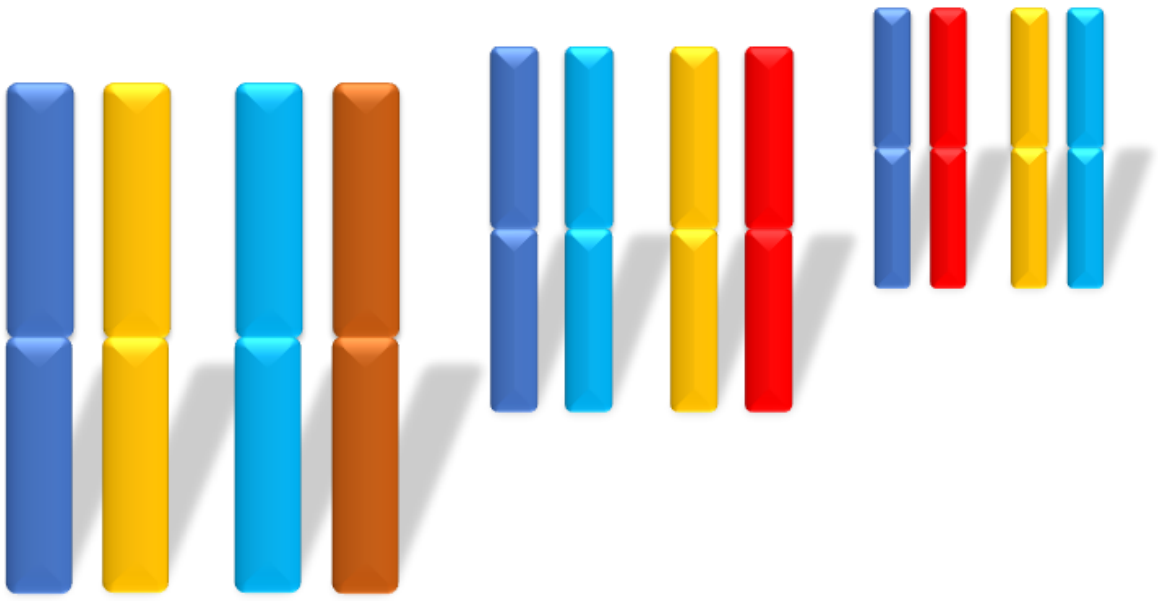
Testing robustness of polyploid preferential pairing detection

A simulation study.



Author: Ronald Nieuwenhuis
Student number: 1015294
Supervisor(s): Dr P.M. Bourke, Dr C.A. Maliepaard
Course code: PBR80436
Credits: 36 ECTS

Abstract

Finding a low-cost, reliable method to skim polyploid populations for preferential pairing between chromosomes during meiosis can help to select candidate populations for in-depth analysis of preferential pairing mechanisms. In this study we found minimal parameters for such a preferential pairing screening by testing decreasing population sizes and decreasing numbers of markers. We further tested the robustness of a computational haplotype reconstruction tool that is commonly applied to polyploid F1 populations by introducing marker skewness and dosage errors using simulated data. The applied simulations revealed that the tool tested is robust to marker skewness and dosage errors and that there are clear lower bounds for population size and numbers of markers. . Given the solid signals this probably translates well into real populations used for genetic studies. The obtained knowledge can be useful for designing a low-resolution assay to identify candidate F1 populations of polyploid species for full genetic screening.

Contents

1. Introduction

1.1 Meiosis as a fundamental process in breeding

Meiosis is the key biological process that ensures genetic diversity over time through inheritance within populations. Recombination of parental chromosomes into offspring causes reshuffling of alleles and together with independent assortment it results in a diversified set of allele combinations (Figure 1). Such combinations may or may not be of advantage in certain environments or under changing conditions and the process of meiotic recombination is thus fundamental in evolution. For crop breeders, fully understanding and eventually controlling meiosis promises to be a powerful tool that will help in their breeding programs already at an early stage. These programs aim to tap into reservoirs of genetic and phenotypic diversity. By means of introgression breeding, breeders try to introduce a desired trait into one of their breeding lines. This is usually introduced by producing an F1 cross of a breeding line with a trait-of-interest containing accession, followed by inbreeding (selfing), or backcrossing into subsequent generations to stabilize the trait. Ideally, a marker tightly linked to the desired trait is developed/found to rapidly screen offspring for the presence of the desired certain allele. Linking of a marker to a trait is called quantitative trait loci (QTL) mapping. Suitable markers are nowadays often single nucleotide polymorphisms (SNPs) but have often been repeat copies (SSR) or DNA digestion (RFLP, AFLP) products as well. In polyploid species, which have more than two homologous chromosomes, this method of breeding is often more difficult than in diploid species because the underlying alleles of a trait can recombine in more combinations and specific markers are more difficult to obtain. This emphasizes the need to understand and monitor meiosis in the parental lines, especially in polyploid species and outcrossing species.

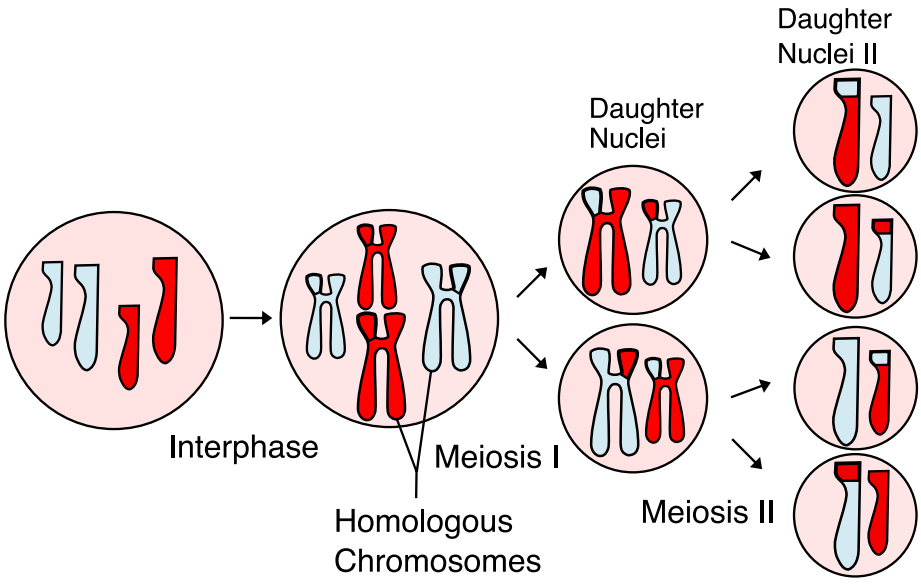


Figure 1: General overview of meiosis in a diploid species. Two main stages of meiosis can be seen. Meiosis I following interphase and separating homologous chromosomes (independent assortment). Meiosis II separating sister chromatids. The result is four haploid gametes. Recombination (crossing-over) happening during prophase of meiosis I. (Rdbickel, 2016, CC BY-SA 4.0)

1.2 Meiosis in diploids and polyploids

Meiosis at a general level is a well understood process. Early models have described and shown how crossovers/chiasmata happen and how a crossover is resolved by the recombination machinery into two recombined chromosomes, and how two homologous pairs of sister chromatid are reduced to produce a tetrad (Morgan, 1911; Thompson & Schild, 2001). Though there is still much knowledge to be obtained as to why recombination occurs at certain positions while not at others and how to control it. Meiotic recombination in polyploids however, is a not yet completely understood process. While allopolyploids, hybridized from two closely related species, can cause problems in processes like genome assembly and variant calling because of local similarity between *homoeologous* chromosomes, autopolyploids derived from whole genome duplication events are easier to assemble into some collapsed form. When looking at meiotic recombination however, autopolyploids are more difficult to study than allopolyploids, because allopolyploid genomes often behave similar to diploid genomes. The difficulty in autopolyploid mapping is caused by the increasing complexity of combinations between chromosomes in metaphase I when compared to diploid genomes (Soares et al., 2021). During this stage, the autopolyploid species can form bivalents, multivalents or a combination of both. This results in disomic, polysomic or mixosomic inheritance. Even within this category of multivalents, different structures occur during synapsis, resulting in different outcomes of recombination for different multivalents. Even locally, along a chromosome, different forms of inheritance can occur (segmental allopolyploidy) (Mason & Wendel, 2020; Sybenga, 1996). An example of an outcome of multivalent pairing is double reduction, where gametes end up with homologues that are partially identical and thus showing a deviating segregation pattern (Bourke, 2018). The underlying mechanism of what type of inheritance occurs under which circumstances and how this is affected is yet unclear. Possible explanations are (local) structural incompatibility between homologous chromosomes, as has been shown in diploid species. Certain structural rearrangements can block recombination in areas of the genome with genes linked to interesting traits for breeders. Other possibilities are environmental factors such as stress (temperature, disease, drought), that allow certain parts of the genome to become more open for recombination, and recent hybridization/polyploidization events that cause genetic and structural genomic instability. Finally, there are certain genes involved that affect the type of inheritance (Soares et al., 2021)Overall, studying what kind of structures form in autopolyploid species during meiosis can help explain certain breeding outcomes and inform breeders about compatibilities between parental lines.

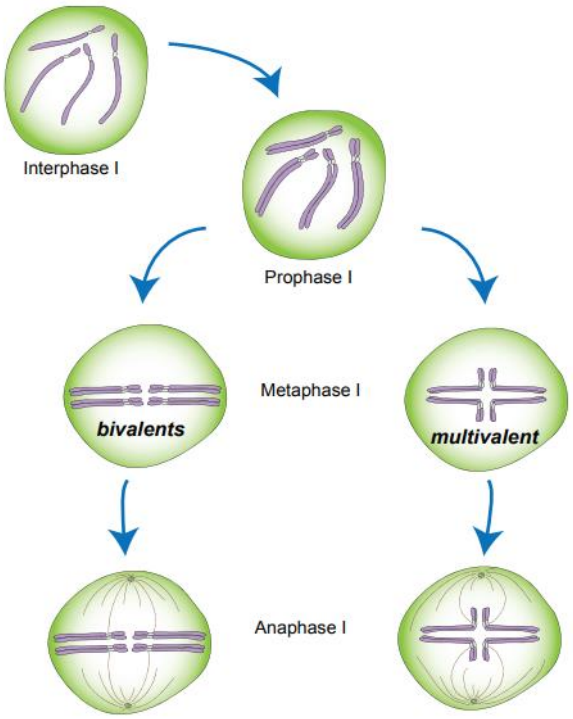
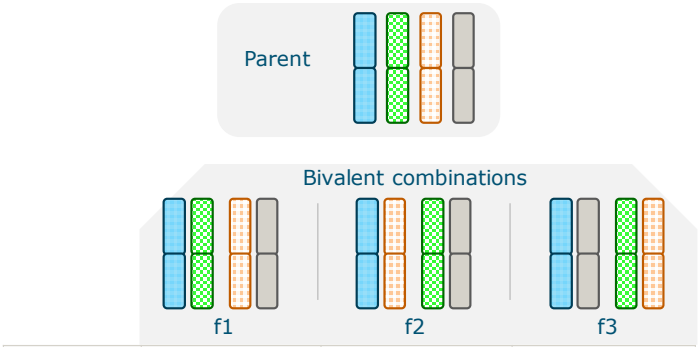


Figure 2: Partial overview of meiosis in an autopolyploid species. In metaphase I four homologous chromosomes may pair in either pairs (bivalents) or together (multivalent) and preferential pairing may occur. (Bourke, 2018)

1.3 Preferential pairing

During the recombination process in autopolyploids displaying polysomic inheritance, the homologues that pair up as bivalents to exchange alleles may show a preference for one over another. This so-called preferential pairing can lead to unexpected segregation patterns in the offspring for certain chromosomes when random pairing is assumed and a deviation from polysomic inheritance (Figure 3). When there is a full preference, the inheritance follows a disomic model instead of a polysomic model. A practical implication of this phenomenon for breeders can be unsuccessful introgression of a trait of interest into their breeding lines or too limited numbers of offspring to observe the trait of interest due to the rarity of pairing homologues. For geneticists, knowing what type of inheritance model to use is important to correctly call dosages and calculate genetic distances between markers as the models are built with a certain inheritance model assumed. Having a refined genetic map with correct distances is necessary to properly link a QTL signal to its real genetic position and find, eventually, the causative allele (region).



	f1	f2	f3
Random	.33	.33	.33
Preferential	.63	.23	.13

Figure 3: Relative frequencies of bivalent pairing combinations under preferential pairing and random pairing.

1.4 Detection of preferential pairing

A major step in observing the process of recombination in autopolyploids has been the development of software that can produce phased genetic maps, such as polymapR, MAPpoly and polyOrigin (Bourke et al., 2018; Mollinari & Garcia, 2019; Zheng et al., 2021). This has been successfully demonstrated in rose, (sweet) potato, and chrysanthemum (Bourke et al., 2015, 2017; Mollinari et al., 2020; van Geest et al., 2017). Given there are enough high-quality markers and sufficient offspring to genotype, genotyping an F1 population can give insight in the recombination frequency across each homologue in the parents of that F1 population. Genotyping experiments inherently contain information about chromosome pairing behavior and preferential pairing, but it may be hard to extract this information due to model assumptions and different biological phenomena displaying similar effects in mapping populations. Two methods of preferential pairing detection are discussed in Bourke et al. (2017): single-point and multi-point methods. One single point method is a diagnostic feature of polymapR to correct the recombination frequencies of the maps created under the assumption of random pairing. The detection method focuses on deviations from random pairing in repulsion-phased *simplex* x *nulliplex* markers that can be unambiguously linked to one parental homolog using a likelihood estimation. The used markers need to be located in close proximity to each other to ensure there is no interference of recombination and markers can considered to be on the same locus. This distance threshold for proximity has been set to 1 centimorgan (cM), based on an observed error rate in duplicate individuals in the K5 rose population described in (Bourke et al., 2017). With a Chi-squared test the distribution of homolog combinations can be tested against the null hypothesis that the frequency of pairing is 1/3, as would be expected under random pairing. If the marker pair does show a different distribution this can be used as evidence of (local) preferential pairing. Combined evidence of multiple such marker pairs can be used to draw a conclusion on preferential pairing. Furthermore, a preferential pairing estimate ρ can be calculated. This estimator ranges from 0 to 0.66 and described

how big the deviation from random pairing is. A value for ρ of 0 means there is no preferential pairing, a value of 0.66 means full preferential pairing. Another, more robust method is implemented in polyOrigin, where the pairing combinations are obtained from the reconstructed haplotypes. The haplotypes are inferred by walking back through a hidden Markov model (HMM) that is built using the (reconstructed) parental genotypes. Subsequently, a Chi-squared test is done to assess whether the frequency of homologue pairs in the offspring deviates from an even distribution amongst combinations. This multi-point approach too yields estimate ρ of preferential pairing and promises to be more robust compared to the single marker pair based method described here because it uses the genotype information across the entire chromosome and includes more marker types

1.5 Previously reported results

While both single-point and multi-point based methods have provided several insights already, it is currently unknown exactly how robust they are (Ahmed et al., 2020; Song & Endelman, 2023). In Bourke et al. (2017), differences and similarities were found between both described methods. Marker panels are inherently only covering part of the genome and contain dosage errors. The screened populations might also contain errors like contamination of parental pollen from not intended parents or labeling errors introduced while maintaining the population. Besides errors, detection of preferential pairing will also be affected by panel sizes, e.g., number of markers, number of screened offspring and marker types. Distance between markers and their place on the chromosome can also influence the detection of preferential pairing. Furthermore, preferential pairing may show a similar pattern as selection. When an allele is deleterious it will cause skewness of markers, as it skews the segregation pattern away from fully tetrasomic. This is sometimes called segregation distortion. Because both preferential pairing and marker skewness may appear in a similar way, it can be hard to distinguish these phenomena. Besides the effects on detection of preferential pairing by the aforementioned properties, the quantification of the underlying structures (in case of multivalent pairing) could maybe be improved upon. In this study the focus will be on the robustness of preferential pairing detection with a multi-point HMM based method. Using simulated data, the effect of different parameters will be assessed. Combining the obtained results and the published *Rosa x hybrida* linkage map, a recommendation for a broad screening of different polyploid accessions can be made.

2. Methods

2.1 General overview

To test robustness of preferential pairing detection, the general workflow consisted of simulation of populations with a known preferential pairing rate and varying variables of interest. Test sets were simulated using PedigreeSim and sampled using custom code. PedigreeSim (Voorrips & Maliepaard, 2012) is a versatile Java based tool that can produce F1 offspring population based on a genetic map of the parents and several user-specified options. Produced offspring were then analysed using PolyOrigin to reconstruct the haplotypes and determine the most likely meiotic configurations in the parents. By varying certain parameters of interest, like *population size*, *number of markers*, *marker skewness* or *error rate*, the detected rate of preferential pairing can be compared to the simulated rate.

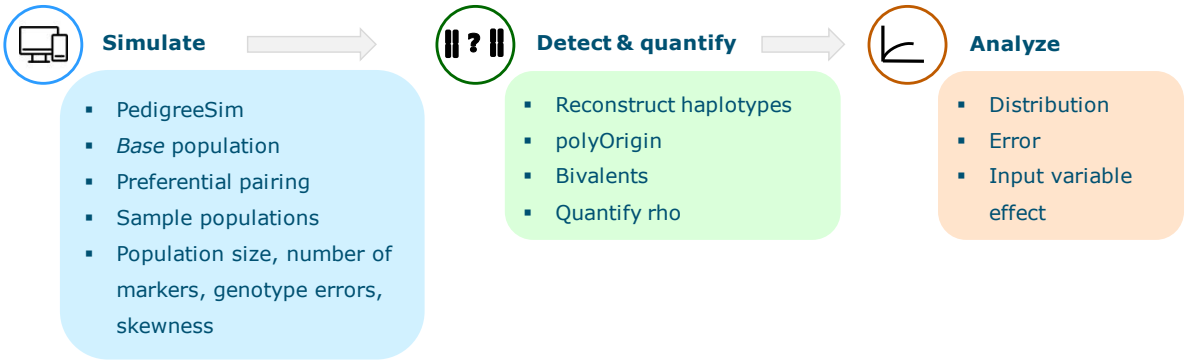


Figure 4: General workflow for testing parameter effect on preferential pairing

The basis for each experiment was a simulation without preferential pairing, from which individual offspring were sampled to generate a test set with a specific rate of preferential pairing. Determination of meiotic configuration group sizes was based on the proposed likelihood function (Equation 1) as described in Bourke et al. (2017). The likelihood model is described such that a scenario where bivalent formation with random pairing would yield $\hat{\rho} = 0$ and full preferential pairing with all offspring stemming from one meiotic configuration would result in $\hat{\rho} = 0.66$, this would be complete disomic behaviour.

Equation 1: Maximum likelihood estimate under a model of preferential pairing in bivalent context.

$$\rho = \frac{\frac{2}{3}\, n_1 - \frac{1}{3}\,(n_2 + n_3)}{n_1 + n_2 + n_3}$$

From Equation 1 follows Equation 2 used for calculation of the sample group sizes for a given combination of ρ and population size $n_{tot} = n_1 + n_2 + n_3$, with $n_2 = n_3$.

Equation 2: Logic used for calculation of meiotic configuration sample sizes.

$$n_1 = \rho \cdot n_{tot} + \frac{1}{3}\, n_{tot}$$

Finally, a linear relation between ρ and the PedigreeSim $prefPairing$ parameter is described as Equation 3 based on observed results in Supplementary S1.

Equation 3: Linear transformation between rho and PedigreeSim input parameter prefPairing

$$\rho = \frac{2}{3} \cdot prefPairing$$

Base simulations were all done with PedigreeSim v2.2 . This version produces an extra output file with the suffix "_out_meioticconfigs.dat" that contains for each individual offspring the pairing configuration during the crossover simulation. Input files for PedigreeSim were generated on a per experiment basis using R functions provided by P.M. Bourke. Further settings used were: map length 100 cM, centromere position 50 cM, Haldane mapping function, ALLOWNOCHIASMATA set and a ploidy of 4, and 0 quadrivalents unless specifically stated in the sections below for each experiment.

Offspring haplotypes were reconstructed using polyOrigin v0.5.10 using the default settings. Further downstream analysis was done using R v4.2.2 and the packages listed in Supplementary section C, page 28.

polyOrigin output consisted of log files, "_polyancestry.csv" and "_postdoseprob.csv". Log files were parsed using command line processing (Supplementary section D, page 29) and the polyancestry files were parsed using the import_IBD() function of polyqtlR. The specific version of the function is not publicly available. It functions similarly except it does not filter the input for a certain probability threshold.

2.2 Population size

To test the effect of population size on the accuracy of preferential pairing detection a base population of 1,000 individuals was simulated from using 100 markers evenly spread with over 100 cM distance. The parental marker types were evenly divided over the 9 fundamental segregating marker types in tetraploid species as shown in Table 1. Ten samples were drawn for each combination of population size and preferential paring rate. Levels for population size were {25, 50, 75, 100, 125, 150, 200, 250} and for ρ {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.66}.

Table 1: Numbers of markers per marker type for testing population size effect

P1 dosage	P2 dosage	Number of markers
1	0	11
2	0	11
0	1	11
1	1	11
2	1	11
1	3	11
0	2	11
1	2	11
2	2	12
	Total:	**100**

2.3 Number of markers

To test the effect of the number of markers used in screening, multiple populations were simulated using PedigreeSim. Numbers of markers tested were {10, 20, 40, 60, 80, 100} and the markers were evenly spaced across a 100 cM map. The distribution of marker types was drawn from the genetic map of the K5 population as published by Bourke et al. (2017) and displayed in Figure 5. Raw dosage data

(screened_data_PQcombined.csv) from the K5 population was processed using polymapR *checkF1()* and *convert_marker_dosages()*, followed by a call on *parental_quantities()* as advised in the vignette of polymapR. Simulated populations contained 1,000 individuals each from which 10 samples of 250 individuals were drawn with a given value of ρ from {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.66}. The influence of the polyOrigin *--chrpairing* parameter on the preferential pairing detection was examined by redoing the experiment with a setting of "22" instead of the default "44".

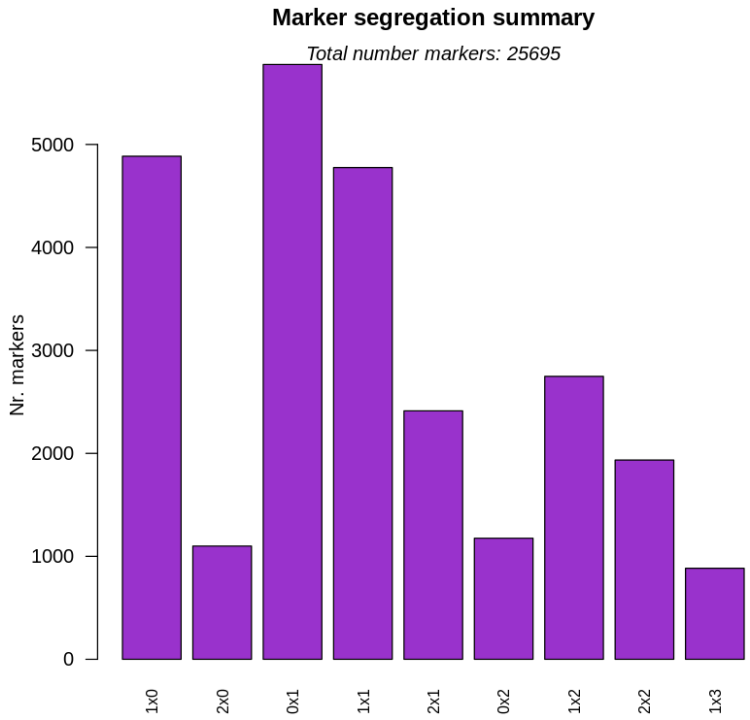


Figure 5: Marker type distribution of the rose K5 mapping population after conversion to fundamental segregating types (Bourke et al., 2017).

2.4 Error rate

The effect of error rate on the preferential pairing detection was explored by creating a genetic map with 50 markers evenly spaced on a 100 cM map. The amount of fundamental marker types was evenly distributed (Table 2, column 3) over the types themselves and each marker was randomly assigned a position as to generate uneven distances between the markers of the same type. Base simulation consisted of 1,000 offspring individuals without preferential pairing. From this population 10 samples were drawn of 250 individuals for each combination of ρ and error rate. Levels for ρ were {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.66} and for the error rate percentages {0, 5, 10, 15, 20}. After sampling, dosage errors were introduced by randomly selecting markers and individuals (by matrix coordinates) until the desired number of dosage errors was reached in accordance with the aforementioned error rates.. The newly assigned dosage was randomly picked from the values {0, 1, 2, 3, 4} with the original value removed.

Table 2: Distribution of marker types for error rate robustness test

P1 dosage	P2 dosage	Number of markers
1	0	5
2	0	5
0	1	5
1	1	5
2	1	6
1	3	6
0	2	6
1	2	6
2	2	6
	Total:	**50**

2.5 Marker skewness

Marker skewness tests implied the removal of individuals with a certain dosage obtained from P1. A marker matrix was produced again with 100 markers uniformly spaced. The distribution of marker types was based on the K5 rose mapping population as displayed in Figure 5. The base simulation with PedigreeSim produced 2,000 individuals to make sure there were enough individuals left after removing selection candidates. From this pool, 10 samples were drawn of 250 individuals for each combination of skewness and ρ. Skewness was tested for values in {0, 0.25, 0.5, 0.75, 1} as a portion of the individuals with dosage 1 for marker “LG1_22_1x0_h4” to be removed from the sampling population and being replaced with individuals having a dosage of 0 for the specified marker. Preferential pairing strength comprised the same values as other experiments listed above.

3. Results

The obtained results from our simulations can be roughly divided in two parts: parameter exploration and robustness testing. The parameters explored focused on the two most important concepts of screening mapping populations: population size and number of markers to test. Minimizing these parameters might come with a cost benefit when scaling up the screening of many small population. The robustness tests might expose limits of the detection methods used.

3.1 Parameter exploration

When minimizing population size stepwise from 250 to 25 individuals, the effect on the detection of preferential pairing, is rather small (Figure 6). There is only a slight decrease in the slope of the regression line through the average detected preferential pairing when comparing a population size of 250 to 25. Notably, the detection seems to be less accurate at higher rates of preferential pairing, whatever the population size. The most accurate results are obtained for ρ between 0 and 0.2, with values very close to the simulated ρ. When no preferential pairing is simulated, a small overestimation of ρ becomes apparent. At lower population sizes this overestimation of ρ seems to increase (0.043 ± 0.024, for n = 25). Finally, with lower population sizes the sample standard deviation increases for all levels of preferential pairing, suggesting the smaller sample sizes do have a noticeable effect on correctly determining the meiotic configuration in parental meiosis.

The measured effect of decreasing the number of markers is much more discernible than the population size effect. The control with 100 markers and 250 individuals confirms the previous finding for the population size experiment with slope of 0.86. By decreasing the number of markers, a deteriorating effect on haplotype reconstruction already shows for 80 markers and a decreasing precision of correct meiotic configurations can be observed (Figure 7). The difference between the simulated and detected values of ρ reaches a maximum of 0.532 ± 0.06 when relying on 10 markers and the maximum degree of preferential pairing in the population. This would mean only one configuration is present, but faulty configurations were reported for a majority of individuals. Testing the polyOrigin software by switching the possibility of multivalents in meiosis off yielded highly similar results (Supplementary S2). The poor performance with regards to preferential pairing detection and hence the underlying reconstruction of haplotypes, remains.

Under a changing population size with 100 markers PolyOrigin never assigns the highest probability to a multivalent configuration, opposed to a scenario with 10 markers and strong preferential pairing. Shown in Figure 8 are combined results for all replicates of a population of 250 offspring with a simulated ρ of 0.66 (left) and one of 250 offspring based on 10 markers and simulated ρ = 0.5 (right). Some probability to the multivalent configuration 1:2:3:4, as denoted in the output of polyOrigin, is generally assigned when allowing for multivalents. Looking at normalized p-values produced by polyOrigin for such a parental meiotic configuration does show an increasing signal coinciding with decreasing accuracy of meiotic configuration as presented when using decreasing numbers of markers (Figure 9). At lower numbers of markers, increasing probabilities for multivalents are reported, ranging from 0.21 for 10 markers to around 0.10 for 100 markers So, multivalent probabilities decrease with increasing numbers of markers. At any given number of markers, the effect of preferential pairing on multivalent probabilities seems rather small, even though some interesting and opposite trends appear. In agreement with the limited effect of population size on preferential pairing detection, the probability of P1 multivalent origin of an individual does not depend on the population size (Supplementary S3).

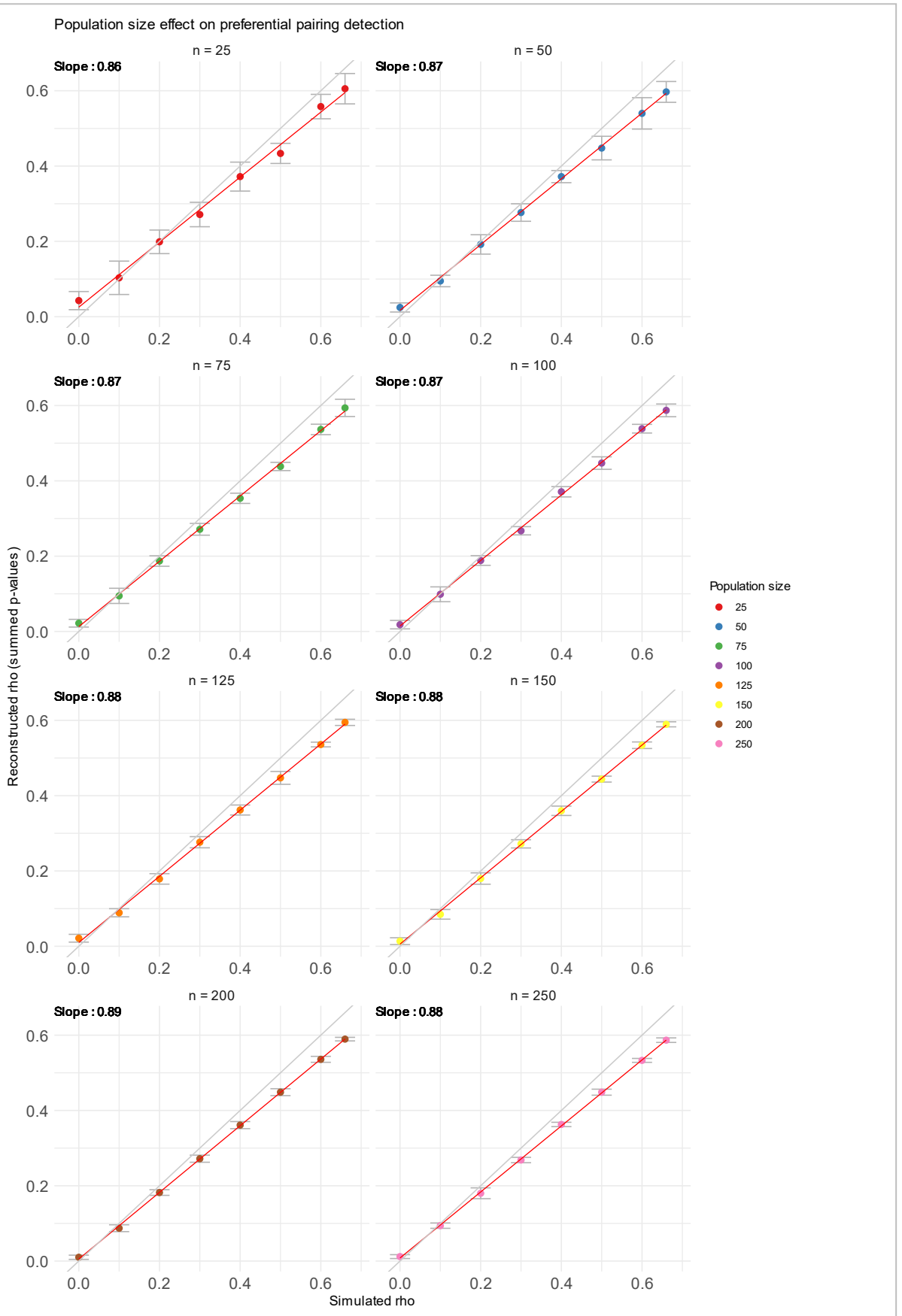


Figure 6: Simulated versus reconstructed preferential pairing strength for different population sizes. Grey line shows diagonal (Slope = 1)

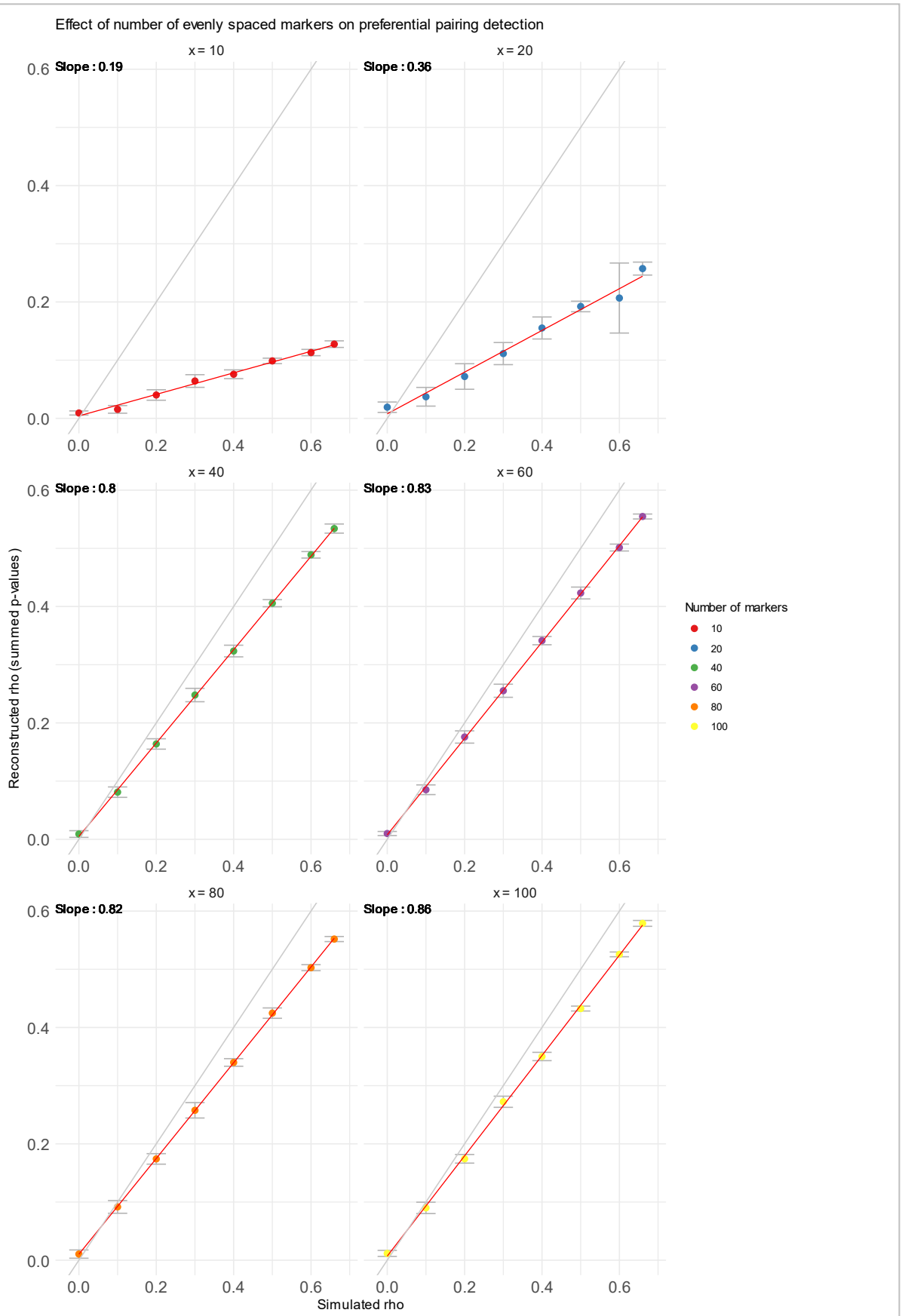


Figure 7: Simulated versus reconstructed values for rho, given different numbers of markers. Grey line describes diagonal (slope = 1)

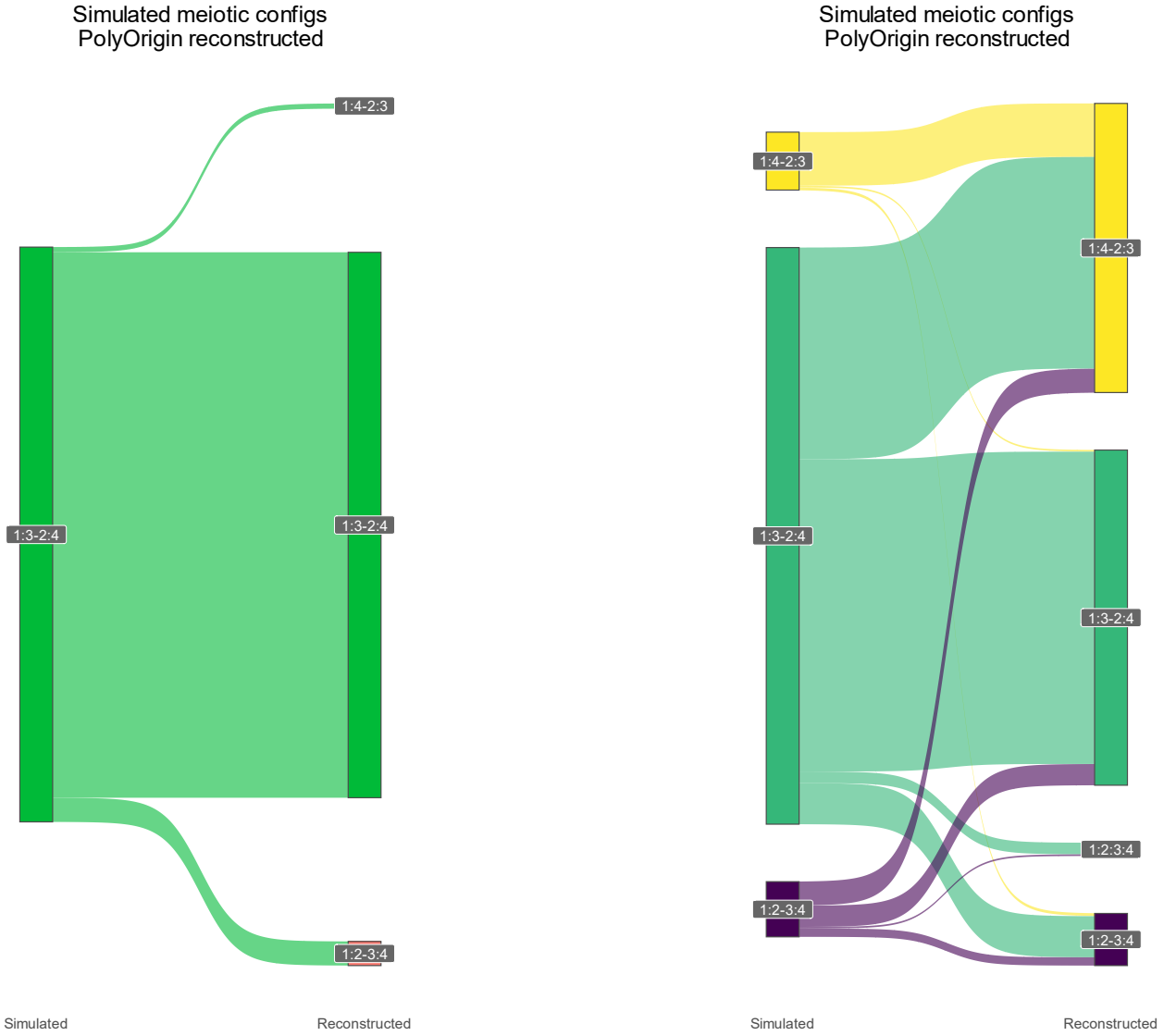


Figure 8: Simulated meiotic configs versus inferred configs by polyOrigin for all replicates of described scenario. PolyOrigin notation used describes multivalents as 1:2:3:4 and different bivalent pairing configurations as i.e. 1:2-3:4, where the numbers refer to parental homologs. Left: polyOrigin results for a population of 250 individuals with one hundred markers and the maximum strength of preferential pairing shows some incorrect meiotic pairing configurations, but no multivalents. Right: Based on ten markers, ρ of 0.50 and an offspring of 250, polyOrigin infers wrong pairs abundantly and even some multivalents.

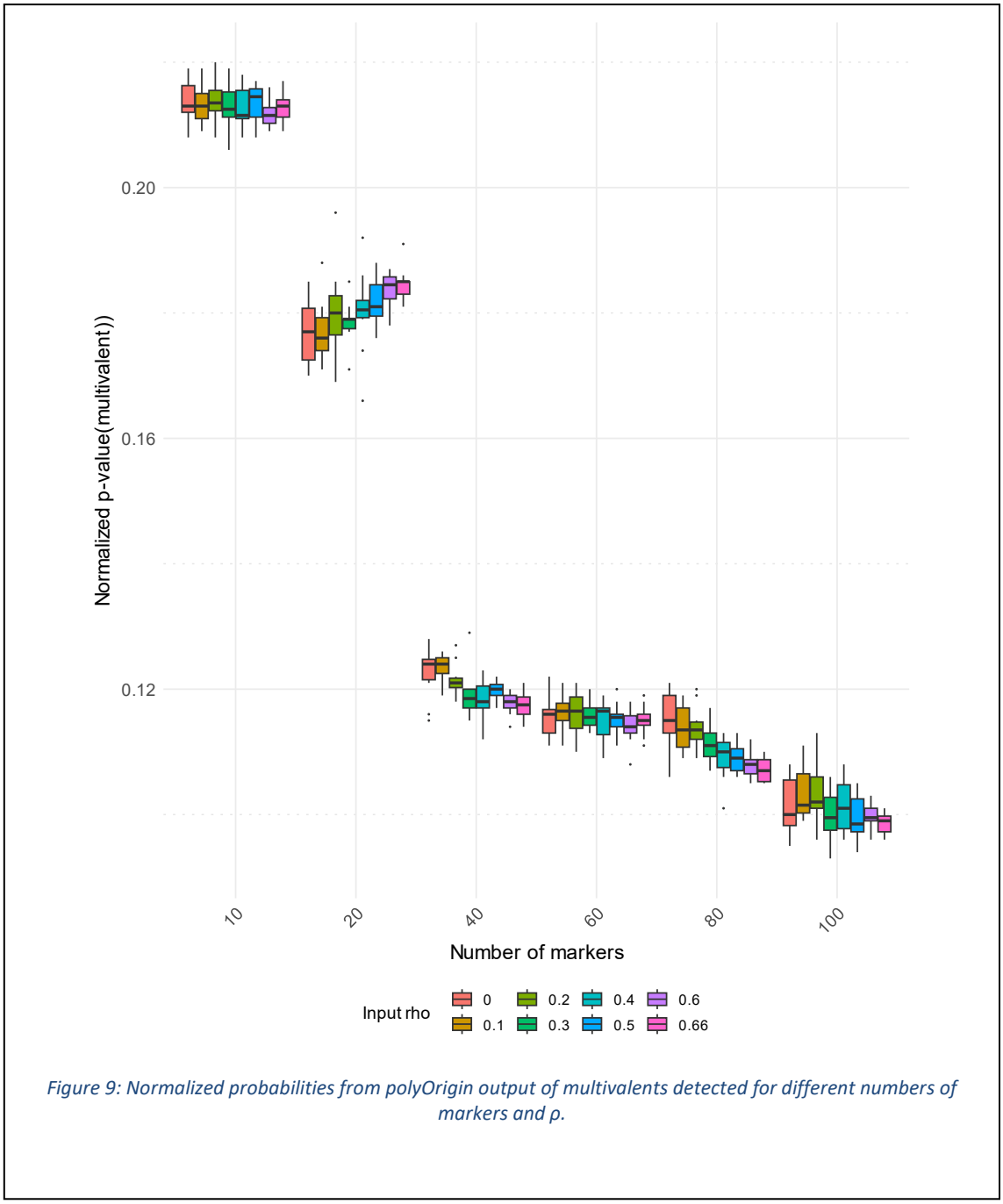


Figure 9: Normalized probabilities from polyOrigin output of multivalents detected for different numbers of markers and ρ.

3.2 Robustness tests

The robustness tests for preferential pairing detection focused on two important parameters, dosage error rate and marker skewness. Dosage errors can be caused by noise and incorrect conversion from raw signal to dosage call. It is naturally a part of each marker screening and it can cause a mismatch between parental dosage and individual offspring dosage. When an abundance of markers is available, if too many individuals show a deviation from the expected segregation pattern based on the parental genotype such markers can be removed from the dataset. When applying minimum numbers of markers as per the goal of this study however, the relevance of dosage error becomes clear. Because we found a detrimental effect on preferential paring detection when using fewer than 40 markers, with some built in safety this test was done using 50 markers. Figure 10 shows that an increasing error rate causes a less steep slope and thus a larger difference between simulated ρ and reconstructed ρ. Again, higher degrees of preferential pairing lead to a more severe underestimation of ρ. While at first glance it would be expected that the results for absent error rate would coincide with the results of the population size experiment, the distribution of marker types between the two experiments is different. The design of the array used to screen the K5 population is optimized to simplex x nulliplex (SxN) markers while the population size experiment is designed with equal numbers of marker types. These SxN markers yield the highest information content as they are absolutely haplotype specific. Because of this, polyOrigin reconstructs haplotypes better in the error rate experiment than the equivalent population size experiment. The effect on preferential pairing is apparent. Additionally, the number of multivalents is not affected by error rate as for all rates steady values in the range of 15%-17% are found.

Marker skewness is an effect on the dosage values across the entire population, skewing segregation ratios and distorting the complete dataset because of linkage to other alleles. To increase effect for this experiment we picked 100 markers with the K5 population. By implementing a deleterious effect on one allele the dataset was skewed (Supplementary S4). The effect of skewness is clearly unnoticeable, as the measured reconstructed ρ values are similar to the population size experiment (Figure 11). The control, without skewness shows similar results to the corresponding population size experiment and is a positive control. Based on the limited effect on estimated ρ, the haplotype reconstruction seems very robust to marker skewness.

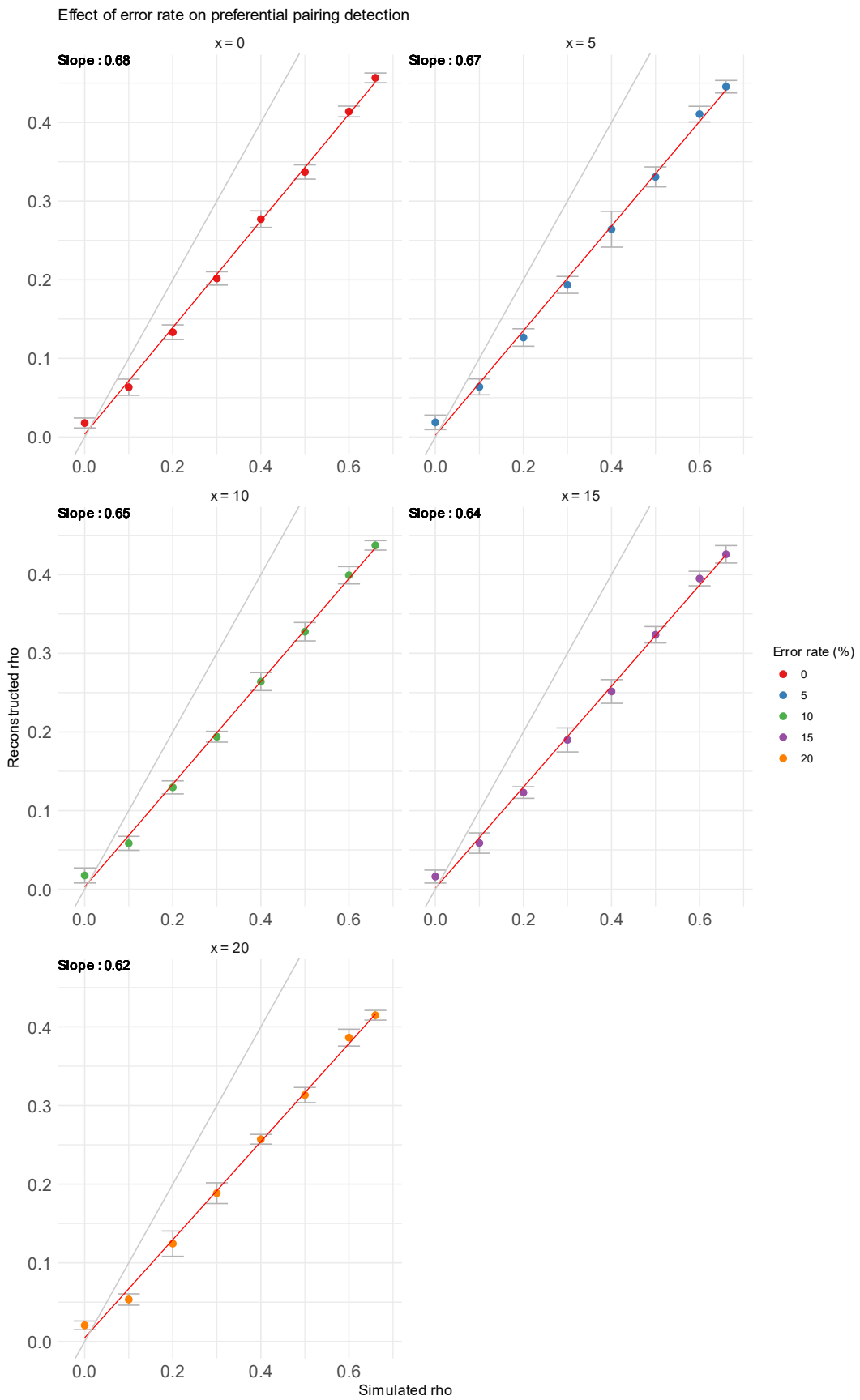


Figure 10: Simulated ρ versus reconstructed ρ for different error rate. The grey diagonal displays a correlation of one.

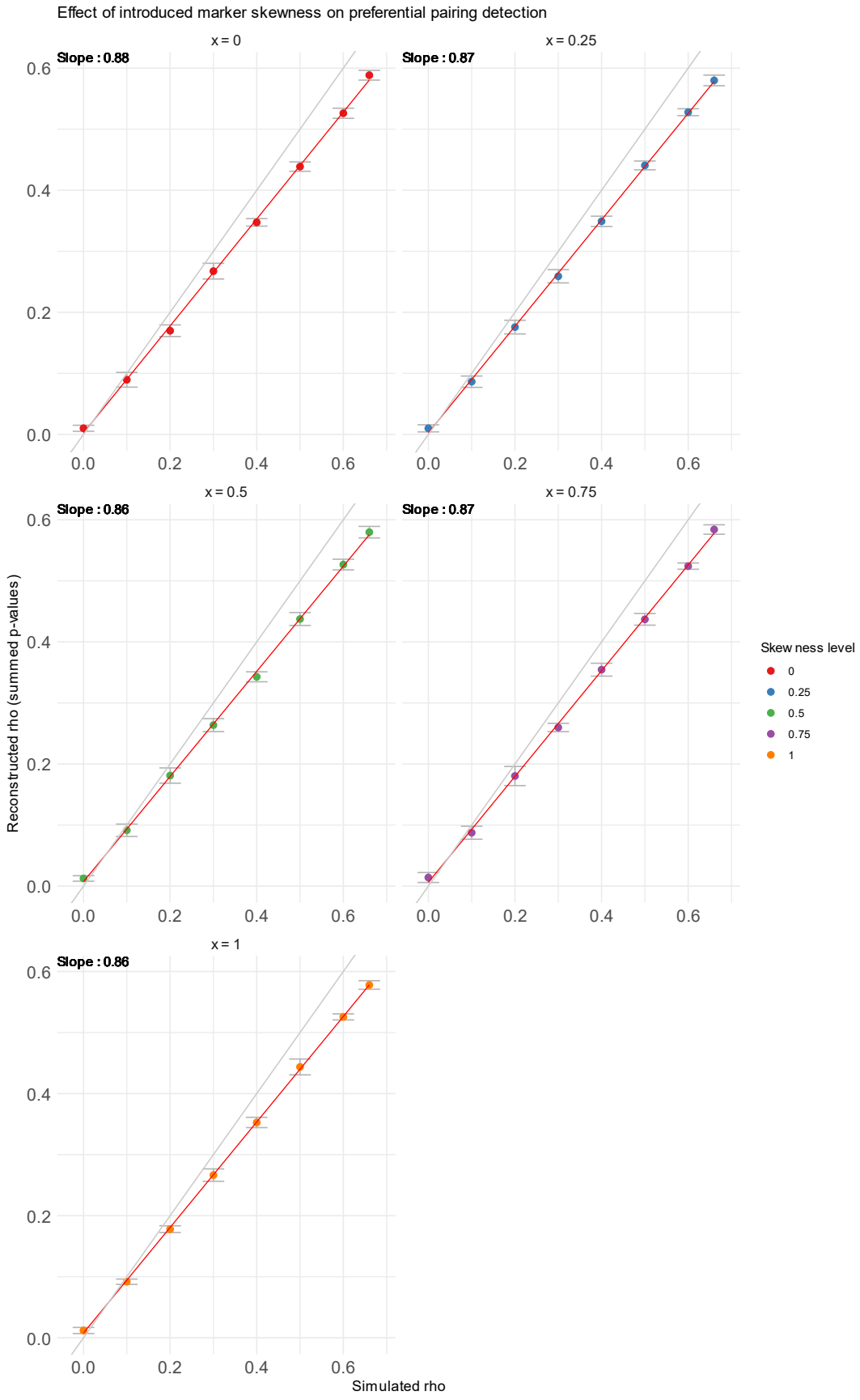


Figure 11: Simulated ρ versus reconstructed ρ for different degrees of marker skewness. The grey diagonal displays a correlation of one.

4. Discussion

Preferential pairing is an interesting phenomenon occurring during meiosis and understanding all details about this pairing behaviour is of fundamental importance of plant breeders and geneticists interested in polyploid species. More insights into preferential pairing can come from harvesting much more data. Such data currently comes from either chromosome staining microscopy studies or population genotyping studies. Population studies using marker assays are mainly used to generate linkage maps of said populations and inherently contain information about preferential pairing. They are however expensive to develop and maintain. Screening many smaller populations with a minimal assay may reduce costs and deliver higher throughput information on homologue conformation during metaphase I of meiosis. This allows for more data to be obtained from more F1 populations and gives geneticists the opportunity to better answer questions surrounding the presence of preferential pairing and underlying causes.

Screening of smaller populations with a minimal marker assay for the purpose of preferential pairing detection might run into limits as a certain amount of information is necessary to properly resolve the parental phasing needed for proper identity-by-descent estimates. Here, we find that population size does not severely affect the detection of preferential pairing as done in PolyOrigin. This is in agreement with results presented by the authors of polyOrigin (Zheng et al., 2021), who show that polyOrigin has a highly consistent ancestral inference error of 0.1 with very narrow 95% confidence intervals for population sizes down to 30 individuals and worse performance below a population size of 20. Because the aforementioned authors presumably measured their ancestral inference accuracy on locus level instead of individual level the values cannot be directly compared though. Given this reported performance, we found testing below 25 individuals to be of no direct added value. While we measure strength of preferential pairing under a null hypothesis of random bivalent pairing, the underlying measurement is in fact reflecting the capacity to infer the correct homologues of P1 that lined up for crossover during gamete simulation. This specific point of view reveals that ρ itself affects the correct inference of homologues negatively and thus the meiotic configuration as well. Evidence for this can be found in the consistent lower accuracy for higher values of ρ. So, population size does not show severe detrimental effects on preferential pairing detection when there are sufficient markers, but strong preferential pairing itself does reduce accuracy of its detection. Finally, at lower population sizes a real preferential pairing effect becomes harder to discern from population sampling variance, but we did not look into statistical evaluation. It would however be possible with simulated data to generate samples with skewed pairing configurations and determine a lower boundary where the sampling effect causes no more than the a certain allowed false negative number of rho measurements. If that lower boundary for population size would prove to be higher than the boundary concluded from our experiment, that would be reason to increase the advised minimum population size.

Decreasing the number of markers has a considerable effect on preferential pairing detection. Less than 80 markers already result in less accurate ancestral homologue inference. Further decreasing the number of markers to lower than 40 results in a large drop that could be considered unusable. In order to estimate the two parental homologues that formed the two inherited offspring homologues there has to be an identifiable portion of each homologue present. When a crossover happens close to the distal part of the chromosome, the resulting inherited chromosome consists of a very large part of one homologue and a small part of the other. When there is no specific marker information in this small portion, inference of the correct parental homologue becomes more difficult. This is also true for chromosomes that have had no exchange with a homolog. If two of such chromosomes end up in the

same gamete, it is impossible for the software to obtain pairing configurations. This would only be possible by having tetrad information.

Robustness to marker skewness is highly relevant when studying preferential pairing, as marker skewness manifests itself in the same way as preferential pairing by showing (local) deviations from the expected segregation rate. This makes it hard to distinguish one from another. Marker skewness is usually found as a result of selection, either because certain alleles or combinations of alleles prove to be deleterious or because of breeding for certain traits. When the idea is to screen many populations of rose and maybe including company germplasm, chances are there will be marker skewness present in these populations. In our experiment we find that marker skewness does not impact the detection of preferential pairing and we show that polyOrigin is robust for populations under strong selection, when one allele is completely deleterious. This may be attributed to the implementation of a Hidden Markov Model that can deal well with local skewness. Even though selection on one specific allele will affect other linked markers, the effect size is determined by genetic distance (linkage). We showed that the majority of markers along a map will show little effect, while closely linked markers do show the selection effect,. As the introduced skewness shows mostly locally, it does not show a bias in our preferential pairing estimator ρ. This robustness of polyOrigin to skewness with regards to preferential pairing detection has not been described before by the authors. They mention specifically in their model design: “We have also assumed implicitly that there is no selection and thus no segregation distortion. This assumption is used as a prior in TetraOrigin ... that might result in incorrect estimation of parental origin and QTL mapping” (Zheng et al., 2016).

Dosage errors do not seem to have a big effect on the detection of preferential pairing as even 20% of dosages containing an erroneous dosage does only show a decrease in the regression slope of 0.05. This is in full agreement with findings of the developers of PolyOrigin, who state that it is highly robust to dosage errors. Overall, the preferential pairing detection is less accurate compared to the simulated 40 marker set used in the experiment focused on number of markers. While parental marker maps for both experiments were designed using the same marker type distribution, they are not exactly the same as for both maps, types were assigned a position randomly. The differences in preferential pairing detection between two similar but not identical maps for our experiments may point to a possible high variability in preferential pairing detection based on the parental maps. As the effect does seem to be consistent across all tested sets and in agreement with literature, no further search on methodological errors was done. Preferential pairing detection using PolyOrigin is robust to dosage errors.

Simulations like the ones in our described experiments will always be lacking in some regard to real biological populations and this will undoubtedly influence the outcome of parameter and robustness testing. For instance, the tested dosage errors and marker skewness will be both present in varying degrees and may have an interaction effect when it comes to estimating parental homologs an thus preferential pairing. Even marker skewness caused by selection will be widely present but its effect can differ per allele and will be dependent on selection pressure and genetic stability. By separating different parameters however, it is possible to isolate and observe the effect and remove interaction effects like we did in this study. While these interaction effects are very valuable as they will be present in real data, it would require a significantly more thorough exploration of parameter space of the used software and within the experiment. The amount of time and compute power for such a study does surpass the resources available for this study. Using a real population for testing of preferential pairing detection could be beneficial to see if the values detected in this experiment hold up. Creating and maintaining such a population, if not publicly available, is expensive though. The model used in this study is a relatively simple one that assumes bivalent pairing. Under this assumption, when sampling

any relevant population size there will be always $\rho > 0$, as there will almost always be some imbalance between parental homologues sampled. This means there is some overlap between real biological preferential pairing and sampling variation. In previous studies a chi-squared test was applied to determine whether there was a significant deviation from an equal distribution of bivalent pairing configurations. We did not focus on this statistical testing in this study, but the software used does readily provides a chi-squared test. Obtaining ρ and its variance might be enough for downstream analysis, when, for a given value of ρ, a correction can be applied, or a different model can be chosen for a certain linkage group.

Next to the bivalent assumption in our calculation for ρ, other assumptions might have influenced results. One being the model and our estimator itself: ρ comes from a model that assumes bivalent pairing and we used it to sample our groups. This is a possible introduction of bias into our experiment. Even though we can see the linear relationship between the PedigreeSim prefPairing parameter and ρ, a more unbiased way would have been to introduce preferential pairing using PedigreeSim, even if that would have introduced more variation due to the sampling. When generating a linkage map or reconstructing haplotypes of a tetraploid species, the most logical choice is to apply a model that allows multivalents because there is most probably no prior information on pairing configurations. In this study we applied this model for PolyOrigin and found that it does not affect the estimations of ρ, as the multivalent probability is separated in parsing of the output. We do find that in more extreme scenarios some multivalent configurations are reported as most likely. Thus, the assigned probabilities do show effects of the tested parameters and our method of calculation of rho does not seem to reflect the exact situation when going to those extremes. This may also be a partial explanation as to why preferential pairing itself seems to have the biggest effect on correct detection. As stronger preferential pairing implies a stronger deviation from a polysomic inheritance model to a disomic model. Another possible weakness in our tests is the use of PolyOrigin with reference haplotypes of the parents. This allowed us to skip the reconstruction of the parental map but may have been too far from reality as in practice these exact haplotypes would not be known and creating the parental map would already build upon assumptions on the mode of inheritance. Nonetheless, the trends described here seem solid and informative and can be used to further study this subject.

Because the topic of polyploid meiosis and inheritance is complex we could not address all issues that would be relevant for development of a tool to determine inheritance modes easily. We restricted this study to bivalents while literature describes multivalents as not uncommon. For instance, in potato, Bourke et al. (2015) report 10-12 percent of pairings is multivalent. Which is the basis for double reduction, that can sew segregation patterns too. We also did not address the issue of what marker types to look for and at what positions they should cover. Based on published resources a lot of information to answer these questions is around already.

Successful development of a low-cost and easy to apply generic method to quickly screen experimental populations can help to select populations that show strong preferential pairing, or maybe exactly the opposite depending on the questions that one would like to answer. This tool could help reduce costs and help by applying extensive genotyping to the most interesting experimental populations. The information can be used to correctly set the inheritance models for downstream analysis such as production of (phased) linkage map and QTL mapping. In that sense it may reduce time needed for such experiments and improve results. It would allow also plant breeders to screen more of their populations or explain why certain introgression targets will not be realized. Given this reasoning and our results, pre-detection of preferential pairing could be useful and would be worth developing, but it would need a more detailed template for design depending on the requirements set.

5. Conclusion

In this study we isolated several parameters important to population mapping and studied the effects on detection of preferential pairing. This is of relevance to gain knowledge on minimal parameters for studying pairing behaviour in a screen of many populations. We found that population sizes down to 25 individuals do not affect correct inference of meiotic configurations. Additionally, we found that using 40 markers seems to be the lower bound for reliably testing for preferential pairing. These results are in agreement with reports of population size effects for PolyOrigin. The same can be concluded for PolyOrigin's robustness to dosage error rate. PolyOrigin also displays robustness to marker skewness even though the design did not take the effect into account and was not tested.

Given the possible improvements stated in the Discussion section, further testing of the software parameter space might be interesting as there is quite some ground not covered yet. However, with the clear trends described here, a low-resolution screen may already be designed for e.g., rose. This can be done with the genomic resources available from other projects or public data repositories. Selection for the most informative marker types can be done by applying genomics tools such as mapping resequencing data to a reference and variant calling. This would allow for a quick screen of allele frequencies across species and physical locations on the chromosomes. After design of such a low-resolution screen, many populations can be screened using a cheap but still high-throughput screening method such as a PCR-based KASP™ assay. The selected populations should then ideally be screened using a long-read sequencing approach (ideally Oxford Nanopore Technologies). While this is more expensive than applying the developed rose SNP array, it does contain much more information about the exact locations of recombination and opens up possibilities for linking recombination to other genome features such as genes, variants, epigenetics, transposable elements and structural differences. The obtained knowledge from this can help answer remaining questions about preferential pairing and can be applied to fine tune inheritance models. and possibly for future enhancement of recombination and introgression prediction.

6. References

Ahmed, D., Curk, F., Evrard, J. C., Froelicher, Y., & Ollitrault, P. (2020). Preferential Disomic Segregation and C. micrantha/C. medica Interspecific Recombination in Tetraploid 'Giant Key' Lime; Outlook for Triploid Lime Breeding. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00939

Bourke, P. M. (2018). *Genetic mapping in polyploids* [Wageningen University]. https://doi.org/10.18174/444415

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., & Maliepaard, C. (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal*, *90*(2), 330–343. https://doi.org/10.1111/TPJ.13496

Bourke, P. M., Van Geest, G., Voorrips, R. E., Jansen, J., Kranenburg, T., Shahin, A., Visser, R. G. F., Arens, P., Smulders, M. J. M., & Maliepaard, C. (2018). PolymapR - Linkage analysis and genetic map construction from F 1 populations of outcrossing polyploids. *Bioinformatics*, *34*(20). https://doi.org/10.1093/bioinformatics/bty371

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., & Maliepaard, C. (2015). The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, *201*(3). https://doi.org/10.1534/genetics.115.181008

Mason, A. S., & Wendel, J. F. (2020). Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution. In *Frontiers in Genetics* (Vol. 11). https://doi.org/10.3389/fgene.2020.01014

Mollinari, M., & Garcia, A. A. F. (2019). Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden Markov models. *G3: Genes, Genomes, Genetics*, *9*(10). https://doi.org/10.1534/g3.119.400378

Mollinari, M., Olukolu, B. A., Da Pereira, G. S., Khan, A., Gemenet, D., Craig Yencho, G., & Zeng, Z. B. (2020). Unraveling the hexaploid sweetpotato inheritance using ultra-dense multilocus mapping. *G3: Genes, Genomes, Genetics*, *10*(1). https://doi.org/10.1534/g3.119.400620

Morgan, T. H. (1911). Random segregation versus coupling in Mendelian inheritance. *Science*, *34*(873). https://doi.org/10.1126/science.34.873.384

Rdbickel. (2016). *Overview of Meiosis*. Https://Commons.Wikimedia.Org/Wiki/File:Meiosis_Overview_new.Svg.

Soares, N. R., Mollinari, M., Oliveira, G. K., Pereira, G. S., & Vieira, M. L. C. (2021). Meiosis in polyploids and implications for genetic mapping: A review. In *Genes* (Vol. 12, Issue 10). https://doi.org/10.3390/genes12101517

Song, L., & Endelman, J. B. (2023). Using haplotype and QTL analysis to fix favorable alleles in diploid potato breeding. *Plant Genome*, *16*(2). https://doi.org/10.1002/tpg2.20339

Sybenga, J. (1996). Chromosome pairing affinity and quadrivalent formation in polyploids: Do segmental allopolyploids exist? *Genome*, *39*(6). https://doi.org/10.1139/g96-148

Thompson, L. H., & Schild, D. (2001). Homologous recombinational repair of DNA ensures mammalian chromosome stability. In *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* (Vol. 477, Issues 1–2). https://doi.org/10.1016/S0027-5107(01)00115-4

van Geest, G., Bourke, P. M., Voorrips, R. E., Marasek-Ciolakowska, A., Liao, Y., Post, A., van Meeteren, U., Visser, R. G. F., Maliepaard, C., & Arens, P. (2017). An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis. *Theoretical and Applied Genetics*, *130*(12). https://doi.org/10.1007/s00122-017-2974-5

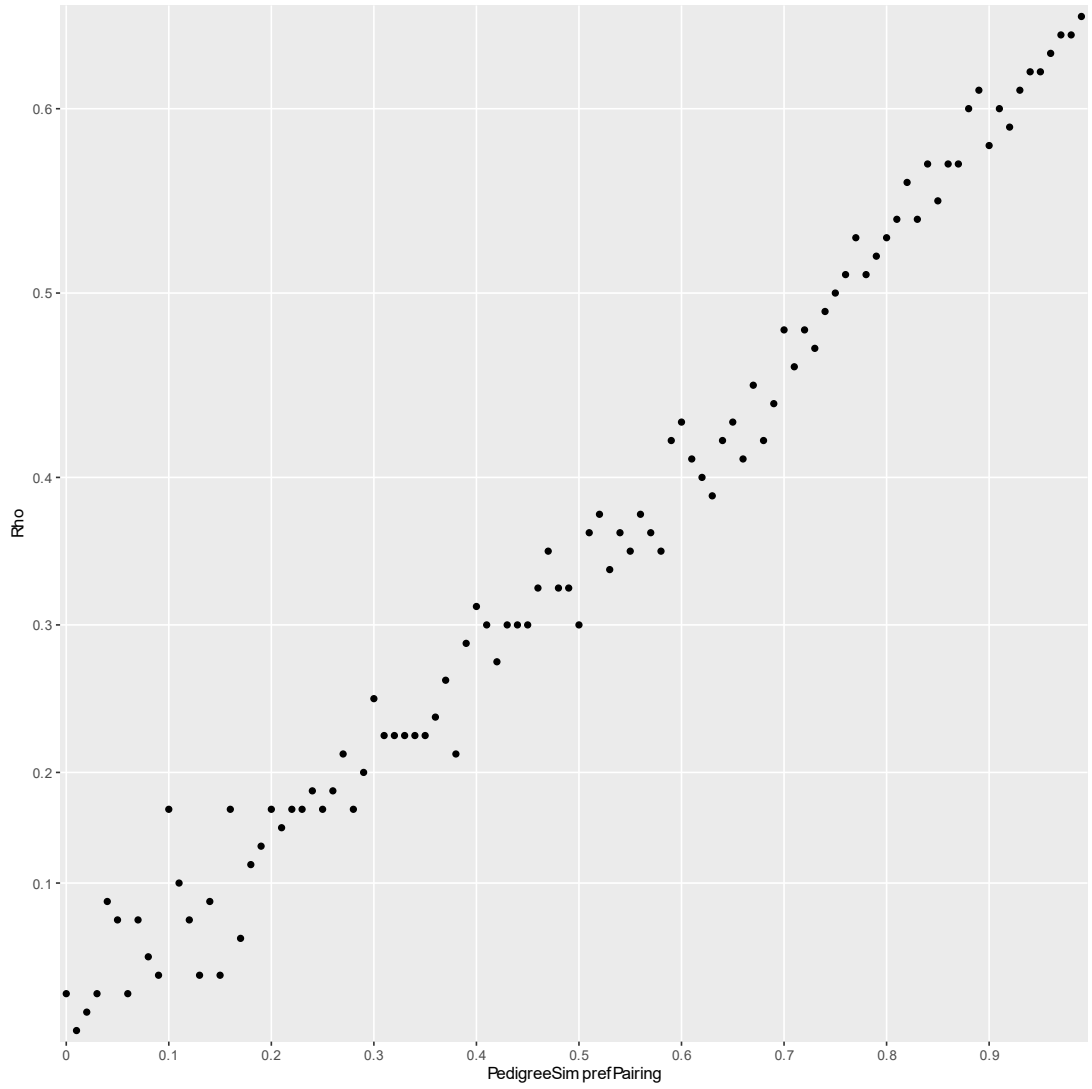
Voorrips, R. E., & Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics*, *13*(1), 1–12. https://doi.org/10.1186/1471-2105-13-248

Zheng, C., Amadeu, R. R., Munoz, P. R., & Endelman, J. B. (2021). Haplotype reconstruction in connected tetraploid F1 populations. *Genetics*, *219*(2). https://doi.org/10.1093/genetics/iyab106

Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J., & Bink, M. C. A. M. (2016). Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics*, *203*(1). https://doi.org/10.1534/genetics.115.185579

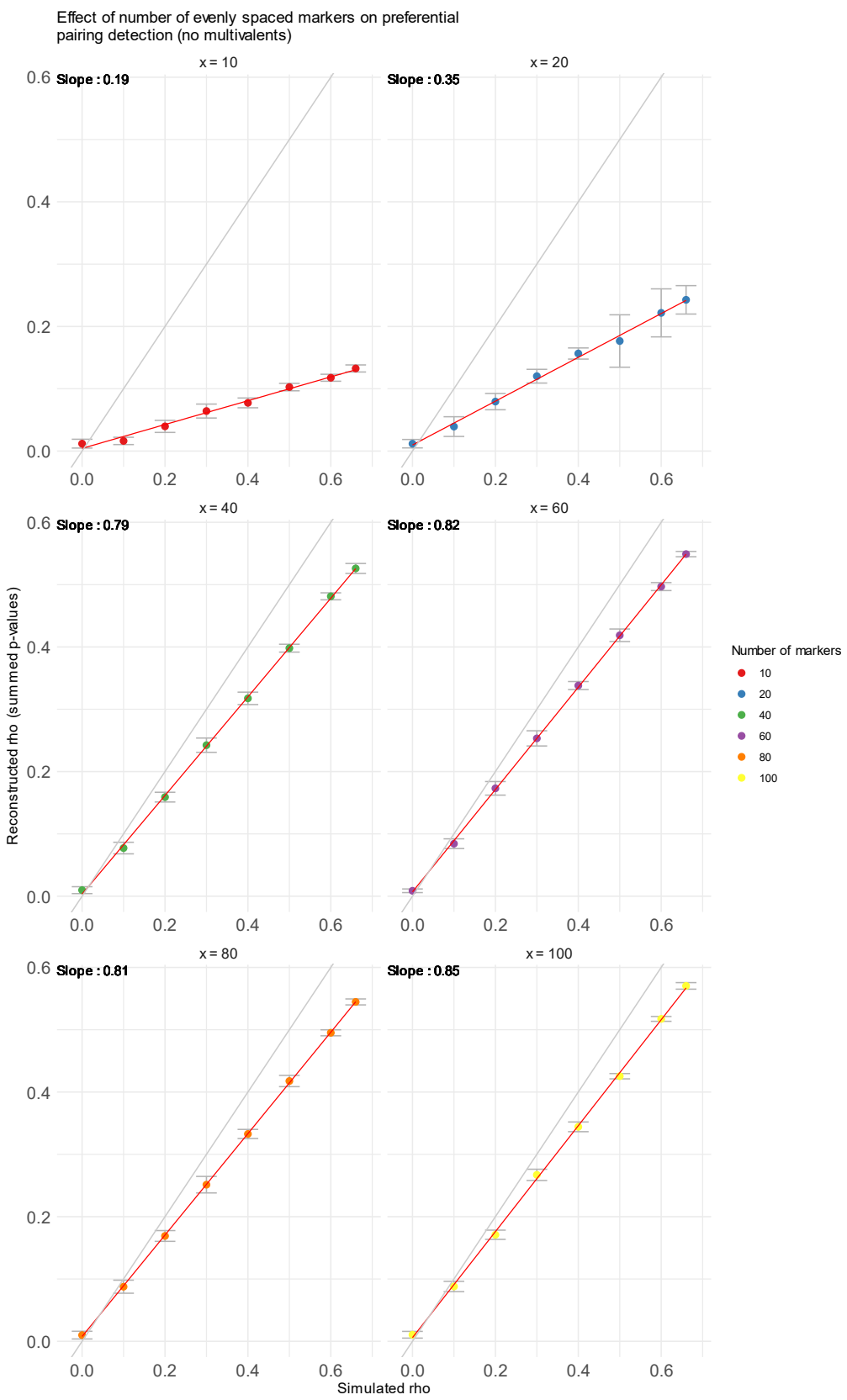
7. Supplementary

A. PedigreeSim prefPairing to rho conversion



Supplementary S1: Linear relationship between ρ and PedigreeSim prefPairing parameter and visible variation introduced by PedigreeSim.

B. Rerun polyOrigin modified settings



Supplementary S2: Effect of number of markers on preferential pairing detection when polyOrigin models bivalents in offspring only. This modified setting causes all reconstructed haplotypes to stem from a bivalent configuration but does not improve preferential pairing detection.

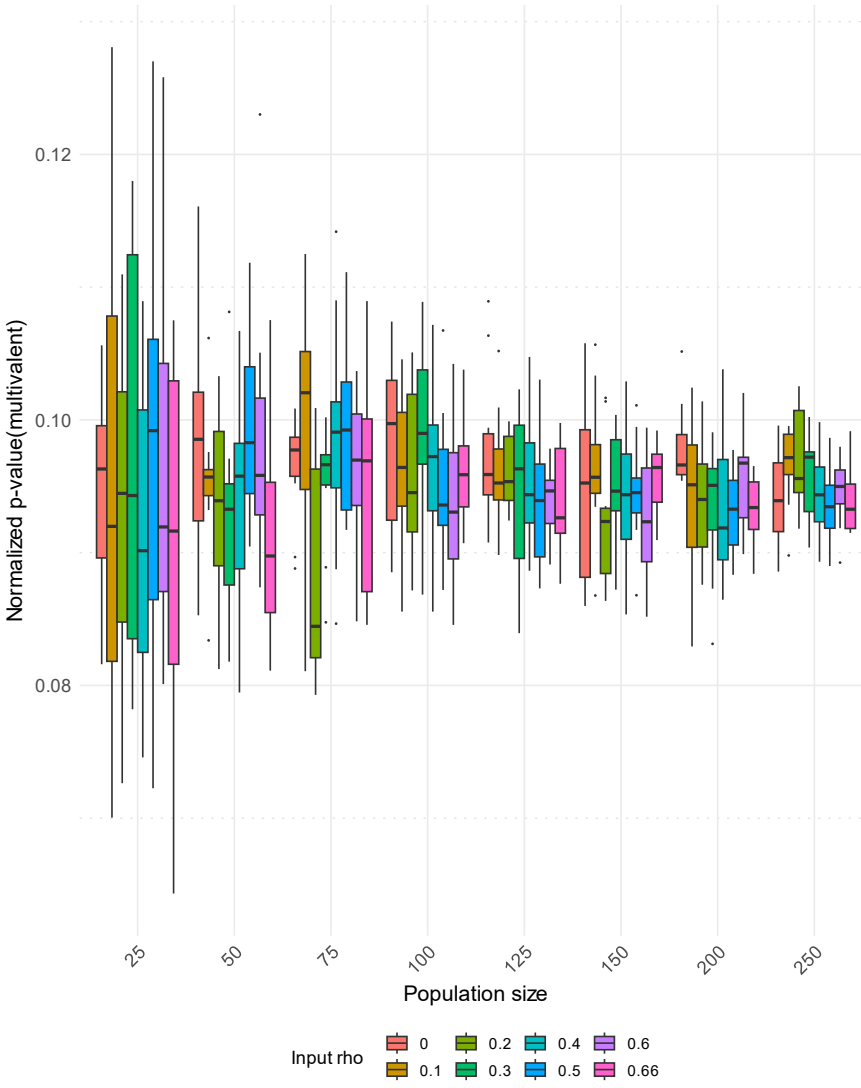
C. Packages and modules

```
R version 4.2.2 (2022-10-31)
Platform: x86_64-conda-linux-gnu (64-bit)
Running under: Ubuntu 20.04 LTS

Matrix products: default
BLAS/LAPACK: /home/WUR/nieuw133/miniconda3/envs/R_kernel/lib/libopenblasp-
r0.3.21.so

locale:
 [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
 [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
 [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
[10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] data.table_1.14.8  ggpubr_0.6.0       lubridate_1.9.2    forcats_1.0.0
 [5] stringr_1.5.0      dplyr_1.1.2        purrr_1.0.2        readr_2.1.4
 [9] tidyr_1.3.0        tibble_3.2.1       tidyverse_2.0.0    ggsankey_0.0.99999
[13] ggthemes_4.2.4     ggplot2_3.4.2      digest_0.6.31      polyqtlR_0.0.10
[17] polymapR_1.1.3

loaded via a namespace (and not attached):
 [1] pbdZMQ_0.3-9     tidyselect_1.2.0 xfun_0.39        repr_1.1.6
 [5] carData_3.0-5    colorspace_2.1-0 vctrs_0.6.3      generics_0.1.3
 [9] htmltools_0.5.5  base64enc_0.1-3  utf8_1.2.3       rlang_1.1.1
[13] pillar_1.9.0     glue_1.6.2       withr_2.5.0      uuid_1.1-0
[17] foreach_1.5.2    lifecycle_1.0.3  ggsignif_0.6.4   munsell_0.5.0
[21] gtable_0.3.3     codetools_0.2-19 evaluate_0.21    knitr_1.42
[25] tzdb_0.4.0       fastmap_1.1.1    fansi_1.0.4      broom_1.0.4
[29] IRdisplay_1.1    Rcpp_1.0.10      backports_1.4.1  scales_1.2.1
[33] IRkernel_1.3.2   jsonlite_1.8.4   abind_1.4-5      hms_1.1.3
[37] stringi_1.7.12   rstatix_0.7.2    grid_4.2.2       cli_3.6.1
[41] tools_4.2.2      magrittr_2.0.3   car_3.1-2        crayon_1.5.2
[45] pkgconfig_2.0.3  timechange_0.2.0 iterators_1.0.14 R6_2.5.1
[49] compiler_4.2.2
```

D. polyOrigin log parsing

```
#!/usr/bin/env bash
for I in *.log
do
        REAL_POPSIZE=$(grep '^data:' $I | cut -f3 -d',' | \
                sed 's/ #offspring=//' -)
        RHO=$(echo "${I}" | cut -f7 -d'_')
        REP=$(echo "${I}" | cut -f9 -d'_' | sed 's/.log//' - )
        MARKERS=$(grep '^#chr=1, #marker=' $I | cut -f2 -d',' | \
                sed 's/ #marker=//')
        NUM_DELETED=$(grep '^delete ' $I | cut -f2 -d' ');
        if [[ "${NUM_DELETED}" == "" ]]
        then
                NUM_DELETED=0
        fi
        MISMATCH_PHASES=$(grep 'comparison: #mismatch phases' $I | \
                cut -f2 -d'=' | sed 's/^ \[//; s/\]$//' - )
        MISMATCH_DOSAGES=$(grep 'comparison: #mismatch dosages' $I | \
                cut -f2 -d'=' | sed 's/^ \[//; s/\]$//' - )
        POLYPHASE_TIME=$(grep 'seconds by polyPhase$' $I | cut -f3 -d',' | \
                sed 's/ time elapsed = //' - | \
                sed 's/ seconds by polyPhase$//' -)
        EPS=$(grep '<eps>=' $I | cut -f3 -d',' | sed 's/ <eps>=//' - )
        OUTLIER_OFFSPRING=$(grep '^no outlier offspring$' $I)
        if [[ "${OUTLIER_OFFSPRING}" == "no outlier offspring" ]]
        then
                OUTLIER_OFFSPRING=0
        else
                OUTLIER_OFFSPRING=$(grep ' outlier offspring:' $I | \
                        cut -f1 -d' ')
        fi
        POLYRECONSTRUCT_TIME=$(grep ' seconds by polyReconstruct!$' $I | \
                cut -f3 -d',' | sed 's/ time elapsed = //' - | \
                sed 's/ seconds by polyReconstruct!$//' -)
        TOTAL_TIME=$(grep ' seconds by polyOrigin$' $I | cut -f3 -d',' | \
                sed 's/ total time used = //' - | \
                sed 's/ seconds by polyOrigin$//' -)
        echo -e "${REAL_POPSIZE}\t${RHO}\t${REP}\t${MARKERS}\t${NUM_DELETED}\t\
                ${MISMATCH_PHASES}\t${MISMATCH_DOSAGES}\t${POLYPHASE_TIME}\t\
                ${EPS}\t${OUTLIER_OFFSPRING}\t${POLYRECONSTRUCT_TIME}\t\
                ${TOTAL_TIME}"
done > ../Exp_1_results_base.tsv
```

E. Multivalent 1:2:3:4 probability by population size



Supplementary S3: Normalized p-values from polyOrigin output for different strengths of preferential pairing and population sizes.

F. Allele frequencies of simulated skewness sets

Supplementary S4: Allele frequency for simplex x nulliplex and nulliplex x simplex markers for different levels of introduced skewness. The red line highlights the specific marker upon which selection was simulated. Linked markers show a similar but more moderate effect, while unlinked markers show an opposite effect.