**BMC Biology**
*incorporating* **Journal of Biology**

CORRESPONDENCE      Open Access

# Polyploidization increases meiotic recombination frequency in *Arabidopsis*: a close look at statistical modeling and data analysis

Lin Wang[1] and Zewei Luo[1,2]*

### Abstract

This paper is a response to Pecinka A, Fang W, Rehmsmeier M, Levy AA, Mittelsten Scheid, O: **Polyploidization increases meiotic recombination frequency in *Arabidopsis***. *BMC Biology* 2011, **9**:24.
See research article at http://www.biomedcentral.com/1741-7007/9/24

## Background

Many species, particularly flowering plants, have usually experienced a state of polyploidy in their evolutionary history. Meiotic recombination creates novel configurations of genetic variants maintained in the genome of a species for facilitating natural and artificial selection. Thus, to understand how a diploid species differs in frequency of meiotic recombination from its polyploid ancestor has a significant impact in both evolutionary biology and plant and animal genetic breeding. It has been well established that the evolution of polyploid genomes is an extremely dynamic process compared to that of diploids, characterized by extensive genetic and epigenetic changes occurring in the nuclear genome following polyploidization [1-3]. Little is known about the mechanism underpinning the genetic changes. To address this fundamental question, Pecinka and colleagues described in a recently published paper by *BMC Biology* a direct comparison in frequency of meiotic recombination of the diploid and tetraploid genomes of *Arabidopsis* [4]. One of the most striking methodological challenges to the study is to properly evaluate the recombination parameter in populations of the species at different levels of polyploidy, particularly in autopolyploids. In fact, linkage analysis with autotetraploids has been a historical problem that can be traced back to the pioneering works of the prominent mathematical geneticists [5-7].

There are at least two major challenges to the tetrasomic linkage analysis. Firstly, double reduction, the most distinct characteristic of polysomic inheritance, allows sister chromatids to enter into the same gamete during meiosis and thus cause systematic distortion in allele segregation [8]. Secondly, multiplex allele segregation makes it almost impossible to infer the underlying genotype directly from PRC-based phenotype data even for codominant markers such as a single nucleotide polymorphism or simple sequence repeats [9]. To avoid these difficulties, these authors firstly developed a seed-based assay by creating transformants with green and red fluorescent markers expressed under a seed-specific promoter in *Arabidopsis thaliana* [10,11]. In parallel, they created diploids, allotetraploids and autotetraploids which carry these fluorescent markers. After carrying out a series of backcross and selection breeding, they were able to create the diploid and tetraploid lines which bear only a single copy of the marker alleles linked on the *Arabidopsis* chromosome III. These lines were used to create the segregation populations from which marker phenotype data were collected and used for estimation of recombination frequency between the markers [4].

The method these authors implemented for modeling and analyzing the marker data from diploid and tetraploid (allo- and autotetraploid) populations needs to be formulated on the basis of the disomic and tetrasomic inheritance models. This paper presents statistically appropriate and mathematically rigorous methods for modeling and re-analyzing the datasets.

* Correspondence: z.luo@bam.ac.uk
[1]Laboratory of Population & Quantitative Genetics, Institute of Biostatistics, Fudan University, Shanghai 200433, China
Full list of author information is available at the end of the article

**BioMed** Central

### Notation, model and analysis

We consider segregation of alleles at the two fluorescent marker loci on the *Arabidopsis* chromosome III in a $F_2$ family from crossing two parental lines at the marker loci. Following notations of Pecinka *et al.* [4], parental genotypes at the markers can be denoted by GR/GR and BC/BC for diploid, GR/BC/DE/DE and BC/BC/DE/DE for allotetraploid, and GR/BC/BC/BC and BC/BC/BC/BC for autotetraploid. The parental lines were crossed to generate offspring populations of diploids, allotetraploids and autotetraploids accordingly. Regardless of polyploidy, the offspring populations from mating these parents can be grouped into four phenotypes: yellow (carrying both red and green marker alleles), green (green allele only), red (red allele only) and grey (none of the marker alleles). The number of individuals for each of the four phenotype classes is denoted by $n_1, n_2$, $n_3$ and $n_4$ respectively. Let $r$ be recombination frequency between the two markers and $\alpha$ be the coefficient of double reduction at the green marker, which is nearer to the centromere than the red marker locus. The probability of observing each of the four phenotypes in the diploids and allotetraploids depends on only one parameter, $r$ ($f_i(r)$, $i = 1,...,4$), but characterization of the phenotypic distribution in the autotetraploids needs the two parameters, $r$ and $\alpha$ ($f_i(\alpha,r)$, $i = 1,...,4$).

The logarithm of the model parameter(s) given the observations $n_1, n_2, n_3$ and $n_4$ can be written as:

$$L(r|n_1, n_2, n_3, n_4) \propto \sum_{i=1}^{4} n_i \log[f_i(r)]$$

for the diploid and allotetraploid populations or:

$$L(\alpha, r|n_1, n_2, n_3, n_4) \propto \sum_{i=1}^{4} n_i \log[f_i(\alpha, r)]$$

for the autotetraploid population where $f_i(r)$ ($i = 1,...,4$) can be worked out following the principle of two-locus disomic linkage analysis and listed in Table 1. It needs to be pointed out that the phenotypic distribution is common between the diploid and allotetraploid populations. This is because homoeologous pairing was completely excluded in meiosis of the synthesized allotetraploids of *Arabidopsis* [10,11], thus the allotetraploids show strict disomic inheritance. However, calculation of phenotypic distribution in the autotetraploid segregation population must follow the principle of tetrasomic linkage analysis as is detailed in [11]. In the present context, the phenotypic distribution can be worked out as:

$$f_1(\alpha, r) = g_1(\alpha, r)[2 - g_1(\alpha, r)] + 2g_2(\alpha, r)g_3(\alpha, r)$$

$$f_2(\alpha, r) = g_2(\alpha, r)[g_2(\alpha, r) + 2g_4(\alpha, r)]$$

$$f_3(\alpha, r) = g_3(\alpha, r)[g_3(\alpha, r) + 2g_4(\alpha, r)]$$

$$f_4(\alpha, r) = g_4(\alpha, r)^2$$

where $g_i(\alpha,r)$ ($i = 1, 2, 3, 4$) are as given below:

$$g_1(\alpha, r) = (2 - \alpha)(1 - r)^2/4 + r(1 - r)/2 + (1 - \alpha)r^2/6$$

$$g_2(\alpha, r) = (1 - \alpha)r(1 - r)/3 + (10 - \alpha)r^2/36$$
$$g_3(\alpha, r) = r(1 - r)/2 + (8 + \alpha)r^2/36$$
$$g_4(\alpha, r) = (2 + \alpha)(1 - r)^2/4 + (2 + \alpha)r(1 - r)/3 + (2 + \alpha)r^2/6$$

The maximum likelihood estimate (MLE) of the recombination frequency in diploids and allotetraploids can be calculated from solving the normal equations:

$$\partial L(r|n_1, n_2, n_3, n_4)/\partial r = 0$$

which is a quadratic equation of $r$. In the diploid population where $n_1= 2805$, $n_2= 322$, $n_3= 333$ and $n_4= 791$, the quadratic equation has only one real root filling in [0.0, 0.5], which is the MLE $\hat{r} = 0.1643$. Based on the likelihood function, one can calculate the standard

**Table 1 Distribution of seed phenotype and the underlying genotype at the two fluorescence markers in $F_2$ diploid and autotetraploid populations and estimates of the model parameters**

| Phenotype | Diploids/Allotetraploid | | | | Autotetraploids | | |
| | Genotype[a] | Frequency ($f_i(r)$) | Obs ($n_i$) | Obs ($n_i$) | Genotype[b] | Frequency[c] | Obs ($n_i$) |
|---|---|---|---|---|---|---|---|
| Yellow | G_R_ | $3(1-r)^2/4+r(1-r)+r^2/2$ | 2,805 | 1,484 | $G^{(i)}B^{(4-i)}R^{(j)}C^{(4-j)}$ | $f_1(\alpha,r)$ | 12,707 |
| Green | G_CC | $r(2-r)/4$ | 322 | 275 | $G^{(i)}B^{(4-i)}C^{(4)}$ | $f_2(\alpha,r)$ | 1,868 |
| Red | BBR_ | $r(2-r)/4$ | 333 | 298 | $B^{(4)}R^{(i)}C^{(4-i)}$ | $f_3(\alpha,r)$ | 2,216 |
| Grey | BBCC | $(1-r)^2/4$ | 791 | 320 | $B^{(4)}C^{(4)}$ | $f_4(\alpha,r)$ | 3,098 |
| Estimates | $\hat{r} \pm s.d.$ | Log-likelihood | $\hat{r} \pm s.d.$ | Log-likelihood | $\hat{\alpha} \pm s.d.$ | $\hat{r} \pm s.d.$ | Log-likelihood |
| Present study | 0.1643 ± 0.0062 | -4,177.79 | 0.2770 ± 0.0110 | -2,553.43 | 0.0676 ± 0.0121 | 0.3048 ± 0.0051 | -20815.3 |
| Pecinka *et al.* | 0.154 ± 0.009 | -4,179.20 | 0.241 ± 0.018 | -2,559.21 | - | 0.205 ± 0.011 | -21010.0 |

[a]Stands for any other alleles; [b]$G^{(i)}B^{(4-i)}R^{(j)}C^{(4-j)}$ is a tetraploid genotype containing $i$ 'G' alleles, 4-$i$ 'B' alleles, $j$ 'R' alleles and 4 - $j$ 'C' alleles; [c]forms of $f_i(\alpha,r)$ ($i = 1, 2, 3, 4$) are presented in context. s.d.: standard deviation.

deviation of the estimate from the Fisher's information measure for MLE. In the present context, it equals:

$$\sqrt{-1/\left[\partial^2 L(r|n_1, n_2, n_3, n_4)/\partial r^2\right]}_{r=\hat{r}} = 0.0062 \text{ (Table 1).}$$

In their original report [4], Pecinka *et al.* estimated the recombination frequency by equating the probability of the recombinant individuals, in other words, those displaying green and red seeds, to the observed proportion of these individuals. In the present notations, the probability has a form of $2r-r^2 = 2(n_2+n_3)/n$ where $n = n_1+n_2+n_3+n_4$. They provided an estimate of $\hat{r} = 0.154$ with a standard deviation of 0.009. It is not clear how the standard deviation was calculated. The method for calculating the recombination frequency may not be statistically appropriate in two aspects. Firstly, the calculation did not use the full information of the data. For example, the individuals with yellow seeds were not taken into consideration when counting for recombination events. In fact, there is a proportion of $[r(1-r)+r^2/2]/[3(1-r)^2/4+r(1-r)+r^2/2]$ among this group of individuals which carry recombinant gametes. Secondly, the MLE of *r* obtained from the present analysis is four times as likely as that provided by the original report, that is:

$$\exp[L(r = 0.1643) - L(r = 0.154)] = 4.096.$$

In the allotetraploid population where $n_1= 1484$, $n_2= 275$, $n_3= 298$ and $n_4= 320$, the MLE and corresponding standard deviation are calculated as 0.2770 ± 0.0110. The estimate is in contrast to 0.241 with standard deviation = 0.018 in the original report.

To analyze the likelihood model for the autotetraploid marker data, we firstly noticed that the parameter α involves information of allele segregation at the green marker only. By setting $r = 0$ in the phenotypic probabilities given above, we can work out:

$$p_g(\alpha) = f_1(\alpha, 0) + f_2(\alpha, 0) = (12 - 4\alpha - \alpha^2)/16$$

$$p_0(\alpha) = f_3(\alpha, 0) + f_4(\alpha, 0) = (2 + \alpha)^2/16$$

which represent the probability of observing an individual carrying or not carrying the green fluorescent marker respectively. Let $n_g = n_1+n_2$ be the number of individuals carrying the green marker allele and $n_0 = n_3 +n_4$ be the number of individuals not carrying the allele. The log-likelihood function has the form:

$$L(\alpha|n_g, n_0) = n_g Log[p_g(\alpha)] + n_0 Log[p_0(\alpha)]$$

Solving the following equation:

$$\partial L(\alpha|n_g, n_0)/\partial\alpha|_{n_g=14575, n_0=5314} = 0$$

results in the MLE of $\hat{\alpha} = 0.0676$ with a standard deviation of 0.0121 (Table 1).

Incorporating $\hat{\alpha} = 0.0676$ into the likelihood function, we found that the equation:

$$\partial L(\hat{\alpha}, r)/\partial r|_{n_1=12707, n_2=1868, n_3=2216, n_4=3098} = 0$$

has only one real root, $\hat{r} = 0.3048$, which is the MLE of the recombination frequency under the tetrasomic model. The estimate has a standard deviation of 0.0051, which was calculated from the second derivative of the likelihood function at the MLE. Based on the estimates of α and *r*, we can predict the coefficient of double reduction at the red marker from:

$$\hat{\beta} = [\hat{\alpha}(3 - 4\hat{r})^2 + 2\hat{r}(3 - 2\hat{r})]/9 = 0.1857 \text{ [12]}$$

which is below the upper bound of the maximum value of double reduction ¼ [12]. The estimates of α and β strongly suggest multivalent pairing among the homologous chromosomes and, in turn, double reduction occurring in the genome region flanked by the fluorescent markers (Table 1).

## Discussion

This article presents likelihood-based methods for estimating the recombination frequency and other relevant parameters from segregating populations of diploids, allotetraploids and autotetraploids. We demonstrate the methods by re-analyzing the datasets published in Pecinka *et al.* [4]. Re-analysis of the datasets with the methods developed here reveals quantitative and qualitative differences from the original analysis. Firstly, unlike Pecinka *et al.*, the present analysis enables estimation of double reduction at the marker loci under study and discovers significant double reduction at the markers, emphasizing the necessity of taking the tetrasomic nature into consideration in the data analysis. Secondly, the present study differs from the original analysis in the inferred order in frequency of meiotic recombination of allotetraploid and autotetraploid *Arabidopsis*. Pecinka *et al.* concluded that meiotic recombination was more frequent in the allotetraploids than in the autotetraploids whilst our results predict differently. Rationale for the present prediction is supported by several aspects. In fact, allotetraploids show strictly disomic inheritance particularly in the present instance where homoeologous chromosome pairing was excluded. In contrast, homologous chromosomes in autotetraploids have a substantially higher chance to pair, which is essential for recombination to occur among the chromosomes [13]. In addition, it has been theoretically demonstrated that autotetraploids have a much higher upper bound value of recombination frequency in autotetraploids (¾) than that in diploids and allotetraploids (1/2)

[14]. Thirdly, the estimates obtained from the present analysis are significantly highly supported by the observed experimental data when compared with those from the original analysis in term of likelihood values of these estimates.

## Computer program

Formulation and numerical analysis represented above were programmed in Mathematica [15] and the programs will be made available from the corresponding author after publication of this paper.

### Author details

[1]Laboratory of Population & Quantitative Genetics, Institute of Biostatistics, Fudan University, Shanghai 200433, China. [2]School of Biosciences, University of Birmingham, Birmingham B15 2TT, UK.

### Authors' contributions

ZWL conceived the research and formulated the statistical analysis. LW and ZWL analyzed the data and wrote the paper. Both authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

### References

1. Soltis DE, Soltis PS: **The dynamic nature of polyploid genome.** *Proc Nat Acad Sci USA* 1995, **92**:8089-8091.
2. Song K, Lu P, Tang K, Osborn TC: **Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution.** *Proc Natl Acad Sci USA* 1995, **92**:7719-7723.
3. Osborn T, Pires J, Birchler J, Auger D, Chen Z: **Understanding mechanisms of novel gene expression in polyploids.** *Trends Genet* 2003, **19**:141-147.
4. Pecinka A, Fang W, Rehmsmeier M, Levy AA, Mittelsten Scheid O: **Polyploidization increases meiotic recombination frequency in Arabidopsis.** *BMC Biology* 2011, **9**:24.
5. Haldane JBS: **Theoretical genetics of autotetraploids.** *J Genet* 1930, **22**:359-372.
6. Mather K: **Segregation and linkage in autotetraploids.** *J Genet* 1936, **30**:287-314.
7. Fisher RA: **The theory of linkage in polysomic inheritance.** *Philos Trans R Soc Lond B Biol Sci* 1947, **23**:55-87.
8. Barley NTJ: **Introduction to Mathematical Theory of Genetic Linkage.** London: Oxford University Press; 1961.
9. Luo ZW, Hackett CA, Bradshaw JE, McNicol JW, Milbourne D: **Predicting parental genotypes and gene segregation for tetrasomic inheritance.** *Theor Appl Genet* 2000, **100**:1067-1073.
10. Melamed-Bessudo C, Yehuda E, Stuitje AR, Levy AA: **A new seed-based assay for meiotic recombination in Arabidopsis thaliana.** *Plant J* 2005, **43**:358-484.
11. Luo ZW, Zhang R, Kearsey MJ: **Theoretical basis for genetic linkage analysis in autotetraploid species.** *Proc Natl Acad Sci USA* 2004, **101**:7040-7045.
12. Cormai L: **The advantages and disadvantages of being polyploidy.** *Nat Rev Genet* 2005, **6**:836-646.
13. Luo ZW, Zhang Z, Leach L, Zhang RM, Bradshaw JE, Kearsey MJ: **Constructing genetic linkage maps under a tetrasomic model.** *Genetics* 2006, **172**:2635-2635.
14. Wolfram S: *Mathematica: A System for doing Mathematics by Computer.* 5 edition. Addison-Wesley Publishing Company, Inc. USA; 2000.