bio-101

Likelihood-based Modeling of Brassiceae Ancient Whole-genome Triplication with POInT

Yue Hao[1, *] and Gavin C. Conant[2, 3, 4]

[1]Translational Genomics Research Institute (TGen), Phoenix, AZ, USA
[2]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA
[3]Program in Genetics North Carolina State University, Raleigh, NC, USA
[4]Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA
*For correspondence: yhao@tgen.org

Abstract

Ancient whole-genome duplication and triplication events have a profound impact on present-day plant genomes. However, sub genome assignments in allopolyploids can be technically challenging. The present study used a likelihood-based tool, POInT (the Polyploid Orthology Inference Tool) to model the resolution of the ancient whole-genome triplication events in Brassiceae. This protocol will showcase the application of POInT for polyploid comparative genomics studies and its potential use for future studies.

Keywords: Allopolyploidy, Whole-genome triplication, Biased fractionation, Synteny, Orthology, Brassiceae

Background

Ancient polyploidy events are prevalent throughout the evolutionary history of flowering plants (One Thousand Plant Transcriptomes Initiative, 2019) and have contributed to the diversity of complexity of present-day species (van de Peer et al., 2009). When the sub genomes were derived from genetically distant parents, such events are termed as allopolyploidy. Post-polyploidy gene loss in allopolyploids are often unbalanced across different sub genomes (Cheng et al., 2012). The biased fractionation reshaped genomic landscape, and the gene retention/loss pattern has important implications in crop development (Qi et al., 2021). However, sub genome assignment in allopolyploids is still challenging (Edger et al., 2018).
POInT (the Polyploid Orthology Inference Tool) is a likelihood-based tool for modeling ancient polyploidy events (Conant and Wolfe, 2008; Emery et al., 2018; Hao and Conant, 2022; Conant, 2023). POInT supports user-defined ploidy levels, namely whole-genome duplication (Conant and Wolfe, 2008; Emery et al., 2018; Conant, 2020), triplication (Schoonmaker et al., 2020; Hao et al., 2021), or quadruplication (Hao et al., 2022). Using POInT, we have modeled the whole-genome triplication (WGT) events shared by the members of the tribe Brassiceae (Hao et al., 2021). The synteny-based and phylogenetically aware analysis of gene loss across different sub genomes provides statistical support for the two-step hypothesis of hexaploidy and biased fractionation across the LF, MF1, and MF2 sub genomes after the ancient WGT (Cheng et al., 2012; Tang et al., 2012). Here, we present a step-by-step guide to recreate the WGT model comparisons in Hao et al. (2021).

Software and datasets

Software

1. POInT (Conant and Wolfe, 2008; Emery et al., 2018; Hao and Conant, 2022; version 1.55; http://conantlab.org/POInT/POInT.html)

Input data

1. The input files for this case study are available at https://doi.org/10.6084/m9.figshare.12277832 (Hao et al., 2021).
2. Input files include:
   a. Gene orders from four individual polyploid genomes:
      i. Brassica_oleracea_POInT_geneorders.txt
      ii. Brassica_rapa_POInT_geneorders.txt
      iii. Crambe_hispanica_v3_POInT_geneorders.txt
      iv. Sinapis_alba_POInT_geneorders.txt
   b. Inferred ancestral pillar order: FourSpp_M2Opt3.txt
   c. Optimal phylogenetic topology: BrBoSaCh_WGT_3rate_G1Dom_M2Opt3_Top3.tre
   d. WGT models for use with POInT:
      - WGT Null: WGT_Null_model.txt
      - WGT 1Dom: WGT_2rate_G1Dom_model.txt
      - WGT 1Dom$_{G3}$: WGT_3rate_G1Dom_model.txt
      - Root-spec. WGT 1Dom$_{G3}$: WGT_3rate_G1Dom_brnspec_model.txt
      - Root model: WGT_RootModel.txt

Procedure

A. Software download and installation

The latest version of the POInT software is available on GitHub (https://github.com/gconant0/POInT) or on the lab website (http://conantlab.org/POInT/POInT.html). For a detailed installation guide, please go to the software installation page (http://conantlab.org/POInT/INSTALL). To download and compile the software:

```
wget http://conantlab.org/POInT/POInT.tar
tar xvf POInT.tar
cd POInT
# To compile the OpenMP parallel version
./configure.pl -omp
Make
```

Create a symbolic link in the home directory bin:

```
# User should check ~/.bash_profile to see how PATH was defined.
# Add "export PATH=$PATH:$HOME/bin" if not already.
cd ~/bin
ln -s /where_POInT_was_downloaded/POInT/POInT
```

Test to see if the software was successfully compiled:

```
cd
POInT
```

You should see the following:

```
Using 16 threads for this run
Usage: POInT -g:<genome file> -g:<genome file> -o:<ortholog file> -
m:<Model file> (-r:<Root model file>) (-t:treefile) (-
p:<posteriortrackprobs file>) (-c:<conditional probabilities file> (-
no_opt) (-s:<start>:<end>) (-zerolengthfixed) (-x:#TreestoSave)
```

B. Basic usage

A basic POInT run requires a minimum of three types of files and a specification of the polyploidy type:
-d: 2 for WGD, 3 for a WGT, and 4 for an octaploidy.
-g: genome files, listing the order of double/triple/quadruple conserved synteny genes in the extant genomes.
-o: order file, an inferred ancestral order of the conserved synteny blocks.
-m: model file, a phylogenetic model of post-polyploidy gene loss.
Additionally, it is very useful to specify an assumed phylogenetic topology using -t. If no tree file is given, POInT will try to search all possible topologies. However, for larger datasets with five or six genomes, this step will be intractable. Here, the optimal topology for the four species in the case study is provided.

C. Testing different WGT models

The analyses showcased here were performed on the Linux system. Twenty-four cores were requested when submitting the job to the high-performance computing cluster. It takes approximately 17 min to complete a

single iteration of likelihood estimation using 96 threads on a 24-core node, and several hours or longer to reach the optimum.

```
export OMP_NUM_THREADS=96
```

1. WGT Null model scenario: Null model with no biased fractionation.

```
POInT -d:3 -g:Brassica_rapa_POInT_geneorders.txt /
-g:Brassica_oleracea_POInT_geneorders.txt /
-g:Sinapis_alba_POInT_geneorders.txt /
-g:Crambe_hispanica_v3_POInT_geneorders.txt -o:FourSpp_M2Opt3.txt /
-m:WGT_Null_model.txt /
-t: BrBoSaCh_WGT_3rate_G1Dom_M2Opt3_Top3.tre
```

2. WGT 1Dom model scenario: MF1 and MF2 sub genomes are more fractionated than the LF sub genome, but the fractionation rates for MF1 and MF2 are the same.

```
POInT -d:3 -g:Brassica_rapa_POInT_geneorders.txt /
-g:Brassica_oleracea_POInT_geneorders.txt /
-g:Sinapis_alba_POInT_geneorders.txt /
-g:Crambe_hispanica_v3_POInT_geneorders.txt -o:FourSpp_M2Opt3.txt /
-m:WGT_2rate_G1Dom_model.txt /
-t: BrBoSaCh_WGT_3rate_G1Dom_M2Opt3_Top3.tre
```

3. WGT 1Dom$_{G3}$ model scenario: MF2 is more fractionated than MF1, and MF1 is more fractionated than LF.

```
POInT -d:3 -g:Brassica_rapa_POInT_geneorders.txt /
-g:Brassica_oleracea_POInT_geneorders.txt /
-g:Sinapis_alba_POInT_geneorders.txt /
-g:Crambe_hispanica_v3_POInT_geneorders.txt -o:FourSpp_M2Opt3.txt /
-m:WGT_3rate_G1Dom_model.txt /
-t: BrBoSaCh_WGT_3rate_G1Dom_M2Opt3_Top3.tre
```

4. Root-spec. WGT 1Dom$_{G3}$ is similar to model 3, but with two sets of parameters: one for the root branch and the other for the remainder of the branches in the phylogenetic tree, modeling the scenario of shifted fractionation rates from root branch to later branches.

```
POInT -d:3 -g:Brassica_rapa_POInT_geneorders.txt /
-g:Brassica_oleracea_POInT_geneorders.txt /
-g:Sinapis_alba_POInT_geneorders.txt /
-g:Crambe_hispanica_v3_POInT_geneorders.txt -o:FourSpp_M2Opt3.txt /
-m:WGT_3rate_G1Dom_brnspec_model.txt /
-t: BrBoSaCh_WGT_3rate_G1Dom_M2Opt3_Top3.tre
```

5. WGT 1Dom$_{G3}$ + Root is modeling the two-step hexaploidy scenario (Cheng et al., 2012; Tang et al., 2012), in which the MF1 and MF2 merged first following an initial round of gene loss, and the LF sub genome joined later, with more subsequent gene loss.

```
POInT -d:3 -g:Brassica_rapa_POInT_geneorders.txt /
-g:Brassica_oleracea_POInT_geneorders.txt /
-g:Sinapis_alba_POInT_geneorders.txt /
-g:Crambe_hispanica_v3_POInT_geneorders.txt -o:FourSpp_M2Opt3.txt /
```

```
-m: WGT_3rate_G1Dom_model.txt -r:WGT_RootModel.txt /
-t: BrBoSaCh_WGT_3rate_G1Dom_M2Opt3_Top3.tre
```

Data analysis

Result interpretation

The likelihoods for different WGT models were compared using likelihood ratio tests (LRTs), shown in Figure 2 of Hao et al. (2021). The significant likelihood increases from model 1 WGT Null to model 2 WGT 1Dom ($P < 10^{-10}$), and from model 2 WGT 1Dom to model 3 WGT $1Dom_{G3}$ ($P < 10^{-10}$), showed strong evidence of biased fractionation in the three sub genomes. We also found strong support ($P < 10^{-10}$) for the two-step hexaploidy model WGT $1Dom_{G3}$ + Root, suggesting that the first two sub genomes, MF1 and MF2, merged first, and the third sub genome LF joined after a certain amount of time. All of the completed ancient polyploidy analyses can be visualized using an online tool called the $POInT_{browse}$ (http://wgd.statgen.ncsu.edu/; Conant and Wolfe, 2008; Siddiqui and Conant, 2023). Users will be able to browse the gene retention/loss patterns in different sub genomes with simple clicks.

Acknowledgments

G.C.C. is supported by U.S. National Science Foundation (NSF) Division of Integrative Organismal Systems (IOS) grant NSF-IOS-1339156. This protocol was derived from Hao et al. (2021) and the POInT website (http://conantlab.org/POInT/POInT.html).

Competing interests

The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G. and Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PLoS One* 7(5): e36442.

Conant, G.C. (2023). POInT: Modeling Polyploidy in the Era of Ubiquitous Genomics. In Y. Van de Peer (Ed.). *Polyploidy* (pp. 77–90). Methods in Molecular Biology. Humana, New York.

Conant, G. C. (2020). The lasting after-effects of an ancient polyploidy on the genomes of teleosts. *PLoS One* 15(4): e0231356.

Conant, G. C. and Wolfe, K. H. (2008). Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics* 179(3): 1681–1692.

Edger, P. P., McKain, M. R., Bird, K. A. and VanBuren, R. (2018). Subgenome assignment in allopolyploids: challenges and future directions. *Curr. Opin. Plant Biol.* 42: 76–80.

Emery, M., Willis, M. M. S., Hao, Y., Barry, K., Oakgrove, K., Peng, Y., Schmutz, J., Lyons, E., Pires, J. C., Edger, P. P., et al. (2018). Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS Genet.* 14(3): e1007267.

Hao, Y. and Conant, G. C. (2022). POInT: A Tool for Modeling Ancient Polyploidies Using Multiple Polyploid Genomes. In Pereira-Santana, A., Gamboa-Tuz, S. D. and Rodríguez-Zapata, L. C. (Eds.). *Plant Comparative Genomics* (pp. 81–91). Methods in Molecular Biology. Humana, New York.

Hao, Y., Fleming, J., Petterson, J., Lyons, E., Edger, P. P., Pires, J. C., Thorne, J. L. and Conant, G. C. (2022). Convergent evolution of polyploid genomes from across the eukaryotic tree of life. *G3 Genes|Genomes|Genetics* 12(6): e1093/g3journal/jkac094.

Hao, Y., Mabry, M. E., Edger, P. P., Freeling, M., Zheng, C., Jin, L., VanBuren, R., Colle, M., An, H., Abrahams, R. S., et al. (2021). The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. *Genome Res.* 31(5): 799–810.

One Thousand Plant Transcriptomes, I. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574(7780): 679–685.

Qi, X., An, H., Hall, T. E., Di, C., Blischak, P. D., McKibben, M. T. W., Hao, Y., Conant, G. C., Pires, J. C., Barker, M. S., et al. (2021). Genes derived from ancient polyploidy have higher genetic diversity and are associated with domestication in*Brassica rapa*. *New Phytol.* 230(1): 372–386.

Schoonmaker, A., Hao, Y., Bird, D. M. and Conant, G. C. (2020). A Single, Shared Triploidy in Three Species of Parasitic Nematodes. *G3 Genes|Genomes|Genetics* 10(1): 225–233.

Siddiqui, M. and Conant, G. C. (2023). POInT$_{browse}$: orthology prediction and synteny exploration for paleopolyploid genomes. *BMC Bioinform* 24: 174.

Tang, H., Woodhouse, M. R., Cheng, F., Schnable, J. C., Pedersen, B. S., Conant, G., Wang, X., Freeling, M. and Pires, J. C. (2012). Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. *Genetics* 190(4): 1563–1574.

Van de Peer, Y., Maere, S. and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10(10): 725–732.

Supplementary information

Data and code availability: all data and code have been deposited into GitHub: https://github.com/Bio-protocol/POInT-WGT.