The Poplar Genome Project

Lieven Sterck, Stephane Rombauts, Pierre Rouzé, and Yves Van de Peer

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Gent, Belgium

Introduction

Poplar trees have been used all over the world to produce a large variety of wood-based products such as timber, pulp, and paper. Besides their great economical value, poplars are also rapidly becoming the model organism for forest biology and tree biotechnology. It is therefore not surprising that, in 2001, poplar was selected as the first woody plant to have its genome sequenced.

Now that the poplar genome is publicly available, genome-wide bioinformatics analyses of the genome sequence can take a start.

Genome annotation

The annotation of the genome was performed by 3 different labs: 2 in the USA (using FgenesH and Grail_EXP) and 1 in Belgium (using Eugene). From these 3 different annotations one consensus annotation was constructed using a rule-based system that for a given locus selected the most plausible gene-model. Before the post-processing steps (like removing putative transposons) we ended up with +- 58000 genes.

method	gene count	
Eugene	25018	43.11%
FgenesH	22546	38.85%
Grail_exp	7352	12.66%
genewise	3120	5.38%

Table 1. Contributions of the different gene finders to the final annotation

Table 1 gives an overview of the selection of genes from the different gene finders. It is clear that our tool performs better than all others used in this annotation effort.

Although the genome is not yet published, the annotations can be viewed and queried on the JGI Genome Browser.

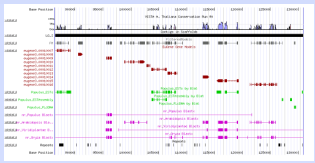


Figure 1. Screenshot of the Poplar Genome Browser

The Genome Browser is available at : http://genome.jgi-psf.org/Poptr1/Poptr1.home.html

Exploring the duplication past of poplar

For all paralogous genes in the poplar genome a Ks-estimation was performed. After corrections and filtering a histogram was constructed by plotting the number of paralogs against their age (Ks). As can be seen on the histogram there are many paralogs with the similar Ks-values (0.20-0.30) which indicate the fact that a large number of genes were created in a short period of time. The most plausible explanation for this observation is that a genome duplication has occurred in the poplar genome.

We performed the same procedure on ESTs of 7 different Poplar species and came to the same conclusion for all of them. This suggests that the genome duplication has occurred before the divergence of all poplar species.

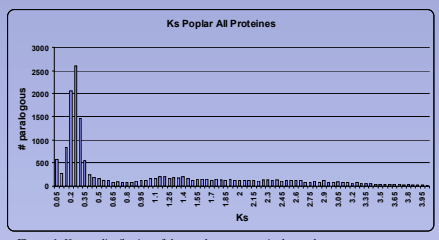


Figure 1. Ks age distribution of the paralogous genes in the poplar genome

The most striking result of this analysis is not the fact that there was a genome duplication, but rather when it has happened. When we infer the time in My from the Ks distribution we dated the duplication event some 8 My ago, which clearly contradicts with the fossil records on poplar.

This indicates that the average Ks-substitution rate used for plants is probably not applicable to poplar. Due to its much longer generation time it is very likely that poplar has a much smaller Ks-substitution rate than all other plants studied so far.

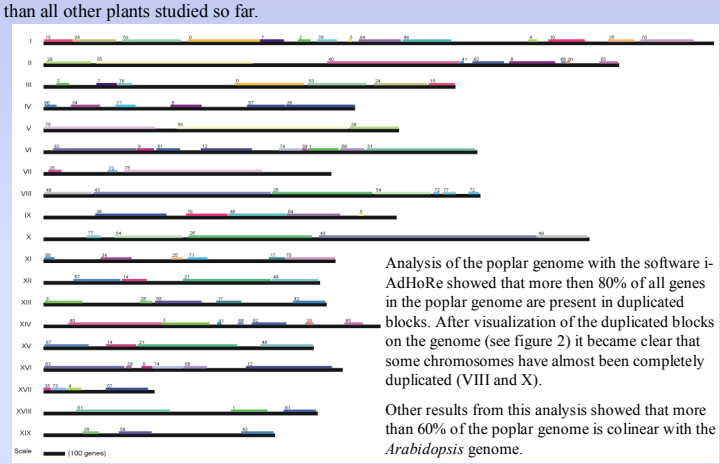
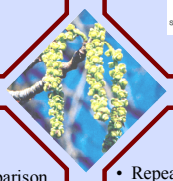


Figure 2. Map with duplicated blocks in the poplar genome

Analysis of the poplar genome with the software i-AdHoRe showed that more then 80% of all genes in the poplar genome are present in duplicated blocks. After visualization of the duplicated blocks on the genome (see figure 2) it became clear that some chromosomes have almost been completely duplicated (VIII and X).

Other results from this analysis showed that more than 60% of the poplar genome is colinear with the *Arabidopsis* genome.



Gene family comparison

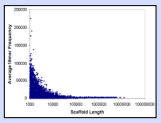
TF_IPR_ID	ipr_description	poplar	Arath
IPR001841	Zn-finger, RING	416	362
IPR001005	Myb DNA-binding domain	416	283
IPR007087	Zn-finger, C2H2 type	297	162
IPR001092	Basic helix-loop-helix dimerization domain bHLH	202	166
IPR001471	Pathogenesis-related transcriptional factor and ERF	197	143
IPR003441	No apical meristem (NAM) protein	153	90
IPR001356	Homeobox	134	91
IPR001965	Zn-finger-like, PHD finger	115	131
IPR003447	Myb-like DNA-binding domain, SHAQKYF class	104	70
IPR001789	Response regulator receiver	103	38
IPR003202	GRAS family transcription factor	103	34
IPR002100	Transcription factor, MADS-box	92	92
IPR004827	Basic-leucine zipper (bZIP) transcription factor	90	73
IPR002057	DNA-binding WRKY	89	68
IPR003571	Zn-finger, C-x8-C-x5-C-x3-H type	89	66
IPR003340	Transcriptional factor B3	78	69
IPR001214	Nuclear protein SET	54	40
IPR004883	Protein of unknown function DUF260	42	37
IPR003851	Zn-finger, Dof type	39	33
IPR003315	Zn-finger, B-box	37	29
IPR003658	Transcription factor CBF/NF-Y/archaeal histone	35	25
IPR003311	AUX/IAA protein	34	32
IPR000679	Zn-finger, GATA type	34	25
IPR003323	TCP family transcription factor	33	23
IPR000047	Helix-turn-helix motif, lambda-like repressor	33	21
IPR003252	Heat shock factor (HSF) type DNA-binding domain	32	21
IPR003347	Transcription factor jumonji, JmjC	26	18
IPR004333	SBP	23	17
IPR000007	Tubby	19	15
IPR003253	Forkhead-associated (FHA)	18	15
IPR003910	HMG/I (I) hook mobility group) box	15	14
IPR003335	Plant regulator RWP-RK	15	13
IPR001606	ARID (AT-rich interaction domain) protein	13	9
IPR004022	CDT domain	12	8
IPR003349	Transcription factor jumonji, JmjN	12	7
IPR006780	YABBY protein	9	5
IPR005172	Tesmin/TSO1-like CXC domain	8	7
IPR006511	LRP1, C-terminal	7	9
IPR007918	Protein of unknown function DUF702	6	9
IPR006510	Putative zinc finger domain, LRP1	6	8
IPR003657	Ethylene insensitive 3	6	6
IPR003316	Transcription factor E2F/dimerisation partner (TDP)	6	4
IPR002910	Floricaula/leafy protein	3	1
IPR006779	DNA binding protein S1FA	2	2

Table 1. No. of proteins belonging to a specific TF-class.

Table 1 shows a comparison between Poplar and *Arabidopsis* of the number of transcription factors. For most of the TFs the numbers are comparable although there are some differences, eg. the Myb TFs, which seem to have undergone lineage specific expansion in poplar. Also the pathogen-related TFs are more abundant in Poplar, which is perhaps correlated with the perennial nature of Poplar.

Repetitive sequences and transposons

- Repeats occupy 28% of the assembled genome and more than 50% of the unassembled scaffolds.
- Small scaffolds and unassembled reads contained substantially more repetitive sequence than large scaffolds.



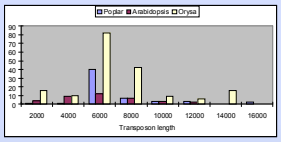


Figure 1. Comparison of length transposons

When we compare the length and frequency of LTR-transposons in *Arabidopsis*, poplar and rice it is clear that in rice there are more and longer transposons, although the longest ones are identified in poplar.

This comparison included 37 Ath, 58 poplar and 181 rice transposons.

Conclusions

- Outstanding international collaboration has placed poplar on equivalent footing with other model species
- Assembly was heavily complicated by repeats, but facilitated by genetic maps
- Unprecedented opportunities to explore molecular basis of commercial and environmental traits
- Most of the hard work and many exciting discoveries are still to come!

References:

- Salamov A., Solovyev V. (2000) **FgenesH**: Ab initio gene finding in Drosophila genomic DNA. Genome Res., 10,516-522
- Schiex,T., Moisan , A. and **Rouzé, P.** (2001). **EuGène** : An Eucaryotic Gene Finder that combines several sources of evidence. Computational Biology, Eds. O. Gascuel and M-F. Sagot, LNCS 2066, pp. 111-125
- **Simillion, C., Vandepoele, K., Saeys, Y. & Van de Peer, Y.** (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res. 14, 1095-1106
- **Sterck, L., Rombauts, S.**, Jansson, S., Sterky, F., **Rouzé, P. & Van de Peer, Y.** (2005) EST data suggest that poplar is an ancient polyploid. New Phytol. (In Press)
- Ying Xu, Manesh Shah, Doug Hyatt, Richard Mural, and Edward C. Uberbacher, **GRAIL-EXP**: Multiple Gene Modeling Using Pattern Recognition and Homology, *The Seventh Department of Energy Contractor and Grantee Workshop, January 1999*