Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

FractBias: a graphical tool for assessing fractionation bias following polyploidy

Blake L. Joyce[1], Asher Haug-Baltzell[2], Sean Davey[1], Matthew Bomhoff[1], James C. Schnable[3] and Eric Lyons[1,2*]

[1]BIO5 Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

[2]Genetis GIDP, University of Arizona, Tucson, AZ 85721, USA

[3]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Following polyploidy events, genomes undergo massive reduction in gene content through a process known as fractionation. Importantly, the fractionation process is not always random, and a bias as to which homeologous chromosome retains or loses more genes can be observed in some species. The process of characterizing whole genome fractionation requires identifying syntenic regions across genomes followed by post-processing of those syntenic datasets to identify and plot gene retention patterns. We have developed a tool, FractBias, to calculate and visualize gene retention and fractionation patterns across whole genomes. Through integration with SynMap and its parent platform CoGe, assembled genomes are pre-loaded and available for analysis, as well as letting researchers integrate their own data with security options to keep them private or make them publicly available.

Availability and implementation: FractBias is freely available as a web application at https://genomevolution.org/CoGe/SynMap.pl. The software is open source (MIT license) and executable with Python 2.7 or iPython notebook, and available on GitHub (https://goo.gl/PaAtqy). Documentation for FractBias is available on CoGepedia (https://goo.gl/ou9dt6)

Contact: ericlyons@email.arizona.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Polyploidy events such as whole genome duplications (WGDs) create two or more copies of a genome within a single organism. Entire sets of homeologous chromosomes derived from the duplication (*subgenomes*) undergo gene loss in a process called *fractionation* whereby genes are deleted from one of the homeologous chromosomes (Langham *et a*l., 2004). Fractionation eventually reduces the numbers of genes in the genome near to the original diploid state before duplication (diploidization). Interestingly, it is not always the case that either copy of a duplicated gene pair is equally likely to be lost; bias towards deletion of genes from particular subgenomes has been observed in several species including: Arabidopsis (Thomas *et a*l., 2006), maize (Schnable *et a*l., 2011), and *Brassica rapa* (Tang *et a*l., 2012).

While fractionation is a primary evolutionary mechanism following WGDs in plants (Freeling *et a*l., 2012), the effects of WGDs and resulting *fractionation bias* has not been studied in full, mostly due to the lack of easy to use tools for characterizing fractionation bias. Here we describe **FractBias**, a web-based automated tool for calculating and visually assessing biased fractionation between duplicated syntenic regions after polyploidy. FractBias is available for use locally or online through the Comparative Genomics (CoGe) platform.

2 Methods

FractBias calculates the syntenic genes retained on every query chromosome for each target chromosome, and generates graphical representations of the pattern of fractionation. The fractionation bias calculation ons 1 and 2:

$$f(T_a, Q_b, x_j) = \frac{\sum_{i=1}^{w} S}{w} \quad (1)$$

$$S = t_i \in T_{a_{x_j}} \cap Q_b \quad (2)$$

where

$a,b = T \times Q$ are the ordered pairs of ch$|Ta| - w$ues of T and Q

j represents the series of windows = 1 ..

S evaluates to be 1 or 0

T = all chromosomes in the target genome

Q = all chromosomes in the query genome

$|T_a|$ = number of genes on the a^{th} chromosome of T

t_i = i^{th} gene on T_a; x = series of genes in window

w = window size, units = number of genes

2.1 User Input

The web-based version of FractBias is integrated into the synteny analysis tool SynMap (https://goo.gl/VhlvmF; Lyons *et a*l., 2008).

(1) Navigate to SynMap and select two genomes to compare.

(2) Under the 'Analysis Options' tab, set the following options:

1.1 *Syntenic Depth Algorithm*: 'Quota Align'

1.2 *Ratio of Coverage Depth*: the ratio of polyploidy between genomes.

1.3 Select 'Run Fractionation Bias'

1.4 *Sliding Window Size*: gene number considered in the sliding window (default 100)

1.5 *Fractionation bias calculation*: All genes (Sup. Fig 1A) or only syntenic genes (Sup. Fig 1B).

2.2 FractBias analysis

After configuring and running, SynMap identifies syntenic regions between the two genomes and screens those results based on the depth of syntenic coverage. Syntenic gene pairs are then automatically imported into FractBias, assigned by target or query genome, and ordered according to their chromosome and start sites. The analysis identifies all syntenic pairs across every chromosome for both genomes, so inversions and genome structure changes do not affect the analysis.

A sliding window analysis is run along each chromosome of the target genome (See Methods; Sup. Fig. 2A & 2B). For each iteration, the percentage of syntenic genes present within that window is calculated for all chromosomes in the query genome. The sliding window analysis ends when the right most side of the window reaches the last gene on that particular chromosome, such that only full-length windows are considered. Example analyses are available in Supplementary Table 1 for the "all genes" and the "only syntenic genes" options using a variety of organisms and syntenic depth ratios. Runtimes for FractBias are in the 10s of seconds range, depending on the size of the data and the speed of the computer.

2.3FractBias output

The graphical output of FractBias is composed of subplots for each target genome chromosome (Fig. 1). X-axes correspond to the sliding window iteration, analogous to the ordered number of genes on the target genome. Y-axes indicate the percentage of genes retained at syntenic locations within that window iteration for each query genome chromosome. In addition to the graphical output, raw data files are available for download for use in downstream analyses. CoGe also designates a unique URL to each analysis that can be used to regenerate and share the analysis, or to embed in manuscripts ensuring reproducibility.

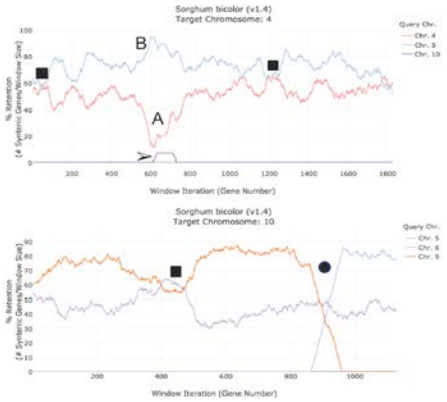

Figure 1. Comparison of *Sorghum bicolor* chromosomes 4 and 10 (target) and *Zea mays* (query) genomes using FractBias on the "only syntenic genes" setting. "A" indicate areas of overfractionation (homeologous genomic regions with more gene loss), whereas "B" indicate areas of underfractionation (homeologous genomic regions with less gene loss). Squares denote areas without bias, circles denote chromosome fusion events, and arrow-heads denote ancient polyploid events that were not filtered by the Quota Align algorithm (https://genomevolution.org/wiki/index.php/Quota_align; Tang *et a*l., 2012). Note that these special characters are added by hand and not generated by FractBias. These results mirror Figure 2 of Schnable *et al.* 2011. Results can be regenerated at https://genomevolution.org/r/k7j3.

3 Results and discussion

To validate FractBias, the 1 sorghum : 2 maize comparison was analyzed and compared against a previous study (Schnable *et a*l., 2011; Fig. 1 & Sup. Fig. 4). The FractBias analysis parallels the prior results, supporting the reliability and reproducibility of the program. While FractBias may be used on any two genomes or in a self-self comparison, having an unduplicated outgroup is important for fractionation assessment. Example analyses from various organisms with other syntenic depth ratios are available in Supplementary Table 1.

4 Conclusions

We developed FractBias to make investigation of the fractionation bias following polyploidy accessible to researchers of all computational skill levels. Online FractBias is integrated into the CoGe platform giving it access to pre-loaded assembled genomes, the ability for researchers to upload (private or public) genomes, and access to additional genome analysis and comparison tools. On a local computer, it can be run via Python 2.7 or an iPython notebook (Shen, 2014). In both use cases, FractBias produces easy-to-interpret graphical output as well as raw output files that can be annotated to test whether fractionation bias likely occurred in a species which can lead to future hypotheses about how fractionation bias affects genome evolution, making it a keystone tool for polyploidy analyses.

Acknowledgements

We would like to thank CyVerse (formerly iPlant) for sharing best practices and providing data and computational infrastructure for CoGe (NSF DBI – 1265383).

Funding

This work has been supported by the U.S. National Science Foundation (IOS – 1339156 and IOS – 1444490).

Conflict of Interest: none declared.

References

Freeling,M. *et al.* (2012) Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Genome Stud. Mol. Genet.*, **15**, 131–139.

Langham,R.J. *et al.* (2004) Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics*, **166**, 935–945.

Lyons,E. *et al.* (2008) The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop. Plant Biol.*, **1**, 181–190.

Schnable,J.C. *et al.* (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci.*, **108**, 4069–4074.

Shen,H. (2014) Interactive notebooks: Sharing the code. *Nature*, **515**, 151–152.

Tang, H. *et al.* (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC bioinformatics*, ***12*(1)**, 1.

Tang,H. *et al.* (2012) Altered Patterns of Fractionation and Exon Deletions in Brassica rapa Support a Two-Step Model of Paleohexaploidy. *Genetics*, **190**, 1563–1574.

Thomas,B.C. *et al.* (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.*, **16**, 934–946.