

Reuse of transcriptome data: how to create and process large scale data to understand biological processes.

Thomas Dugé de Bernonville

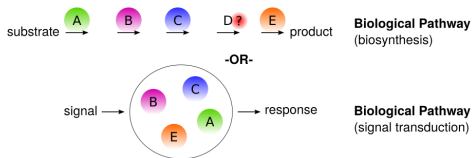
Université de Tours, EA2106 Biomolécules et Biotechnologies Végétales
Thèse en cours: **Franziska Liesecke**

June 2017



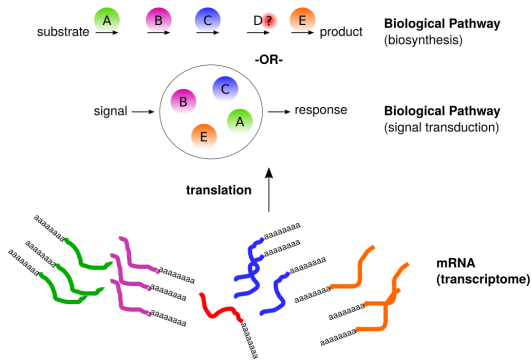
Transcriptome = sum of all transcripts

DNA is **transcribed** into RNA which is in turn **translated** into proteins



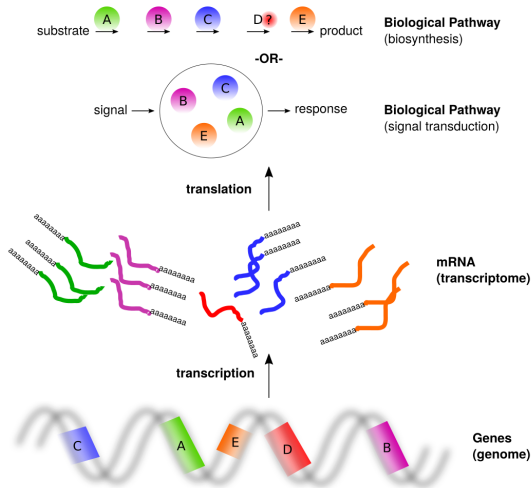
Transcriptome = sum of all transcripts

DNA is **transcribed** into RNA which is in turn **translated** into proteins



Transcriptome = sum of all transcripts

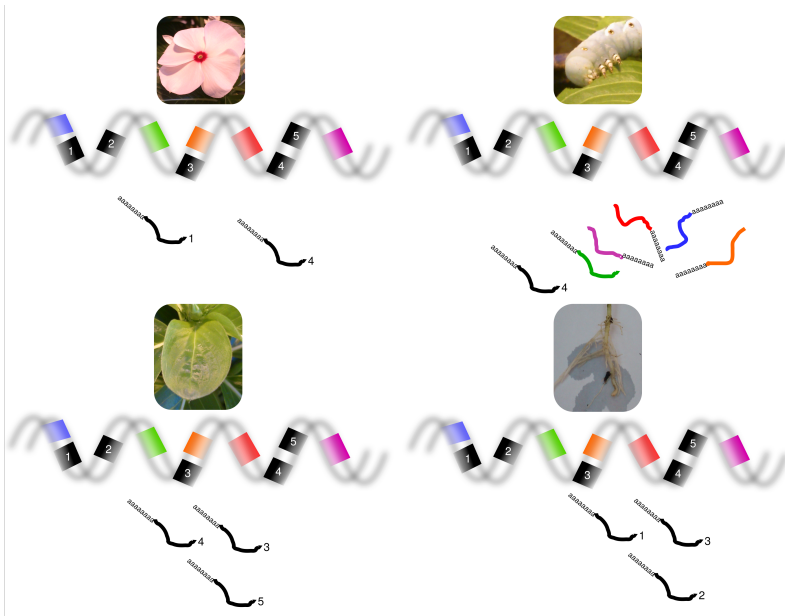
DNA is **transcribed** into RNA which is in turn **translated** into proteins



A dynamic profiling of gene expression



A dynamic profiling of gene expression



Use raw gene expression data to reconstruct and complete *in silico* biological pathways

How to measure gene expression levels?

- ▶ Hybridization (base pair complementarity)
- ▶ Sequencing (determine each base)



TOPIC PAGE

Transcriptomics technologies

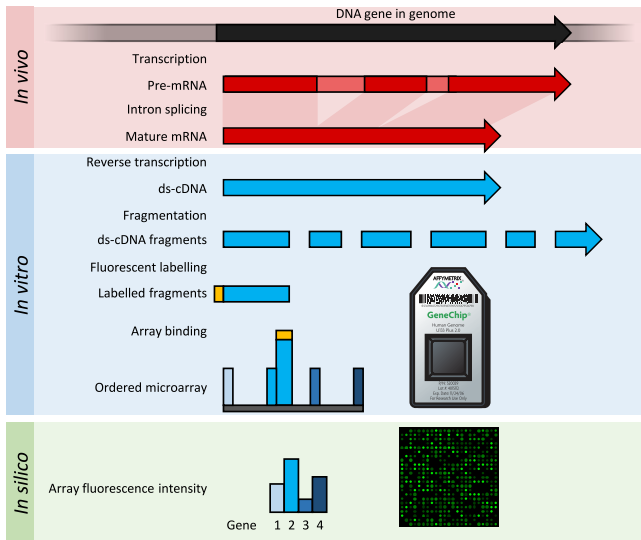
Rohan Lowe¹, Neil Shirley², Mark Bleackley¹, Stephen Dolan³, Thomas Shafee^{1*}

1 La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Australia, **2** ARC Centre of Excellence in Plant Cell Walls, University of Adelaide, Adelaide, Australia, **3** Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

* T.Shafee@LaTrobe.edu.au

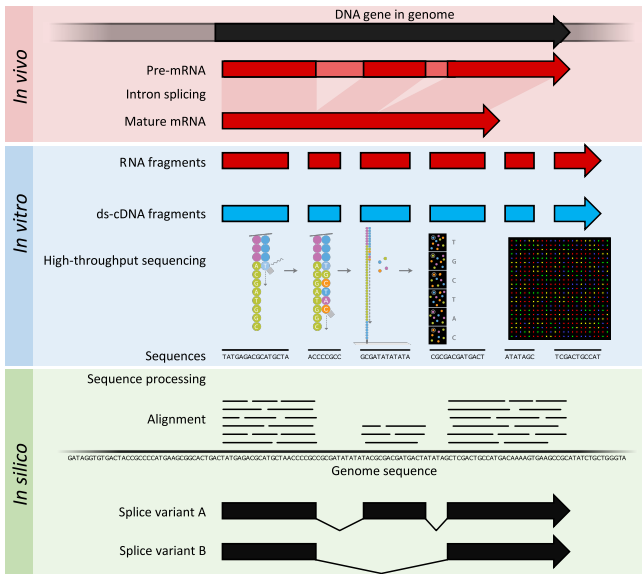
How to measure gene expression levels?

Microarrays = Hybridization-based

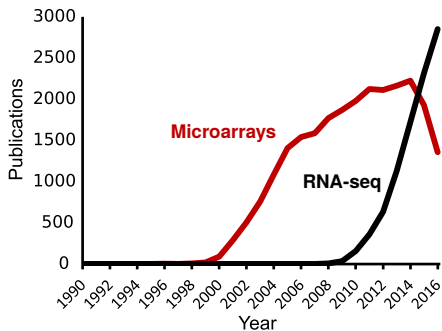


How to measure gene expression levels?

RNA-sequencing = Sequencing-based



How to measure gene expression levels?



Method	RNA-Seq	Microarray
Throughput	High	Higher
Input RNA amount	Low ~ 1 ng total RNA	High ~ 1 µg mRNA
Labour intensity	High (sample preparation and data analysis)	Low
Prior knowledge	None required, though genome sequence useful	Reference transcripts required for probes
Quantitation accuracy	~90% (limited by sequence coverage)	>90% (limited by fluorescence detection accuracy)
Sequence resolution	Can detect SNPs and splice variants (limited by sequencing accuracy of ~99%)	Dedicated arrays can detect splice variants (limited by probe design and cross-hybridisation)
Sensitivity	10^{-8} (limited by sequence coverage)	10^{-3} (limited by fluorescence detection)
Dynamic range	$>10^3$ (limited by sequence coverage)	10^2 – 10^4 (limited by fluorescence saturation)
Technical reproducibility	>99%	>99%

Using gene expression data

Gene co-expression analysis

Use transcriptome analysis to capture relationships between transcripts

Correlation of gene expression levels

= *Comparison of their expression profiles*

	Tissue1	Tissue2	Tissue3	Tissue4
GeneB				
GeneC				
GeneD				
GeneE				
GeneG				

Using gene expression data

Gene co-expression analysis

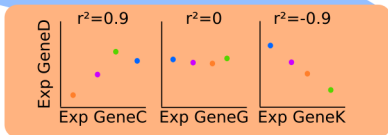
Use transcriptome analysis to capture relationships between transcripts

Correlation of gene expression levels

= *Comparison of their expression profiles*

	Tissue1	Tissue2	Tissue3	Tissue4
GeneB				
GeneC				
GeneD				
GeneE				
GeneG				

Expression matrix



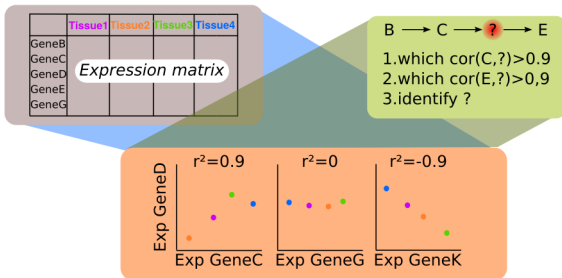
Using gene expression data

Gene co-expression analysis

Use transcriptome analysis to capture relationships between transcripts

Correlation of gene expression levels

= *Comparison of their expression profiles*



Getting gene expression data

► Generate new data



Characterization of new model herbivory model on C. roseus by RNA-seq allowed the discovery of new P450

► Reuse previously published data

Public databases with raw and/or processed expression data

NCBI Resources How To | L.dugedebeonville@gmail.com My NCBI Sign Out

GEO DataSets GEO DataSet Search Help

GEO DataSets

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

Getting Started

- [GEO Documentation](#)
- [GEO FAQ](#)
- [About GEO DataSets](#)
- [Construct a Query](#)
- [Download Options](#)

GEO Tools

- [Submit to GEO](#)
- [Advanced Search](#)
- [DataSet Browser](#)
- [Programmatic Access](#)
- [GEOQR](#)

More Resources

- [GEO Home](#)
- [GEO Profiles](#)
- [SRA](#)

Example Searches

Keywords and species	<code>(smok* OR died) AND !mammals[organism] NOT human[organism]</code>
Study type	<code>"expression profiling by high throughput sequencing"[DataSet_Type]</code>
Studies with CEL files	<code>cel[Supplementary_File]</code>
DataSets that have 'age' as an experimental variable	<code>age[Subset_Variable_Type]</code>
Studies with between 100 and 500 samples	<code>100-500[Number_of_Samples]</code>
Author	<code>smith[Author]</code>

Services Research Training About us EMBL-EBI

EMBL-EBI

The home for big data in biology

Our unique Search service helps you explore dozens of biological data resources. [More about EBI Search >](#)

Find a tool for your data analysis. [Find a tool >](#)

Share your scientific data with the world. [Deposit data >](#)

Find a name, member or channel

Public databases with raw and/or processed expression data

[ERX1583400](#): Illumina HiSeq 2500 paired end sequencing

1 ILLUMINA (Illumina HiSeq 2500) run: 34.6M spots, 6.9G bases, 2.8Gb downloads

Submitted by: Universite-Francois-Rabelais

Study: Transcriptomics of Catharanthus roseus upon challenge with Manduca sexta in local and distal leaves.

[PRJEB14626](#) • [ERP016279](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample: Catharanthus roseus leaves during folivory

[SAMEA4058182](#) • ERS1229292 • [All experiments](#) • [All runs](#)

Organism: [Catharanthus roseus](#)

Library:

Name: A4

Instrument: Illumina HiSeq 2500

Strategy: RNA-Seq

Source: TRANSCRIPTOMIC

Selection: unspecified

Layout: PAIRED

Construction protocol: TruSeq_v3

Runs: 1 run, 34.6M spots, 6.9G bases, [2.8Gb](#)

Run	# of Spots	# of Bases	Size	Published
ERR1512373	34,619,625	6.9G	2.8Gb	2017-01-01

Public databases with raw and/or processed expression data

- ▶ SRA: 5 petabases in 2017 (all origins combined)
- ▶ More than 610,000 RNA-seq accessions (ca. 1 Million accessions of DNA high throughput sequencing)
- ▶ 1 sequencing run: an average of 30 Million of bases, file size ca 3 Go
- ▶ Storage: >100 000 000 Go

Public databases with raw and/or processed expression data

PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens¹, Skylar Y. Lee¹, Faraz Faghri², Roy H. Campbell², Chengxiang Zhai³, Miles J. Efron⁴, Ravishankar Iyer¹, Michael C. Schatz^{5*}, Saurabh Sinha^{3*}, Gene E. Robinson^{6*}

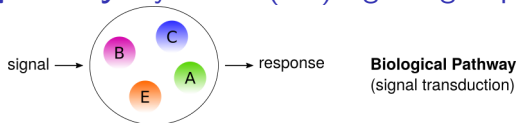
Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

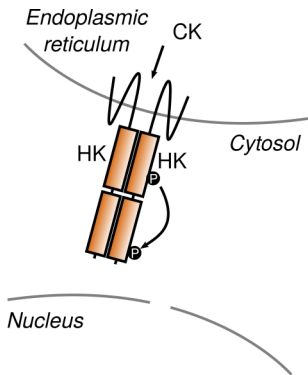
Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

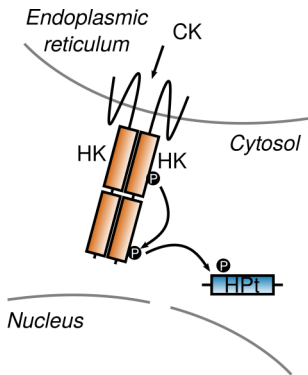
Work in progress @ EA2106 BBV: complete knowledge on biological pathways

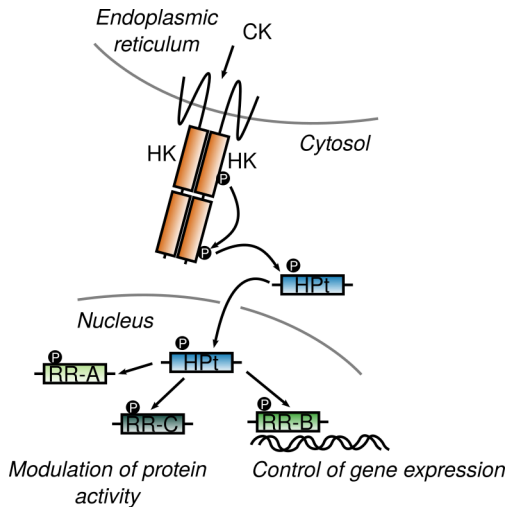
Metabolic pathway: monoterpene indole alkaloids in *Catharanthus roseus*

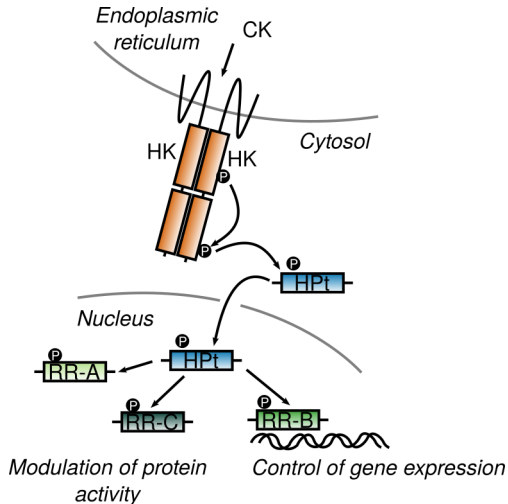
Signaling pathway: cytokinin (CK) signaling in plants












Association within levels? several proteins in each level
Extract associations with query genes from a global network

Construction of co-expression networks in *Arabidopsis thaliana*

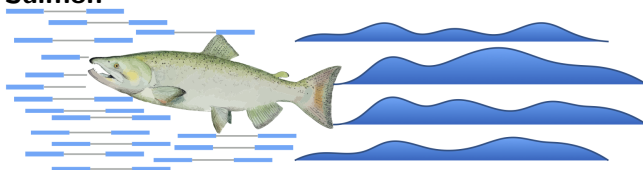


Reuse of expression data

- ▶ Microarrays: more than 10,000 samples (obtained and processed with “ArrayExpress”  package)
- ▶ RNA-seq: more than 1,600 samples downloaded from EBI

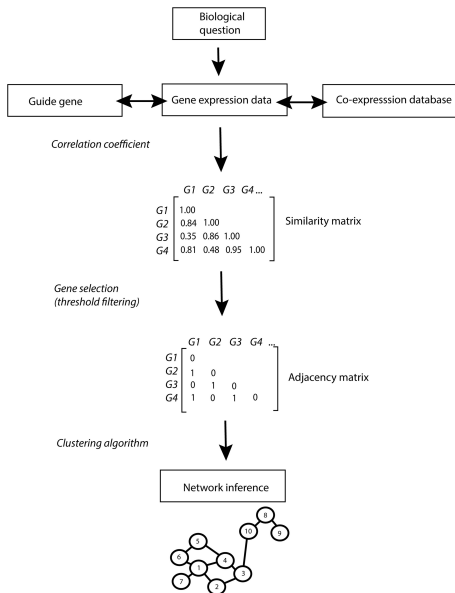
Processing of RNA-seq expression data

1. Identify accessions (DRR, ERR or SRR numbers)
2. Download raw Fastq files
3. Remove low-quality reads with **Trimmomatic**
4. Pseudo-align reads to Arabidopsis reference transcriptome with **Salmon**



5. ca. 1h for a 15 Million of paired-end reads with 5 threads (parallelization with subsets of accessions and array jobs)

Construction of co-expression networks



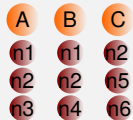
$n \times p$ expression matrix
 n genes
 p samples

$n \times n$ correlation or rank matrix

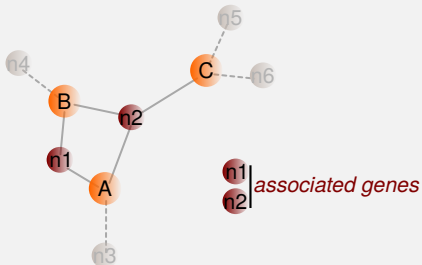
extract a $n \times g$ guide matrix from correlation or rank matrix



define best coexpressed genes
($PCC > x$ or $rank < y$) with each gene in g



group genes in g according to overlaps between best coexpressed genes



Construction of co-expression networks in *Arabidopsis thaliana*

Calculate all pairwise correlations between each pairs of genes

- ▶ 33,600 predicted transcripts in *A. thaliana*
- ▶ 33,600 × 33,600 correlation matrix
- ▶ **Which distance estimator? and for which data?**

Construction of co-expression networks in *Arabidopsis thaliana*

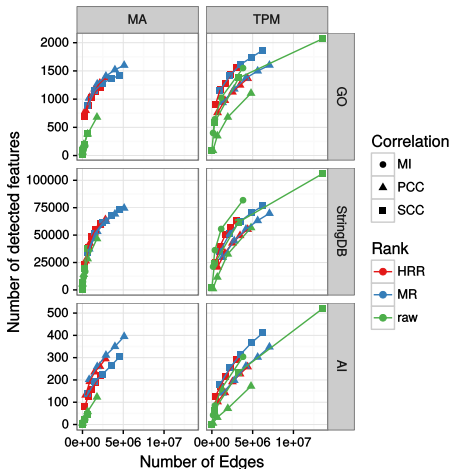
Calculate all pairwise correlations between each pairs of genes

- ▶ Development of parallelized tool (mpich2) written in C to compute all pairwise correlations (<15 minutes on 50 cpus)
- ▶ Pearson Correlation Coefficient, Spearman
- ▶ Ranked Correlation Coefficients (Mutual ranks or Highest Reciprocal ranks)
- ▶ Mutual Information (with "Parmigene" R package, in multicore mode)

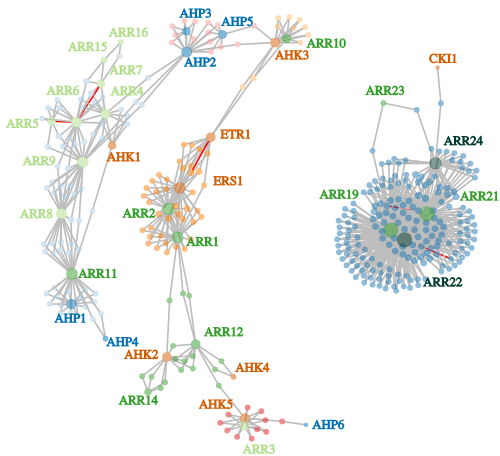
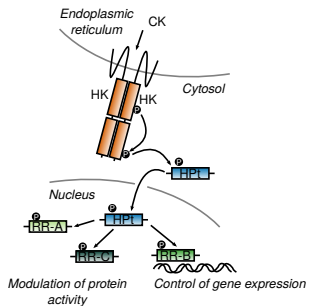
How to estimate the quality of the resulting networks?

Compare transcriptional associations with biologically known associations

- ▶ Gene Ontology
- ▶ Protein-Protein Interactions



A Cytokinin co-expression network

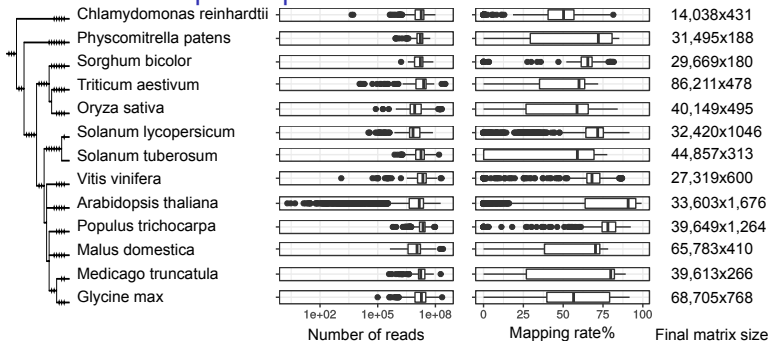


Prospects (1)

Are predicted associations conserved between plant species?

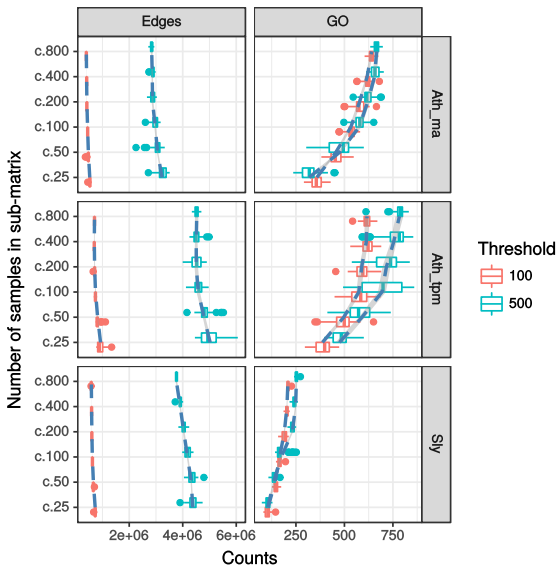
- ▶ Identify available data
- ▶ Construction of orthology groups
- ▶ Comparison of co-expression networks

Extend to other plant species



Prospects (2)

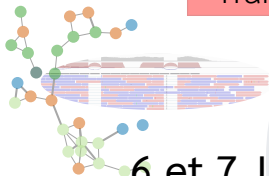
How much information do we need to capture a relevant transcriptional relationship?



Thank you for your attention

Acknowledgements

- ▶ EA2106
- ▶ Projets Région Abyssal et InsectEffect; ARD2020 BioProPharm
- ▶ Fédération Cascimodot pour l'accès à Artemis
- ▶ Yann Jullian pour l'aide sur Neptune



Traitement informatique de larges données en biologie

6 et 7 Juillet 2017

Stratégies d'analyse données omiques

Ateliers introductifs  et  Cytoscape

Jeudi 6 Juillet

Christophe Antoniewski

Institut de Biologie Paris

Yves Bigot

INRA Nouzilly

Nicolas Daccord

IRHS Angers

Pierre Nicolas

INRA Jouy-en-Josas

Thibault Guinoiseau

INSERM/PST ASB Tours

Présentation Galaxy

EA2106

Université de Tours

Vendredi 7 Juillet

Olivier Poch

CNRS Strasbourg

Laurence Liaubet

INRA Toulouse

Marc Clastre

Université de Tours

Véronique Brunault

INRA Paris

Présentation Cytoscape

Yves Vandembrouck

CEA Grenoble

Inscriptions et détails sur:

tild.sciencesconf.org

UFR Sciences Pharmaceutiques
Université François Rabelais, Tours
Amphi A50