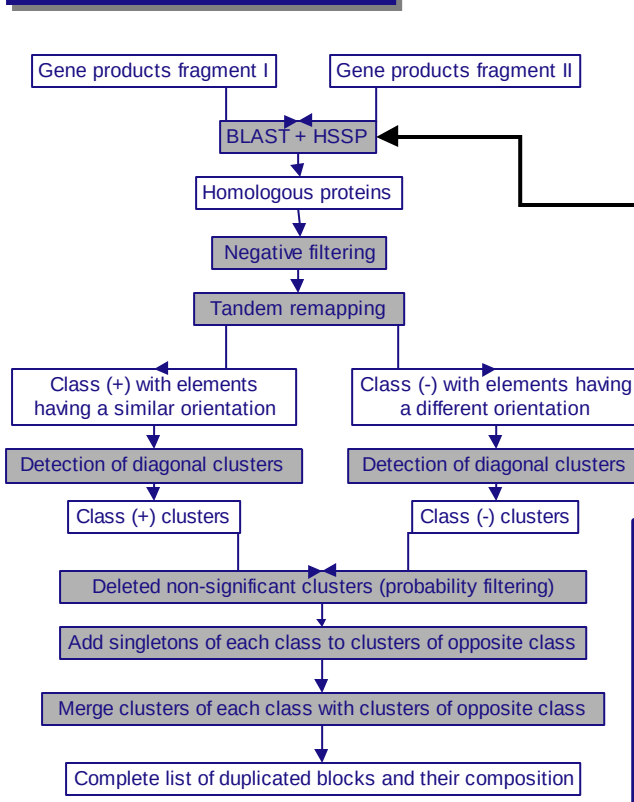
Reconstructing the duplication past of *Arabidopsis thaliana*

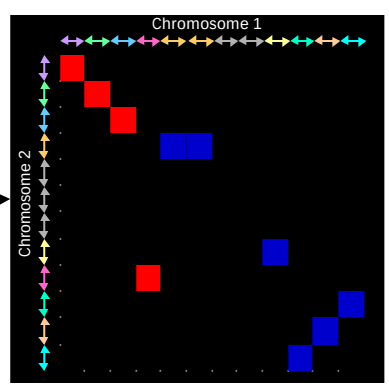
Cedric Simillion, Klaas Vandepoele, Pierre Rouzé and Yves Van de Peer
BioInformatics Research Group, Flanders Interuniversity Institute for Biotechnology (VIB)
Department of Plant Genetics, K.L. Ledeganckstraat 35, 9000 Gent - Belgium
contact: cesim@gengenp.rug.ac.be

Abstract

During the past years it has become clear that large-scale gene and genome duplications play an important role in the evolution of most eukaryotic organisms. Several models suggest that genomic duplications tend to increase the total amount of 'raw' genetic material from which novel gene functions can arise through functional divergence. We have developed a software tool (**ADHoRe**: **A**utomated **D**etection of **Ho**mologous **Re**gions) that allows the automated creation of detailed and complete overviews of segmental duplications within genomes. We applied this tool to study the duplication landscape of Arabidopsis thaliana. Our study reveals that about 80% of this genome is duplicated. This is significantly more than previous studies indicated.
In addition our study also shows that heavily degenerated block duplications that cannot be observed by directly comparing both segments involved, can still be detected through indirect comparison with other segments. Adding these so called 'hidden'-duplications to the global duplication landscape of Arabidopsis thaliana sheds a new light on the number of large-scale duplications that this genome has undergone in its evolutionary past.

ADHoRe Algorithm



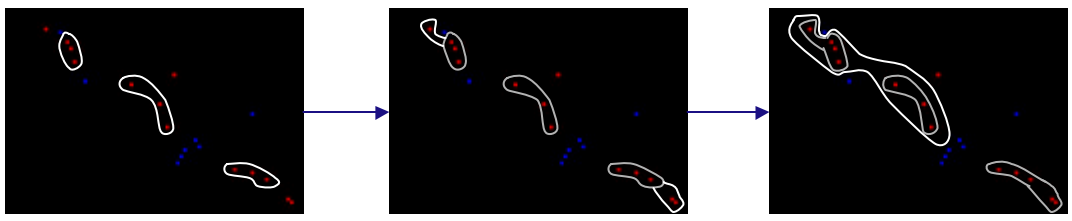


For all gene products on both genomic fragments, for which we want to find gene colinearity, initially an **all-against-all sequence similarity search** is performed, using BLASTP. In a second step, all these results are converted into identity scores (given a certain alignable region) between query and hit sequence. Next, the HSSP identity cut-off curve (Rost, 1999) is used to **select only hits with a similar secondary structure**. With this procedure, we obtain all couples of homologues proteins between both genomic fragments.

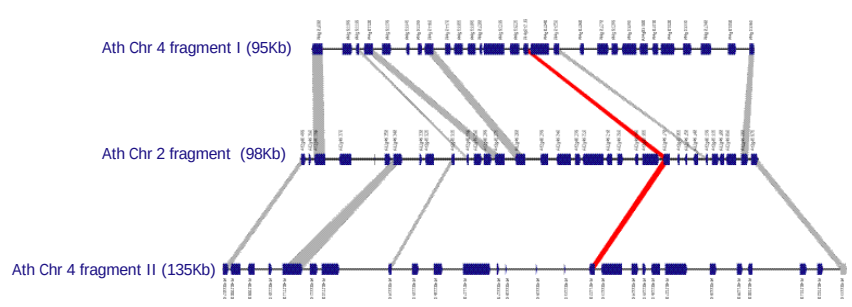
This information is then stored in a matrix of (*m*.*n*) elements - *m* and *n* being the total number of genes on each genomic fragment - each element (*x*,*y*) being an couple of homologous gene products (*x* and *y* being the coordinates of these genes), also denoted as **singletons**. The matrix is then subject to a number of steps (*see diagrams*), which in the end returns all duplicated regions present between both genomic fragments, denoted as **clusters**. The **only two parameters** required for the user to specify are the **gap size G** (which describes the maximum number of intervening, non-homologues genes tolerated between two pairs of homologous genes within a duplicated block) and a parameter describing the **quality of a cluster** of colinear genes.

Cluster Detection Algorithm

Foreach g (3 to **G**)



'ghost' duplications



Both segments of chromsome 4 (upper and lower fragments) show colinearity with a segment of chromosome 2. **Colinearity between both chromosome 4 segments cannot be observed directly** since they share only 1 gene (marked in red). However, since both fragments are colinear with the same segment of chromosome 2, they must share a common ancestor and **thus are descendents of the same duplication event**. Thus, both segments are duplicates despite the fact that direct **colinearity between them can no longer be observed**. We call such a pair of homologous segments a '***hidden***' duplication.

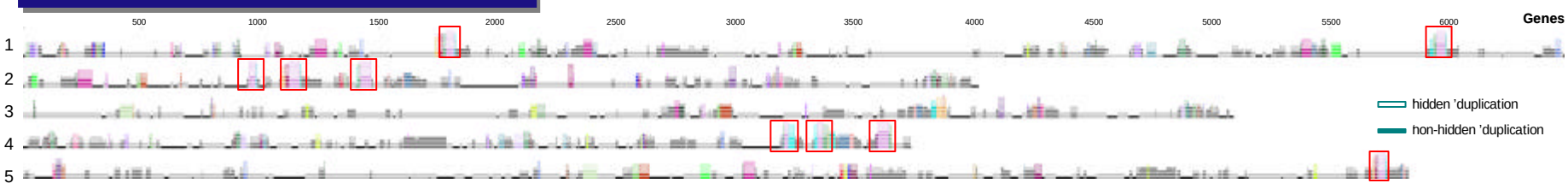
Duplicated portion of Arabiopsis

Chromosome	Genes in duplicated regions	Total Genes	% of genes in duplicated regions	bp in duplicated areas	Total bp	% bp in duplicated regions
1	5409	6488	83,37%	24155029	29640317	81,49%
2	2928	4023	72,78%	12296249	19643621	62,60%
3	4222	5096	82,85%	18958794	23333883	81,25%
4	2846	3738	76,14%	12563813	17549528	71,59%
5	4365	5832	74,85%	19327049	26269328	73,57%
Total	19770	25177	**78,52%**	87300934	116436677	**74,98%**

Genome duplication events

Adding the detected 'ghost' duplications to the global duplication landscape of Arabidopsis reveals that **numerous segments appear in multiple duplications** (see diagram below). Several segments appear in **5 to 8**-fold (coloured stacks), which can only be explained by **at least 3 large-scale duplication events**, probably genome duplications (or polyploidisation events).
Only 1 segment on chromosome 1 appears in 9-fold (red boxes), probably due to an extra block duplication event.

Multiply duplicated segments



References:1) Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A.99,* **13627-32 (2002). 2) K. Vandepoele, Y. Saeys, C. Simillion, J. Raes and Y. Van de Peer: The Automatic Detection of Homologous Regions (ADHoRe) and its Application to Microcolinearity between Arabidopsis and Rice.** *Genome Res.* **12, 1792-801 (2002). 3) The Arabidopsis Genome Initiative:** Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000) 4) **T. Vision, D. Brown and S. Tanksley**: The Origins of Genomic Duplications in Arabidopsis. *Science* **290**, 2114-6 (2000) 5) **J. Raes, K. Vandepoele, C. Simillion, Y. S aeys and Y. Van de Peer:** Investigating ancient duplication events in the Arabidopsis genome *Functional and structural genomics, In press*