The Trees in the Peaks

David Sankoff[1(✉)], Chunfang Zheng[1], Eric Lyons[2], and Haibao Tang[3]

[1] Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa K1N 6N5, Canada
{sankoff,czhen033}@uottawa.ca
[2] School of Plant Science, Bio5 Institute, University of Arizona, Tucson, AZ 85721, USA
ericlyons@email.arizona.edu
[3] Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou 350002, China
tanghaibao@gmail.com
http://albuquerque.bioinformatics.uottawa.ca

Abstract. We suggest a gene-tree/species-tree approach to speciation and whole genome duplication (WGD) to resolve the occurrence of these events in phylogenetic analysis. We propose a more principled way of estimating the parameters of gene divergence and fractionation than the standard mixture of normals analysis. We formulate an algorithm for resolving data on local peaks in the distributions of duplicate gene similarities for a number of related genomes. Illustrating with a comprehensive analysis of WGD-origin duplicate gene data from six members of the family Brassicaceae, we discuss the effects of variable evolutionary rates and data degradation due to fractionation. We introduce the notion of peak tree, as a template for all gene trees evolving by speciation, WGD and fractionation.

Keywords: Gene tree · Species tree · Whole genome duplication · Algorithms · Mixture of distributions · Brassicaceae

1 Introduction

The investigation of gene trees and species trees furnishes a genomic perspective on evolution insofar as it requires a complete inventory of the paralogs of the orthologously related genes in the species under study. This line of study also requires a different king of algorithm than those familiar from traditional single-gene based phylogenetics, or even the so-called "phylogenomics" based on large numbers of concatenated genes using what is basically traditional methodology. However, gene trees and species trees are each based on a tiny portion of the genome. In the context of whole genome duplication (WGD) in flowering plants, we can take the gene-tree/species tree approach to a more comprehensive kind of genomic data than the usual one-gene-at-a-time focus.

Specifically, we will study the set of $\binom{N}{2} + N$ gene similarity distributions within and across N species where WGD has affected one or more of these

M. Botón-Fernández et al. (Eds.): AlCoB 2016, LNBI 9702, pp. 3–14, 2016.
DOI: 10.1007/978-3-319-38827-4_1

species. This typically involves many thousands of genes. This paper raises more technical problems than it solves, but its goal is to show how concepts from gene-tree theory enable us to better understand genomic history.

We first sketch out a model of gene similarity distribution under random sequence divergence, speciation and fractionation, leading to a principled treatment of the statistical inference of divergence and fractionation rates and to speciation and WGD times.

Still lacking an implementation of this methodology, we can nonetheless proceed with our gene-tree approach by simply identifying local modes or "peaks" in all the similarity distributions, and translating these into phylogenetically related paralogous and orthologous entities. We present a rapid algorithm to resolve these in the case of ideal instances where no data are missing and all data are mutually compatible.

Finally, we illustrate our approach with six species spanning three genera of the family Brassicaceae.

2 Distributions of Gene Similarity

2.1 Background

We will discuss the distribution of similarities between homologous genes, according to a simple model that takes into account only

- gene mutation by random substitution of nucleotides independently at each position, and
- random duplicate gene loss after whole genome duplication (WGD).

Moreover, to simplify we treat all genes as having length l, i.e., l positions each containing one nucleotide.

After speciation, the genes in the two new species diverge independently according to a rate parameter λ. The simplest model for this divergence is based on binomial trials for change of nucleotide at each of the l positions of the gene. A success in the binomial trial at a position is the event that the nucleotide is the same at time t in both species. The similarity of the pair of orthologous genes at time t is binomially distributed $B[l, p(\lambda, t)]$, where $p(\lambda, t) = e^{-2\lambda t} + \frac{1}{4}(1 - e^{-2\lambda t})$. As time elapses, $p \to \frac{1}{4}$, so that the similarity between genes becomes indistinguishable from "noise", since $p = \frac{1}{4}$ is characteristic of pairs of random sequences.

Since we treat all n genes as having l nucleotides, the predicted frequency distribution of successes is given by $\Phi[n, \lambda, t] = nB[l, p]$. Inference under this model is simple. If the empirical frequency distribution of similarities (number of successes across l trials) is nF, where the mean of F is m, then $\hat{p} = \frac{m}{l}$ and $\hat{\lambda}t = -\frac{1}{2}\log\frac{4\frac{m}{l}-1}{3}$. If t is known, this gives us an estimate of the mutation rate constant λ, while if λ is known, this gives us an estimate of the divergence time t.

When a genome undergoes whole genome duplication (WGD), each gene is duplicated, creating one pair of "paralogous" genes. Over time, the frequency

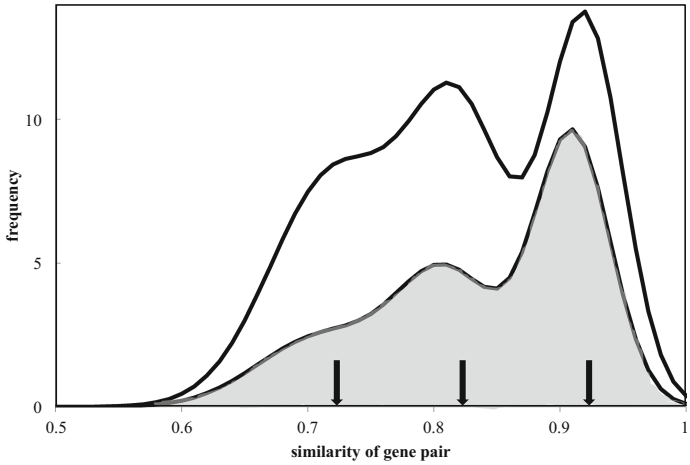

Fig. 1. Idealized gene similarity distribution between two species represented by three events, two WGD and a speciation. The arrows indicated $p = exp{-}\lambda t$ for the individual events. The number of gene pairs throughout is 4000. The upper curve represents the situation where the fractionation rate ρ is zero; the broadening of the part of the distribution reflecting the earliest event is due to increasing variance with greater age. The lower curve bounding the shaded area adds the effect of non-zero ρ so that the earliest event is increasingly hidden by subsequent events. The x-axis in this type of diagram is often shown in a log scale, scaling linearly with time, so that very early events appear much farther to the left.

distribution of the similarities between paralogous genes becomes $e^{-\rho t}\Phi[n, \lambda, t]$, where the rate parameter ρ accounts for process of losing (deleting or inactivating) one of each duplicate gene pair. Note that paralogous genes begin diverging at a duplication event, while orthologous genes begin diverging at a speciation event. Inference of ρ and λ, or of t, from frequency data is once again straightforward.

Speciation and WGD events can combine in any number of ways in the history of an evolutionary domain. For example, after speciation, one of the two sister genomes may undergoes WGD at time t_1, where $0 \leq t_1 \leq t$. Here, the similarities between the $2n$ pairs of homologs in the two species at time t, will be distributed as $2e^{-\rho[t-t_1]}\Phi(\lambda, t)$. Again there is no difficulty in inferring the parameters.

If however after WGD at time 0, a genome undergoes speciation at time t_1, where $0 \leq t_1 \leq t$, then the similarities between the $4n = 2 \times 2n$ pairs of homologs in the two species at time t, will be distributed as $2e^{-\rho t}[\Phi(\lambda, t) + \Phi(\lambda, t - t_1)]$. (This distribution is bi-modal when the parameters λ and ρ are suitably small with respect to t_1 and $t - t_1$.) In this model there is no closed form for the maximum likelihood estimators of the parameters or times. It is the usual practice to resort to numerical procedures embodied in software such as Emmix [10] for resolving mixtures of normal distributions.

Similarly, if after whole genome triplication (WGT), a genome undergoes speciation at time t_1, where $0 \leq t_1 \leq t$, the similarities between the $9n = 3 \times 3n$ pairs of homologs in the two species at time t, will be distributed as $3e^{-\rho t}[2\Phi(\lambda, t) + \Phi(\lambda, t - t_1)]$. The same kind of logic applies to a speciation after a higher degree polyploidization (whole genome quadruplication, etc.) These models all have the same inferential complications as the previous one. They can all also result in bimodal distributions, In the case a genome undergoes two successive WGD at time 0 and t_1, then a speciation at time t_2, so that $0 \leq t_1 \leq t_2 \leq t$, the $16n = 4 \times 4n$ pairs of homologs in the two species at time t, will be distributed as $4e^{-\rho t}[2\Phi(\lambda, t) + \Phi(\lambda, t - t_1) + \Phi(\lambda, t - t_2)]$. This can be a trimodal distribution.

It is important to note that many species in a phylogeny may be related by the same WGD and speciation events, and the times estimated for these events should be constrained to be equal. Other events should be constrained to occur in an order compatible with the phylogeny. Such constraints are not available in standard statistical mixtures of distributions software.

Our theoretical considerations pertain to the simple model assumed at the beginning of this section. In practice, various other processes affect the distribution of similarities so that the number of gene homologs between and within genomes may be severely reduced from those expected from the model. Within a group of related organisms, however, the parameter λ tends to have a constant value, although there are particular cases where it may be substantially lower [11] or higher [1]. The hypothesis of constant ρ has been investigated in [12].

The broadening of effects of event age and of fractionation on similarity distributions as time elapses are illustrated in Fig. 1. Eventually, all events become indistinguishable from noise caused by random gene resemblances, widespread domain sharing, tandem and near-tandem duplications, gene-order rearrangements, gene conversion and other processes.

It is important to note that methods like EMMIX, powerful and flexible as they may be, are not tailored to the problem of detecting speciation and WGD in a *set* of related similarity distributions. For any mixture of normals, EMMIX will identify these components as long as there is enough data. But not every mixture of normals credibly reflects some sequence of genomic events. More important, among the $\binom{N}{2} + N$ gene similarity distributions within and across N species, there are many constraints that are not handled by software packages, such as requiring $\hat{t}$ to be the same for an event in all the distributions that are affected by it.

Despite these problems with speciation- or WGD-event detection, in this paper, we will assume constant λ, and we will assume that we can infer p, and hence the age of an event, simply by identifying the mode, or "peak" of the similarity distribution, without recourse to other estimation procedures. This unfortunately foregoes any attempt at present to pick out events visible as "shoulders" of other events on the similarity distribution, but it will allow us to validate the notion that $\hat{t}$ should be the same for an event for all the distributions in which it plays a role.

2.2 From Peaks to Species Trees and Duplication Gene Trees

There are two observations underlying our method for reconstructing the species tree and "peaks tree" from a perfect set of inter- and intra-genome comparisons:

- each intra-genome distribution of similarities only has peaks due to all the WGD in its direct lineage, and
- each inter-genome distribution may contain many peaks due to WGDs, but only one peak due to speciation, i.e., at the date of the most recent common ancestor of the two species.

We use these principles one after the other to produce our results. The pseudocode below assumes a perfect set of inter- and intra-genome comparisons, namely that all events affecting a between-genome or within-genome comparison are detected by the kind of inferential statistics mentioned in Sect. 2, and these comparisons are found in all the genomes affected by the event, and only these, according to the above principles.

For an event i at time t_i, we write $(time, genome_1, genome_2)$. Each event time is associated with two genomes, which may be distinct or identical.

Algorithm 1. Construct the tree

Input: A set of genomes $G = \{g_1, g_2, \ldots, g_m\}$,
A set of event times $E = \{t_1, t_2, \ldots, t_n\}$,
Output: A speciation tree in Newick format with duplication nodes

for $i \leftarrow 1$ ***to*** m **do**
  get all the duplicate time(s) Dt_i for genome i by **Algorithm 2**
for $i \leftarrow 1$ ***to*** $(m-1)$ **do**
  for $j \leftarrow i+1$ ***to*** m **do**
    get the speciation time St for g_i and g_j by **Algorithm 3**
split G into G_{left} and G_{right} by **Algorithm 4**
while G_{left} *or* G_{right} *contains* ≥ 2 *genomes* **do**
  recursively apply **Algorithm 4** to G_{left} or G_{right}
return G *in the Newick format*

Algorithm 2. Get duplication time for a genome

Input: A genome g_i
A set of event times $E = \{t_1, t_2, \ldots, t_n\}$,
Output: duplication event(s) Dt_i for g_i

$Dt_i \leftarrow \emptyset$
for $j \leftarrow 1$ ***to*** n **do**
  if $t_j : genome1 = g_i$ *and* $t_j : genome2 = g_i$ **then**
    add $t_j : time$ to Dt_i
return Dt_i

Algorithm 3. Get speciation time for g_i and g_j

Input: Two genomes g_i and g_j,
Dt_i and Dt_j
A set of event times $E = \{t_1, t_2, \ldots, t_n\}$,
Output: Speciation time St for g_i and g_j

```
for k ← 1 to n do
    if t_k : genome1 = g_i and t_k : genome2 = g_j then
        if t_k : t ∉ Dt_i and t_k : t ∉ Dt_j then
            St : time ← t_k : t
            St : genome1 ← g_i
            St : genome2 ← g_j
return St
```

Algorithm 4. Split a group of genomes into two groups by a SpeciationNode

Input: A set of genomes ψ, can be G or subset of G
a set of speciation times $\{St_1, St_2, \ldots, St_r\}$, for all pairwise genomes in ψ
Output: A speciationNode and two subsets of ψ, ψ_{left} and ψ_{right}.
$\psi_{left} \cup \psi_{right} = \psi$

```
ψ_left ← ∅
ψ_right ← ∅
leftGenome = 0
rightGenome = 0
speciationNode = 0
duplicationNode = 0
for k ← 1 to r do
    if St_r : time > speciationNode then
        speciationNode = St_r : time
        leftGenome = St_r : genome1
        rightGenome = St_r : genome2
for all the duplication times for each genome in ψ do
    if ∃ a duplication time dt <speciationNode AND all the genomes in ψ
    have this duplication time then
        duplicationNode = dt for each genomes in ψ do
            remove dt from Dt of this genome
for k ← 1 to r do
    if St_r : time = speciationNode and
    St_r : genome1(/genome2) = leftGenome then
        Add St_r : genome2(/genome1) to ψ_right
    if St_r : time = speciationNode and
    St_r : genome1(/genome2) = rightGenome then
        Add St_r : genome2(/genome1) to ψ_left
return duplicationNode, speciationNode,ψ_left, ψ_right
```

3 The Brassicaceae

To illustrate our discussion, we draw on six published genomes in the Brassicaceae family, three in the genus *Brassica*: *B. rapa* (turnip, Chinese cabbage) [13], *B. oleracea* (cabbage, cauliflower) [7] and *Raphanus sativus* (radish) [6], two in the genus *Arabidopsis*: *A. lyrata* (rock cress) [3] and *A. thaliana* (thale cress, mouse-ear cress) [4] and one in the genus *Sisymbrium*: *S. irio* (London rocket) [2]. Figure 2 shows the phylogenetic relationship among the six species:

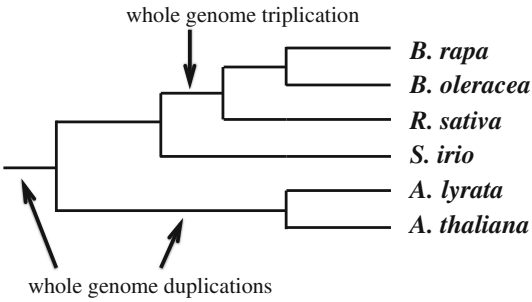

Fig. 2. Phylogenetic relationship of six species in the family Brassicaceae, showing lineages affected by WGD and WGT events.

We extracted genomic data from these species using the database in CoGe [8,9]. We then used the SynMap routine (with default parameters) on this platform to compare the gene orders of each of the $\binom{6}{2} = 15$ pairs of genomes. This procedure implicitly validates the identification of orthologs produced by speciation by detecting collinear arrays of several duplicate pairs in two species with approximately the same divergence: "syntenic blocks". Similarly, we did a self-comparison of five of the six genomes; the sixth one, the *Sisymbrium* genome, did not have enough closely spaced duplicate pairs for SynMap to produce paralogous syntenic blocks. The distributions of similarities calculated are shown in Fig. 3. The peaks found in each genome are tabulated in Table 1.

From Fig. 3 and Table 1, we note that the data are not quite "perfect"; the earliest duplication, detected at 79–80 % in the *Arabidopsis* self-comparisons, shows no peaks in the other self-comparisons – there is a shoulder or heavy tail in the appropriate place in the *Brassica* self-comparisons, but this is swamped by the later triplication. The triplication itself is visible in all three *Brassica* self-comparisons and in the comparison of *B. oleracea* and *B. rapa*, but not in the weaker signals involving *Raphanus*. Most of these missing data could be recovered using statistical means such as those discussed in Sect. 2 involving constraints instead of relying on identification of peaks.

More interesting is that the peaks at 90 % reflecting the *Sisymbrium* speciation, known to occur before the *Brassica* triplication, suggest that speciation is more recent, since the triplication peak is at 89 %. This apparent conflict is

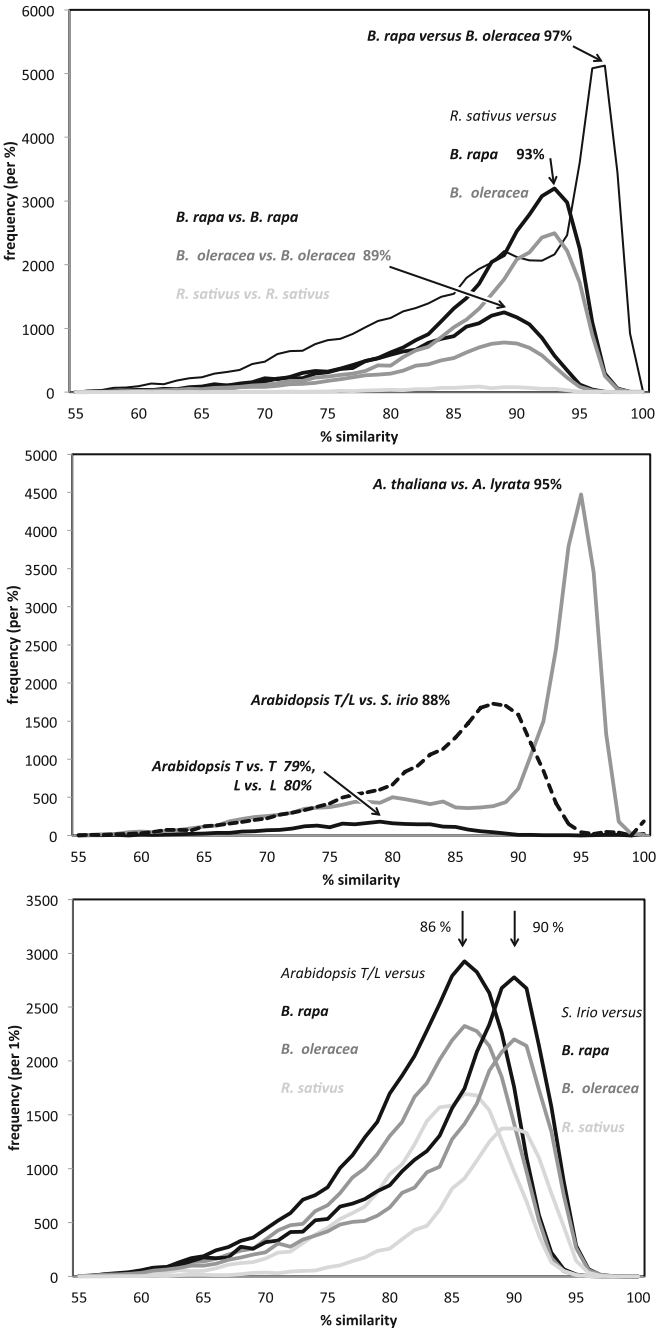

Fig. 3. Gene similarity distribution between 15 pairs of genomes in the Brassicaceae and 5 self comparisons. Local modes ("peaks") are indicated. Only one of each comparison is shown for *Arabidopsis*, the other is superimposed and indistinguishable.

Table 1. Peak similarity level, by genome. np: no peak, but one could be found by mixtures of distribution methods. - : no peak expected. Note peak 3 occurring before peak 4 due to slow evolutionary rate (λ) of *Sisymbrium*.

Peak number	Description	Genome					
		BR	BO	RS	SI	AL	AT
1	Alpha duplication [5]	np	np	np	np	80	80,79
2	Divergence of genus *Arabidopsis*	86	86	86,87	88	88-86	88-86
3	Whole genome triplication	89	89	87	-	-	-
4	Divergence of genus *Sisymbrium*	90	90	90	90	-	-
5	Divergence of genus *Raphanus*	93	93	93	-	-	-
6	Speciation of *Arabidopsis T & L*	-	-	-	-	95	95
7	Speciation of *B. rapa & B. oleracea*	97	97	-	-	-	-

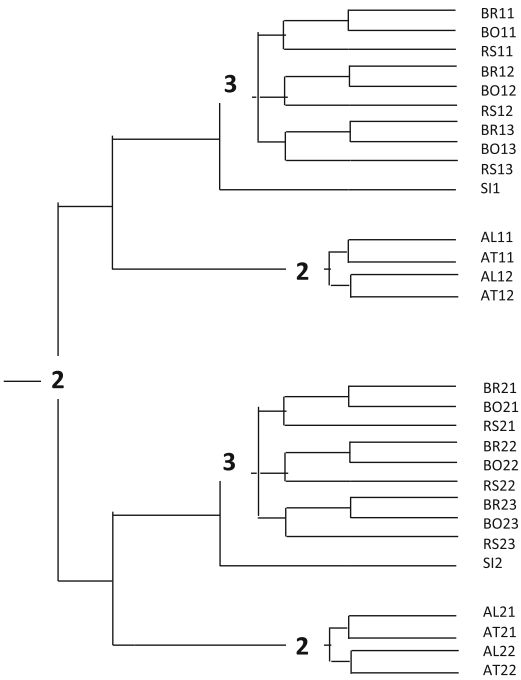

Fig. 4. Peak-tree for the Brassicaceae. Boldface numbers indicate WGD or triplication. All families of genes descended from the various genome WGD or WGT without any additional duplications must be formed from this tree with truncations of appropriate lineages.

clearly ascribable to a slower rate of evolution (lower λ), since the divergence of *Arabidopsis* from *Sisymbrium* also seems to occur more recently (88 %) than the divergence of *Arabidopsis* from the *Sisymbrium* sister genus *Brassica* (86 %). Note that the small differences between peak similarities are not insignificant, given the many thousands of gene pairs involved in these comparisons.

Were we to fill in the missing peaks, and correct the *Sisymbrium* times to account for slower evolution, the data set would be perfect and Algorithm 1 would convert it to a species tree with duplication times indicated. This could be then displayed in the form of Fig. 4. This "peaks tree" represents a general template for gene families evolving through WGD and fractionation-based gene loss only. The gene tree for any particular gene family would have exactly the same form, but with losses of various lineages.

4 Conclusion

We have pointed out connections between gene-tree/species-tree theory and the study of whole genome duplications in a phylogeny. The "peaks" tree should be a template for all the gene families proliferating through WGD and speciation only, where each gene family would simply require pruning of some of the branches of the tree, due to fractionation of duplicate genes. Despite the shortcomings of our Brassicaceae analysis, in the ideal case, the peaks tree itself would fill out the template completely, although no individual gene family is likely to be complete.

Our model and methodology is simplified. We have seen that λ may vary somewhat for individual lineages, and ρ is probably even more variable. Genomic processes such as chromosomal rearrangements disrupt gene order and degrade the recovery of synteny blocks and duplicate gene pairs. These issues should all be addressed in future work.

Our simplest DNA substitution model assumes equal base frequency and equal mutation rates. DNA substitution models with more parameters and rate variation among sites can be readily applied here. For example, one commonly used distance metric K_s (substitutions per synonymous sites) is typically calculated using more specific codon substitution models. The K_s distance scales linearly with time and $\log p$.

Despite the need for inference procedures focusing on the parameters λ and ρ (rather than μ and σ) jointly estimated for a compete set of similarity distributions among N genomes (rather than just one distribution), we have not yet implemented one, and have resorted to a primitive procedure of peak recognition in illustrating our model. Nevertheless, applying our concepts to six genomes in the family Brassicaceae illustrates the potential usefulness of our approach in understanding multiple WGD in a phylogenetic context.

Competing Interests
The authors declare that they have no competing interests.

Acknowledgements. Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics.

References

1. Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.M., Da Silva, C., Brinkmann, H., Mikhaleva, J., Olsen, L.C., Jubin, C., Cañestro, C., Bouquet, J.M., Danks, G., Poulain, J., Campsteijn, C., Adamski, M., Cross, I., Yadetie, F., Muffato, M., Louis, A., Butcher, S., Tsagkogeorga, G., Konrad, A., Singh, S., Jensen, M.F., Cong, E.H., Eikeseth-Otteraa, H., Noel, B., Anthouard, V., Porcel, B.M., Kachouri-Lafond, R., Nishino, A., Ugolini, M., Chourrout, P., Nishida, H., Aasland, R., Huzurbazar, S., Westhof, E., Delsuc, F., Lehrach, H., Reinhardt, R., Weissenbach, J., Roy, S.W., Artiguenave, F., Postlethwait, J.H., Manak, J.R., Thompson, E.M., Jaillon, O., Du Pasquier, L., Boudinot, P., Liberles, D.A., Volff, J.N., Philippe, H., Lenhard, B., Crollius, H.R., Wincker, P., Chourrout, D.: Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science **330**, 1381–1385 (2010)
2. Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., Dewar, K., Stinchcombe, J.R., Schoen, D.J., Wang, X., Schmutz, J., Town, C.D., Edger, P.P., Pires, J.C., Schumaker, K.S., Jarvis, D.E., Mandakova, T., Lysak, M.A., van den Bergh, E., Schranz, M.E., Harrison, P.M., Moses, A.M., Bureau, T.E., Wright, S.I., Blanchette, M.: An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. **45**, 891–898 (2013)
3. Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J.D., Ossowski, S., Ottilar, R.P., Salamov, A.A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M.E., Bergelson, J., Carrington, J.C., Gaut, B.S., Schmutz, J., Mayer, K.F.X., Van de Peer, Y., Grigoriev, I.V., Nordborg, M., Weigel, D., Guo, Y.L.: The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat. Genet. **43**, 476–481 (2011)
4. The Arabidopsis Genome Initiative: Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**, 796–815 (2000)
5. Kagale, S., Robinson, S.J., Nixon, J., Xiao, R., Huebert, T., Condie, J., Kessler, D., Clarke, W.E., Edger, P.P., Links, M.G., et al.: Polyploid evolution of the brassicaceae during the cenozoic era. Plant Cell **26**, 2777–2791 (2014)
6. Kitashiba, H., Li, F., Hirakawa, H., Kawanabe, T., Zou, Z., Hasegawa, Y., Tonosaki, K., Shirasawa, S., Fukushima, A., Yokoi, S., Takahata, Y., Kakizaki, T., Ishida, M., Okamoto, S., Sakamoto, K., Shirasawa, K., Tabata, S., Nishio, T.: Draft sequences of the radish (*Raphanus sativus* l.) genome. DNA Res. **21**(5), 481–490 (2014)
7. Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I.A.P., Zhao, M., Ma, J., Yu, J., Huang, S., Wang, X., Wang, J., Lu, K., Fang, Z., Bancroft, I., Yang, T.J., Hu, Q., Wang, X., Yue, Z., Li, H., Yang, L., Wu, J., Zhou, Q., Wang, W., King, G.J., Pires, J.C., Lu, C., Wu, Z., Sampath, P., Wang, Z., Guo, H., Pan, S., Yang, L., Min, J., Zhang, D., Jin, D., Li, W., Belcram, H., Tu, J., Guan, M., Qi, C., Du, D., Li, J., Jiang, L., Batley, J., Sharpe, A.G., Park, B.S., Ruperao, P., Cheng, F., Waminal, N.E., Huang, Y., Dong, C., Wang, L., Li, J., Hu, Z., Zhuang, M., Huang, Y., Huang, J., Shi, J., Mei, D., Liu, J., Lee, T.H., Wang, J., Jin, H., Li, Z., Li, X., Zhang, J., Xiao, L., Zhou, Y., Liu, Z., Liu, X., Qin, R., Tang, X., Liu, W., Wang, Y., Zhang, Y., Lee, J., Kim, H.H., Denoeud, F., Xu, X., Liang, X., Hua, W., Wang, X., Wang, J., Chalhoub, B., Paterson, A.H.: The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. **5**, 3930 (2014). http://dx.org/10.1038/ncomms4930

8. Lyons, E., Freeling, M.: How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. **53**, 661–673 (2008)
9. Lyons, E., Pedersen, B., Kane, J., Freeling, M.: The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates rosids. Trop. Plant Biol. **1**, 181–190 (2008)
10. McLachlan, G.J., Peel, D., Basford, K.E., Adams, P.: The Emmix software for the fitting of mixtures of normal and t-components. J. Stat. Softw. **4**(2), 1–14 (1999)
11. Ming, R., VanBuren, R., Liu, Y., Yang, M., Han, Y., Li, L.T., Zhang, Q., Kim, M.J., Schatz, M.C., Campbell, M., Li, J., Bowers, J.E., Tang, H., Lyons, E., Ferguson, A.A., Narzisi, G., Nelson, D.R., Blaby-Haas, C.E., Gschwend, A.R., Jiao, Y., Der, J.P., Zeng, F., Han, J., Min, X.J., Hudson, K.A., Singh, R., Grennan, A.K., Karpowicz, S.J., Watling, J.R., Ito, K., Robinson, S.A., Hudson, M.E., Yu, Q., Mockler, T.C., Carroll, A., Zheng, Y., Sunkar, R., Jia, R., Chen, N., Arro, J., Wai, C.M., Wafula, E., Spence, A., Han, Y., Xu, L., Zhang, J., Peery, R., Haus, M.J., Xiong, W., Walsh, J.A., Wu, J., Wang, M.L., Zhu, Y.J., Paull, R.E., Britt, A.B., Du, C., Downie, S.R., Schuler, M.A., Michael, T.P., Long, S.P., Ort, D.R., William Schopf, J., Gang, D.R., Jiang, N., Yandell, M., dePamphilis, C.W., Merchant, S.S., Paterson, A.H., Buchanan, B.B., Li, S., Shen-Miller, J.: Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biol. **14**(5), 1–11 (2013)
12. Sankoff, D., Zheng, C., Zhu, Q.: The collapse of gene complement following whole genome duplication. BMC Genomics **11**, 313–313 (2010)
13. Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J.C., Paterson, A.H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A.G., Park, B.S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G.J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I.A.P., Batley, J., Kim, J.S., Just, J., Li, J., Xu, J., Deng, J., Kim, J.A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M.G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P.J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S.J., Choi, S.R., Lee, T.H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z., Zhang, Z.: The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. **43**, 1035–1039 (2011)