrophoregrams are generally interpreted as presence or absence of alleles [5]. This

is also our approach in this article but considering all possibilities of multiplicity, for which we have seen in the previous sections how our model enables to build and probabilize $g(i)$ from $\widehat{g}(i)$. An example of signal is shown in Figure 8. In addition we must not forget that a calibration of the equipment is needed, for practical purposes. In concrete terms, the abscissa of the signals is made of decimal values, which is clearly incompatible with what it is supposed to highlight, namely some *base pairs*. Hence we take rounded values, and an offset of ± 1 for each allele has to be considered. This is the reason why we decided to switch to criterion (2.3) in the real data analysis.

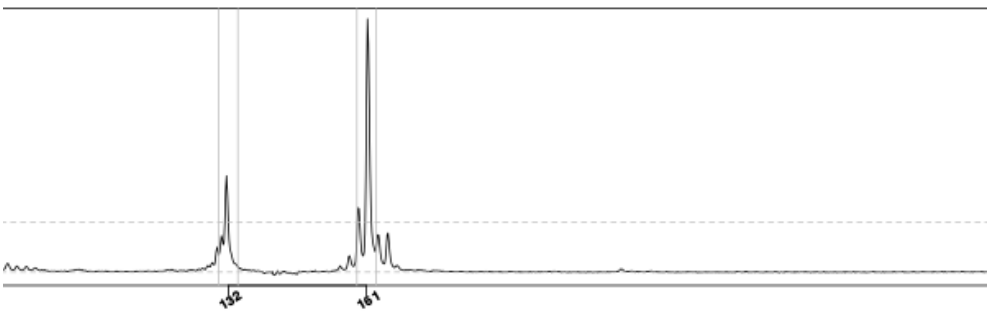

FIGURE 8. Example of signal for a particular microsatellite marker. The individual i is tetraploid and two peaks have been detected. Here $\widehat{g}(i)$ is $\{132, 161\}$ and $g(i)$ is $\{132, 132, 132, 161\}$ with probability π_{31}, $\{132, 132, 161, 161\}$ with probability π_{22} and $\{132, 161, 161, 161\}$ with probability π_{13}. To simplify, scales are deliberately removed.

4.2.1. *Family trees and most likely genealogy.* Now we put aside $n = 116$ rose bushes, selected for the knowledge of their ploidy and for the clarity of their signals, and we look for potential genealogical links among them using the same allelic probabilities as in the simulation study. The whole genealogies are superposed on Figure 9. Even if the graphical representation seems unexploitable, it illustrates the fact that many solutions are conceivable. More than one genealogy maximize the likelihood, for some links have the same probability. An example of most likely genealogy is given on the left of Figure 10, it contains 35 connected components. Within the largest one, a chain of 5 generations is obtained ($9 \to 56 \to 67 \to 59 \to 47$).

4.2.2. *Missing links.* On the right of Figure 10, one of the most likely genealogies is represented when $n_v = 3$ new individuals suggested by the algorithm of Section 3 are added (117, 118 and 119). Again, their role as missing links and their usefulness to connect separated branches of the genealogy are clearly brought to light. Only 32 of them remain, due to the fact that each missing link allowed to connect two components. In particular, we can notice the important intercession of 118, plugging the two largest ones.

4.2.3. *Selected individuals.* To look for selected individuals, the estimated probabilities (2.13) and expectations (2.14) are computed for all $i \in \mathcal{P}$ on the basis of a subset of genealogies made of links whose likelihood is greater than $\pi_{\min} = 0.2$. Indeed, since $\mathrm{Card}(\mathbb{G}(\mathcal{P})) > 10^{28}$ the computation with no threshold is infeasible. It appears that with this choice of threshold, $\mathrm{Card}(\mathbb{G}(\mathcal{P}))$ is in the range of 10^6 which is small enough to proceed to computations and large enough to trust the statistical estimations. Figure 11 contains the empirical expectations of

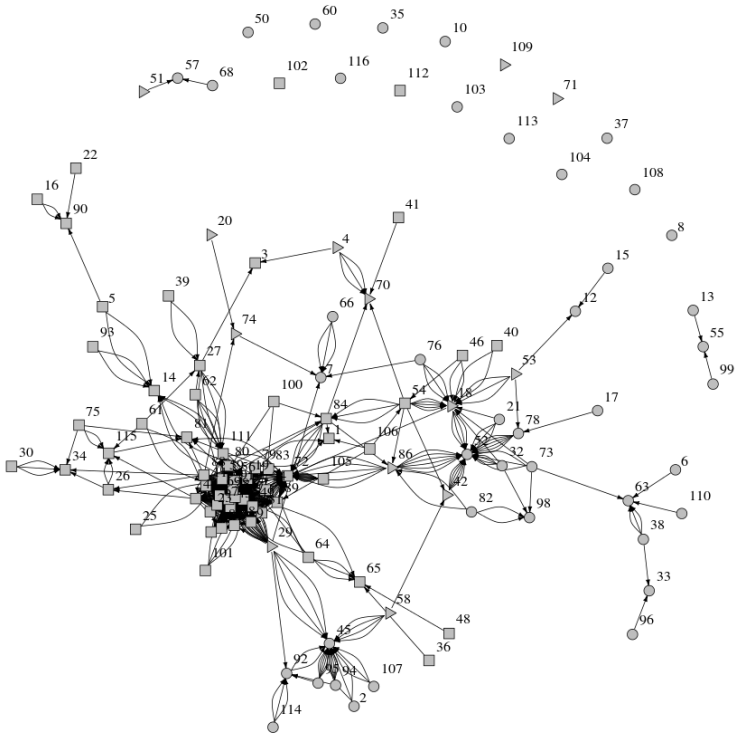

FIGURE 9. Superposition of all genealogies of the rose bushes subpopulation found by the model.

all individuals. The horizontal dotted lines respectively give the mean of the whole results, and an *outlier threshold* set to the standard $q_3 + 1.5\,(q_3 - q_1)$ with q_1 and q_3 the first and third quartiles. Each individual having a higher mean number of offsprings is considered as a potential target for the retrospective selection by breeders, they are 8 in this subpopulation. Amongst all individuals, $i = 88$ has, on average, the largest number of offsprings in the population. Figure 12 shows the empirical distribution of $N(88)$. Concretely,

$$\widehat{\mathbb{P}}(N(88) = 5) \approx 0.770, \quad \widehat{\mathbb{P}}(N(88) = 6) \approx 0.230 \quad \text{and} \quad \widehat{\mathbb{E}}[N(88)] \approx 5.230.$$

The last empirical distribution represented is the one of $N(73)$, chosen to illustrate the fact that an individual may have offsprings in some genealogies and no offspring in the others. Numerically,

$$\widehat{\mathbb{P}}(N(73) = 0) \approx 0.222, \quad \widehat{\mathbb{P}}(N(73) = 1) \approx 0.444, \quad \widehat{\mathbb{P}}(N(73) = 2) \approx 0.278,$$

$$\widehat{\mathbb{P}}(N(73) = 4) \approx 0.056 \quad \text{and} \quad \widehat{\mathbb{E}}[N(73)] \approx 1.167.$$

5. Conclusion

To conclude, we would like to draw the attention of the reader to some weaknesses of the model, essentially relying on the allelic multiplicity. Indeed, our choice of considering each potential multiplicity weighted by a probability, instead of selecting a particular one, may lead to contradictions in the genealogy. Suppose to simplify that the most likely genealogy contains the links $\{(p_1, p_2) \mapsto q_1\}$ and $\{(q_1, q_2) \mapsto e\}$ where p_1 is a tetraploid such that $g(p_1) =$

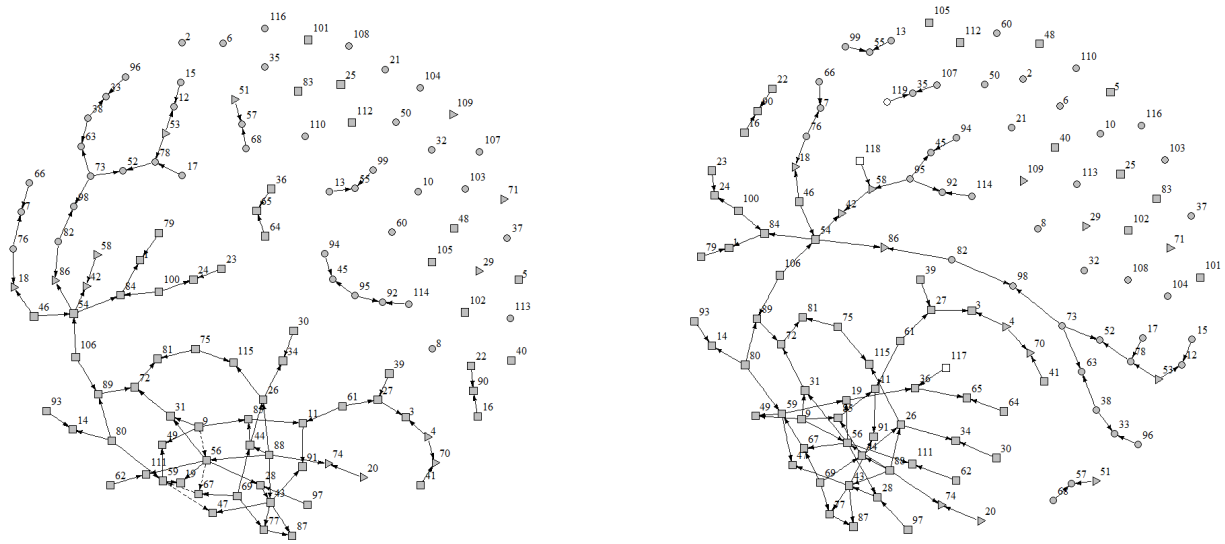

FIGURE 10. Genealogy $\mathcal{T}^*$ maximizing the log-likelihood of the rose bushes subpopulation found by the model (the dotted line highlights a chain of 5 generations), on the left. There are 35 connected components. Genealogy $\mathcal{T}_3$ maximizing the log-likelihood of the rose bushes subpopulation enhanced with three individuals (117, 118 and 119) found by the model, on the right. There are 32 connected components.

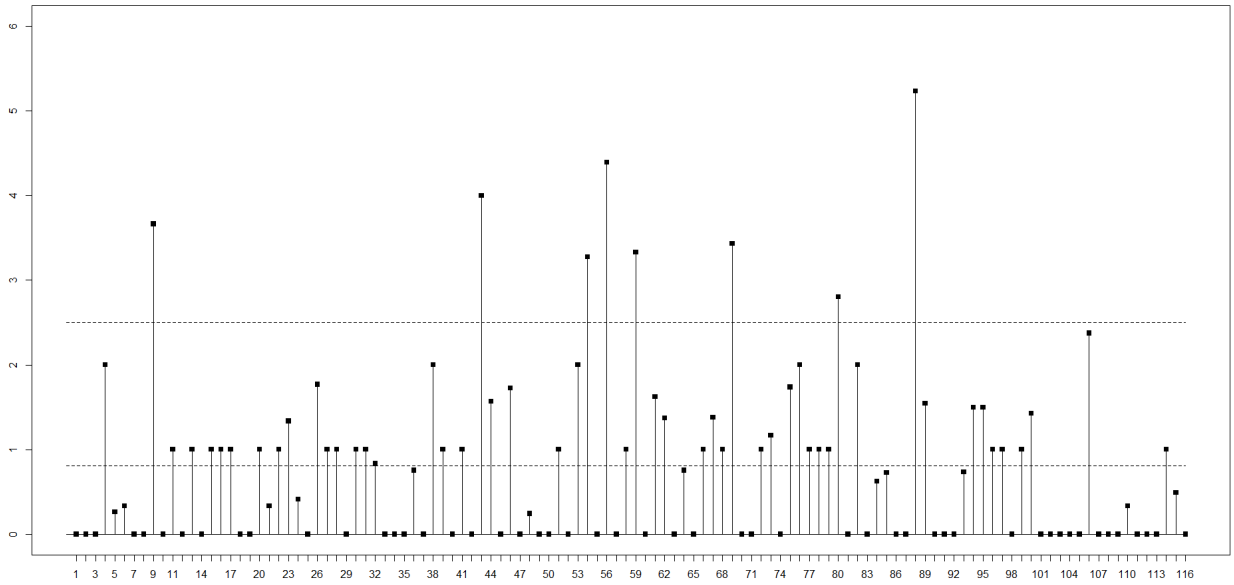

FIGURE 11. Mean number of offsprings for each individuals. The abscissa displays the individuals $i \in \mathcal{P}$ and the ordinate represents the estimated expectation of $N(i)$. Horizontal dotted lines give the mean and the outlier threshold, respectively. There are 8 probably favored individuals.

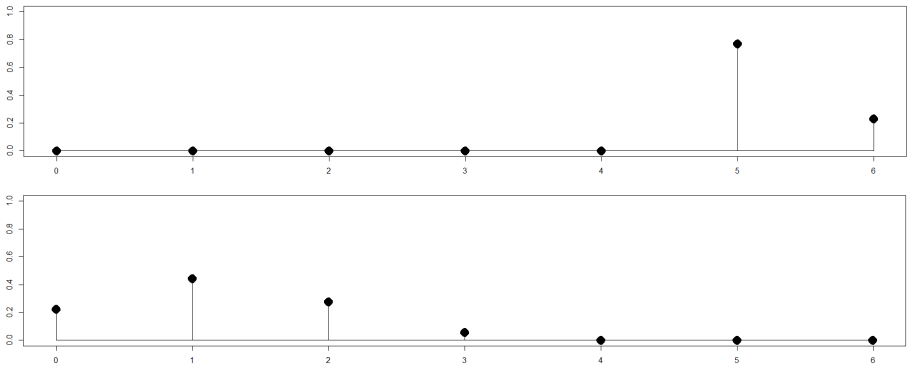

FIGURE 12. Empirical distribution of the random variable $N(88)$, at the top. The abscissa represents the number k of offsprings, the ordinate is the estimated probability associated with the event $\{N(88) = k\}$. At the bottom, empirical distribution of the random variable $N(73)$.

$\{a, a, a, a\}$, and p_2 is a diploid such that $g(p_2) = \{b, b\}$. Both of them are homozygous, so there is no allelic uncertainty derived from their observed genotypes, but $\widehat{g}(q_1) = \{a, b\}$ for the triploid q_1 can only match with $\{(p_1, p_2) \mapsto q_1\}$ in case of $g(q_1) = \{a, a, b\}$. Suppose now that q_2 and e are tetraploids, having $g(q_2) = \{c, c, c, c\}$ and $\widehat{g}(e) = \{b, c\}$, respectively. Then, the link $\{(q_1, q_2) \mapsto e\}$ has a nonzero probability only for $g(q_1) = \{a, b, b\}$. In other words, the most likely genealogy treats q_1 as a link between (p_1, p_2) and e, but at the cost of incompatible allelic combinations. This is a trail for future improvements of our model, in particular it seems worth considering an algorithm to detect contradictions and to eliminate such trees from the set of genealogies. Another weakness is the estimation of $\pi_{21}, \pi_{12}, \pi_{31}, \ldots$, namely the probabilities of allelic multiplicity. As we have seen in Section 4.2, we lack information to properly evaluate them. An ambitious track could be the generalization of [4], in which the authors establish the well-known *Hardy-Weinberg equilibrium* to deal with heterozygoty in a diploid population. A challenging study will be to characterize this equilibrium in our polyploid population – if it exists – and to determine its degrees of freedom. This additional information will enable to refine the probabilities of multiplicity, considering that the population has reached its equilibrium. The crossbreeding patterns also have to be enhanced with double reductions and preferential matches, both of them easily treated on a theoretical point of view but difficult to estimate. Finally, it is important to insist upon the fact that this work is mainly theoretical and that the application of our model on a real population of rose bushes is only relevant in order to show that coherent and interpretable results are obtained. Nevertheless, we cannot draw any conclusion from an empirical study relying on $m = 4$ genes. In-depth experiments will be conducted on more genes, and the comparison of any interesting result with available historical sources will constitute strong arguments to understand the breeder's strategies over the past centuries, and also to try to complete the datasets with some lost or missing information.

Acknowledgements. This research was conducted in the framework of the regional programme "Objectif Végétal, Research, Education and Innovation in Pays de la Loire", supported by the French Région Pays de la Loire, Angers Loire Métropole and the European Regional Development Fund, in the framework of the PedRo project. Empirical data were obtained thanks to the support of the Région Pays de la Loire in the framework of the

FLORHIGE project, by the National Institute of Agricultural Research (INRA) and the National Natural Science Foundation of China in the framework of the SIFLOR project, and by the French Ministry of Higher Education and Research. The authors also thank the University of Bretagne Loire, Angers Loire Métropole and the University of Angers for their financial support.

Appendix

This Appendix is devoted to the precise description of the simulated population appearing in Section 4.1. All useful information are given in Tables 1, 2 and 3, displaying the composition of the successive generations. For each individual, the columns indicate an identifier i, the ploidy $x(i)$, the observed genotypes $\widehat{g}(i)$ on the four signals, the couple of parents and the reproduction pattern.

Generation 1							
i	$x(i)$	$\widehat{g}_1(i)$	$\widehat{g}_2(i)$	$\widehat{g}_3(i)$	$\widehat{g}_4(i)$	Par.	Pat.
1	2	10–20	110	210–310	310–320	$\varnothing$	–
2	2	30–40	130–140	220–230	330	$\varnothing$	–
3	2	50	150–160	240–250	340	$\varnothing$	–
4	3	60	170–180–190	260–270	350–360–370	$\varnothing$	–
5	3	70–80	200	280	380–390–400	$\varnothing$	–
6	3	90–100–110	210–220	290–300–310	410	$\varnothing$	–
7	4	120–130–140	230–240–250–260	320–330	420–430–440	$\varnothing$	–
8	4	150–160–170–180	270–280	340	450	$\varnothing$	–
9	4	190–200	290–300	350–360–370	460–470–480–490	$\varnothing$	–
10	4	210–220	310–320	380–390–400	500–510–520	$\varnothing$	–

Generation 2							
i	$x(i)$	$\widehat{g}_1(i)$	$\widehat{g}_2(i)$	$\widehat{g}_3(i)$	$\widehat{g}_4(i)$	Par.	Pat.
11	2	10–40	110–130	210–220	320–330	$(1,2)$	(P_1)
12	2	40–50	140–150	220–250	330–340	$(2,3)$	(P_1)
13	2	20–40	110–130	210–220	310–330	$(1,2)$	(P_1)
14	3	50–60	160–170–180	250–270	340–350–370	$(3,4)$	(P_2)
15	3	40–100–110	140–210	220–290–310	330–410	$(2,6)$	(P_2)
16	2	20–80	110–200	210–280	320–400	$(1,5)$	(P_2)
17	3	50–210–220	150–320	240–380–400	340–520	$(3,10)$	(P_3)
18	4	130–160–180	240–250–270	320–330–340	430–450	$(7,8)$	(P_6)

Table 1. Full description of generations 1 and 2 in the simulated population.

Generation 3							
i	$x(i)$	$\widehat{g}_1(i)$	$\widehat{g}_2(i)$	$\widehat{g}_3(i)$	$\widehat{g}_4(i)$	Par.	Pat.
19	2	20–60	110–180	270–280	350–400	$(4, 16)$	(P_2)
20	2	40	110–150	220	330	$(11, 12)$	(P_1)
21	4	130–180–200	270–290–300	340–350–370	450–480–490	$(9, 18)$	(P_6)
22	3	60–210	180–320	250–390–400	370–520	$(10, 14)$	(P_5)
23	3	90–130–140	220–230–240	300–320	410–420–440	$(6, 7)$	(P_5)
24	3	10–130–160	130–270	210–330–340	330–430–450	$(11, 18)$	(P_3)
25	4	190–200	290–300	350–360–370	410–520	$\varnothing$	–
26	4	130–160–180	240–250–270	320–330–340	410	$\varnothing$	–

Generation 4							
i	$x(i)$	$\widehat{g}_1(i)$	$\widehat{g}_2(i)$	$\widehat{g}_3(i)$	$\widehat{g}_4(i)$	Par.	Pat.
27	4	80–200	200–270–300	280–340	380–400–480–490	$(5, 21)$	(P_5)
28	4	40–100–200	210–290	220–310–350–360	330–410–470–490	$(9, 15)$	(P_5)
29	3	100–200	140–290	310–350–370	410–460–490	$(9, 15)$	(P_5)
30	4	130–200	270	330–340–350	450	$(18, 21)$	(P_6)
31	2	20–210	180–320	280–380	400–520	$(17, 19)$	(P_2)
32	2	40–60	110	220–270	330–350	$(19, 20)$	(P_1)
33	4	10–180–200	130–270–380	210–340–370	430–450–520	$\varnothing$	–
34	4	20–90–200	160–270–330	370	520–530–550	$\varnothing$	–
35	4	130–180–200	270–290–300	340–350–370	410	$(25, 26)$	(P_6)

Generation 5							
i	$x(i)$	$\widehat{g}_1(i)$	$\widehat{g}_2(i)$	$\widehat{g}_3(i)$	$\widehat{g}_4(i)$	Par.	Pat.
36	2	60–100	180–290	270–370	340–490	$(14, 29)$	(P_4)
37	3	50–200	140–160–180	250–270–370	350–370–410	$(14, 29)$	(P_4)
38	4	50–60–100	160–170–290	270–310–350	340–370–410–490	$(14, 29)$	(P_4)
39	2	20–210	110–150	210–380	320–340	$(1, 17)$	(P_2)
40	3	40–200	130–290	210–310–370	330–410–460	$(11, 29)$	(P_2)
41	2	20–110	150–320	260	410–520	$\varnothing$	–
42	3	230	170–390–420	240–340	380–390	$\varnothing$	–
43	4	70–90–100	210–220–270	310–330–340–400	490	$\varnothing$	–

TABLE 2. Full description of generations 3, 4 and 5 in the simulated population.

Generation 6							
i	$x(i)$	$\widehat{g}_1(i)$	$\widehat{g}_2(i)$	$\widehat{g}_3(i)$	$\widehat{g}_4(i)$	Par.	Pat.
44	3	110–230	170–320–390	240–260–340	390–520	$(41, 42)$	(P_2)
45	2	110–230	320–420	260–340	390–520	$(41, 42)$	(P_2)
46	4	130–150–160	270	330–340	450	$(8, 18)$	(P_6)
47	2	90	220	310–320	410–510	$\varnothing$	–
48	3	50–210	180–320	250–400	410–520	$(22, 37)$	(P_4)
49	2	240	320	410	510–520	$\varnothing$	–
50	4	100–200	270	310–330–370	410–490–520	$\varnothing$	–

Generation 7							
i	$x(i)$	$\widehat{g}_1(i)$	$\widehat{g}_2(i)$	$\widehat{g}_3(i)$	$\widehat{g}_4(i)$	Par.	Pat.
51	3	200–240	270–320	310–370–410	410–490–520	$(49, 50)$	(P_3)
52	4	230	170–390–420	240–340	390	$(42, 44)$	(P_4)
53	3	130	230–240	320	410–420–440	$(7, 23)$	(P_5)

Generation 8							
i	$x(i)$	$\widehat{g}_1(i)$	$\widehat{g}_2(i)$	$\widehat{g}_3(i)$	$\widehat{g}_4(i)$	Par.	Pat.
54	4	230	170–390–420	240–340	390	$(42, 52)$	(P_5)

TABLE 3. Full description of generations 6, 7 and 8 in the simulated population.

References

[1] Ackerman, M. W., Hand, B. K., Waples, R. K., Luikart, G., Waples, R. S., Steele, C. A., Garner, B. A., McCane, J., and Campbell, M. R. Effective number of breeders from sibship reconstruction: empirical evaluations using hatchery steelhead. *Evolutionary applications 10*, 2 (2017), 146–160.

[2] Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist 210*, 2 (2016), 391–398. 2015-19414.

[3] Bourke, P. M., Arens, P., Voorrips, R. E., Esselink, G. D., Koning-Boucoiran, C. F. S., van't Westende, W. P. C., Santos Leonardo, T., Wissink, P., Zheng, C., van Geest, G., Visser, R. G. F., Krens, F. A., Smulders, M. J. M., and Maliepaard, C. Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal 90*, 2 (2017), 330–343.

[4] Chaumont, L., Malécot, V., Pymar, R., and Sbai, C. Reconstructing pedigrees using probabilistic analysis of ISSR amplification. *Journal of theoretical biology 412* (2017), 8–16.

[5] Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology 23*, 1 (2014), 40–69.

[6] Esselink, G., Nybom, H., and Vosman, B. Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting–peak ratios) method. *Theoretical and Applied Genetics 109*, 2 (2004), 402–408.

[7] Gudin, S., et al. Rose: genetics and breeding. *Plant breeding reviews 17* (2000), 159–190.

[8] Jian, H., Zhang, H., Tang, K., Li, S., Wang, Q., Zhang, T., Qiu, X., and Yan, H. Decaploidy in *Rosa praelucens* byhouwer (Rosaceae) endemic to Zhongdian plateau, Yunnan, China. *Caryologia 63*, 2 (2010), 162–167.

[9] Jones, A. G., and Ardren, W. R. Methods of parentage analysis in natural populations. *Molecular Ecology 12*, 10 (2003), 2511–2523.

[10] Kong, N., Li, Q., Yu, H., and Kong, L.-F. Heritability estimates for growth-related traits in the pacific oyster (*Crassostrea gigas*) using a molecular pedigree. *Aquaculture Research 46*, 2 (2015), 499–508.

[11] Koning-Boucoiran, C. F. S., Gitonga, V. W., Yan, Z., Dolstra, O., van der Linden, C. G., van der Schoot, J., Uenk, G. E., Verlinden, K., Smulders, M. J. M., Krens, F. A., and Maliepaard, C. The mode of inheritance in tetraploid cut roses. *Theoretical and Applied Genetics 125*, 3 (Aug 2012), 591–607.

[12] Lacombe, T., Boursiquot, J.-M., Laucou, V., Di Vecchi-Staraz, M., Péros, J.-P., and This, P. Large-scale parentage analysis in an extended set of grapevine cultivars (*Vitis vinifera* l.). *Theoretical and Applied Genetics 126*, 2 (2013), 401–414.

[13] Liorzou, M., Pernet, A., Li, S., Chastellier, A., Thouroude, T., Michel, G., Malécot, V., Gaillard, S., Brie, C., Foucher, F., Oghina-Pavie, C., Clotault, J., and Grapin, A. Nineteenth century french rose (*Rosa* sp.) germplasm shows a shift over time from a european to an asian genetic background. *Journal of Experimental Botany 67*, 15 (2016), 4711–4725.

[14] Lucena-Perez, M., Soriano, L., López-Bao, J. V., Marmesat, E., Fernández, L., Palomares, F., and Godoy, J. A. Reproductive biology and genealogy in the endangered iberian lynx: Implications for conservation. *Mammalian Biology 89* (2018), 7–13.

[15] Mable, B., Alexandrou, M., and Taylor, M. Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology 284*, 3 (2011), 151–182.

[16] Maia, N., and Venard, P. Cytotaxonomie du genre *Rosa* et origine des rosiers cultivés. 7-20. p. *Travaux sur rosiers de serre Antibes: FNPHP. Cit.: GUDIN, S.(2000): Rose: Genetics and Breeding. Plant Breeding Reviews 17*, 1 (1976), 159–189.

[17] Oghina-Pavie, C. Rose and pear breeding in nineteenth-century france: the practice and science of diversity. In *New Perspectives on the History of Life Sciences and Agriculture*. Springer, 2015, pp. 53–72.

[18] Otto, S. P., and Whitton, J. Polyploid incidence and evolution. *Annual review of genetics 34*, 1 (2000), 401–437.

[19] Raju, D., Namita, K. P. S., Prasad, K., and Janakiram, T. Self-and cross-incompatibility relationship in rose (*Rosa hybrida*) varieties. *Current Horticulture 1*, 2 (2013), 7–9.
[20] Van Huylenbroeck, J., Leus, L., and Van Bockstaele, E. Interploidy crosses in roses: use of triploids. In *1st International Rose Hip Conference* (2005), vol. 690, International Society for Horticultural Science (ISHS), pp. 109–112.
[21] Wang, J., and Scribner, K. T. Parentage and sibship inference from markers in polyploids. *Molecular ecology resources 14*, 3 (2014), 541–553.

Laboratoire Angevin de REcherche en MAthématiques (LAREMA), CNRS, Université d'Angers, Université Bretagne Loire. 2 Boulevard Lavoisier, 49045 Angers cedex 01, France.
E-mail address: frederic.proia@univ-angers.fr
E-mail address: fabien.panloup@univ-angers.fr
E-mail address: chiraz.trabelsi@univ-angers.fr

IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV, 49071, Beaucouzé, France.
E-mail address: jeremy.clotault@univ-angers.fr