Transcriptomic platform

URGV

Apports du RNA-seq dans l'analyse des génomes orphelins

Véronique Brunaud et Etienne Delannoy

Colloque EPGV

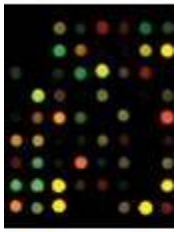
25 juin 2014, Evry

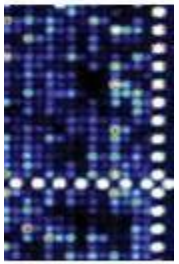
genopole®

INRA SCIENCE & IMPACT

université evry val-d'essonne

cnrs dépasser les frontières

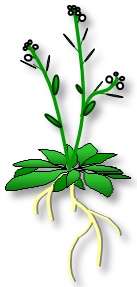




Transcriptomic platform

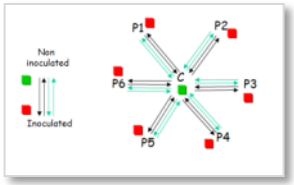
URGV

Manager: Sandrine Balzergue



transcriptome choice (method)

Experimental design



Hybridizations/ library/sequencing

Statistical analysis

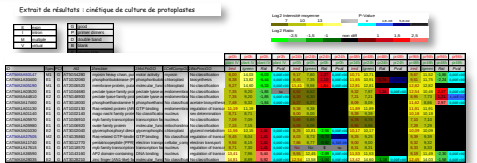
Expertise/help to result interpretation

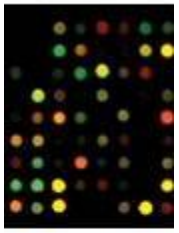
CATdb

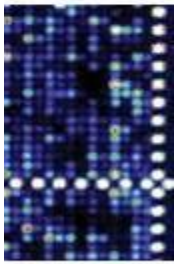
a Complete Arabidopsis Transcriptome database

GEO Gene Expression Omnibus

NCBI

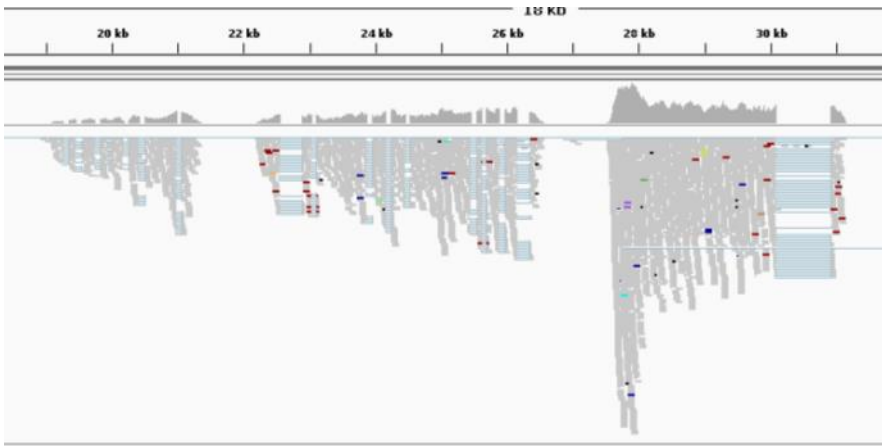



Transcriptome analysis

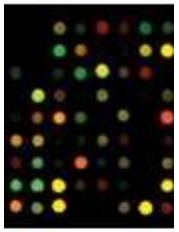
Exploration

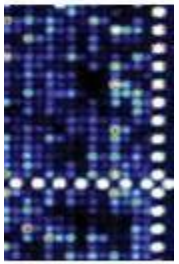


Quantification

	C1		C2		C3		C4	
AGI	Rat	Pval	Rat	Pval	Rat	Pval	Rat	Pval
AT1G18980	1,37	0,00E+00			0,94			
AT1G23720	1,33	0,00E+00	1,18	1,11E-10	1,25	0,00E+00		
AT1G52690	2,53	0,00E+00			1,24	0,00E+00		
AT1G58270	1,52	0,00E+00	0,87		1,20	0,00E+00		
AT1G62570	1,44	0,00E+00			0,96			
AT2G25625	1,45	0,00E+00						
AT2G33790	1,73	0,00E+00			1,10	1,09E-10		
AT2G39800	1,06	0,00E+00					-1,10	0,00E+00
AT2G46680	1,47	0,00E+00			0,76			
AT2G47770	2,06	0,00E+00			0,92			
AT3G02480	2,91	0,00E+00			1,50	0,00E+00		
AT3G15670	1,89	0,00E+00			0,87			
AT3G28550	1,16	0,00E+00			0,86			
AT3G50970	2,06	0,00E+00	1,52	0,00E+00	1,79	0,00E+00	0,82	
AT3G53980	1,07	0,00E+00			0,75		-0,73	
AT3G54580	1,55	0,00E+00	0,99		1,27	0,00E+00		
AT4G11340	1,06	0,00E+00						
AT4G35770	-3,60	0,00E+00	-2,09	0,00E+00	-2,84	0,00E+00		
AT5G06760	2,17	0,00E+00			1,08	2,68E-10		

List of differentially expressed genes


RNA-Seq and transcriptome exploration

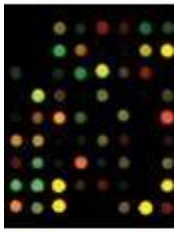
Gene discovery (genome annotation, homologous genes, phylogeny…)

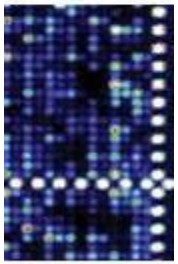
Transcript variants (splicing, editing, SNP…)

No prior genome sequence required

Single nucleotide resolution

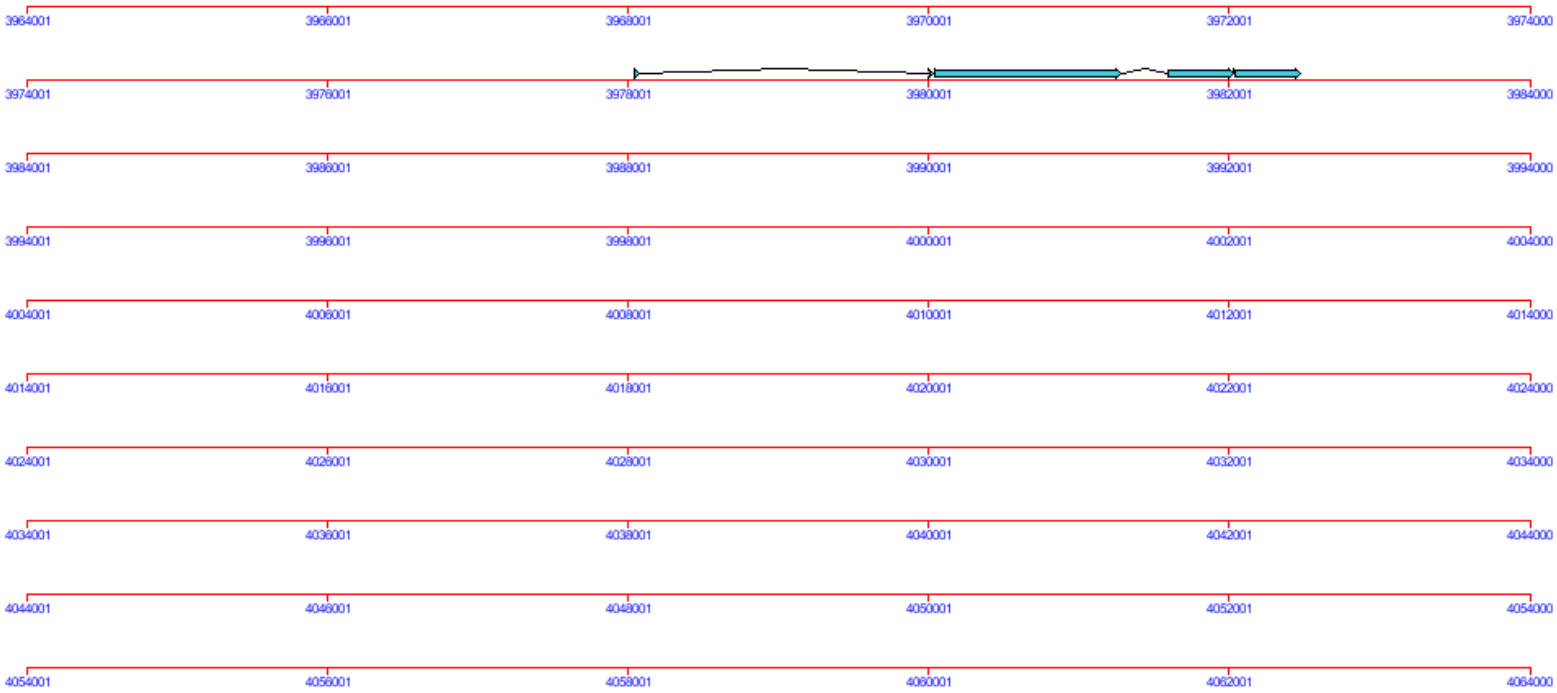
Allele expression

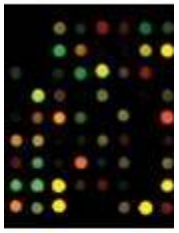

RNA-Seq vs DNA-Seq

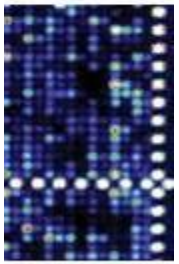
RNA-seq = focused sequencing



Flagdb snapshot: 100kb from chr 9 of *P. trichocarpa*

Small sequencing depth required to cover the transcriptome compared to the full genome

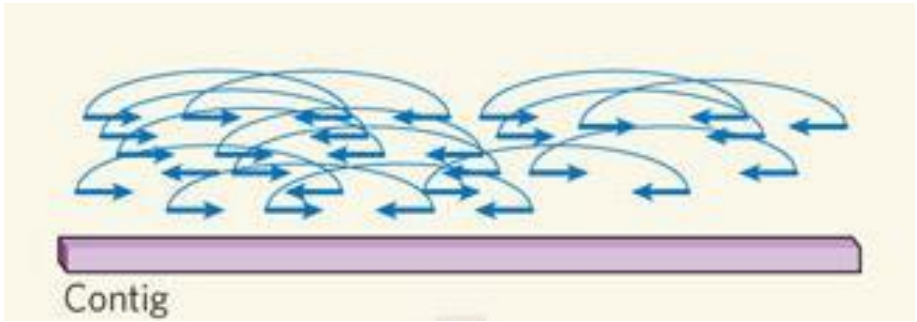




Assembly, key step of the RNA-Seq analysis

RNA

Raw reads

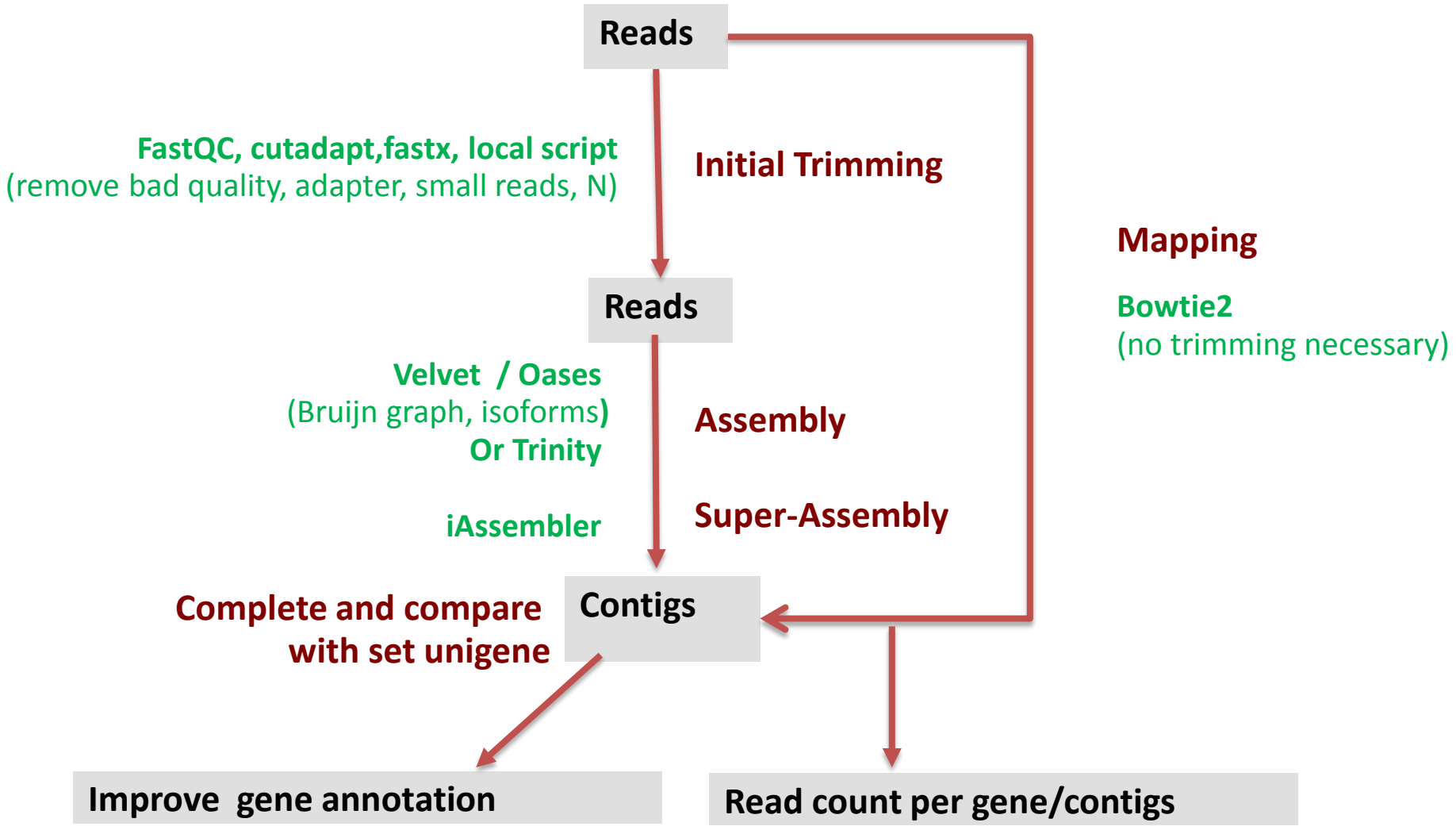
mRNA gene models



Thousands of expected mRNA gene models

Arabidopsis thaliana
Pipeline set up

Quality and efficiency of Assembly on Arabidopsis genome



\+ defined new gene models

\- Assembly: not perfect, defined kmer, time and memory consuming

Assembly method and quality

F1_Mplex	
Nb of PE reads	43 030 388 PE
Nb of contigs	33 736 (length mean 1360)
Nb of mapped contigs Genome TAIR10 (FLAGdb++)	33 072 98%

Data from Illumina HiSeq2000

- Velvet/oases (kmer 61,71)
- iAssembler

Assembly and comparison of annotations TAIR10 versus Contigs

F1_Mplex	
Nb of PE reads	43 030 388 PE
Nb of contigs	33 736 (length mean 1360)
Nb of mapped contigs Genome TAIR10	33 072 98%
Comparaison annotations	
Nb tagged Genes Nb contigs	17 783 32 220 (97%)
Nb of genes with confirmed structure Nb Contigs	15 613 (88%) 20 881 (65%) contigs

Data from Illumina HiSeq2000

- **Velvet/oases (kmer 61,71)**
- **iAssembler**

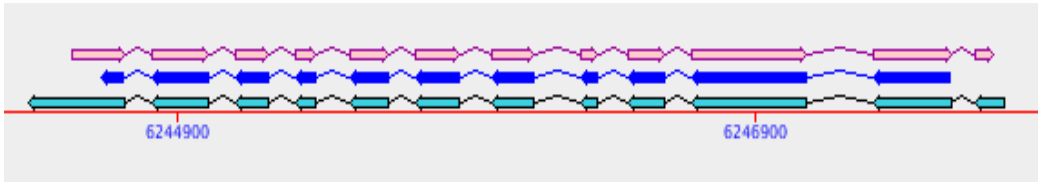
→Gene=Locus

→ Model of genes with confirmed exon/intron structure

→ Good quality : 98% of contig mapped against genome

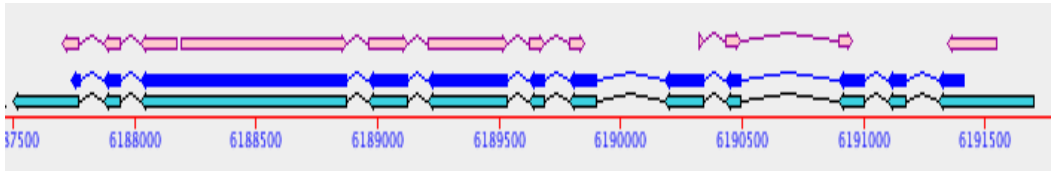
→ 35% “new contigs”

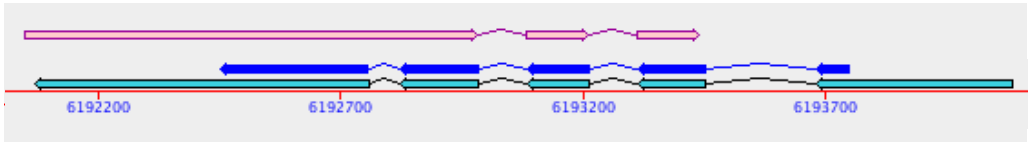
Assembly Challenges



contig
CDS
mRNA

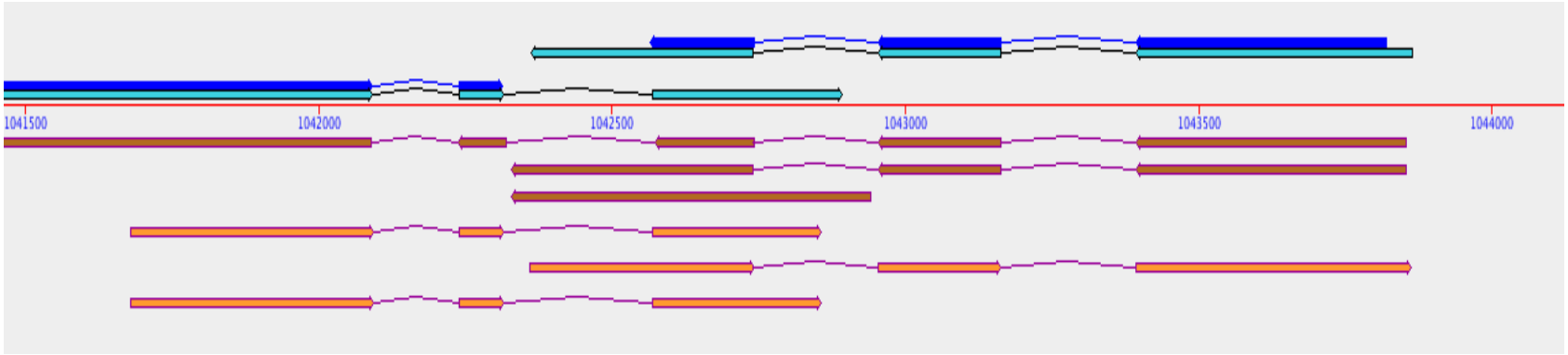
Gene models OK





Splicing alternatif

Oriented /non oriented



Contigs without Gene annotations

Contigs

Annotations
TAIR10

Reads density
CuffLinks prediction

→ 3% of contigs from assembly are not associated with gene annotations

Genes characteristics:

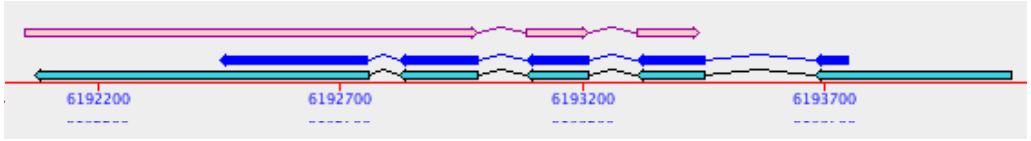
→ genes are shorter (mean contig length : 650 instead of 1360)

→ matching with otherRNA, TE, ESTs

Quality of Assembly : contig versus gene annotation

35% of contigs with other gene models

1 gene – 1 or n contigs
with other gene models



3% of contigs with no annotated genes

Sequencing Depth

- 1776 additional genes tagged with 200 millions of reads can be detected
- But equivalent results with the 40 millions depth with an more efficient assembly (smaller kmer)

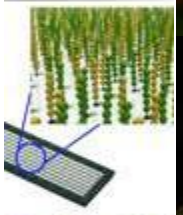
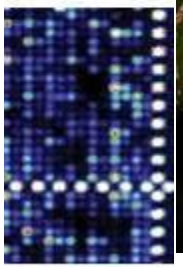
Conclusion on assembly

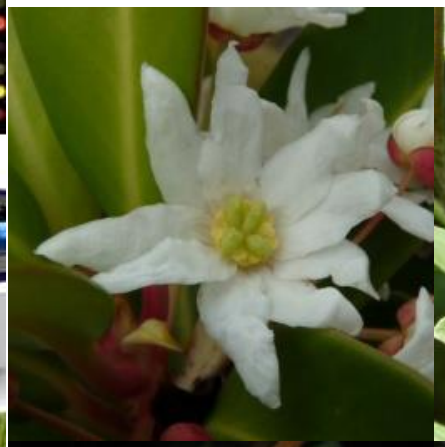
→ A good quality of contigs, efficient to detect new gene models
→ Problems: distinct false/good gene models, chimera that increase with read number
→ Improving Assembly tools (PE, oriented,

Other challenges: getting RNAs

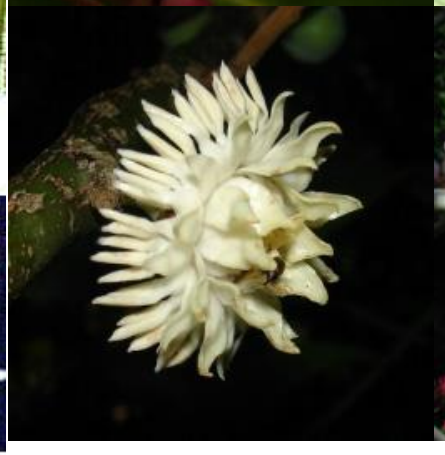
ANR MAGNIPHY (Hervé SAUQUET, Orsay): Floral diversity of Magnoliidae
Magnoliidae = 4 orders / 20 families / 270 genera / 10,000 species



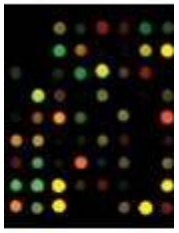


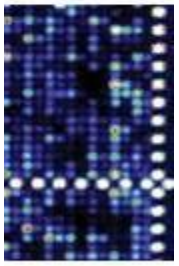





20 floral transcriptomes for phylogenomics and Evo-Devo

Other challenges: getting RNAs

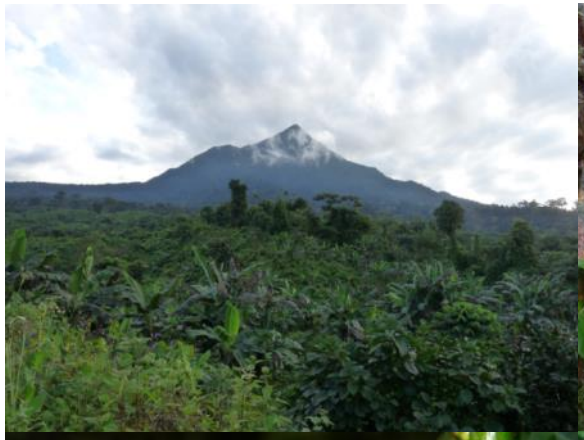

Species	Family	Origin
Chimonanthus praecox	Calycanthaceae	Orsay (360)
Glossocalyx longicuspis	Siparunaceae	**Mt Cameroon**
Aristolochia clematitis	Aristolochiaceae	Orsay (360)
Piper umbellatum	Piperaceae	**Mt Cameroon**
Magnolia maudiae	Magnoliaceae	Orsay (360)
Aristolochia arborea	Aristolochiaceae	BG Vienna

No universal RNA extraction protocol

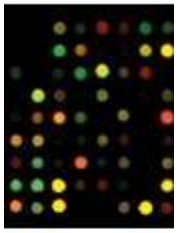
Not always possible to extract RNAs immediately

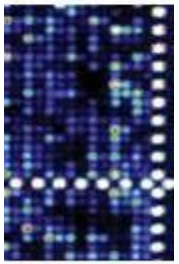
Degraded or partially degraded RNAs…

New sampling maybe necessary….



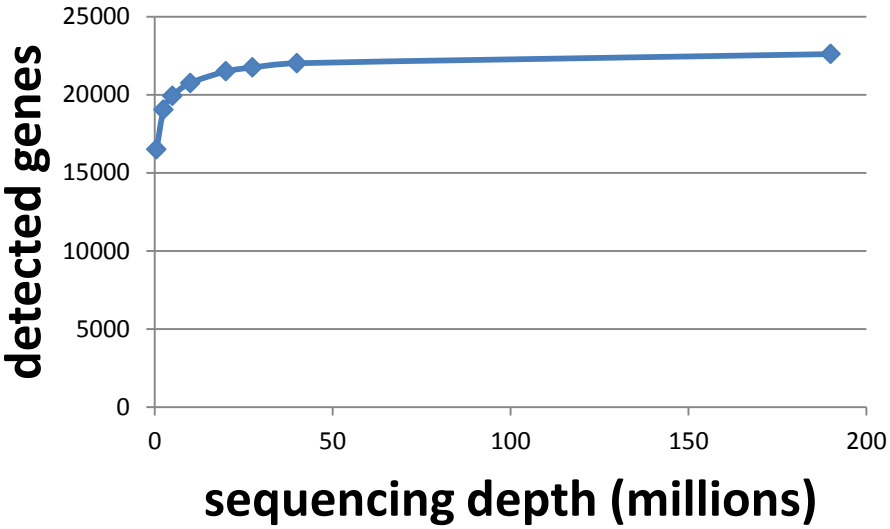





Other challenges: gene expression

In Arabidopsis leaf,

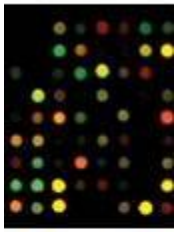


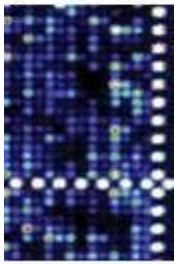
"Only" 22000 detected genes out of 28000

Tissue specific expression

Not compensated by sequencing depth

Assembly → 16000 with corresponding contigs
genes with low expression


Other challenges: gene expression

Common Bean:
annotation of the NBS-LRR genes (Valérie Geffroy, Orsay)

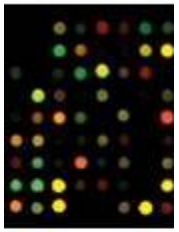
770 M reads from flowers, buds, stem, roots, leaves, seed pods (11 samples)

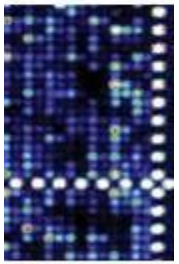
32180 contigs with N50 = 1719bp

Whole transcriptome coverage:
57% of the predicted genes covered at more than 50%

NBS-LRR coverage: 428 predicted genes
22 genes covered at more than 50%
249 covered by at least one contig…..

Very low expression level

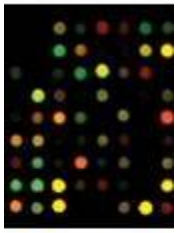

Other challenges: gene expression

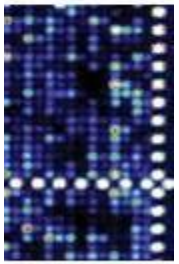
Solutions:

Normalised cDNA libraries,

Choice of the samples,

Multiplying samples…..


Other challenges: polymorphism

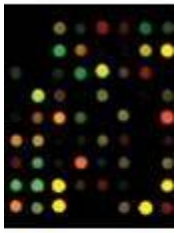
Assembly softwares very sensitive to polymorphism

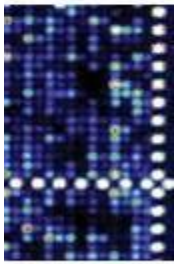
Technical “polymorphism” = sequencing errors

	Proton	Hiseq
read Nb	55384478	52161156
velvet kmer=61		
contig Nb	213178	22316
N50	265	1508
Median cov depth	7.8	23
using reads	10271478 (18%)	39302128 (75%)

Natural polymorphism: heterozygocy and polyploidy

How to adjust assembly parameters??

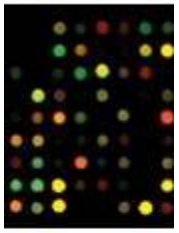



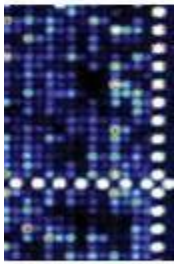

Other challenges: polymorphism

SPAM project (Sophie Nadot, Orsay)
Floral transcriptom of *Grevillea rosmarinifolia*

200M pairs of reads → 48000 contigs with N50 = 750

Identification of “floral morphology” gene homologs:
Agamous, Pistillata, Wuschel, Crabsclaw, Cup-Shaped Cotyledon, Shootmeristemless, Spatula, Bel1, Tousled, Apetala (3 contigs), Sepallata (4 contigs), Cycloidea (2 contigs).
→ successful!

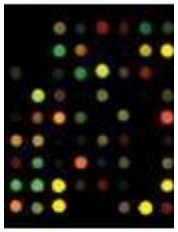



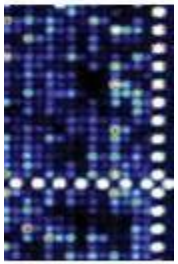

RNA-Seq for the study of orphan genomes

Numerous genomic analyses available through RNA-Seq

Powerful tool through focused sequencing of the “active “ part of the genomes (→ partial analysis of the genome)

The availability of a sequenced genome is not required but it allows a much easier analysis and interpretation of the RNA-Seq data.


Acknowledgments

Transcriptomic platform

URGV

Rémi Bounon
José Caius
Stéphanie Huguet
Cécile Labrune
Stéphanie Pateyron
Ludivine Soubigou-Taconnat
Jennifer Yansouni
Claire Lurin
Sandrine Balzergue

Bioinformatics and predictive genomics

Philippe Grevet
Cécile Guichard
Zakia Tariq
Jean-Philippe Tamby
Rim Zaag
Marie-Laure Martin-Magniette

Common Bean

Vincent Thareau
Valérie Geffroy

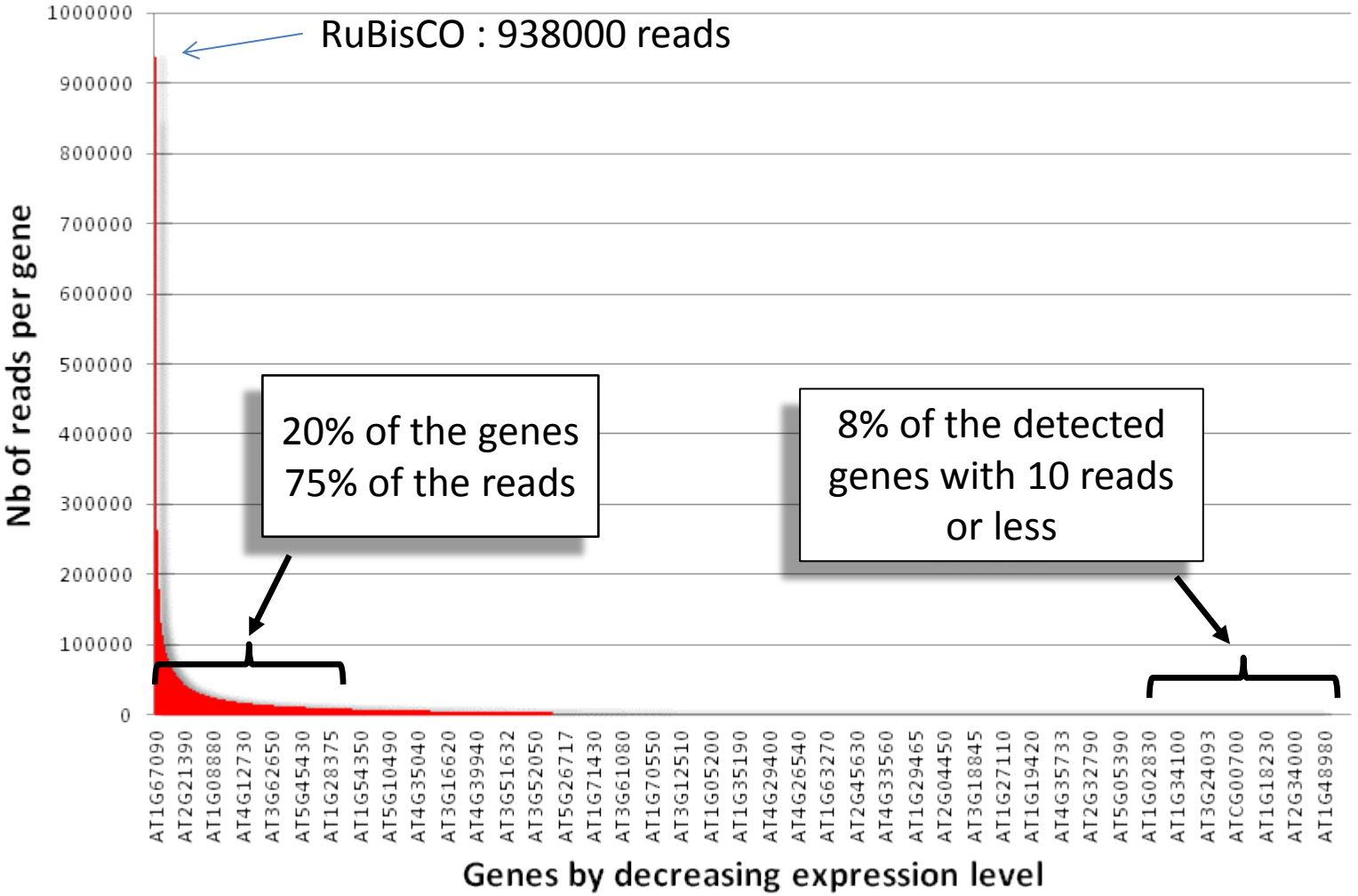
SPAM

Franck Simmonet
Catherine Damerval
Sophie Nadot

Magniphy

Julien Massoni
Véronique Normand
Hervé Sauquet

Other challenges: gene expression



Very biased distribution of the reads

Sample:
Col0
buds
HiSeq2000
PE