Project Report
COMP 7934

A Workflow for Whole Genome Duplication Analysis and Phylogenetic Tree Construction

Muhammad Arsal Asif [1*]

[1] *University of Manitoba, MB, Canada*

Correspondence*:
asifma@myumanitoba.ca

ABSTRACT

Whole genome duplication holds critical significance in the history of evolution. We demonstrate the analysis and detection of WGD events using example workflows. The described workflows use available online tools and the code provided with this paper. Our workflow includes the comparison of syntenic blocks of genomes and the creation of the similarity distributions through them. We conclude with the construction of phylogenetic trees. We demonstrate our approach with genomes from the *Brassica* family.

Keywords: comparative genomics, synteny, whole genome duplication, wgd, phylogeny, similarity distribution, upgma

1 INTRODUCTION

Whole genome duplication is the event that duplicates an entire genome in an organism. The role of WGD in shaping the history of evolutionary lineages has been extensively studied [1]. WGD affects lineages of different classes of species such as parasites, fungus, plants, and animals. It is evident from analyses in the past [2] [3] [4] [5] [6] [7].

It is difficult to identify events of WGD as traces of WGD are erased with time [8]. Fractionation and mutation cause degradation in similarity [9]. Fractionation is the process of gradual gene loss following WGD and mutation comprises of random mutation and rearrangement patterns.

1.1 Similarity Distribution

Syntenic regions are two or more genomic regions from a common ancestral genomic region [10]. Colinear sets of homologous genes can be used to identify syntenic regions. The comparisons can be intra-genome or inter-genome.

Comparisons on syntenic blocks help identify the similarity between two genomes. The percentages of similarities against the number of compared pairs are used to construct the similarity distributions. These distributions are useful in the identification of whole genome duplication, whole genome triplication, and speciation events. It stems from the fact that the duplication events significantly increase the similarity in genomes and a decline in similarity follows them.

1.2 Related Work

Blanc and Wolfe outlined the difficulties in identification of the whole genome duplication events [8]. They introduced an approach using synonymous (K_s) and non-synonymous (K_n) nucleotide substitutions

affecting protein-coding sequences. Nucleotide substitution distances between each pair of genomes can be used to identify relative ages of duplication [9]. The relative ages can then be used to construct age distributions. Age distributions are analogous to similarity distributions. The difference lies in the usage of K_s and K_n distances instead of gene pair similarity.

The complication in using K_s and K_n substitutions for age distributions are emphasized by Vanneste et al. [11]. They simulated the evolution of duplicated sequences to estimate K_s values for inference of WGD events. Their results showed the appearance of artificial peaks when constructing age distributions based on K_s distances.

De Bodt et al. [3] built on the idea of "molecular clocks", or K_s substitution rate estimates [12] in the duplication analysis on Arabidopsis Thaliana genome [13]. They summarized the difficulties in selecting the right estimates for K_s values.

Most of the related work describes using K_s values for the construction of distributions. However, we consider similarity distributions as described in the previous section and ignore the K_s values for our analysis.

1.2.1 Contributions of the Selected Paper

Sankoff et al. [14] described a method to analyze WGD on a set of closely related species. The method involves using similarity distributions to identify WGD and speciation events. In addition, they proposed a phylogenetic algorithm derived from modification of the classic Neighbor-Joining (NJ) approach [15]. The newly proposed algorithm uses WGD and speciation event times to construct a rooted phylogenetic tree with identified whole genome duplications.

1.2.2 More from the Authors

Haibao Tang and Eric Lyons, two of the authors of the selected paper, analyzed genomes of the *Brassica* family [16] [17]. Their analysis is closely related to the one in the selected paper. It included the identification of whole genome triplication event in the selected genomes. They used the CoGe database [18], SynMap [19] and SynFind [20] tools in their analysis. CoGe and SynMap are used in this project as well. The methodology section contains the description of these tools.

1.3 Structure of This Work

In this paper, we present a workflow for the analysis of WGD and the construction of the phylogenetic tree. The outlined workflow is based on the idea described by Sankoff et al. [14]. The key concept is the usage of peaks of similarity in the similarity distributions to recognize WGD and speciation events.

The workflow is divided into multiple parts and numbered as such. Relevant sections contain their separate workflows. The first component of the methodology section outlines the process of collection and comparison of the genomic data. It further describes the generation of similarity distributions through gathered data. The second component describes the use of dot plots and analysis of the similarity distributions in the identification of WGD events. The third component illustrates the phylogenetic tree construction process. The included zip file contains the code required to supplement these workflows.

2 METHODOLOGY

2.1 Genomic Data

In this section, we explain the data collection process. The following sections describe the selected genomes and the tools required for genomic comparisons. Furthermore, we illustrate the usage of genomic comparisons in the generation of similarity distributions.

2.1.1 Selected Genomes

Two sets of genomic data from the *Brassica* family were selected to demonstrate the analysis. Phylogenetic trees from both sets were constructed and shown in section 2.3.

The first set consists of *Brassica Oleracea* (cabbage) [21], *Brassica Rapa* (turnip) [22], *Raphanus Raphanistrum* [23], and *Raphanus Sativus* [24] genomes.

The second set consists of *Arabidopsis Lyrata* [25], *Arabidopsis Thaliana* [13], *Capsella Rubella* [26], and *Leavenworthia Alabamica* [16].

Additionally, the code in this project was also tested to analyze WGD and the resulting phylogeny on a subset of genomes from *Olea Europaea* family. The phylogeny of *Olea Europaea* was described by Julca et al. [27]. However, due to the length of this paper, the analysis is excluded from this report.

Files of all genomic comparisons, including self-comparisons, are provided with the code file. File names contain the full names of genomes. The first names are automatically contracted by code in the output figures. The format for naming files is as follows: $FirstnameLastname_{genome1}\_FirstnameLastname_{genome2}$. An example of a file name is: $BrassicaRapa\_BrassicaOleracea$.

2.1.2 CoGe

Comparative Genomics (CoGe) database [18] is a platform for managing and analyzing genomic data. CoGe's tool can be used to analyze both public and privately available data. Currently, CoGe has over 48,000 publicly listed genomes. It is important to note, however, that the same genomes are often replicated on CoGe, as they are added by different people and through different sources. All tools on the CoGe website are open-source, and their codes are on GitHub [28].

2.1.3 SynMap

SynMap is a one-stop tool that can be used to perform comparisons of syntenic regions between two genomes. It offers many adjustable parameters. It uses cloud-computation for efficiency. It is built upon different algorithms such as BLAST [29], DAGChainer [30], and CodeML [31]. For a complete overview of SynMap and other CoGe tools, the reader is advised to view the tutorials by Castillo et al. [32].

2.1.4 Workflow 1: Using Synmap to Retrieve Genomic Comparison Data

We describe our first workflow using CoGe and SynMap to perform genomic comparisons and generate the input files for the given code. Our workflow assumes that all the required genomes are publicly available on SynMap.

1. Access the SynMap tool through CoGe: https://genomevolution.org/coge/SynMap.pl
2. Select the two genomes as "Organism 1" and "Organism 2" (in case of self-comparisons, select same genome in both sections).

3. If SynMap displays the message "No Coding Sequence in Genome", proceed with either of the following:
   - Select the drop-down of "Genomes" and choose a different instance of the same genome.
   - Find another copy of the same genome listed as a different organism.
4. Click on "Generate SynMap". The comparisons performed by SynMap are computationally extensive. It can take a few hours on genomes with a large number of chromosomes.
5. SynMap displays the status of each underlying process that it runs.
6. After the comparison is complete, SynMap displays a dot plot, along with various other options. It includes the option to download output files.
7. Click on "click to view options..." on the right side of "Download Results".
8. Click on "DAGChainer Output" under "Results".
9. Save the file under "genome_comparisons" folder on your hard drive where the code is located. Rename the file in a format described in section 2.1.1.
10. Download the image file by clicking on "Image file" under "General" for dot plot.
11. Repeat these steps for $\binom{N}{2}$ pair of genomes and N self-comparisons. N is the total number of genomes.

2.1.5 Workflow 2: Generation of Similarity Distribution

The second workflow describes the generation of similarity distributions using the code provided. The input files are genomic comparison files saved from SynMap by following the workflow 1 in section 2.1.4.

1. Install Anaconda with Jupyter IPython [33]. The necessary steps for installation of the required Python packages are well-documented in code.
2. Open the IPython notebook.
3. In cell 3 of the IPython notebook, add the names of genomes in the **genomes** array. Note that full-names of genomes are required.
4. Run cells 1-4 to get the output similarity distributions.
5. All genomic comparisons, including self-comparisons, are plotted in the figure. For an example of output figure, refer to section 2.2.2 in this paper.
   - Tip 1: Hover over any point in the output figure for more details in the plotted figure.
   - Tip 2: Click on any labels shown on the right side of the figure to hide the plot corresponding to that label.

2.2 Analysis of WGD

In this section, we analyze the similarity distributions and syntenic dot plots. As noted before, we illustrate these analyses using a subset of selected genomes.

2.2.1 Syntenic Dot Plot

We demonstrate how to identify WGD and speciation events using syntenic dot plots. Syntenic dot plots of a subset of selected genomes are presented here.

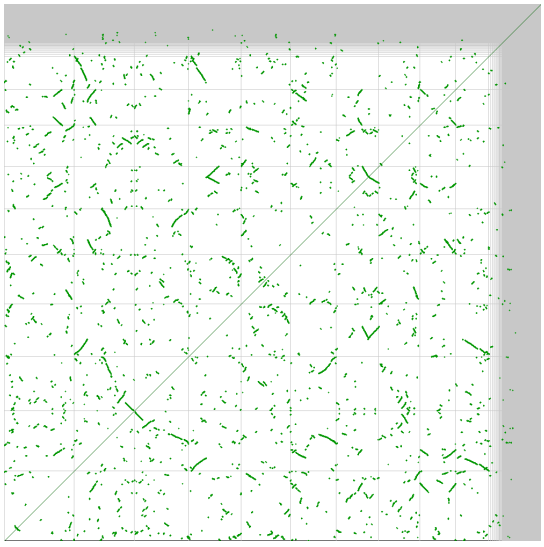

Figure 1. Dot Plot: Self-comparison of Brassica Rapa.

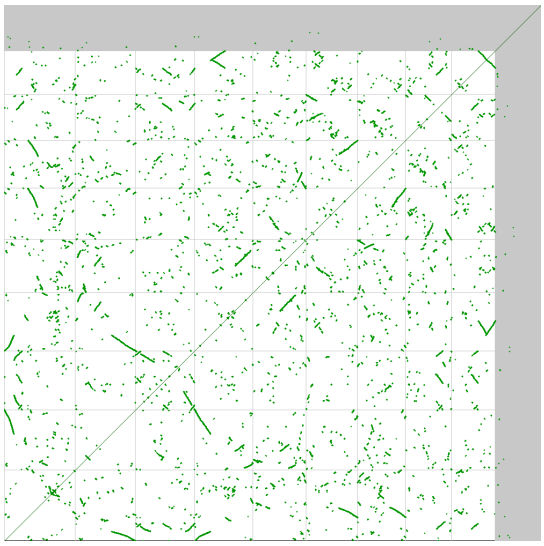

Figure 2. Dot Plot: Self-comparison of Brassica Oleracea.

Figures 1 and 2 show the syntenic dot plots of self-comparisons of the two genomes. The thick green lines represent a WGD event. The full diagonal line represents the perfect match.

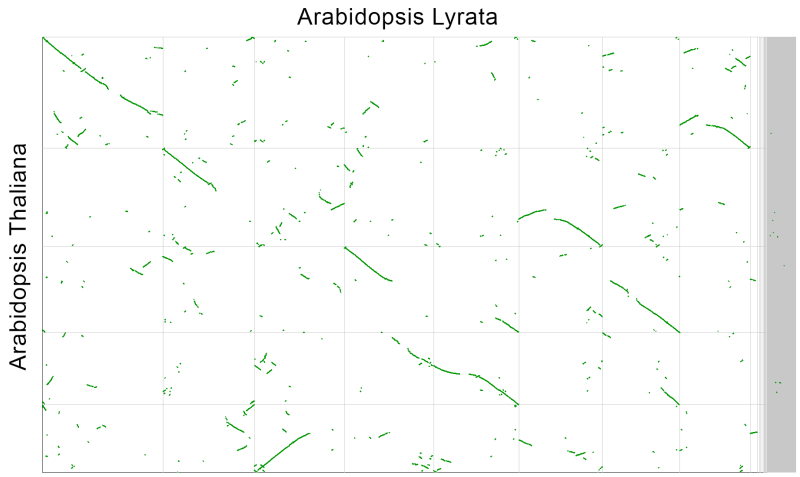

Figure 3. Dot Plot: Comparison of Arabidopsis Thaliana and Arabidopsis Lyrata.

Figure 3 shows the syntenic dot plot of comparison between *Arabidopsis Thaliana* and *Arabidopsis Lyrata*. The thick longer green lines represent the speciation event between the two genomes. The small green lines represent the WGD event shared by these genomes.

2.2.2 Similarity Distributions

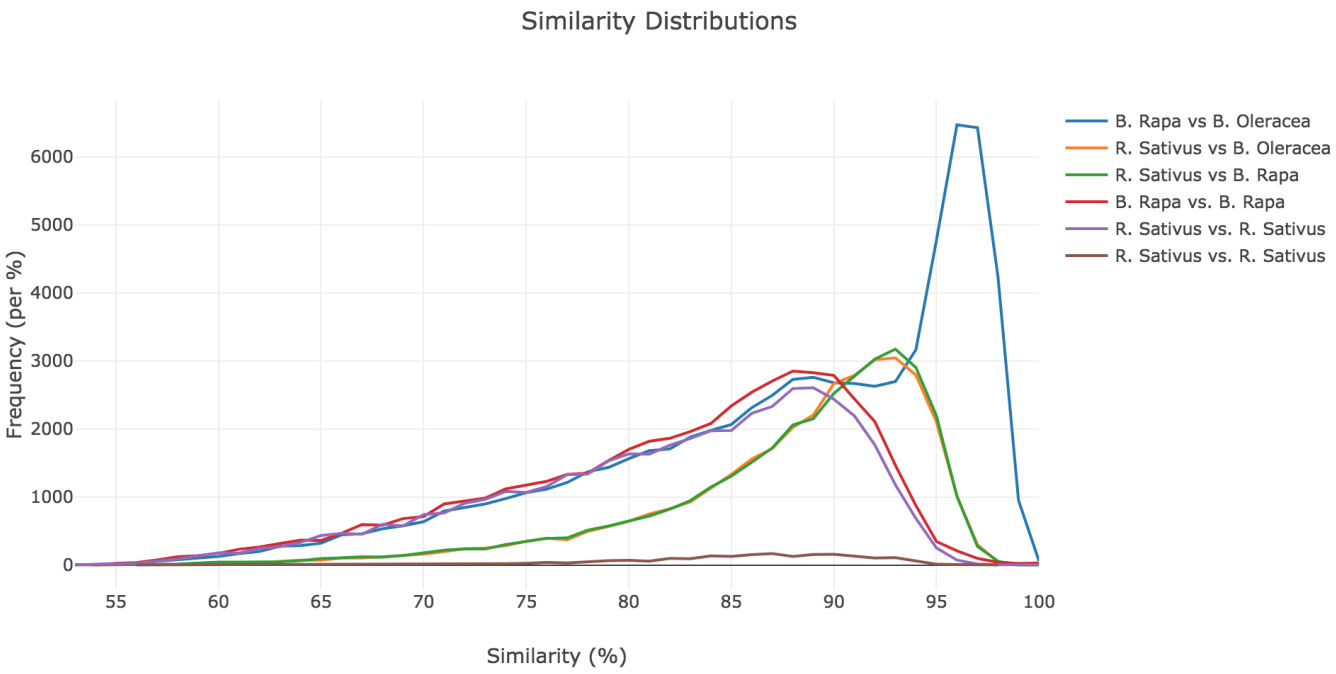

Figure 4. Gene Similarity Distributions: Brassica Rapa, Brassica Oleracea, and Raphanus Sativus genomes.

The genomes *B. Oleracea*, *B. Rapa* and, *R. Sativus* were selected for demonstration of the similarity distributions. Figure 4 shows the similarity distributions of these genomes. The figure closely resembles the one in the selected paper [14]. As described before, we base our analysis on using peaks of similarity from these distributions. The peaks help us identify whole genome duplication and speciation events. A speciation peak follows the peaks of WGD. The most recent peak in a similarity distribution shows the speciation event.

The interface in the IPython notebook can be utilized to see the exact percentages of these peaks. By hovering over any point in the figure, we can see the percentages. A peak at 96% clearly shows the speciation of *B. Rapa* and *B. Oleracea*. The peak at 93% represents the divergence of *Raphanus* genome. At 88%, a peak in self-comparisons of *B. Rapa*, *B. Oleracea*, and *R. Sativus* represent a whole genome tripling. We also saw this duplication event in dot plots of the self-comparisons of *Brassica* genomes. The following WGD and speciation events are evident from the analysis by Sankoff et al. [14].

2.3 Phylogenetic Tree

In this section, we describe the algorithms, and the workflow used to build the phylogenetic tree.

2.3.1 Neighbor Joining (NJ)

The maximum parsimony principle is always to use the simplest solution. In the case of phylogeny, the maximum parsimony problem is to find a phylogenetic tree that is a result of the minimum number of evolutionary events. Parsimony is one of the most extensively studied and used method for phylogeny [34] [35].

Neighbor-joining (NJ) is a phylogenetic tree construction method. It is based on the principle of maximum parsimony and as such, aims to find pairs of neighbors (operational taxonomic units [36]) that minimize total branch length.

In the scope of the phylogeny, species are defined as OTUs. Neighbors are any pair of species connected through a single internal node in an unrooted tree. For N > 3, number of pairs is at least 2 and at most $N/2$ (when N is even) or $(N-1)/2$ (when N is odd).

The resulting output of the algorithm is an unrooted tree topology describing a phylogeny [15].

2.3.2 Hierarchical Clustering (UPGMA)

Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [37] is an agglomerative clustering algorithm. It is one of the most popularly used clustering methods in comparative genomics [38]. UPGMA uses a distance matrix to cluster each point. In the case of the phylogenetic tree construction, UPGMA is utilized to construct rooted phylogenetic trees. For that purpose, UPGMA clusters a set of genomes. The distance matrix contains pairwise distances between each pair of genomes.

2.3.3 Workflow 3: Phylogenetic Tree Construction

As mentioned before, Sankoff et al. [14] proposed a new algorithm for phylogenetic tree construction. However, in this project, we used the classic NJ approach to generate unrooted trees and the UPGMA method for rooted trees.

Biopython library is used for phylogenetic tree construction [39]. It contains implementations of the NJ and the UPGMA methods. For the distance matrix for both methods, the library uses DistanceMatrix class [40]. It implements a 2d matrix in a lower triangular format. High similarity between two genomes equals less distance between them. As such, the inverse of percentages of peak similarities from the similarity distributions were used as distances between each pair of genomes.

The third workflow describes the construction of phylogenetic trees using the code provided. The steps outlined in the workflow 1 in section 2.1.4 and the workflow 2 in section 2.1.5 must be completed before this workflow.

1. Follow the steps in workflow 1 in section 2.1.4.
2. Follow the steps in workflow 2 in section 2.1.5.
3. Open the IPython notebook.
4. Click on ”Cell” in the menu bar. Click on ”Run All”.
5. The output phylogenetic trees from both algorithms can be seen in cell 5.

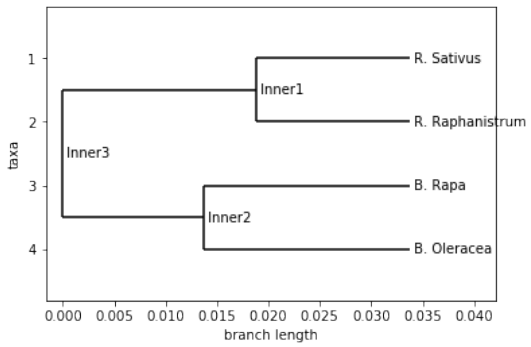

Figure 5. UPGMA method: Phylogeny of Brassica Rapa, Brassica Oleracea, Raphanus Raphanistrum, and Raphanus Sativus genomes.

Figure 5 shows the output using the UPGMA method. It is a rooted tree topology. Names of the inner nodes are automatically assigned. The tree is akin to the one generated by the selected paper's method [14]. As identified before in the section 2.2.2, we can see a WGD event at node Inner3.

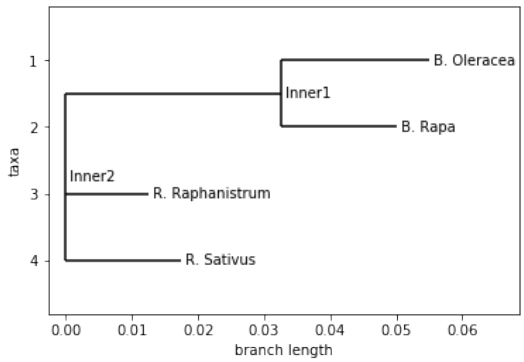

Figure 6. NJ method: Phylogeny of Brassica Rapa, Brassica Oleracea, Raphanus Raphanistrum, and Raphanus Sativus genomes.

Figure 6 shows the output using the NJ method. It is an unrooted tree topology. In comparison to the generated rooted tree, the unrooted tree is harder to understand. However, we can see a WGD event at node Inner2 in this figure.

3 CONCLUSION

We have devised a workflow for the WGD analysis and phylogenetic tree construction. We used available tools to compare syntenic blocks of genomes and retrieved the resultant files and dot plots. Furthermore, we described a workflow that can be followed to create similarity distributions. We showed how to use the dot plots and the created distributions for WGD and speciation analysis. Lastly, we generated rooted and unrooted phylogenetic trees using the peaks of similarities as pairwise distances between genomes.

Our analysis included genomes of the *Brassica* family. The genomes were a subset of the ones presented in the selected paper [14]. We showed how our designed workflows and code achieved the same phylogenetic tree and similarity distributions as the ones in the selected paper.

Different methods result in different phylogenies [41]. Our method generates one phylogeny out of the many possibilities. We demonstrated with a phylogenetic tree which is closely related to the one in literature [16]. However, on a different set of genomes, it is possible that it might result in a tree completely different from the one in literature. For future work, a different set of genomes, from a different genome family, can be analyzed using the described workflow. Distantly related genomes, such as genomes from different families can also be analyzed. Usage of synonymous (K_s) and non-synonymous (K_n) distance is popular in the literature. As such, future experiments can include these distances.

Our approach has several limitations, and they remain as open problems. In reality, various other factors can affect similarity between two genomes. However, our analysis follows a simple approach which does not consider these factors. The identification and inclusion of all the factors affecting similarity is a challenging task. Our approach also requires genomes with protein coding sequence data. However, in practice, it is not always applicable. For instance, we tried to compare the *Saccharomyces bayanus* (yeast) [42] genome from the National Center for Biotechnology Information (NCBI) database [43]. SynMap could not find any coding sequence data.

REFERENCES

[1]Susumu Ohno, Ulrich Wolf, and Niels B Atkin. Evolution from fish to mammals by gene duplication. *Hereditas*, 59(1):169–187, 1968.

[2]Ellen S Martinsen, Susan L Perkins, and Jos J Schall. A three-genome phylogeny of malaria parasites (plasmodium and closely related genera): evolution of life-history traits and host switches. *Molecular phylogenetics and evolution*, 47(1):261–273, 2008.

[3]Stefanie De Bodt, Steven Maere, and Yves Van de Peer. Genome duplication and the origin of angiosperms. *Trends in ecology & evolution*, 20(11):591–597, 2005.

[4]Dharmendar Rathore, Allison M Wahl, Margery Sullivan, and Thomas F McCutchan. A phylogenetic comparison of gene trees constructed from plastid, mitochondrial and genomic dna of plasmodium species. *Molecular and biochemical parasitology*, 114(1):89–94, 2001.

[5]Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617, 2004.

[6]Todd J Vision, Daniel G Brown, and Steven D Tanksley. The origins of genomic duplications in arabidopsis. *Science*, 290(5499):2114–2117, 2000.

[7]Frédéric G Brunet, Hugues Roest Crollius, Mathilde Paris, Jean-Marc Aury, Patricia Gibert, Olivier Jaillon, Vincent Laudet, and Marc Robinson-Rechavi. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular biology and evolution*, 23(9):1808–1816, 2006.

[8]Guillaume Blanc and Kenneth H Wolfe. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The plant cell*, 16(7):1667–1678, 2004.

[9]Michael Lynch and John S Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[10]Asher Haug-Baltzell, Sean A Stephens, Sean Davey, Carlos E Scheidegger, and Eric Lyons. Synmap2 and synmap3d: web-based whole-genome synteny browsers. *Bioinformatics*, 33(14):2197–2198, 2017.

[11]Kevin Vanneste, Yves Van de Peer, and Steven Maere. Inference of genome duplications from age distributions revisited. *Molecular biology and evolution*, 30(1):177–190, 2012.

[12]Marcus A Koch, Bernhard Haubold, and Thomas Mitchell-Olds. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in arabidopsis, arabis, and related genera (brassicaceae). *Molecular biology and evolution*, 17(10):1483–1498, 2000.

[13]Arabidopsis Genome Initiative et al. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *nature*, 408(6814):796, 2000.

[14]David Sankoff, Chunfang Zheng, Yue Zhang, Joao Meidanis, Eric Lyons, and Haibao Tang. Models for similarity distributions of syntenic homologs and applications to phylogenomics. *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[15]Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[16]A. Haudry, A. E. Platts, E. Vello, D. R. Hoen, M. Leclercq, R. J. Williamson, E. Forczek, Z. Joly-Lopez, J. G. Steffen, K. M. Hazzouri, K. Dewar, J. R. Stinchcombe, D. J. Schoen, X. Wang, J. Schmutz, C. D. Town, P. P. Edger, J. C. Pires, K. S. Schumaker, D. E. Jarvis, T. Mandakova, M. A. Lysak, E. van den Bergh, M. E. Schranz, P. M. Harrison, A. M. Moses, T. E. Bureau, S. I. Wright, and M. Blanchette. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet*, 45:891–898, 2013.

[17]Haibao Tang and Eric Lyons. Unleashing the genome of brassica rapa. *Frontiers in plant science*, 3:172, 2012.

[18]CoGe: Comparative Genomics. `https://genomevolution.org`. Accessed: 2018-12-14.

[19]CoGe: SynMap. `https://genomevolution.org/CoGe/SynMap.pl`. Accessed: 2018-12-14.

[20]CoGe: SynFind. `https://genomevolution.org/coge/SynFind.pl`. Accessed: 2018-12-14.

[21]Shengyi Liu, Yumei Liu, Xinhua Yang, Chaobo Tong, David Edwards, Isobel AP Parkin, Meixia Zhao, Jianxin Ma, Jingyin Yu, Shunmou Huang, et al. The brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature communications*, 5:3930, 2014.

[22]Xiaowu Wang, Hanzhong Wang, Jun Wang, Rifei Sun, Jian Wu, Shengyi Liu, Yinqi Bai, Jeong-Hwan Mun, Ian Bancroft, Feng Cheng, et al. The genome of the mesopolyploid crop species brassica rapa. *Nature genetics*, 43(10):1035, 2011.

[23]Gaurav D Moghe, David E Hufnagel, Haibao Tang, Yongli Xiao, Ian Dworkin, Christopher D Town, Jeffrey K Conner, and Shin-Han Shiu. Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish raphanus raphanistrum and three other brassicaceae species. *The Plant Cell*, pages tpc–114, 2014.

[24]HIROYASU Kitashiba, FENG Li, HIDEKI Hirakawa, TAKAHIRO Kawanabe, ZHONGWEI Zou, YOICHI Hasegawa, Kaoru Tonosaki, Sachiko Shirasawa, Aki Fukushima, Shuji Yokoi, et al. Draft sequences of the radish (raphanus sativus l.) genome. *DNA research*, 21(5):481–490, 2014.

[25]Tina T Hu, Pedro Pattyn, Erica G Bakker, Jun Cao, Jan-Fang Cheng, Richard M Clark, Noah Fahlgren, Jeffrey A Fawcett, Jane Grimwood, Heidrun Gundlach, et al. The arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics*, 43(5):476, 2011.

[26]Tanja Slotte, Khaled M Hazzouri, J Arvid Ågren, Daniel Koenig, Florian Maumus, Ya-Long Guo, Kim Steige, Adrian E Platts, Juan S Escobar, L Killian Newman, et al. The capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nature genetics*, 45(7):831, 2013.

[27]Irene Julca, Marina Marcet-Houben, Pablo Vargas, and Toni Gabaldon. Phylogenomics of the olive tree (olea europaea) disentangles ancient allo-and autopolyploidizations in lamiales. *bioRxiv*, page 163063, 2017.

[28]GitHub. `https://github.com`. Accessed: 2018-12-14.

[29]Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[30]Brian J Haas, Arthur L Delcher, Jennifer R Wortman, and Steven L Salzberg. Dagchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20(18):3643–3646, 2004.

[31]Ziheng Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

[32]Andreina I Castillo, Andrew DL Nelson, Asher K Haug-Baltzell, and Eric Lyons. A tutorial of diverse genome analysis tools found in the coge web-platform using plasmodium spp. as a model. *Database*, 2018, 2018.

[33]Anaconda Python . `https://www.anaconda.com`. Accessed: 2018-12-14.

[34]Michael J Sanderson, BG Baldwin, G Bharathan, CS Campbell, C Von Dohlen, D Ferguson, JM Porter, MF Wojciechowski, and MJ Donoghue. The growth of phylogenetic information and the need for a phylogenetic data base. *Systematic Biology*, 42(4):562–568, 1993.

[35]Luay Nakhleh, Guohua Jin, Fengmei Zhao, and John Mellor-Crummey. Reconstructing phylogenetic networks using maximum parsimony. In *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*, pages 93–102. IEEE, 2005.

[36]CoGe: SynFind. `https://en.wikipedia.org/wiki/Operational_taxonomic_unit`. Accessed: 2018-12-14.

[37]Peter HA Sneath, Robert R Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

[38]Zhian N Kamvar, Jonah C Brooks, and Niklaus J Grünwald. Novel r tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, 6:208, 2015.

[39]Biopython. `https://biopython.org`. Accessed: 2018-12-14.

[40]Biopython: Distance Matrix . `http://biopython.org/DIST/docs/api/Bio.Phylo.TreeConstruction.DistanceMatrix-class.html`. Accessed: 2018-12-14.

[41]David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147, 1981.

[42]GI Naumov. Saccharomyces bayanus var. uvarum comb, nov., a new variety established by genetic analysis. *Microbiology*, 69(3):338–342, 2000.

[43]National Center for Biotechnology Information. `https://www.ncbi.nlm.nih.gov`. Accessed: 2018-12-14.