PhygOmics, a phylogenetic pipeline for evolutionary analysis of high throughput sequence data.

Aureliano Bombarely[1], Kieron D. Edwards[2], Juan Sanchez-Tamburrino[2] & Lukas Mueller[1]

1- Boyce Thompson Institute for Plant Research, Tower Rd, Ithaca, NY 14853 USA
2- Advanced Technologies (Cambridge) Ltd, 210 Cambridge Science Park, Cambridge, UK.

ABSTRACT:

Motivation: The development of next generation sequencing methods has produced a rapid growth in the number of publicly available genome and transcriptome datasets. One of the challenges is the integration and analysis of these data from an evolutionary perspective.

Results: PhygOmics is a phylogenetic pipeline designed for the analysis of thousands of genes through gene homology clustering, tree reconstruction for homology groups, tree topology analysis and synonymous-non synonymous ratio analysis. The application of this pipeline to the leaf transcriptomes of *Nicotiana tabacum*, and its closest relatives *N. sylvestris* and *N. tomentosiformis* has identified 1000 homeologous and homologous gene relations between these three species, and possible implications in terms of gene duplication fate in an early polyploid is presented.

Availability: https://github.com/solgenomics/sgn-home/tree/master/aure/scripts/phylo/PhyGomic

Introduction:

Traditional phylogenetic approaches analyze tens of genes of several species to elucidate the phylogenetic relations between these species, but each gene could have a distinct evolutionary story. Different homologous gene groups (HGG) can have different evolutionary rates [1]. The clustering of phylogenetic tree topology and the analysis of the synonymous-non synonymous ratio of thousands of HGG can give some keys about these how these groups have evolved. PhygOmics is a Perl pipeline that integrates several phylogenetic analysis programs such as Phylip[2] or PAML[3] and cluster the phylogenetic trees based in their topologies.

The analysis of the leaf transcriptome of *Nicotiana tabacum*, an allotetraploid and its closest parent relatives, *N. sylvestris* (S-genome donor) [4] and *N. tomentosiformis* (T-genome donor) [5] has identified around 1000 homeologus in *N. tabacum*. Additionally evolutionary ratios as Kn/Ks ratios were calculated and compared between homoelogs (*N. tabacum* S-genome/T-genome comparison) and homologs.

Methodology:

PhygOmics, a Perl pipeline was developed to (1) realign the sequences of homologous gene groups (HGG), (2) calculate distances and create one phylogenetic tree per HGG, (3) cluster them based in its topology and (4) calculate the Kn, Ks and Kn/Ks ratio for each of the members pair. Additionally applied to a polyploid species have a tool to (5) identify possible homeologs based in the tree topology and the alignment properties (Figure 1).

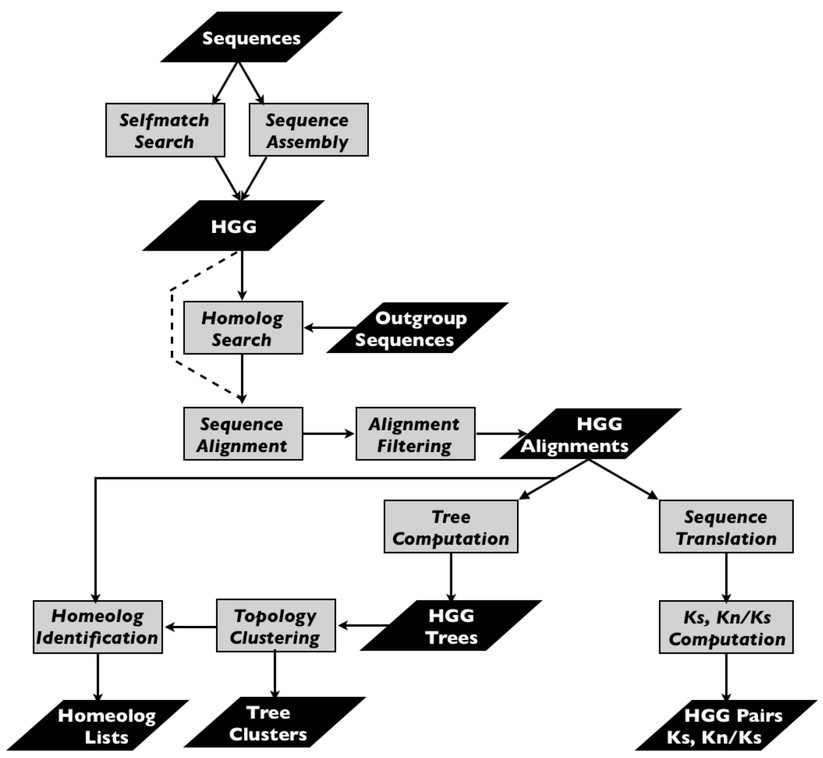

Figure 1: PhygOmics pipeline flow chart. Homolog gene groups (HGG) are created processing sequence assembly files (.ace) or selfmatch search files (.blast8). Alignments are computed and filtered for each HGG. Alignment files (.clustalw) are used as input for tree computation and Ks, Kn/Ks computation

There are two ways to create HGG from gene models or unigene sequences from different species or sources:

a) Parse a selfmatch search (for example a selfblast).
b) Parse an assembly file (.ace) retrieving sequences used.

An HGG consensus sequence can be used to search homolog in an outgroup

(1) PhygOmics can use several tools to align the sequences (Table 1). Alignments are filtered based in identity percentage, alignment length and species composition of the overlapping region. An alternative filtering method based in a computed score (Figure 2) of the overlap ("ovlscore") is available.

$$OS = OL * (OP / 100) ^ 2$$

Figure 2: Algorithm to compute ovlscore (OS). OL is length of the overlarp, OP is identity percentage of the overlap.

(2) Neighbor Joining (NJ) and Maximun Likelihood (ML) methods can be used to compute the trees. Trees are constructed with an outgroup. Tree filtering based in bootstrapping values are available.

(3) Tree topology comparison and clustering are performed using Bio::Tree::TopoType module, integrated as a part of the pipeline. It produce 3 files: 2 column (HGG_id, topology_id) output file and two graphs: topology composition based in the HGG count and tree topology. Topology graphs are created using R and the R-Perl interface R::YapRI.

(4) Kn, Ks and Kn/Ks values are computed using PAML (Table 1) [3]. The results are intergrated with the topology of the HGG and the species/source of each HGG member, producing a 14 column output (HGG_id, topology_id, pair1_source, pair2_source, pair1_ID, pair2_ID, N, S, Kn, Ks, Kn/Ks, kappa, LnL, t).

(5) Homeolog identification is performed using the pipeline module Strain::AlleleIdentification (Figure 3).

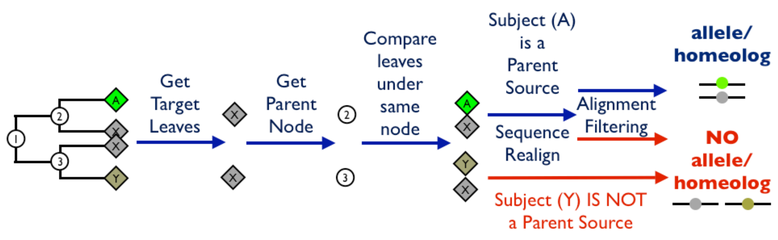

Figure 3: Process of homeolog identification for PhygOmics module: Strain::AlleleIdentification

The pipeline can run different analytical runs ("Paths") and compare between them to optimize parameters (for example using different aligmment software such as ClustalW, Muscle and MAFFT) or to compare different HGG sources (for example HGG with 3 members, A, B, C and with 4 members A1, A2, B, C).

Process	Tool
Selfmatch Search	Blast
Sequence Assembly	CAP3, MIRA, gsAssembler
Homology Search	Blast
Sequence Alignment	ClustalW, Muscle, Mafft, Kalign, T-Coffee
Alignment Filtering	Bio::Align::Overlaps*
Tree Computation	Phylip, Phyml
Topology Clustering	Bio::Tree::TopoType*
Homeolog Identification	Strain::AlleleIdentification*
Sequence Translation	EstScan, Bio::Seq
Kn/Ks computation	PAML
Topology graphs	R::YapRI, and R

Table 1: Tools used by PhygOmics pipeline. Modules specific from this pipeline are indicated with *. These tools are integrated with the pipeline using BioPerl [6].

Results

The pipeline was applied to the analysis of RNA-seq leaves samples from *N.tabacum*, *N. sylvestris* and *N. tomentosiformis*. Sequences were initially assembled by species using gsAssembler with a minimum sequence identity = 97%. The unigenes were used to create the HGG through a reassembly with a minimum sequence identity = 90%. 7,974 HGG were created with at least one member of each specie. *Solanum lycopersicum* gene models (ITAG2.3) were used as outgroup. After realignment and filtering 968 HGG trees were analyzed. 11 topologies were produced using NJ and ML methods (Figure 4), AB_C and AC_B , where only one of the *N.tabacum* homeolog is expressed, were the most abundant. Around 1,000 of *N.tabacum* homeolog genes were identified. There is no any HGG where *N.tabacum* homeologs have Kn/Kn > 1 when *N. sylvestris-N.tomentosiformis* homologs have Kn/Ks < 1.

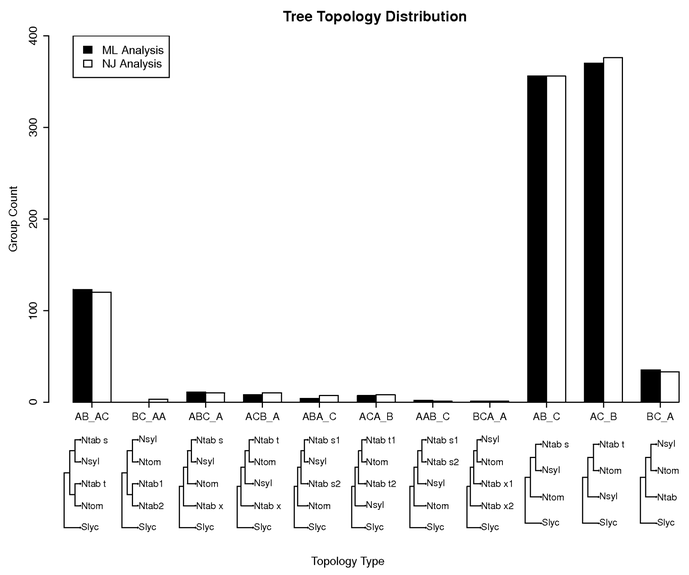

Figure 4: Results for the topology analysis of the Nicotiana HGGs.

Conclusion

PhygOmics is a robust phylogenetic analysis pipeline to extend the classical phylogenetic tools such as Phylip or PAML to thousand of genes grouped by HGG. Applied to a polyploid species has provided a powerful tool to identify alleles, and in a transcriptomic context, relates the expression of each homoelog with its parents closest relatives.

References:

[1] Yang YH, Zhang FM, Ge S. "Evolutionary rate patterns of the Gibberellin pathway genes.", **BMC Evol Biol.** 2009 Aug 18;9:206
[2] Felsenstein J. "Notices". **Cladistics** 1989, 5(2):163-166.
[3] Yang Z: "PAML 4: phylogenetic analysis by maximum likelihood:" ; **Mol Biol Evol.** 2007 Aug;24(8):1586-91
[4] Kenton A, Parokonny AS, Gleba YY, Bennett MD: "Characterization of the Nicotiana tabacum L. genome by molecular cytogenetics". **Mol Gen Genet** 1993, 240(2):159-169.
[5] Gazdova B, et al: "Characterization of a new family of tobacco highly repetitive DNA, GRS, specific for the Nicotiana tomentosiformis genomic component". **Chromosome Res** 1995, 3(4):245-254.
[6] Stajich JE. *et al.* "The BioPerl toolkit: Perl modules for the life sciences" **Genome Res**. 2002 Oct;12(10):1611-8.