



HAL
open science

Analyse du transcriptome

Annabelle Déjardin

► **To cite this version:**

| Annabelle Déjardin. Analyse du transcriptome. Master. Orléans, France. 2018. hal-03289368

HAL Id: hal-03289368

<https://hal.inrae.fr/hal-03289368>

Submitted on 16 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse du transcriptome

Annabelle Déjardin

Chargée de recherches INRA

UMR INRA - ONF BioForA

Biologie intégrative pour la valorisation de la diversité des arbres et de la forêt

annabelle.dejardin@inra.fr



UMR INRA - ONF BioForA

Biologie intégrative pour la valorisation de la diversité des arbres et de la forêt

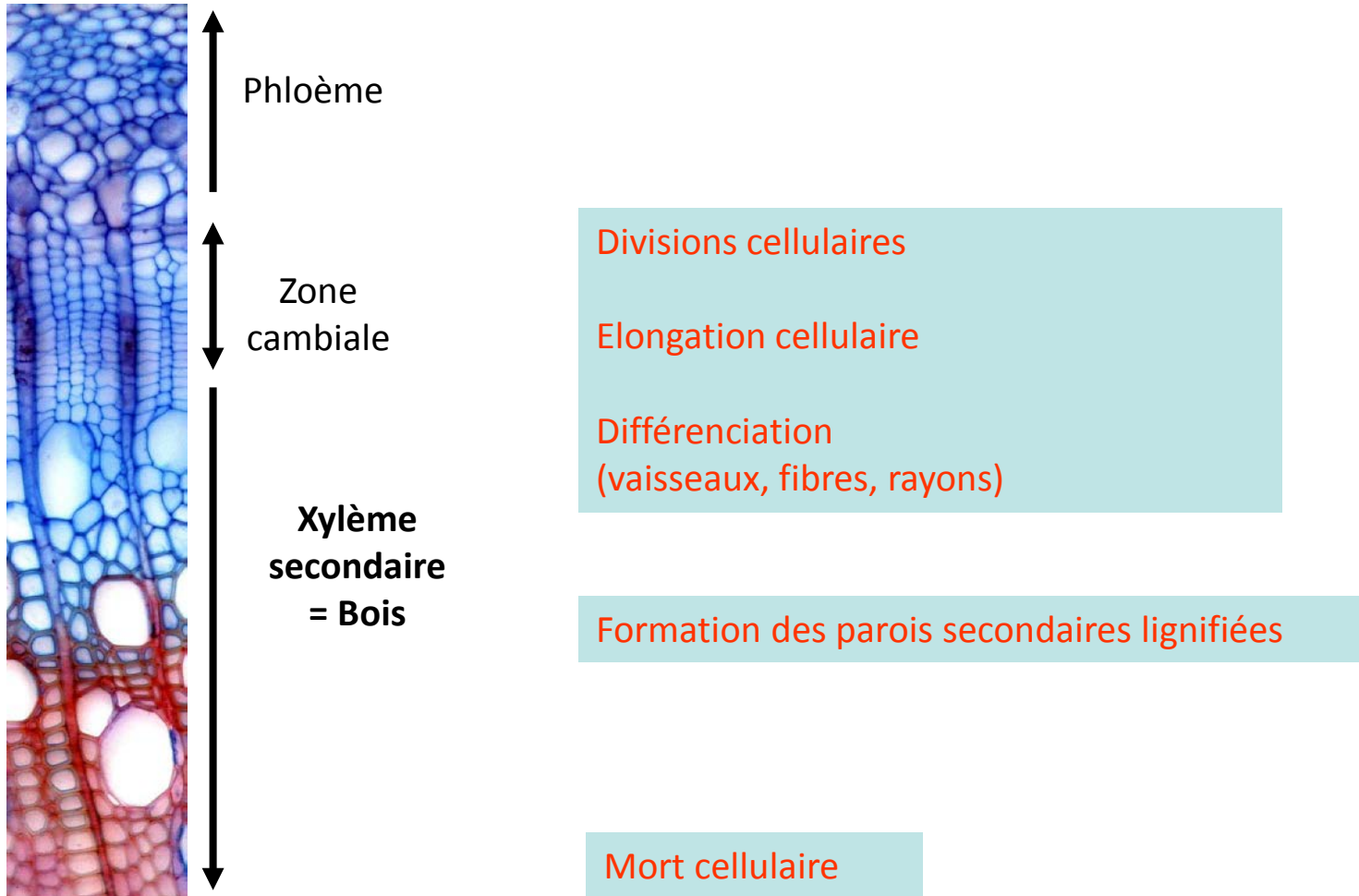
INRA Val de Loire

<https://www6.val-de-loire.inra.fr/biofora>



Equipe « Physiologie moléculaire de la formation du bois »

Objectif général de l'équipe: Elucidation des mécanismes moléculaires impliqués dans la formation du bois et dans la définition de ses propriétés chimiques et mécaniques



Plan du cours

I. Introduction

Objectifs de la transcriptomique?

Définitions / Quelques rappels sur la transcription

II. Méthodes d'analyse du transcriptome

Historique

Evolution des méthodes

Importance de l'échantillonnage

Vers le séquençage Single Cell

III. Applications : apport de la transcriptomique

Expression de gènes

Transcript profiling

Découverte de nouveaux gènes

Apport à la génomique structurale

Détection de polymorphisme de séquences

eQTL

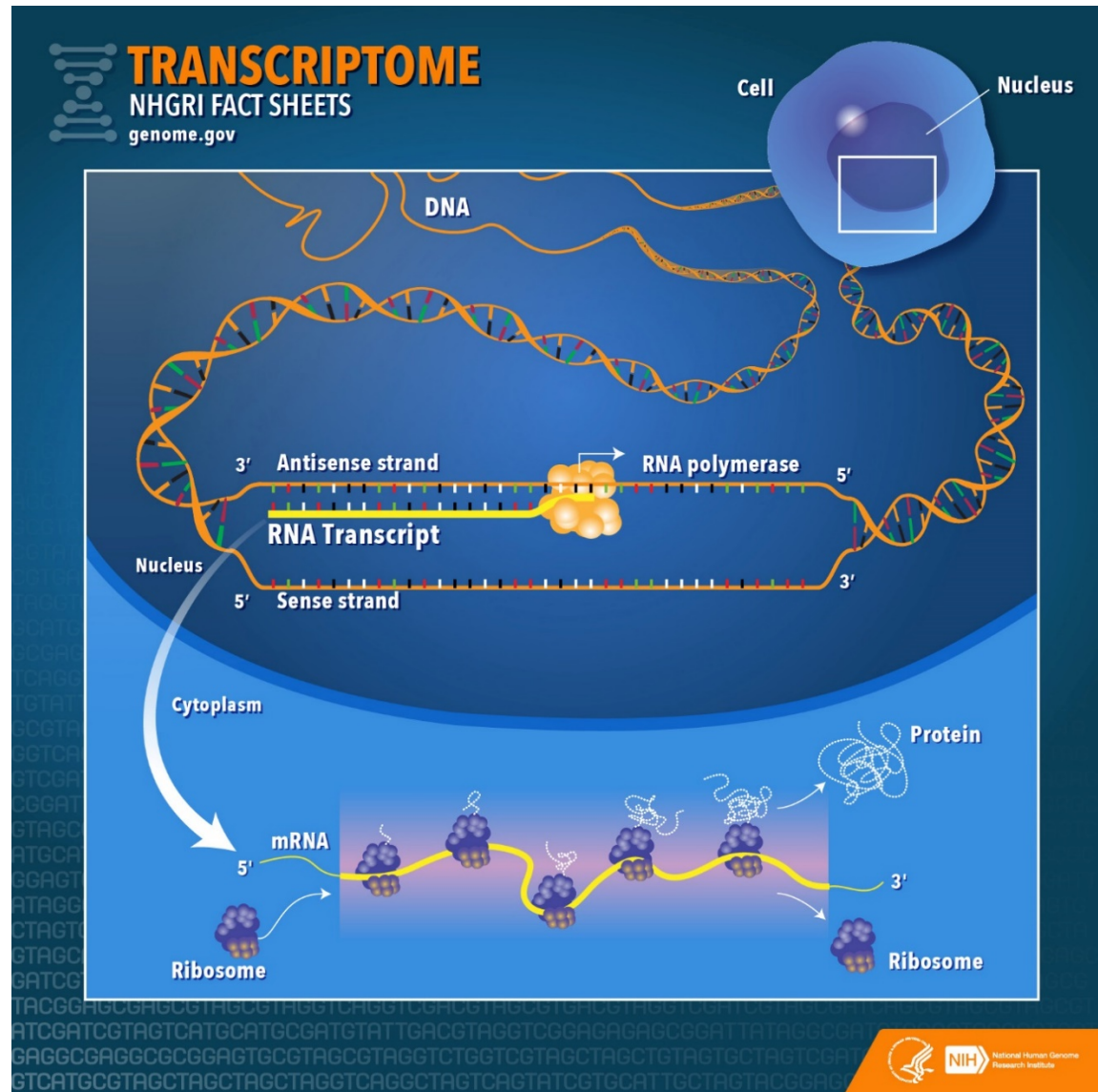
IV. Ressources disponibles

I. Introduction

Transcriptome = ensemble des ARN transcrits présents dans une population de cellules dans une condition donnée

Le transcriptome est multiple

Transcriptome = un instantané des transcrits présents dans l'échantillon considéré au temps t du prélèvement



Objectifs majeurs de la transcriptomique

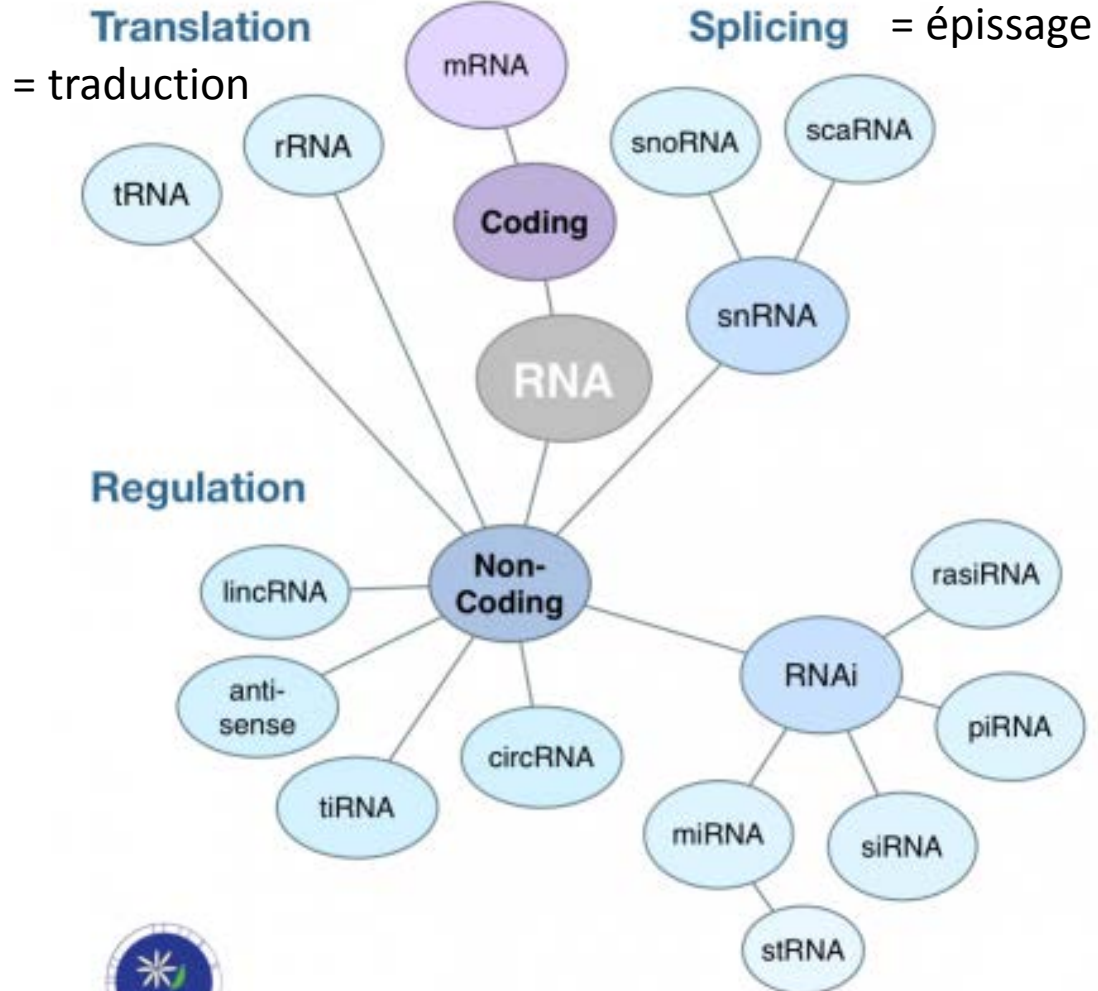
Faire un **catalogue des transcrits** (**ARNm, ARN non codants**), si possible quantitatif, prenant en compte les transcrits alternatifs *Découverte de nouveaux gènes*

Suivre le **changement de niveaux d'expression des transcrits** au cours du développement / en fonction de différentes conditions *Compréhension des processus physiologiques*

Résoudre la **structure des gènes** (identification des régions 5'UTR et 3'UTR, épissage)

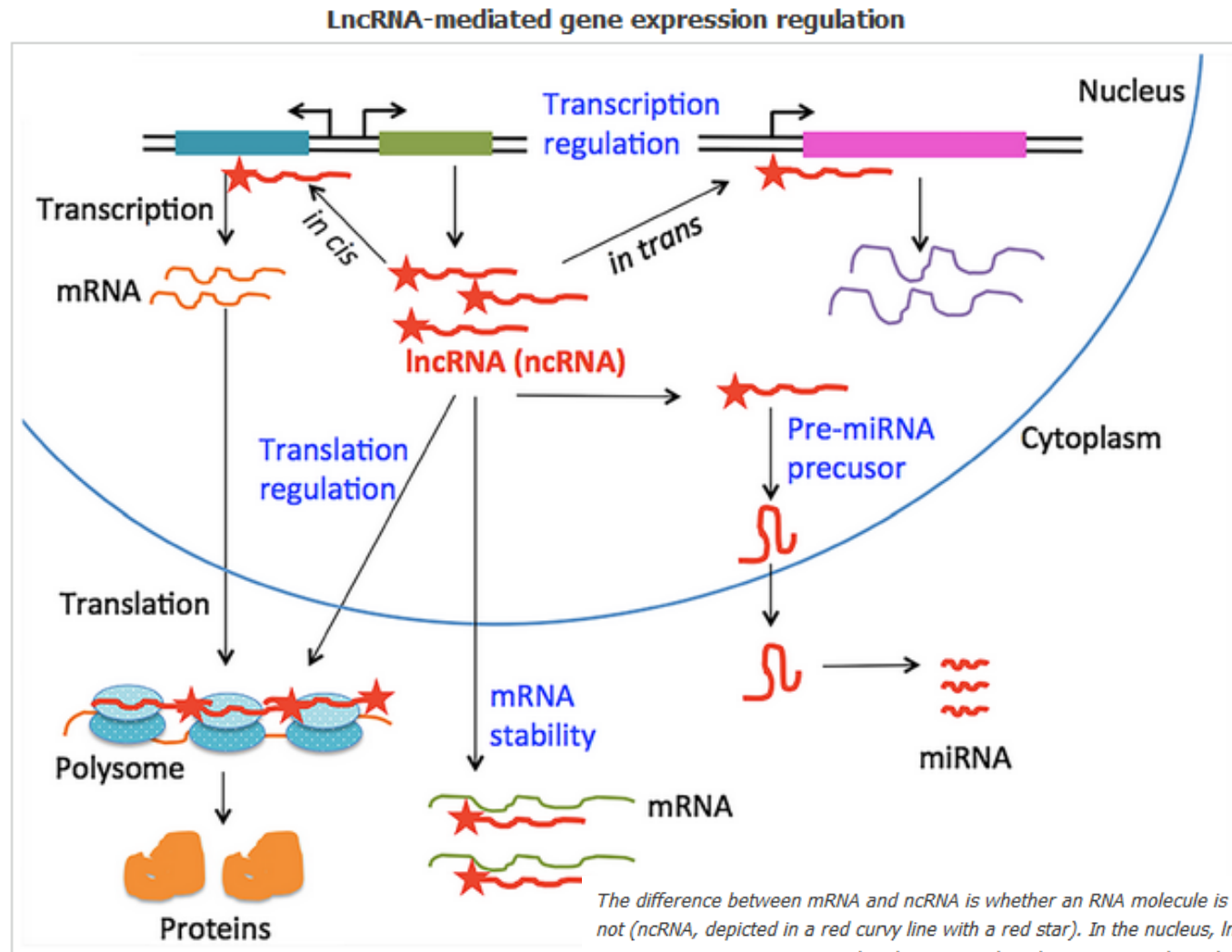
Et beaucoup plus encore....

RNA World



© Digital World Biology 2014

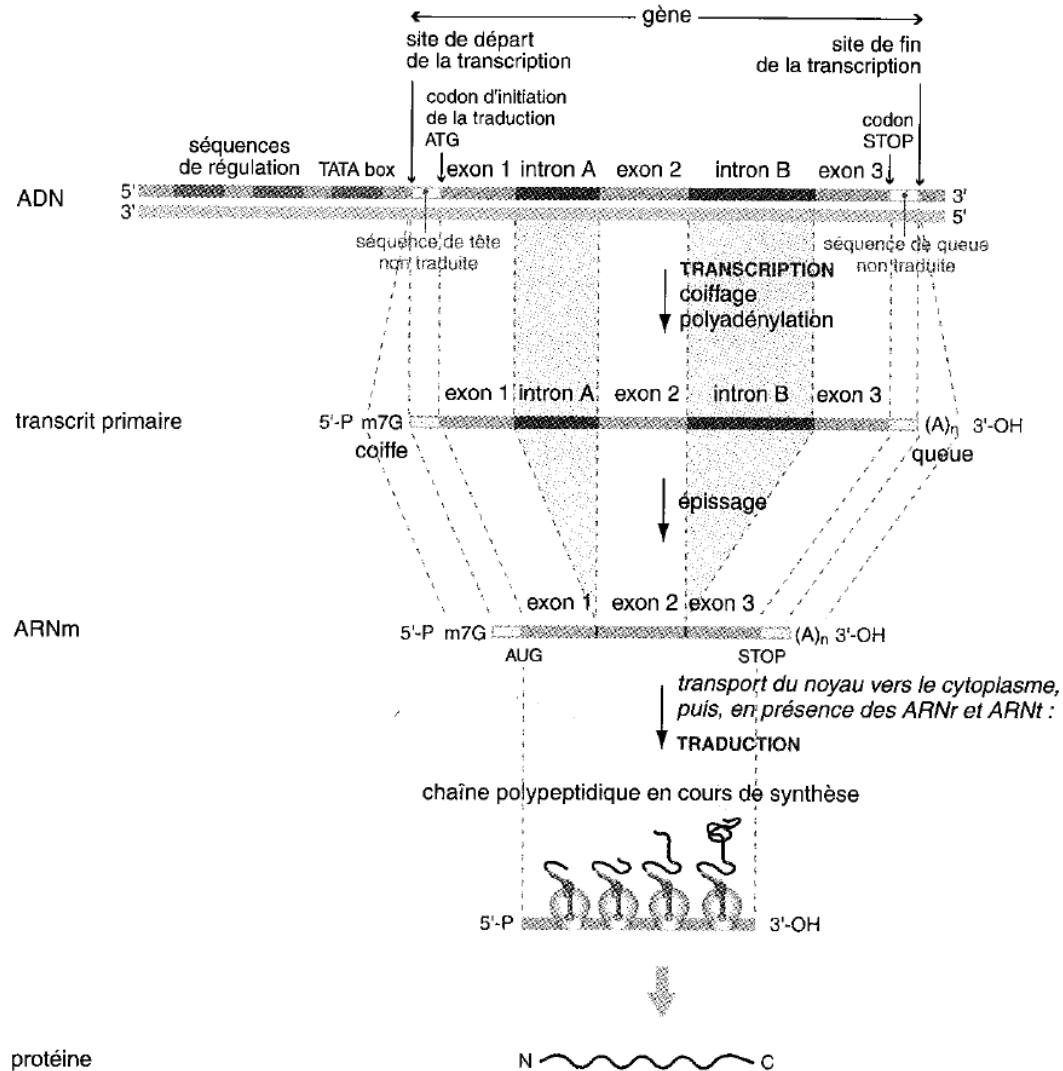
Rôle des ARN non codants – une illustration



The difference between mRNA and ncRNA is whether an RNA molecule is encoded into a protein (mRNA) or not (ncRNA, depicted in a red curly line with a red star). In the nucleus, lncRNAs regulate transcription in cis or in trans. In cis means when lncRNA-mediated transcriptional regulation occurs at a gene(s) on the same chromosome. In contrast, lncRNA-mediated transcription regulation in trans occurs at a gene(s) on a different chromosome. Expressed lncRNAs can stabilize and protect mRNAs from miRNA-induced degradation and also facilitate translation of their client mRNAs in the cytoplasm. Ribosomes are shown in light blue circles aligned on an mRNA in an orange line. In addition, lncRNAs can be processed to generate hairpin-structured pre-miRNAs in nucleus and then transported to the cytoplasm where they become to be matured into miRNAs

Structure et expression d'un gène eucaryote

Tagu et Moussard, 2003



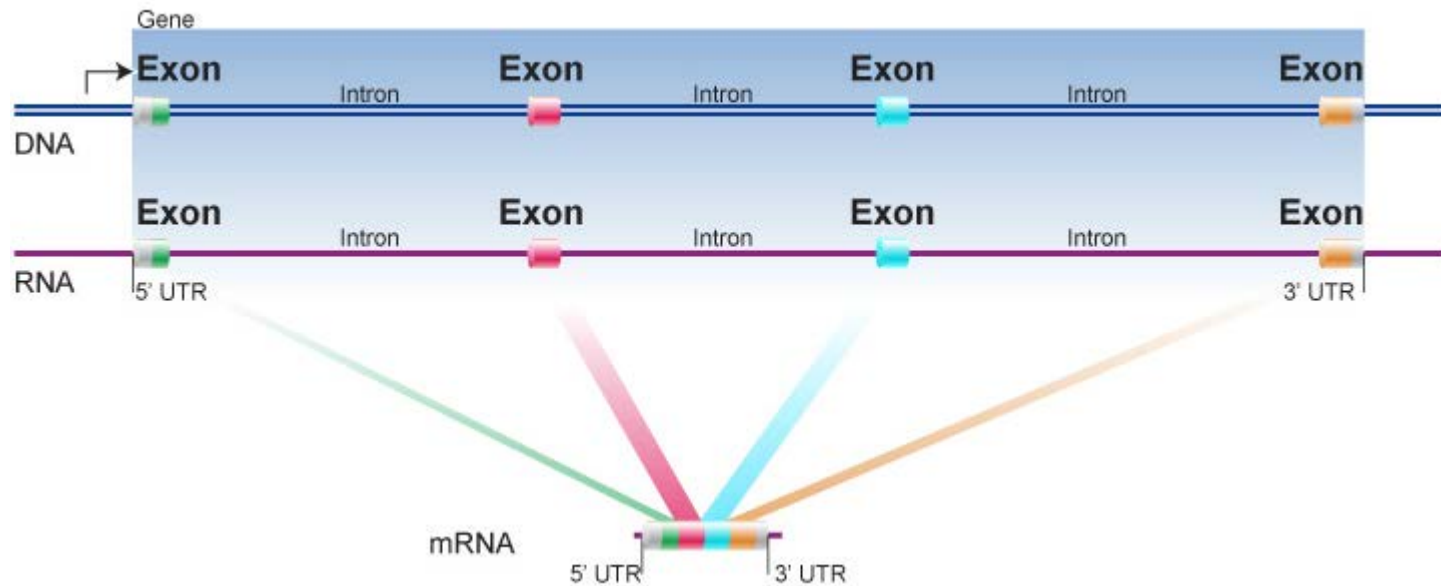
Séquence de tête non traduite = 5'UTR

Séquence de queue non traduite = 3'UTR

Exons = segments d'un précurseur ARN qui sont conservés dans l'ARN après épissage et que l'on retrouve dans l'ARN mature dans le cytoplasme.

Introns = par opposition, les segments du précurseur ARN qui sont éliminés lors de l'épissage s'appellent des introns

Structure et expression d'un gène eucaryote



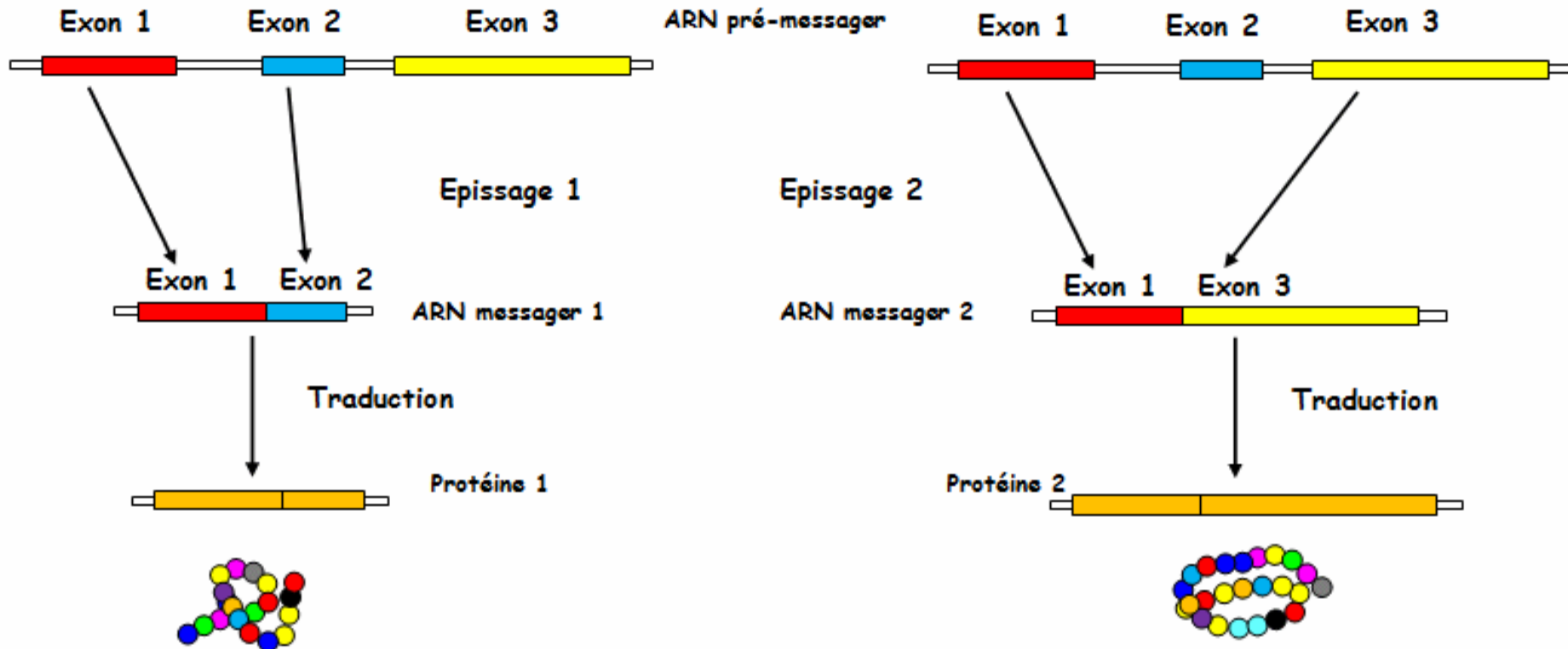
Exons = segments d'un précurseur ARN qui sont conservés dans l'ARN après épissage et que l'on retrouve dans l'ARN mature dans le cytoplasme.

Introns = par opposition, les segments du précurseur ARN qui sont éliminés lors de l'épissage s'appellent des introns

5'-UTR = 5' Untranslated Transcribed Region = correspond à la portion de l'ARN messager (ARNm) placée en amont du codon d'initiation de la traduction.

3'-UTR = en aval du codon stop

Epissage alternatif



1 gène \longrightarrow potentiellement plusieurs transcrits alternatifs

II. Méthodes d'analyse du transcriptome

Figure 1 | A historical timeline of transcriptomics. ▶

Illustrated is the lockstep development of experimental and computational aspects of transcriptomics. Advances in the experimental protocols for the high-throughput profiling of RNA necessitate the development of databases to catalogue the results and trigger curation efforts to define reference transcriptomes. However, these endeavours depend on the development of accurate and scalable computational methods to search, quantify and assemble RNA molecules. Within each field, the most influential, seminal or unique references were selected. AceView, a gene annotation resource²⁶⁷; ArrayDB, a database of microarray gene expression data²⁶⁸; ArrayExpress, a public repository for microarray gene expression data⁷⁸; BLAST, Basic Local Alignment Search Tool⁷³; CAGE, cap analysis of gene expression²⁶⁹; CEL-seq, cell expression by linear amplification and sequencing⁴⁹; CGAP, Cancer Genome Anatomy Project²⁷⁰; CIBERSORT, a tool for estimating the abundances of cell types in a mixed cell population²⁶⁰; dbEST, a database for expressed sequence tags⁵⁸; EdgeR, a package for differential expression analysis¹⁷⁰; EMBL, European Molecular Biology Laboratory; Ensembl, a genome browser for vertebrate genomes⁷⁴; ESTs, expressed sequence tags²⁸; FANTOM5, Functional Annotation of the Mammalian Genome 5 (REF. 271); FASTA, a text format for representing nucleotide or peptide sequences⁷²; GenBank, the US National Institutes of Health (NIH) genetic sequence database; GENCODE, the genome annotation project of the Encyclopedia of DNA Elements (ENCODE)²⁷²;

GenomeSpace, a cloud-based resource for integrative genomics analyses⁸³; GEO, Gene Expression Omnibus⁷⁷; GSEA, gene set enrichment analysis²⁷³; InsilicoDB, a database of microarray and RNA-seq data⁸²; Known Genes, a resource of RNA and protein data²⁷⁴; Limma, Linear Models for Microarray Data¹⁶⁷; MiTranscriptome, a human RNA-seq database⁷⁶; Mitelman, Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer²⁷⁵; MPSS, massively parallel signature sequencing⁴²; Oncomine, a cancer microarray database and integrated data mining platform¹⁸⁰; qPCR, quantitative PCR²⁹; RACE, rapid amplification of cDNA ends²⁷⁶; RefSeq, NCBI Reference Sequence Database⁷⁵; RNAscope, an *in situ* hybridization assay for RNA detection²⁷⁷; RNA-seq, RNA sequencing; RNA-seq 454, RNA sequencing using the 454 (Roche) pyrosequencing platform⁴⁴; RNA-seq SBS, RNA sequencing using sequencing-by-synthesis platforms²⁷⁸; RT-qPCR, reverse transcription quantitative PCR³⁰⁻³²; SAGE, serial analysis of gene expression³⁶; SAGEmap, SAGE tag to gene mapping²⁷⁹; Sailfish, a transcript isoform quantification tool²⁸⁰; SAM, Significance Analysis of Microarrays²⁸¹; Smith–Waterman, a local sequence alignment algorithm⁷⁰; STAR, Spliced Transcripts Alignment to a Reference²⁸²; SymAtlas, gene expression and annotation resource, now superseded by BioGPS²⁸³; TACO, Transcriptome Assemblies Combined into One (a consensus transcriptome tool)²⁸⁴; TopHat and Cufflinks, software tools for RNA-seq alignment and transcriptome assembly⁵; Trans-ABYSS, Transcript Assembly By Short Sequences²⁸⁵; Trinity, a tool for *de novo* assembly of RNA-seq data²⁸⁶; UMI, unique molecular identifier⁴⁸; Xena, a genomic data mining and analysis portal²⁸⁷.

Avant la transcriptomique ... dans les années 80...

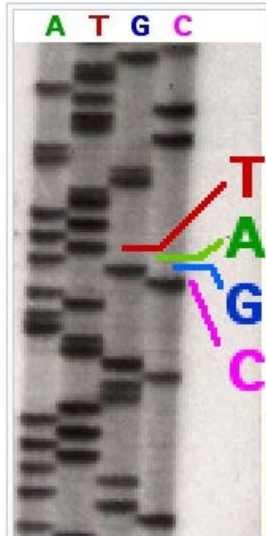
Qques gènes étudiés à la fois

Extraction d'ARNm et Northern blot

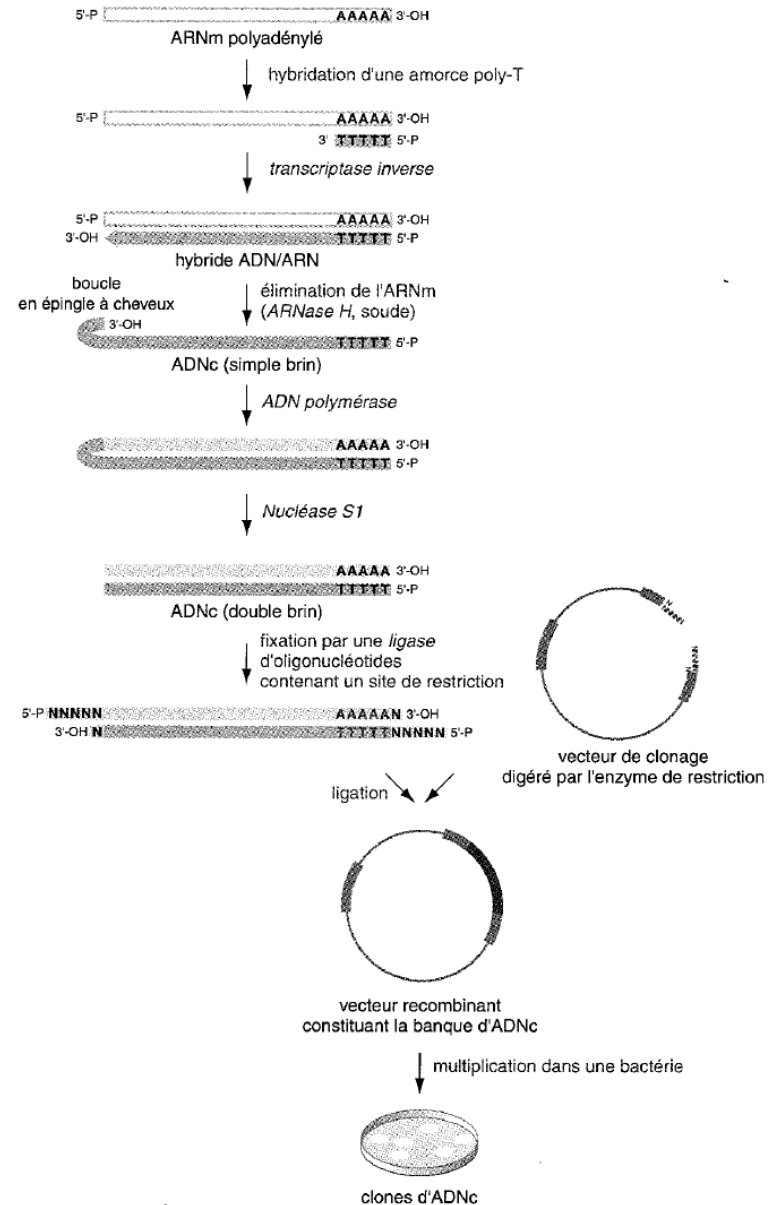
Transcriptase inverse : synthèse d'ADNc à partir d'ARNm, plus grande stabilité

Possibilité de construire des banques d'ADNc

Séquençage de type Sanger



Résultat du séquençage par la méthode de Sanger. L'ordre de chaque bande indique la position d'un nucléotide A,T,C ou G



A partir des années 90

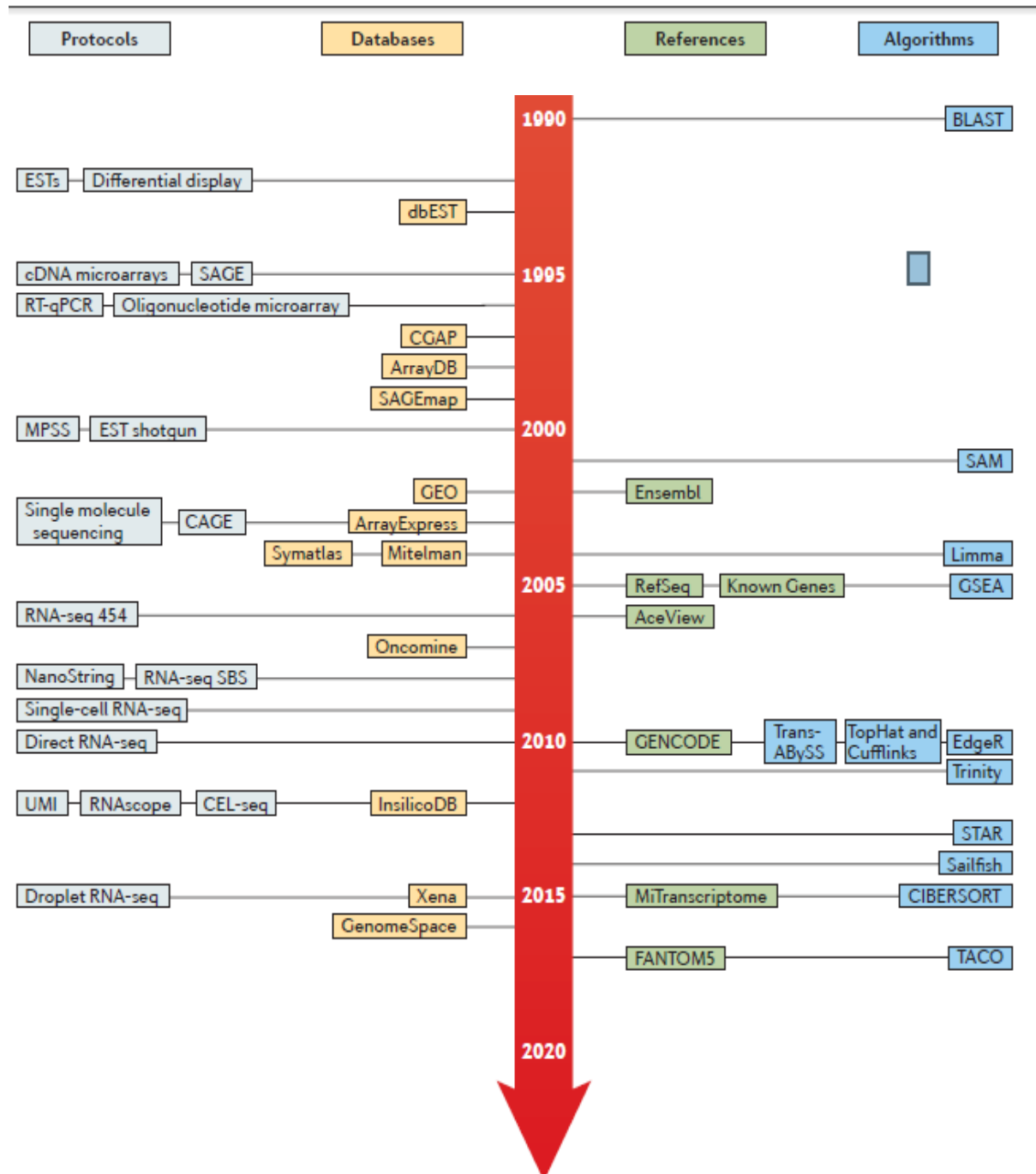
Développement concomittant de méthodes de biologie moléculaire, d'algorithmes pour traiter les données de séquences, de bases de données pour stocker les séquences, de sites conservant des données de référence

Banques d'EST

SAGE / CAGE

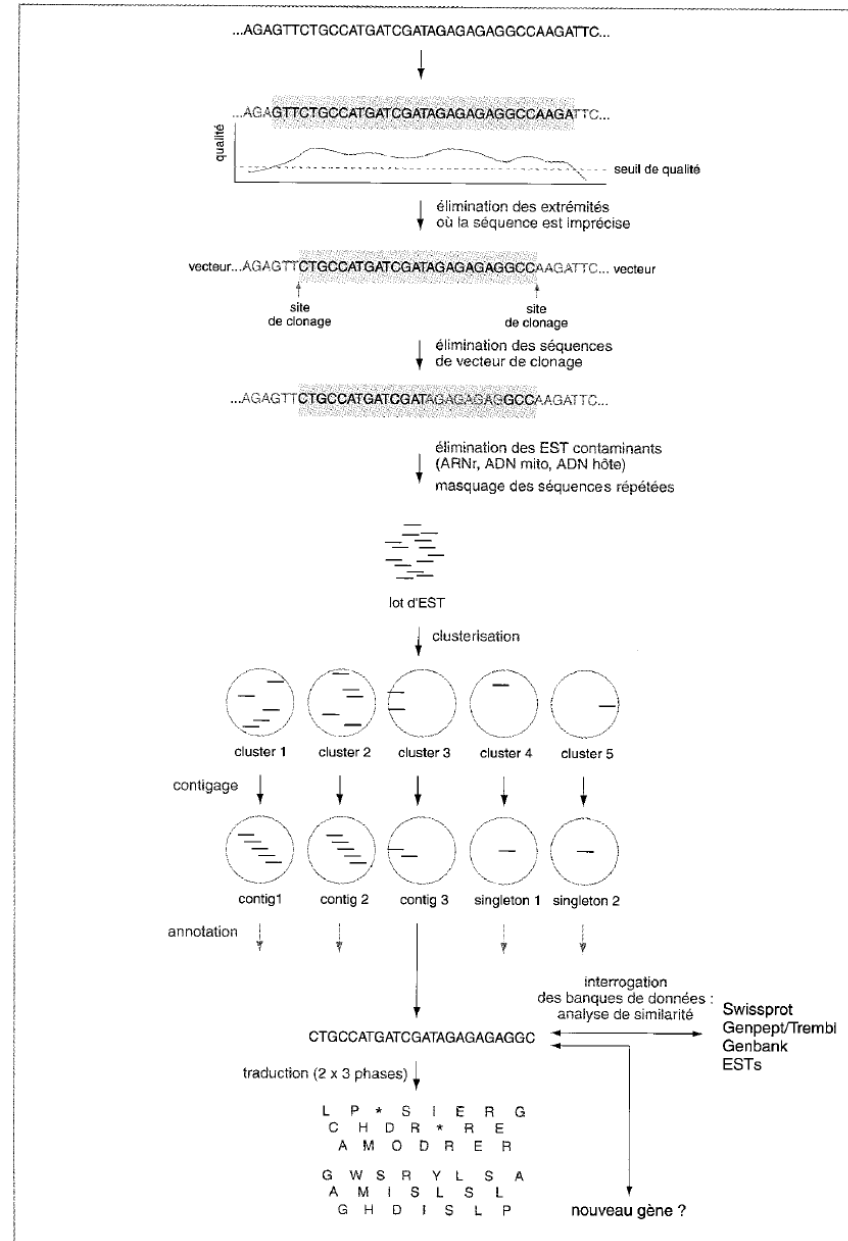
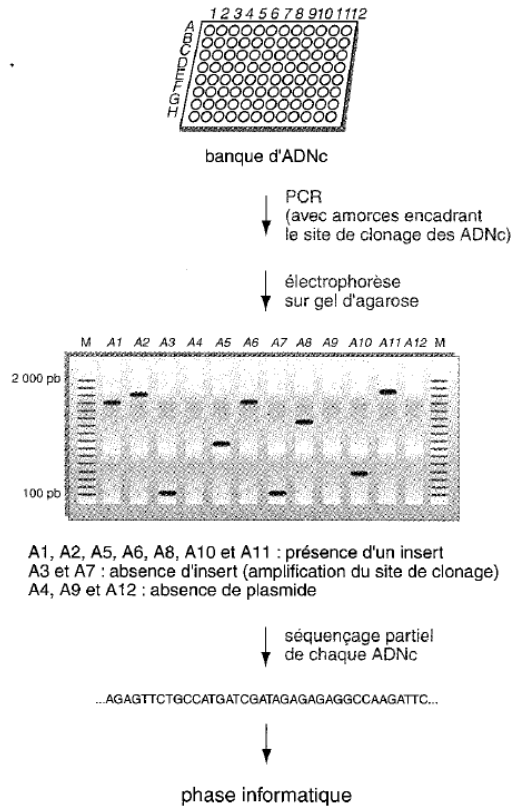
Microarrays

RNAseq



Banques d'EST

Extraction d'ARNm provenant de tissus / traitements différents → ADNc et construction de banques d'ADNc



EST = Expressed Sequence Tag = Etiquette de Séquence Transcrite

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 9693–9698, August 1998
Plant Biology

Analysis of xylem formation in pine by cDNA sequencing

ISABEL ALLONA*[†], MICHELLE QUINN*, ELIZABETH SHOOP[‡], KRISTI SWOPE[‡], SHEILA ST. CYR[‡], JOHN CARLIS[‡], JOHN RIEDL[‡], ERNEST RETZEL[‡], MALCOLM M. CAMPBELL*[§], RONALD SEDEROFF*, AND ROSS W. WHETTEN*

*Forest Biotechnology Group, Department of Forestry, North Carolina State University, Raleigh, NC 27695-8008; and [†]Computational Biology Center, University of Minnesota, Minneapolis, MN 55455-0312

1998

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 13330–13335, October 1998
Plant Biology

Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags

(cambium/forestry/functional genomics/xylem/xylogenesis)

FREDRIK STERKY*[†], SHARON REGAN^{†‡}, JAN KARLSSON[§], MAGNUS HERTZBERG[‡], ANTJE ROHDE[¶], ANDERS HOLMBERG*, BAHRAM AMINI*, RUPALI BHALERAO[§], MAGNUS LARSSON*, RAIMUNDO VILLARROEL[¶], MARC VAN MONTAGU[¶], GÖRAN SANDBERG[‡], OLOF OLSSON[¶], TUULA T. TEERI*, WOUT BOERJAN[¶], PETTER GUSTAFSSON[§], MATHIAS UHLÉN*, BJÖRN SUNDBERG^{‡**}, AND JOAKIM LUNDEBERG*^{**}

*Department of Biotechnology, Kungl Tekniska Högskolan, Royal Institute of Technology, SE-10044 Stockholm, Sweden; [†]Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-90183 Umeå, Sweden; [‡]Department of Plant Physiology, Umeå University, SE-90187 Umeå, Sweden; [§]Laboratorium voor Genetica, Department of Genetics, Flanders Interuniversity Institute for Biotechnology, Universiteit Gent, B-9000 Gent, Belgium; and [¶]Department of Cell and Molecular Biology, Lundberg Laboratory, Göteborg University, Box 462, SE-40530 Göteborg, Sweden

SAGE = Serial Analysis of Gene Expression

Lowe et al., 2017

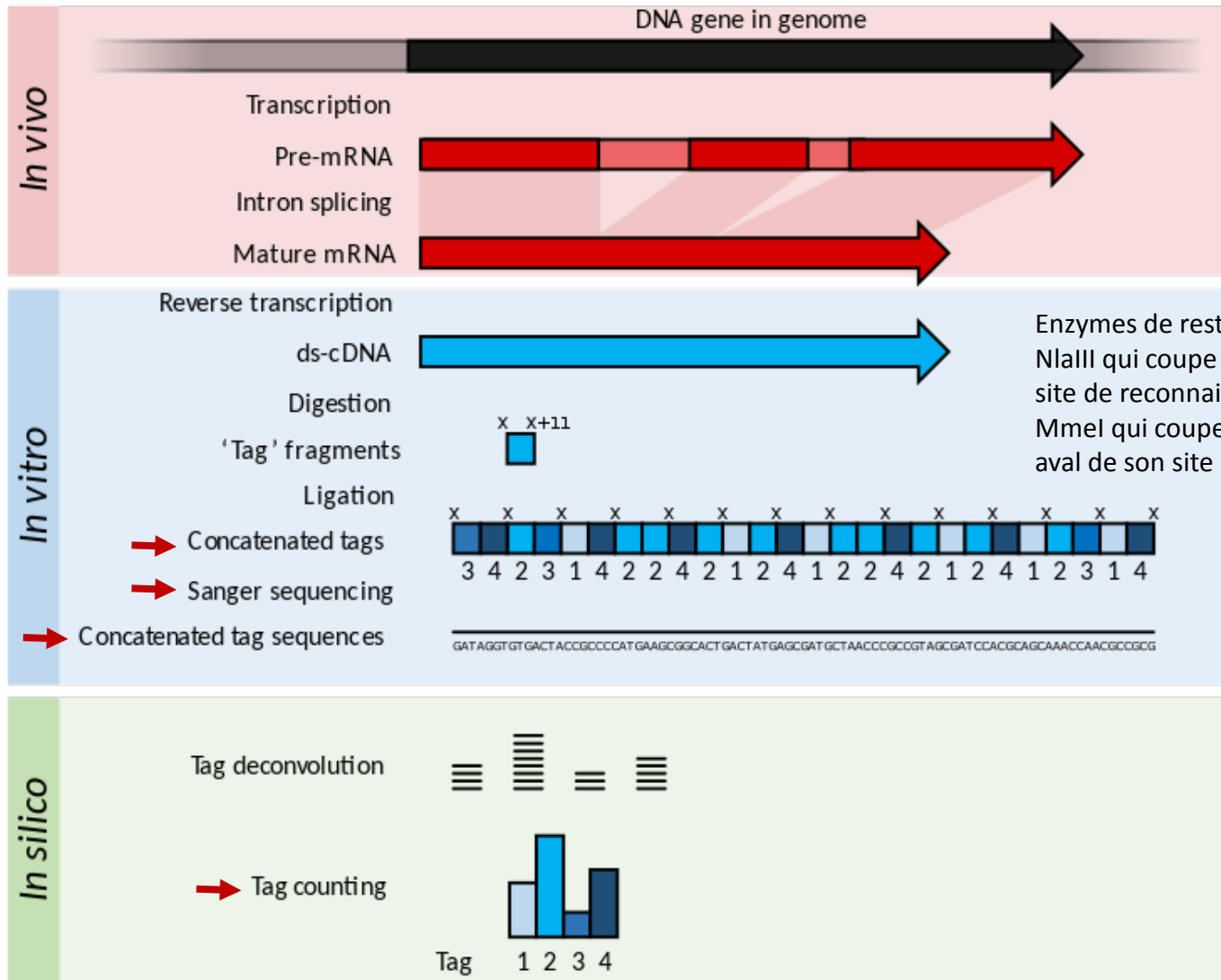
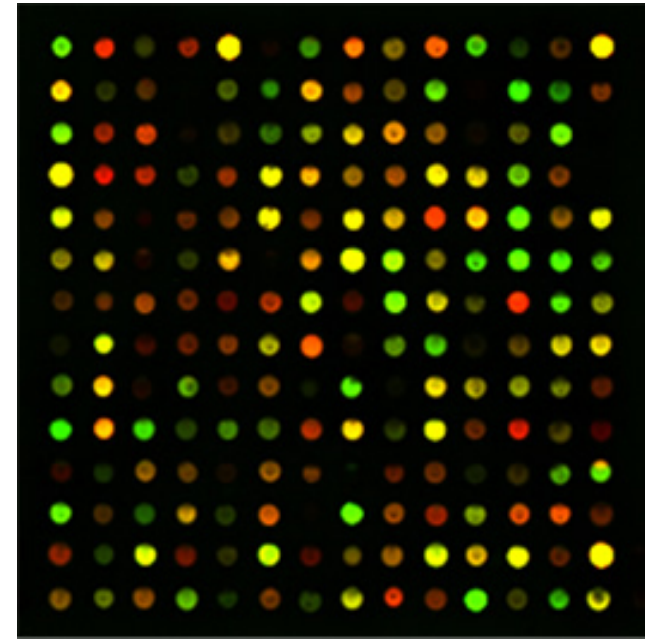
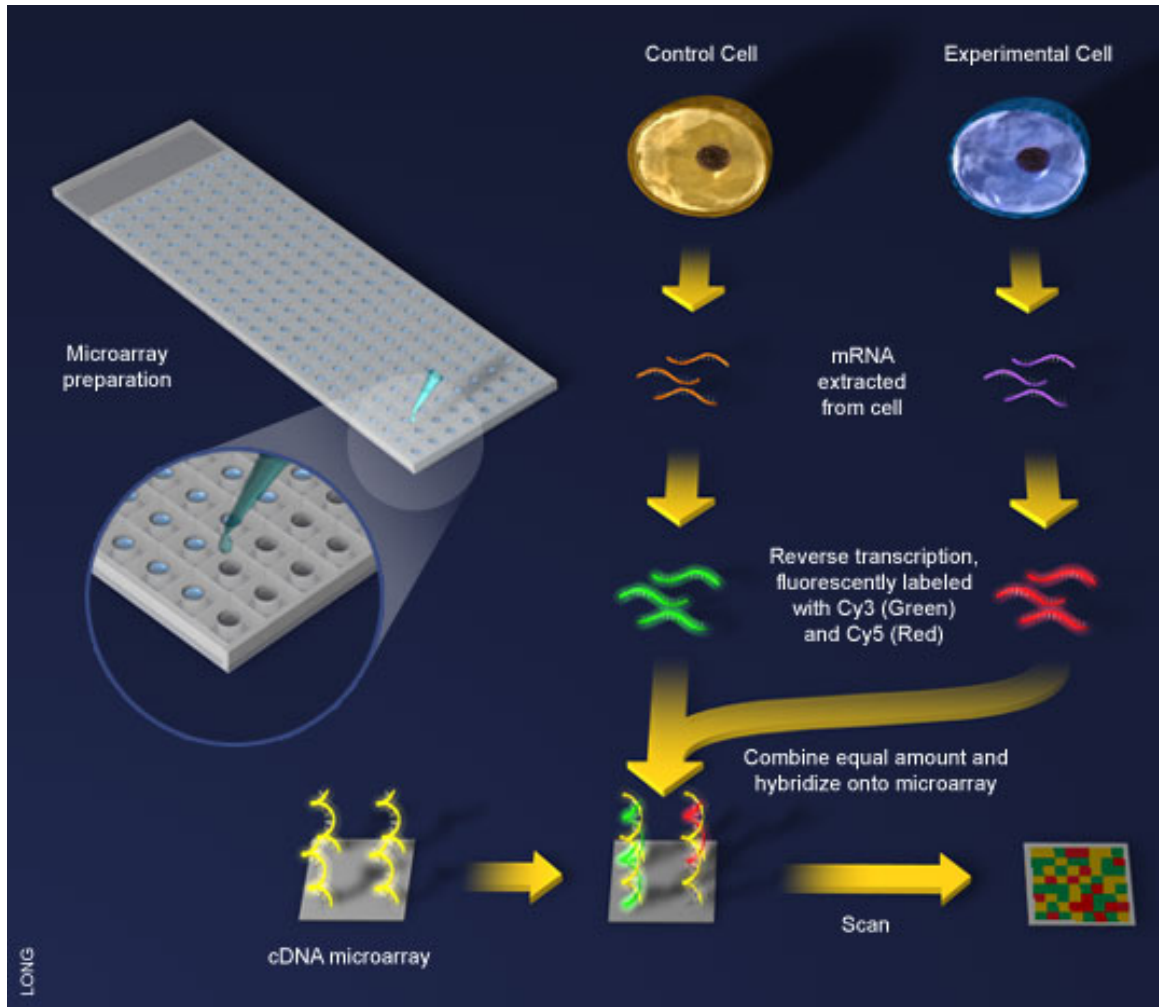


Fig 2. Summary of SAGE. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, and reverse transcriptase is used to copy the mRNA into stable double-stranded-cDNA (ds-cDNA; blue). In SAGE, the ds-cDNA is digested by restriction enzymes (at location “X” and “X”+11) to produce 11-nucleotide “tag” fragments. These tags are concatenated and sequenced using long-read Sanger sequencing (different shades of blue indicate tags from different genes). The sequences are deconvoluted to find the frequency of each tag. The tag frequency can be used to report on transcription of the gene that the tag came from.

Microarrays ou puces à ADN



Microarrays ou puces à ADN

Lowe et al., 2017

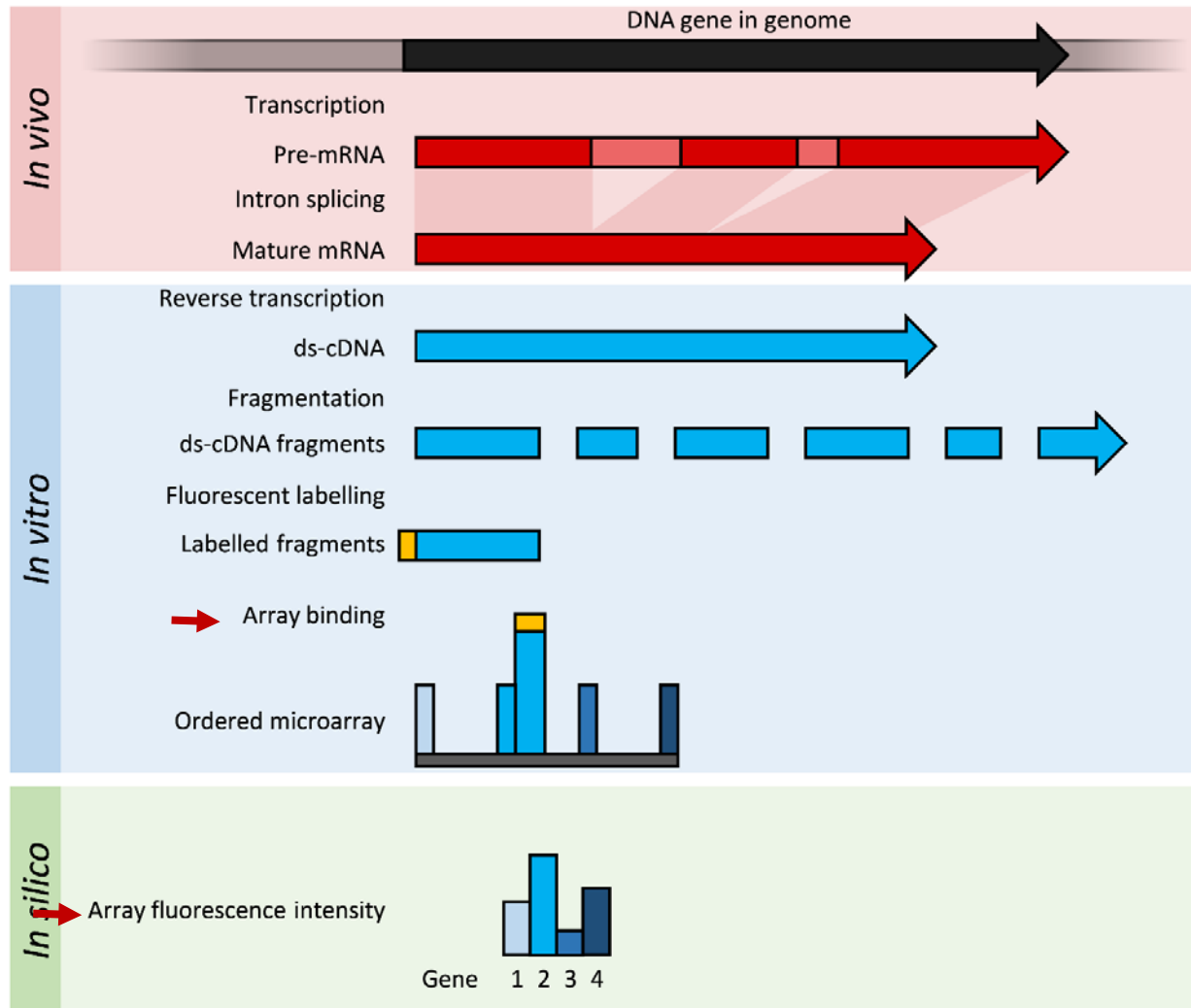
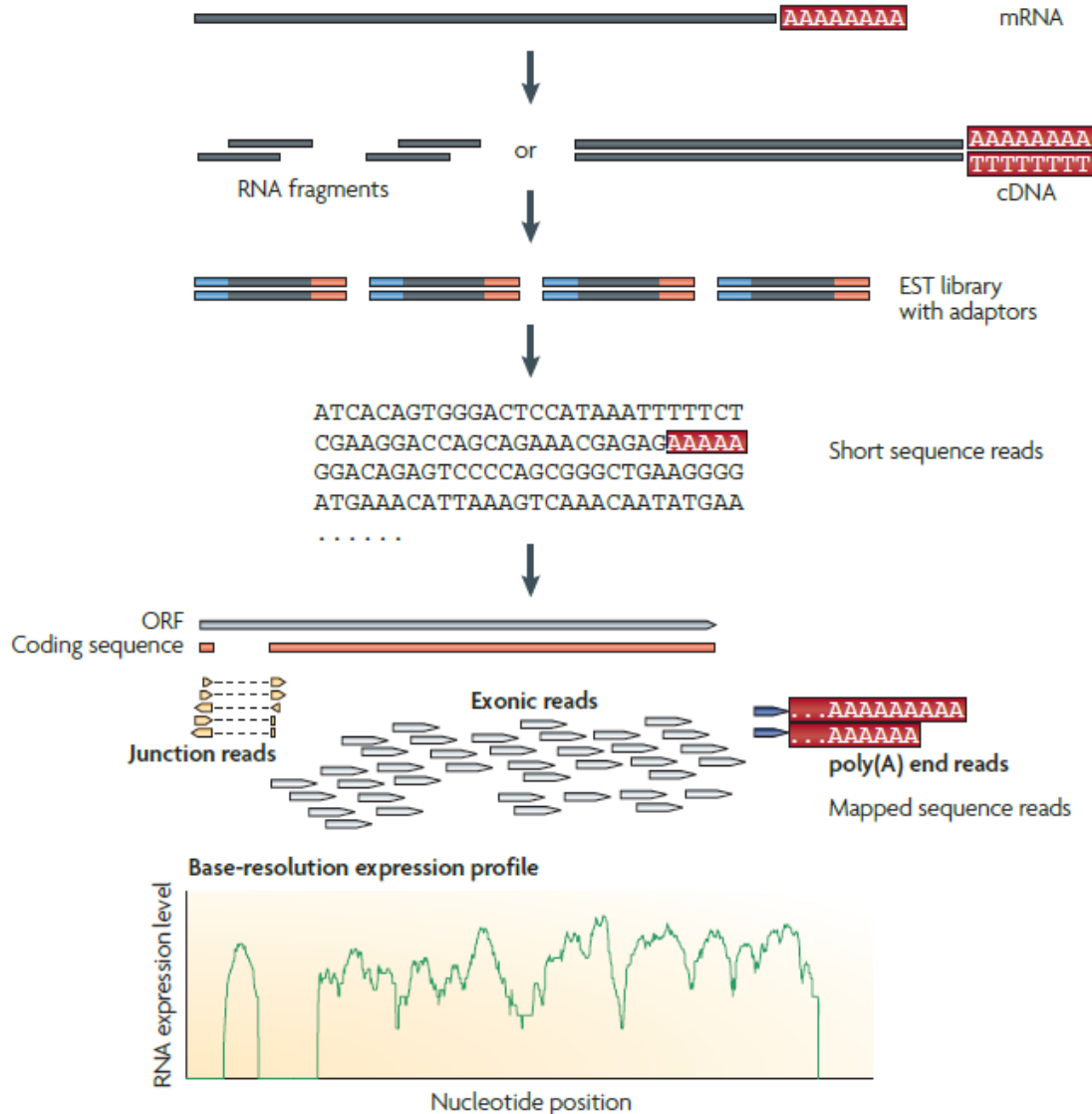


Fig 3. Summary of DNA microarrays. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism and reverse transcriptase is used to copy the mRNA into stable double-stranded-cDNA (ds-cDNA; blue). In microarrays, the ds-cDNA is fragmented and fluorescently labelled (orange). The labelled fragments bind to an ordered array of complementary oligonucleotides, and [measurement of fluorescent intensity](#) across the array indicates the abundance of a predetermined set of sequences. These sequences are typically specifically chosen to report on genes of interest within the organism's genome.

RNA sequencing = RNAseq

Rendu possible par le développement du séquençage NGS = Next Generation Sequencing



Banques ou librairies avec des fragments de taille identique

Séquences = reads = 50 – 150 pb

Alignement sur un génome de référence = mapping

ou assemblage *de novo* si pas de séquence génomique

RNAseq

Lowe et al., 2017

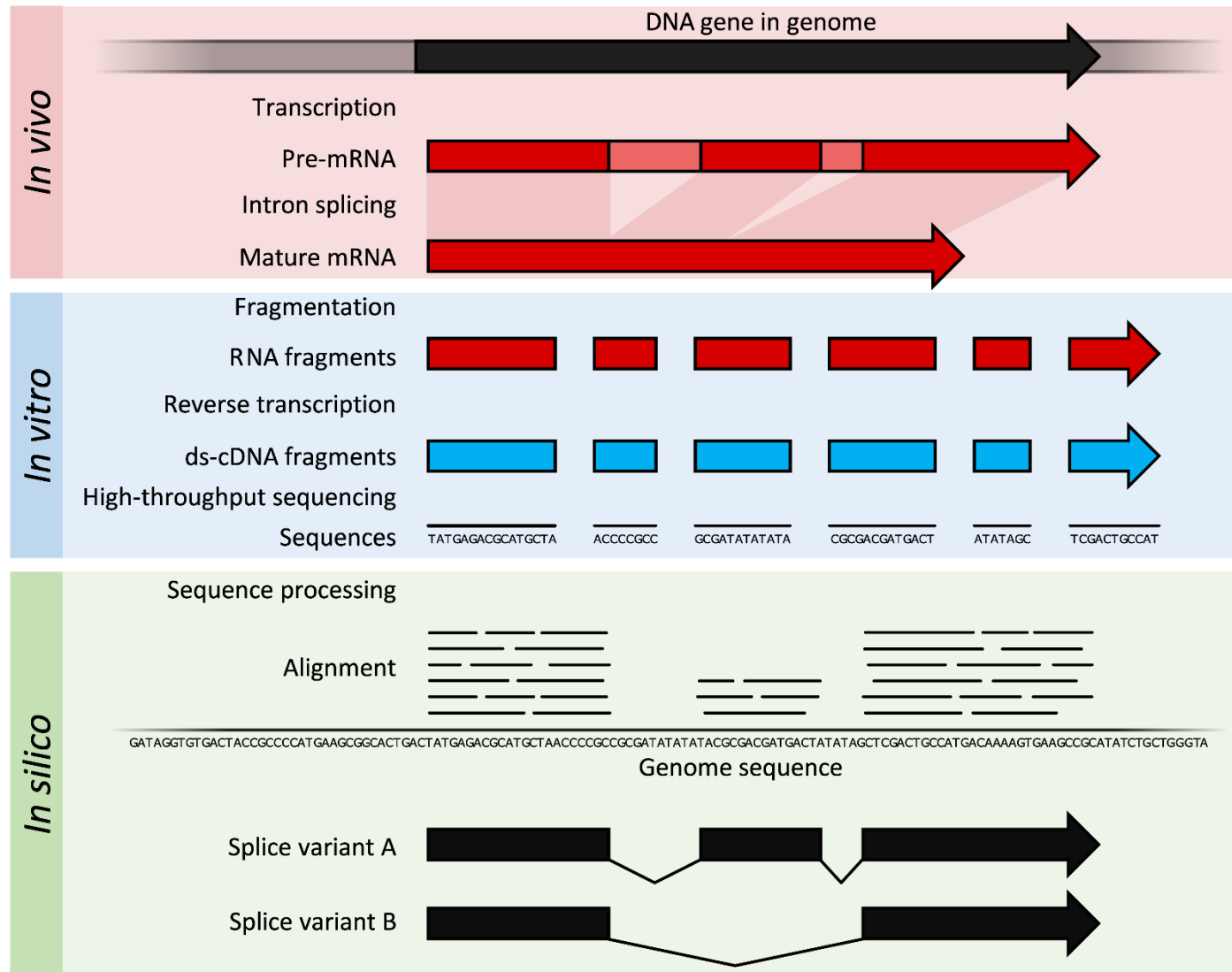
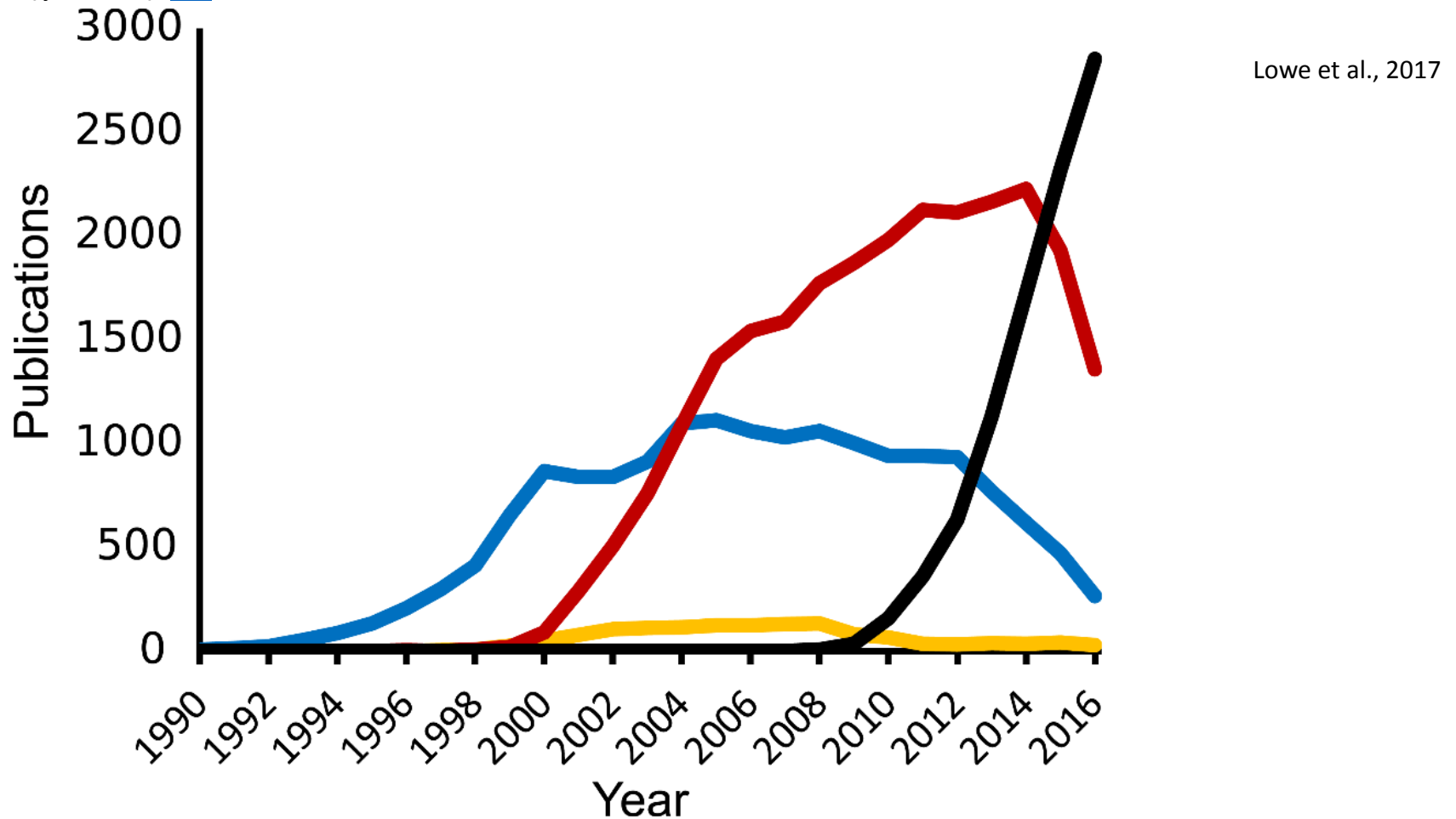


Fig 4. Summary of RNA sequencing. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, fragmented and copied into stable double-stranded-cDNA (ds-cDNA; blue). The ds-cDNA is sequenced using **high-throughput**, short-read sequencing methods. These sequences can then be **aligned** to a reference genome sequence to reconstruct which genome regions were being transcribed. These data can be used to annotate where expressed genes are, their relative expression levels, and any alternative splice variants.

Fig 1

Transcriptomics method use over time.

Published papers since 1990, referring to RNA sequencing (black), RNA microarray (red), expressed sequence tag (blue), and serial/cap analysis of gene expression (yellow)[12].



Activité 1

Analyse de l'introduction de l'article Wang et al., 2009, Nature reviews, 10, 57 - 63

INNOVATION

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, Mark Gerstein and Michael Snyder

Méthodes	Microarrays	Banques EST	SAGE	RNAseq
Principe	hybridation	Séquençage Sanger	Séquençage Sanger	Séquençage NGS
Intérêts	<p>Pas cher</p> <p>Résolution de qqes bases à 100 pb</p> <p>Haut-débit pour applications en génotypage, moyen débit pour applications en physiologie</p>	<p>Méthode sans a priori qui peut permettre de trouver de nouveaux gènes</p> <p>Pas besoin de génome de référence.</p> <p>Assemblage de novo permet d'avoir un transcriptome de référence pour le cas des espèces dont le génome n'est pas séquencé.</p>	<p>Haut débit</p> <p>Résultats précis et quantitatifs</p>	<p>Seule méthode qui donne vraiment accès au transcriptome complet, à haut-débit et de façon quantitative</p> <p>Pas besoin de génome ou transcriptome de référence.</p> <p>Résolution à la base près.</p> <p>A la fois adapté pour gènes faiblement et très fortement exprimés.</p>
Limites	<p>Besoin d'avoir la connaissance d'un génome / transcriptome de référence pour construire la puce : ne permet donc pas d'identifier de nouveaux gènes.</p> <p>Problème de bruit de fond dues aux hybridations croisées, qui limite la détection des gènes faiblement exprimés</p> <p>Problème de saturation du signal fluorescent, non linéaire, pas adapté pour gènes fortement exprimés.</p> <p>La normalisation des expériences est complexe, il est difficile de comparer les expériences entre elles.</p>	<p>Débit faible, cher, pas quantitatif</p>	<p>Moyennement cher</p> <p>Ne permet pas de détecter isoformes ou transcrits alternatifs</p> <p>Besoin d'un génome ou transcriptome de référence</p>	

Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
<i>Technology specifications</i>			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
<i>Application</i>			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
<i>Practical issues</i>			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang et al., 2009

Table 1. Comparison of contemporary methods [23] [24] [10].

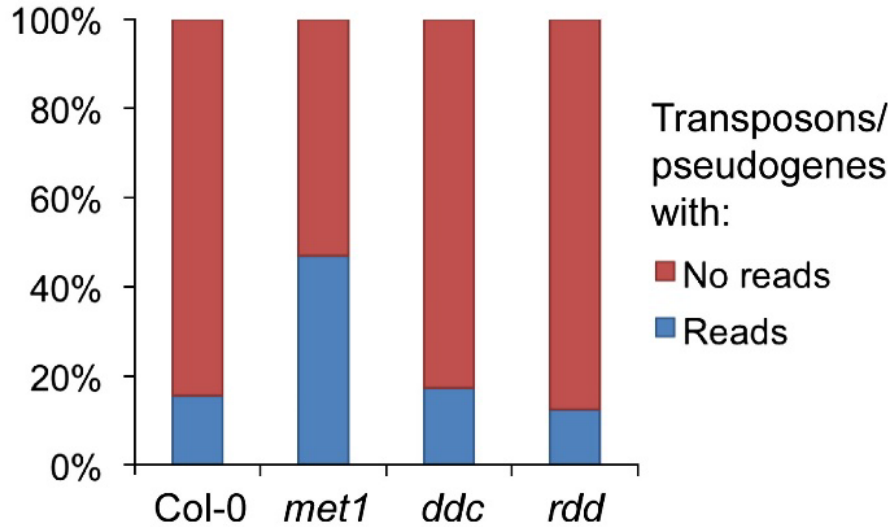
Method	RNA-Seq	Microarray
Throughput	High [10]	Higher [10]
Input RNA amount	Low ~ 1 ng total RNA [25]	High ~ 1 µg mRNA [26]
Labour intensity	High (sample preparation and data analysis) [10][23]	Low [10][23]
Prior knowledge	None required, though genome sequence useful [23]	Reference transcripts required for probes [23]
Quantitation accuracy	~90% (limited by sequence coverage) [27]	>90% (limited by fluorescence detection accuracy) [27]
Sequence resolution	Can detect SNPs and splice variants (limited by sequencing accuracy of ~99%) [27]	Dedicated arrays can detect splice variants (limited by probe design and cross-hybridisation) [27]
Sensitivity	10^{-6} (limited by sequence coverage) [27]	10^{-3} (limited by fluorescence detection) [27]
Dynamic range	$>10^5$ (limited by sequence coverage) [28]	$10^3 - 10^4$ (limited by fluorescence saturation) [28]
Technical reproducibility	>99% [29][30]	>99% [31][32]

RNA-Seq, RNA Sequencing

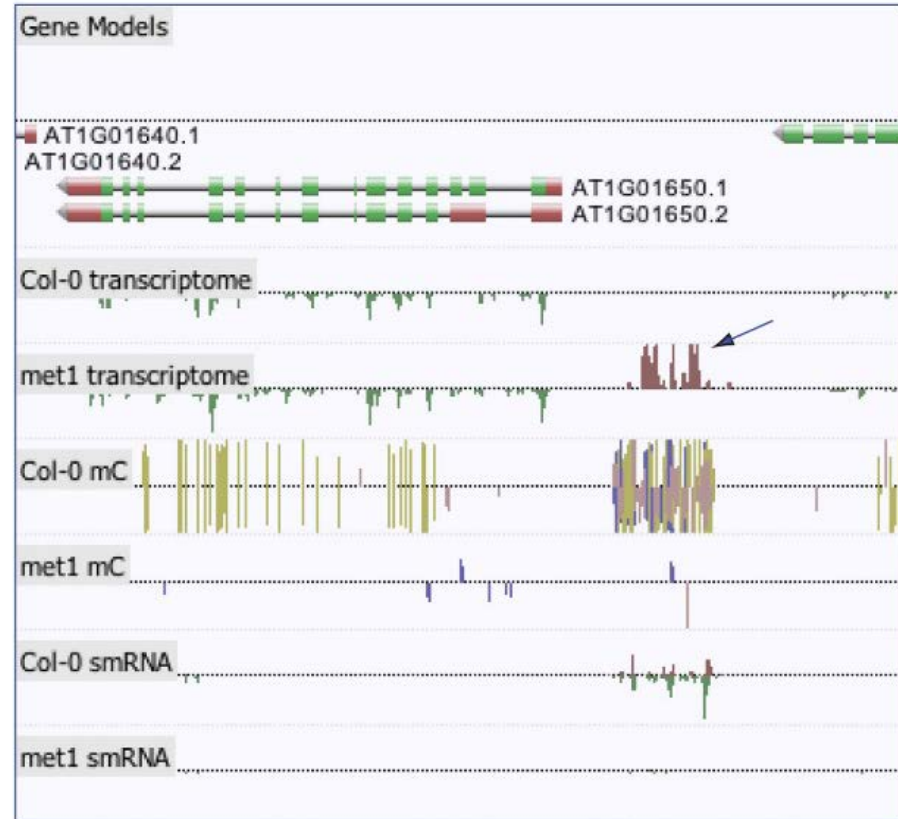
<https://doi.org/10.1371/journal.pcbi.1005457.t001>

Puissance du RNAseq pour découvrir de nouveaux gènes

A



C

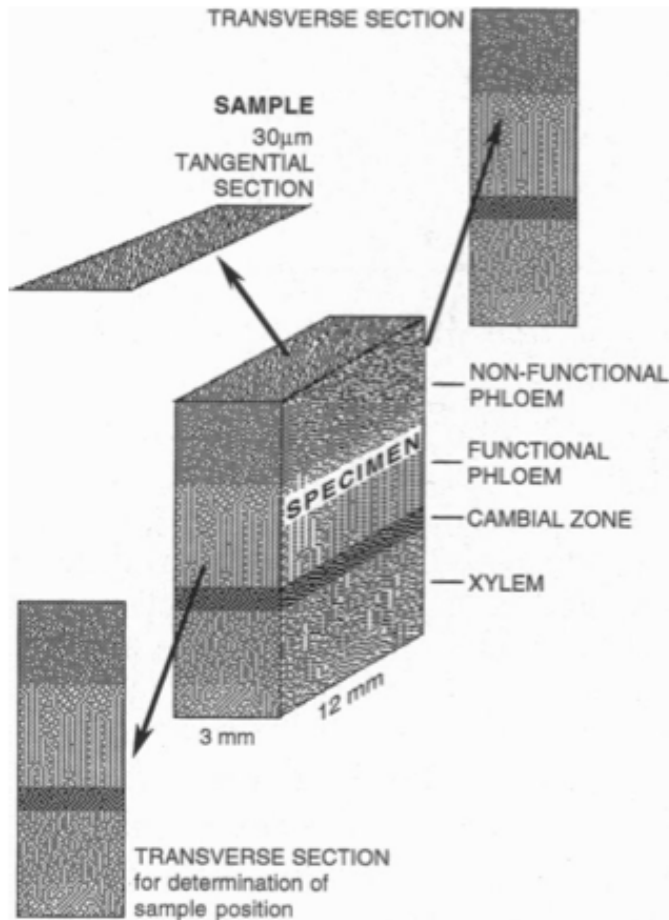


Lister et al., 2008

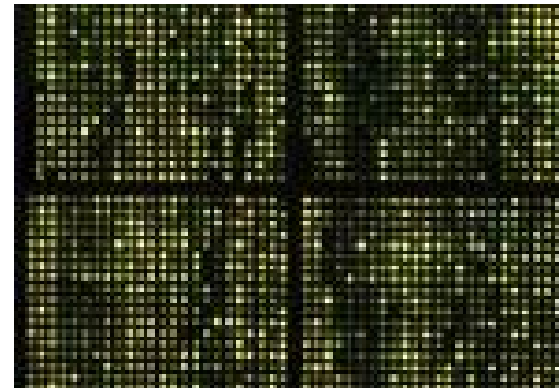
Ex : Arabidopsis

Importance de l'échantillonnage

Gene expression patterning along the developmental stages of xylogenesis

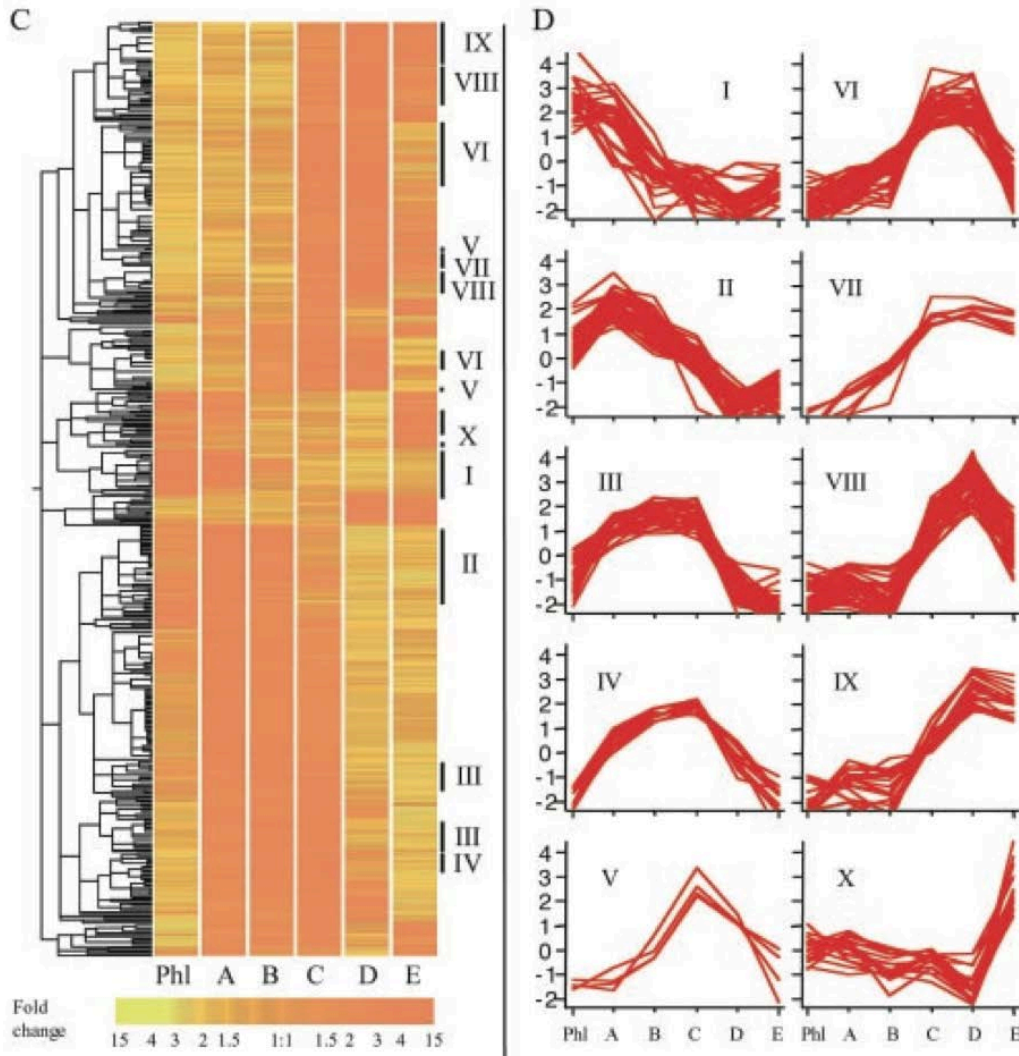


Cryosectioning of the vascular cambium



Microarray based on a unigene set of 2,995 cDNA clones

Hierarchical cluster analysis of 1,791 selected genes with differential expression in the tissues



A & B: cell fate and cell identity (ATHB-8 & ATHB-9 homologues)

A-C:

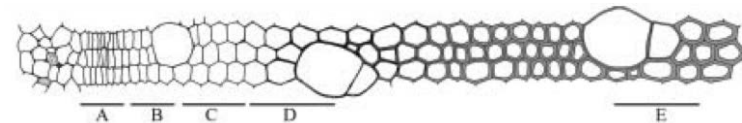
- cell-cycle machinery

- Cyclins, cyclin-dependent kinases (still expressed in late expansion zone)

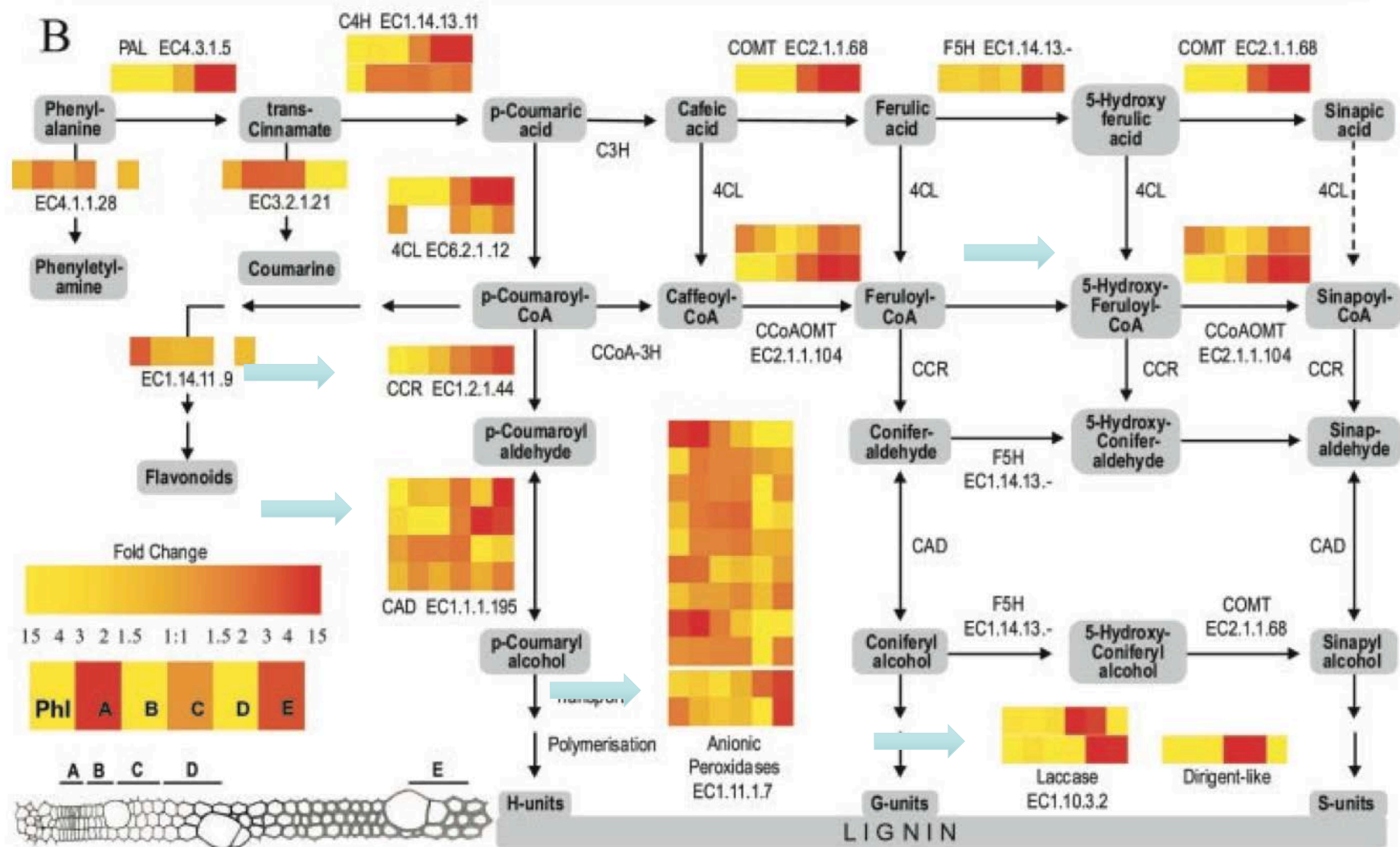
- Cell expansion (pectin metabolism, protein synthesis and 6 MYBs transcription factors)

C-E: biosynthesis of secondary cell wall

- Tubulins, 2 CesA, KOR, polygalacturonase (pectins remodelling), Glycosyl-transferases, a specific SCW UDP glucose dehydrogenase (hemicellulose synthesis)



Lignin biosynthesis

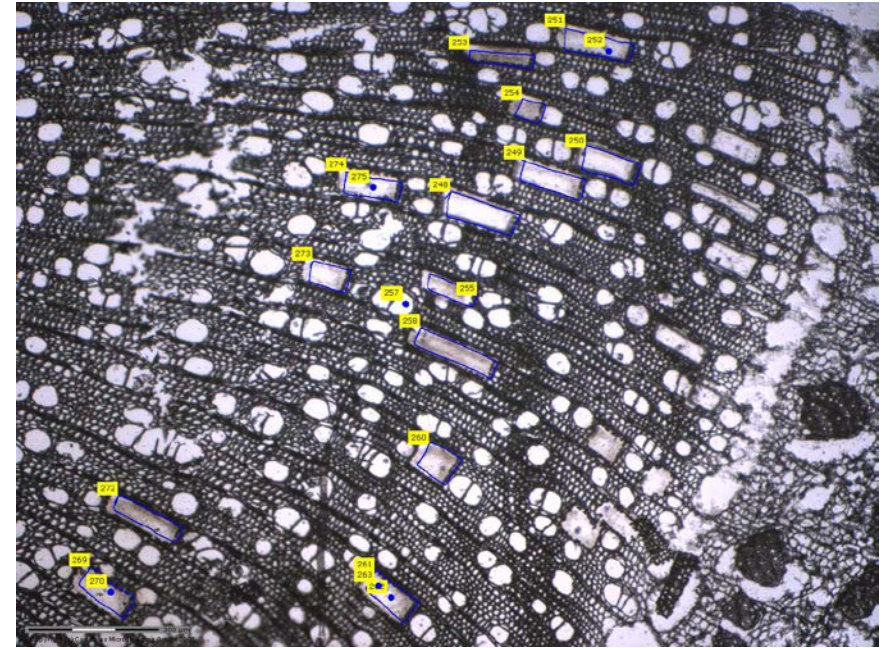
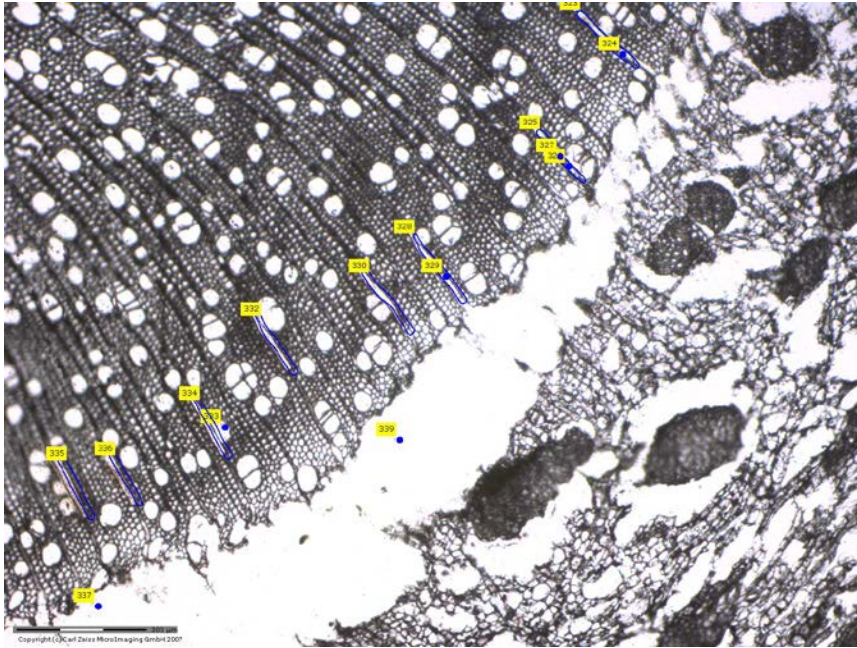
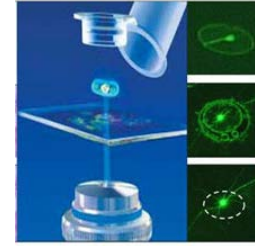


Echantillonnage de types cellulaires par microdissection laser

= Micro - méthode permettant de sélectionner dans un tissu complexe des types cellulaires ou des groupes de cellules.

Plateforme Cytologie et Imagerie Végétale de Versailles

PALM Zeiss Le laser pulsé (355 nm) découpe sur commande la région sélectionnée puis catapulte la zone découpée vers le capuchon d'un tube positionné à quelques millimètres à l'aplomb des cellules.



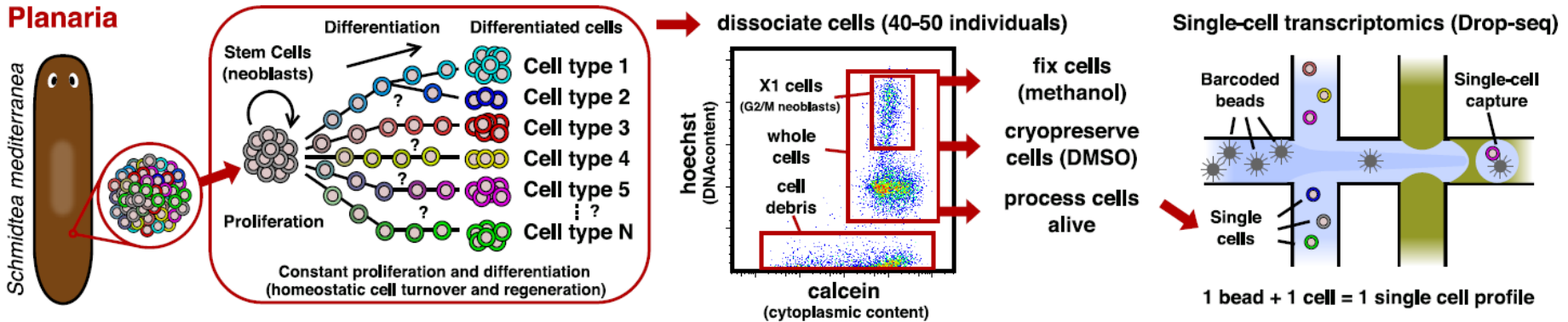
Extraction ARN : mise au point sur des cryocoupes lyophilisées, manuellement macro-disséquées sous stéréomicroscope
RNAseq sur fibres et rayons de bois de peuplier

En cours d'analyse...

Apport du séquençage cellule unique / single cell

Plass et al., 2018

A



B

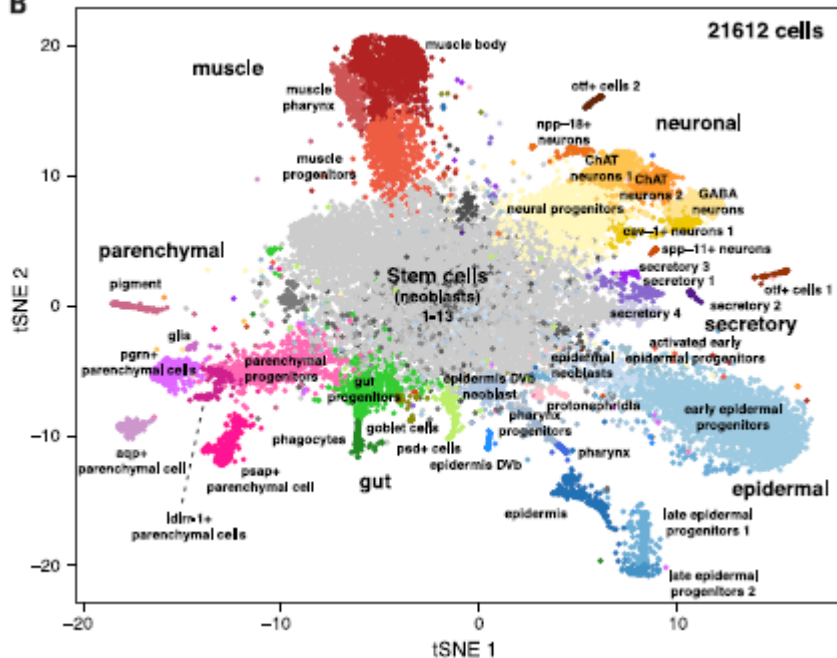
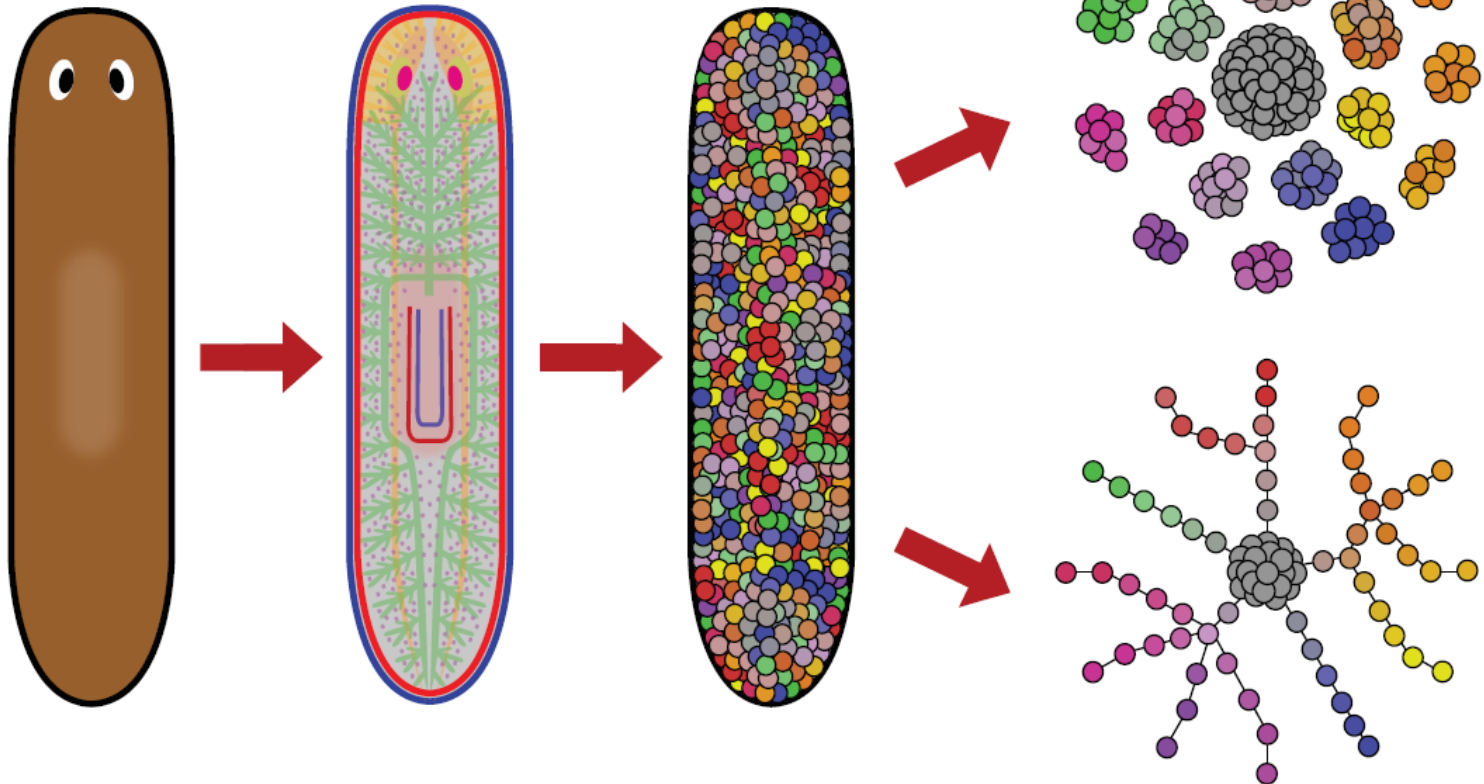


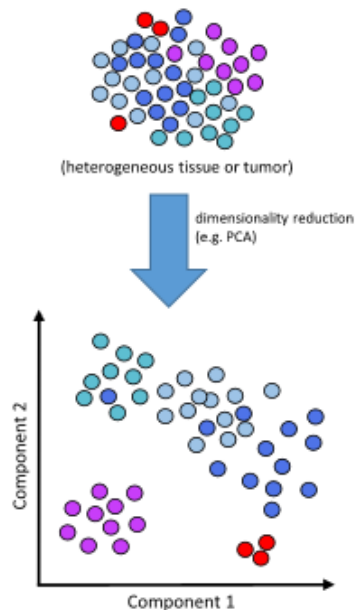
Fig. 1. Cell type atlas by single-cell transcriptomics. (A) Experimental workflow. (B) tSNE representation of the single-cell transcriptomics data with clusters colored according to the expression of previously published marker genes as follows: gray, neoblasts; orange, neuronal lineage; red, muscle; purple, secretory; blue, epidermal lineages; pink, protonephridia; green, gut; magenta, parenchymal lineages. (C) Proportions of cell types identified by Baguñà and

L'algorithme t-SNE (t-distributed stochastic neighbor embedding) est une technique de réduction de dimension pour la visualisation de données = une méthode non-linéaire permettant de représenter un ensemble de points d'un espace à grandes dimensions dans un espace de deux ou trois dimensions, les données peuvent ensuite être visualisées avec un nuage de points.

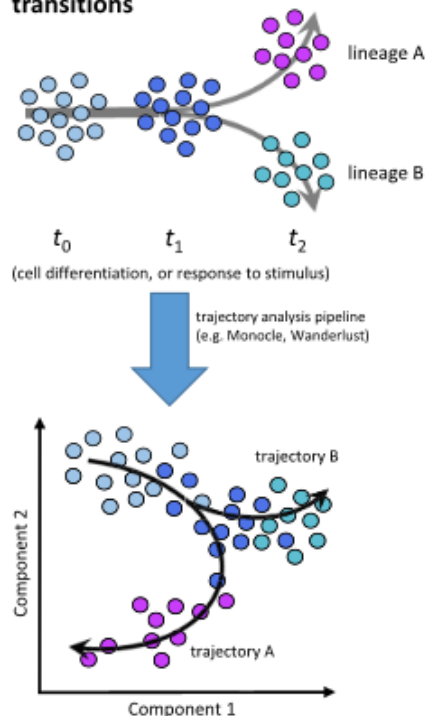


A lineage tree for complex animals from single-cell transcriptomics. Planarians are multicellular organisms. They contain adult pluripotent stem cells that continuously renew all tissues and differentiate into all adult cell types. Using single-cell transcriptomics, we characterized all major mature cell types and many intermediate cellular states. We then derived a lineage tree describing planarian stem cell differentiation into all mature cell types of the animal.

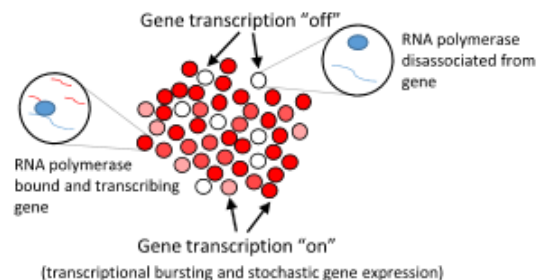
a) Deconvolving heterogeneous cell populations



b) Trajectory analysis of cell state transitions



c) Dissecting transcription mechanics



d) Network inference

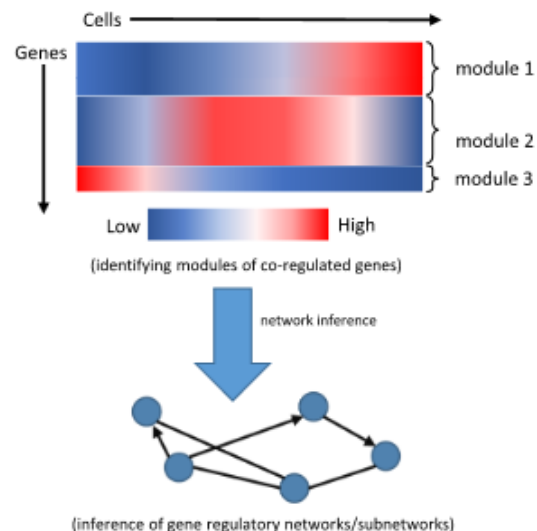


Figure 1. Common applications of single-cell RNA sequencing. (a) Deconvolving heterogeneous cell populations. Clustering by single-cell transcriptomic profiles can reveal population substructure and enable the identification of cell subtypes and rare cell species (e.g. red cells above). Clusters may be tight and well defined (purple, red) or diffuse (blue). (b) Trajectory analysis of cell state transitions. Single-cell RNA sequencing time-series data can be used to map cell developmental trajectories over the course of dynamic processes such as differentiation or signaling responses to an external stimulus. Some computational suites (e.g. Monocle⁶) can also accommodate branching trajectories, enabling identification of lineage-specific gene expression and key genes that drive branching events. (c) Dissecting transcription mechanics. Genes' expression profiles across many cells can be compared to study transcriptional bursting and to model the kinetics of stochastic gene expression. (d) Network inference. Genes can be clustered by expression profile to identify modules of putatively co-regulated genes, and gene-gene covariation relationships can be used to infer gene regulatory networks or subnetworks.

Conclusions partielles / Partie 1

Analyse du transcriptome : génère des données de grandes dimensions, qu'il faut stocker et savoir traiter - Besoins en statistiques et en bioinformatique importants. Besoins de développer également des méthodes pour les représenter (ex: heatmap, voir Partie 2)

Les cellules / le tissu de départ sont / est primordial pour l'interprétation biologique des résultats. Actuellement se développent des approches Single cell pour s'affranchir de l'effet « mélange de cellules » et permettant de prendre en compte les phénomènes de différenciation des cellules.

Ces données sont disponibles et partagées : Openscience / Opendata (voir Partie 2)

Références

Articles de revue

- Bunch, H. (2018). Gene regulation of mammalian long non-coding RNA. *Molecular Genetics and Genomics* 293, 1–15.
- Cieślik, M., and Chinnaiyan, A.M. (2017). Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics* 19, 93–109.
- Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Research.
- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology* 13, e1005457.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10, 57–63.

Articles de recherche

- Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., Bhalerao, R., Uhlen, M., Teeri, T.T., Lundeberg, J., et al. (2001). A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences* 98, 14732–14737.
- Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* 133, 523–536.
- Plass, M., Solana, J., Wolf, F.A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F.J., Kocks, C., and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 360, eaaq1723.

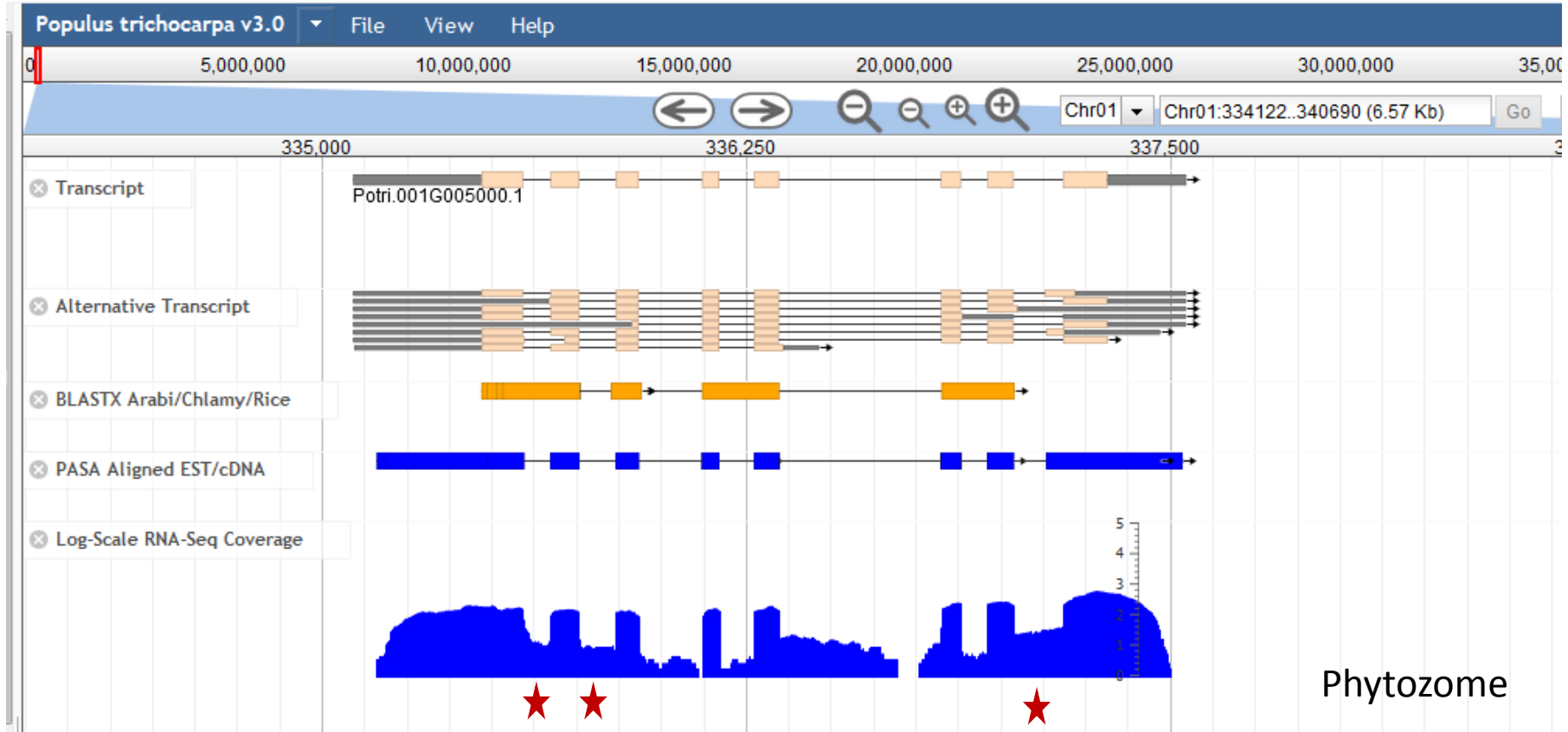
Ouvrage

- Tagu, D., and Moussard, C. (2003). *Principes des techniques de biologie moléculaire* (Paris: Institut National de la Recherche Agronomique).

III. Applications : apport de la transcriptomique

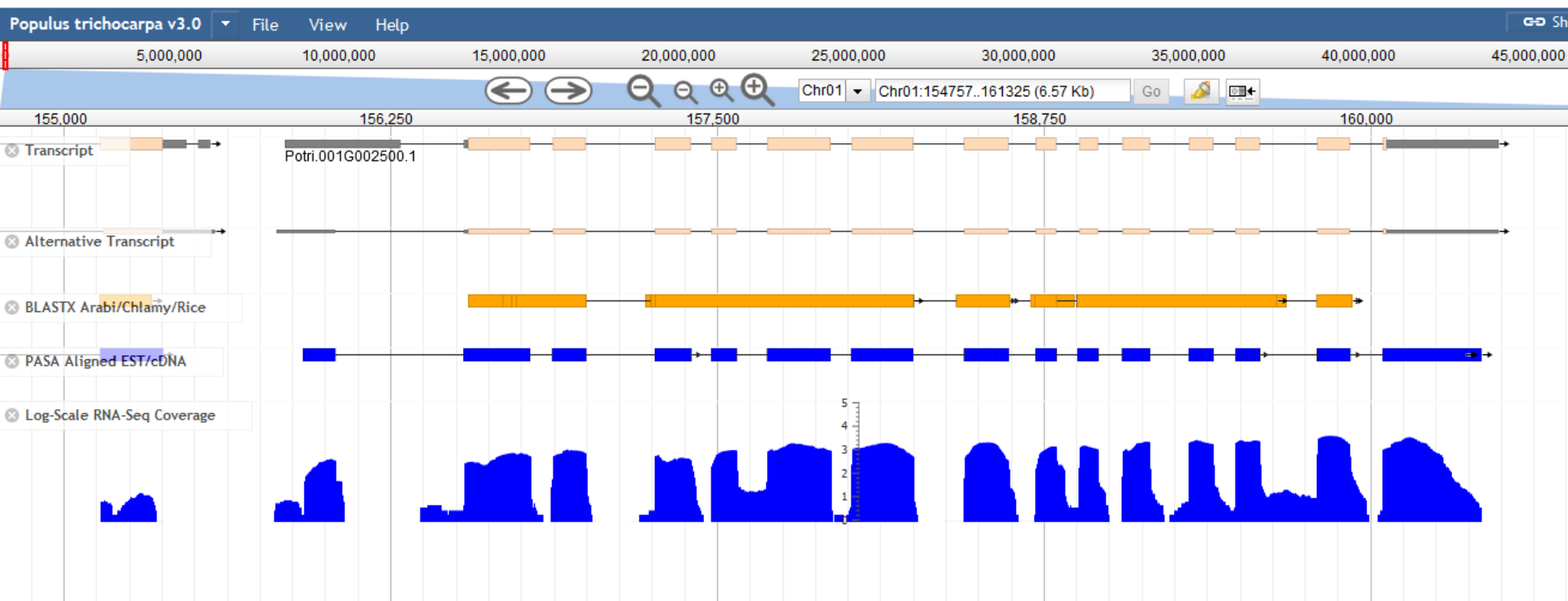
Apport à la génomique structurale

Genome browser : interface graphique qui permet de naviguer, faire des recherches, récupérer de l'information dans un génome donné.

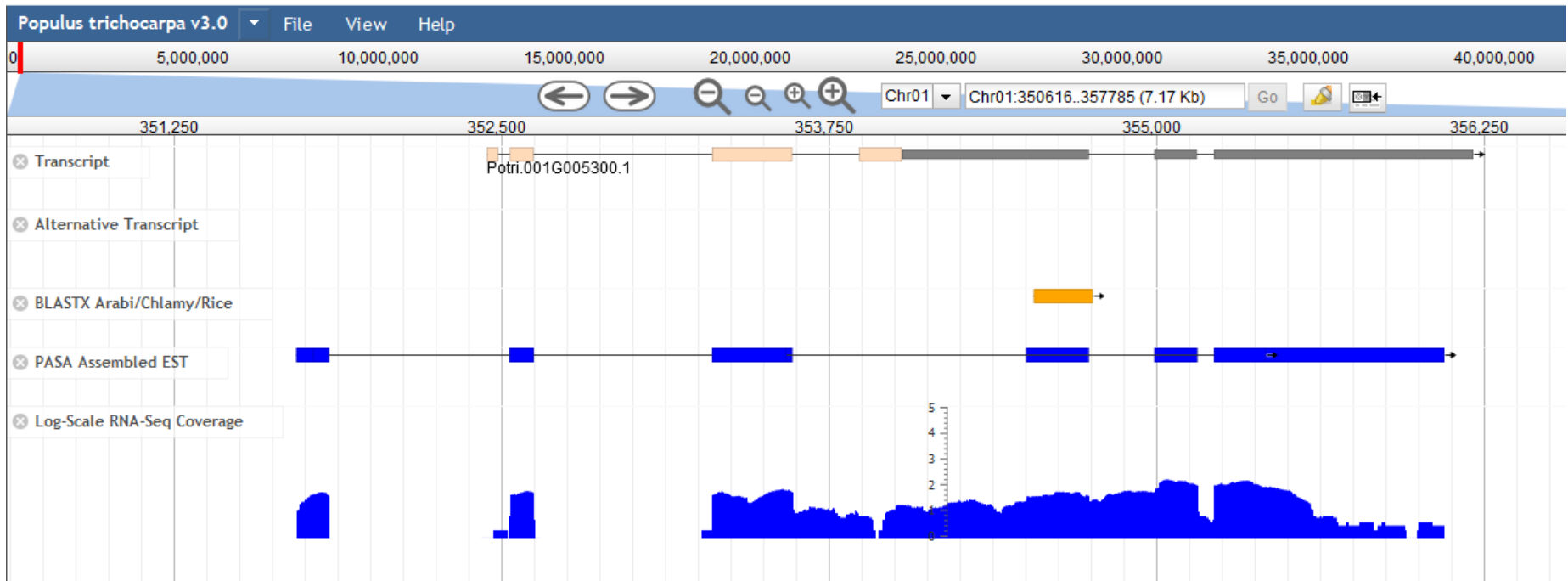


Phytozome

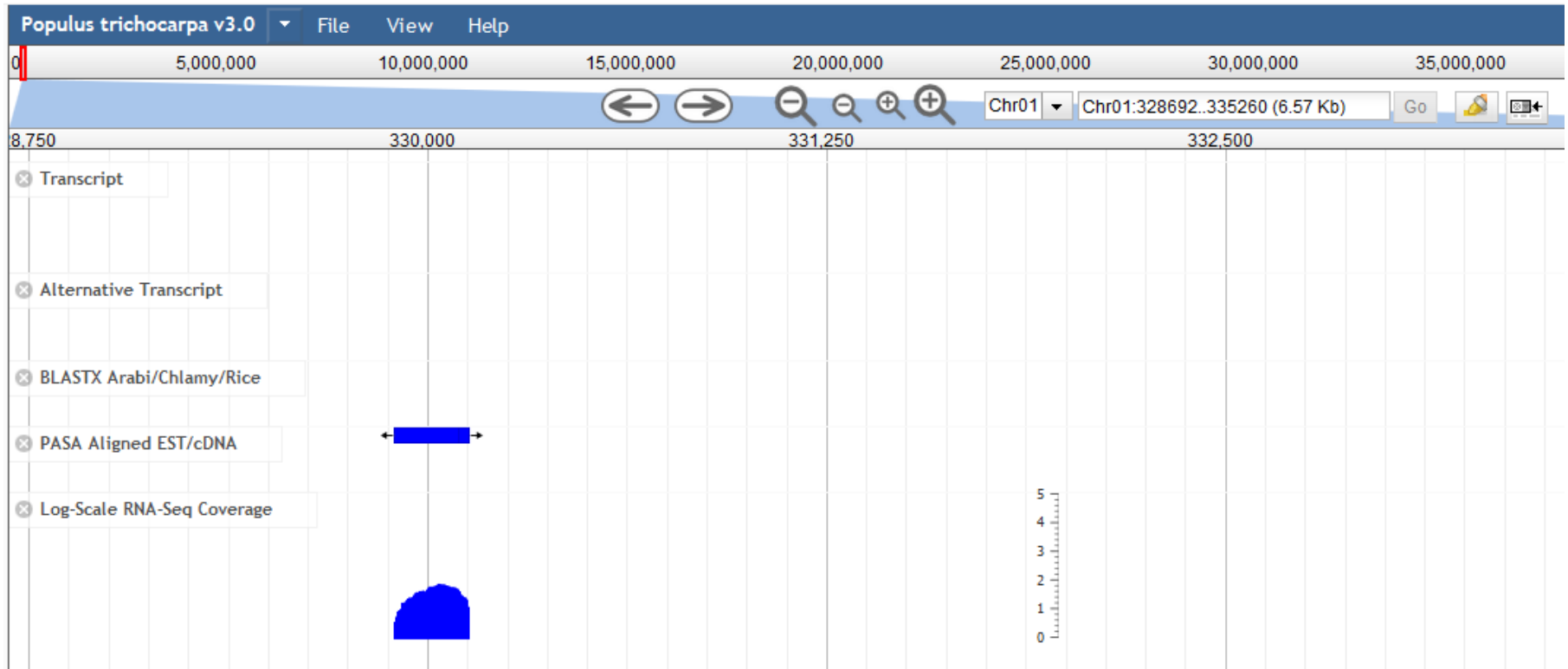
Pour le modèle de gène Potri.001G00500, les données EST/ADNc et RNAseq viennent en appui en grande partie aux différents transcrits proposés.



Qu'apportent les données de transcriptome pour l'annotation structurale de Potri.001G002500?



Qu'apportent les données de transcriptome pour l'annotation structurale de Potri.001G005300?



Et ici?

L'analyse du transcriptome permet de:

- préciser les régions 5'UTR et 3' UTR souvent mal prédites
- de préciser la structure exon/intron, voire de mettre en évidence des transcrits alternatifs
- d'identifier de nouveaux gènes (notamment pour les ARN non codants)

Full Paper

Genome re-annotation of the wild strawberry *Fragaria vesca* using extensive Illumina- and SMRT-based RNA-seq datasets

Yongping Li¹, Wei Wei¹, Jia Feng¹, Huifeng Luo¹, Mengting Pi¹,
Zhongchi Liu^{1,2}, and Chunying Kang^{1,*}

Fraisier sauvage, *Fragaria vesca*, génome séquencé en 2011

Annotation : prediction modèles de gène *ab initio*, seulement les séquences codantes

Amélioration de l'annotation en utilisant les données de sequences d'une banque PacBio (SMRT = Single Molecule Real Time sequencing, longues séquences = long reads)+ 90 banques RNA-seq Illumina (sequences courtes, short reads) + 9 banques pour petits ARN

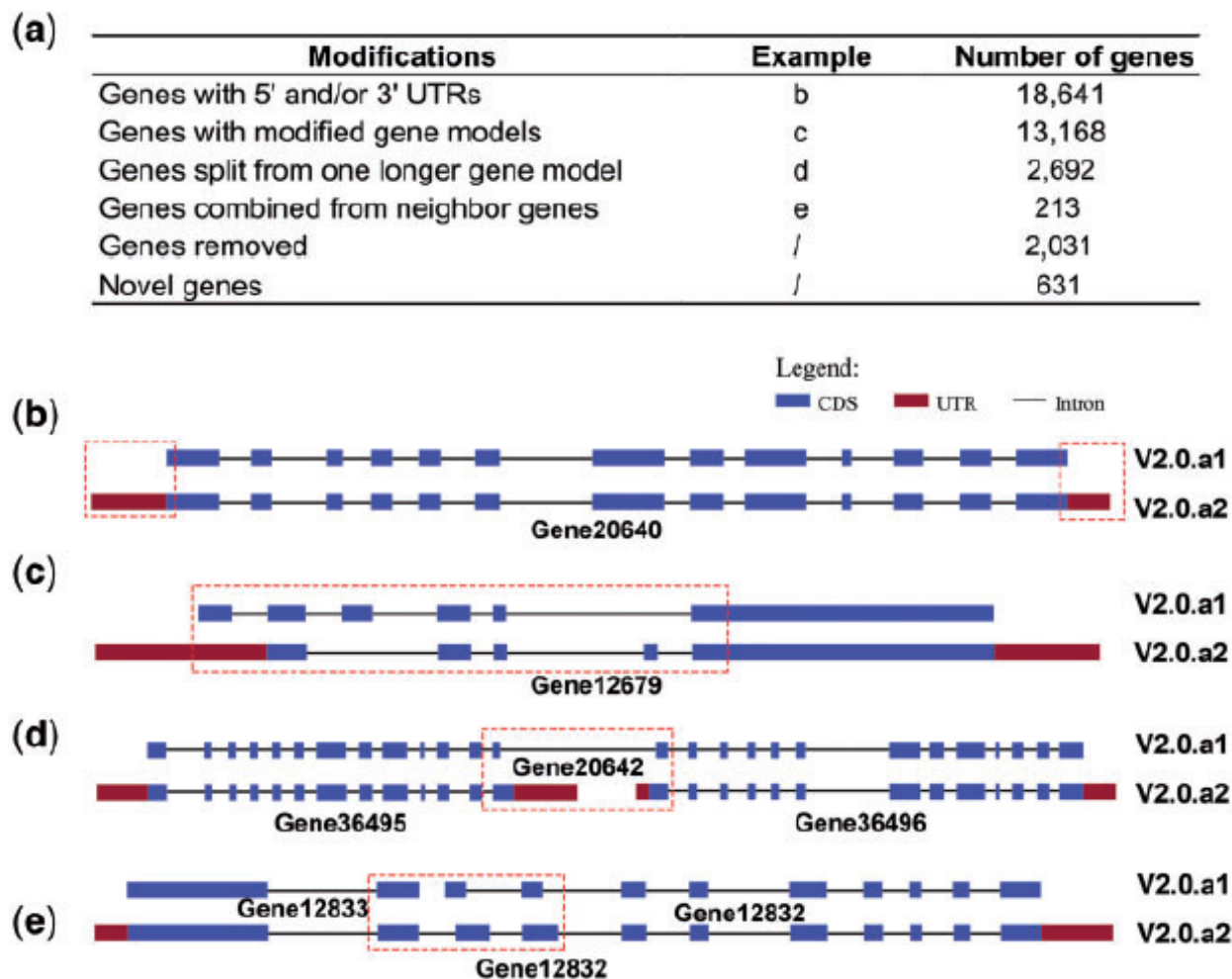


Figure 2. Illustration of the modifications to protein-coding genes in v2.0.a2. (a) Summary of the major types of modifications. The examples of four types are shown in the subfigures b to e. 'Number of genes' indicates the number of genes with each type of modifications in v2.0.a2. (b) 5' and 3' UTRs were added for gene20640. (c) Gene12679 has different and fewer exons than the v2.0.a1 equivalent. (d) Gene20642 in v2.0.a1 was split to create two loci, gene36495 and gene36496. (e) Gene12833 and gene12832 in v2.0.a1 were fused to create a single locus, gene12832. Exons are in blue, untranslated regions (UTRs) are in dark red, and introns are indicated by the thin black lines. The modified regions were highlighted by a dotted rectangle.

Table 1. Summary of the v2.0.a2 annotation

Type	V2.0.a1	V2.0.a2
Protein-coding genes		
Number of genes	33,673	33,538
Mean length of genomic loci	2,863	2,856
Mean exon number	5.1	4.62
Mean CDS length	1,188	1,123
Mean length of introns	408	320
Genes with 5' UTR	–	17,616
Genes with 3' UTR	–	17,971
Genes with both 5' and 3' UTR	–	16,946
Mean 5' UTR length (bp)	–	320
Mean 3' UTR length (bp)	–	513
Number of genes with isoforms	–	7,370
Mean isoform number per gene	1.00	1.51
Genes with functional annotations	27,875	28,798
Genes with GO terms	17,156	22,106
Complete BUSCOs	88.9%	95.7%
Fragmented BUSCOs	5.6%	1.9%
Missing BUSCOs	5.5%	2.4%
Non-coding genes		
microRNAs (miRNAs)	–	171
Small RNA clusters	–	51,714
Long non-coding RNAs (lncRNAs)	–	1,938
Total loci	33,673	87,361

Découverte de nouveaux gènes

Ici, gène = unité fonctionnelle, dont la protéine qui en dérive a un rôle dans la cellule

Méthodologie : RNAseq (ou EST par le passé) car **identification du transcriptome sans connaissance *a priori***

Assemblage *de novo* ou alignement contre un génome de référence

Annotation fonctionnelle : à quoi correspondent les gènes identifiés?

List of bioinformatics tools and databases used for sequence based function annotation

S.no	Software	Function
A Sequence similarity search		
1	Basic local alignment tool (BLAST)	Used for finding similar sequences in protein databases
B Physiochemical characterization		
2	ExPASy – Protparam tool	Used for computation of various physical and chemical parameters like molecular weight, isoelectric point (Pi), amino acid composition, atomic composition, extinction co-efficient, instability index, aliphatic index, and grand average of hydropathy (GRAVY)
C Sub-cellular localization		
3	signalP	Predicts signal peptide cleavage sites.
4	secretomeP	Used for identifying proteins involved in non-classical secretory pathway.
5	PSORT B	Predicts subcellular localization of bacterial proteins.
6	PSLpred	Predicts subcellular localization of proteins from Gram-negative bacteria.
7	CELLO	Assign localization to both prokaryotic and eukaryotic proteins
8	TMHMM	used to authenticate whether the protein is a membrane protein or not.
9	HMMTOP	Predict transmembrane topology.

D Domain analysis and protein		
10	Pfam	Collection of multiple protein sequence alignments
11	SVMprot	SVM (Support vector machine based classification of proteins
12	SYSTEMS	For grouping of proteins on the basis of their functions.
13	SUPERFAMILY	Hierarchical domain classification of PDB structures. NCBI Entrez protein database search of domain architecture
14	CATH (Class, Architecture, Topology, Homology)	Used for finding protein similarities across evolutionary distances based on domain architecture. Classification based on HMM–HMM search. PANTHER is a
15	CDART (The conserved domain architecture retrieval tool)	comprehensively organized database of protein families and sub-families, their evolutionary relationships in the form of phylogenetic trees
16	PANTHER (Protein analysis through evolutionary relationships)	Identification and annotation of protein domains.
17	SMART	Automatic hierarchical clustering of the protein sequences
18	ProtoNet	
E Motif Analysis		
19	InterProScan	Searches interPro for motif discovery. It is the integration of several large protein signature databases.
20	MOTIF	used for Motif discovery.
21	MEME suite	Database searching for assigning function to the discovered motifs.
F Protein–Protein interaction		
22	STRING	Used for predicting protein–protein interactions.

BLAST = Basic Local Alignment Search Tool

Méthode développée par Altschul et al., 1990

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Elle permet de mettre en évidence des régions similaires au sein de séquences biologiques (ADN, protéines). Le programme compare des séquences nucléotidiques ou protéiques à des bases de données et calcule des tests statistiques de significativité.

Différentes variantes selon la nature de la séquence d'entrée, et de la base de donnée utilisée :

- blastn, de nucléotides, séquence nucléotidique contre une base de données de séquences nucléotidiques ;
- blastp, de protéines, séquence de protéine contre une base de données de séquences de protéines ;
- blastx, séquence nucléotidique traduite en séquence de protéine contre une base de données de séquences de protéines ;
- tblastn, séquence de protéine contre une base de données de séquences nucléotidiques traduites en séquences de protéines ;
- tblastx, séquence nucléotidique traduite en séquence de protéine contre une base de données de séquences nucléotidiques traduites en séquences de protéines.

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

[Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear Query subrange

Text input field for query sequence

From To input fields for query subrange

Or, upload file Aucun fichier sélectionné.

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
Nucleotide collection (nr/nt)

Organism Optional Enter organism name or id-completions will be suggested Exclude +
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional Sequences from type material

Entrez Query Optional Enter an Entrez query to limit search [YouTube](#) [Create custom database](#)

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
 Show results in a new window

+ Algorithm parameters

BLAST Results

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

Job title: Potri.001G002500.1##CDS (1824 letters)

RID [WXK5K5NF014](#) (Expires on 10-24 15:24 pm)

Query ID Id|Query_216445

Description Potri.001G002500.1##CDS

Molecule type nucleic acid

Query Length 1824

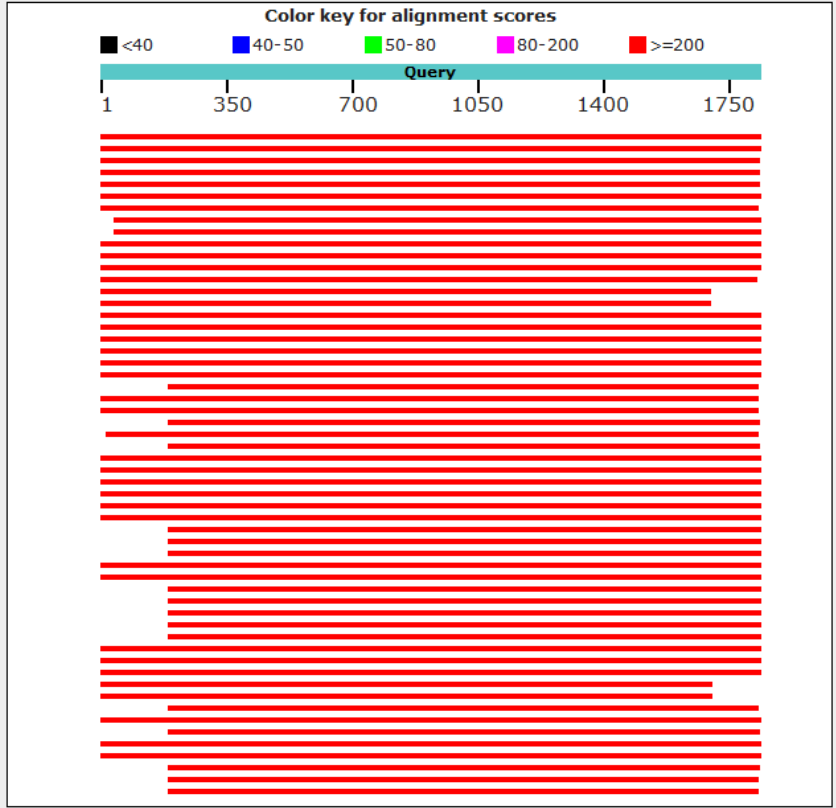
Database Name nr
Description Nucleotide collection (nt)
Program BLASTN 2.8.1+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [MSA viewer](#)

Graphic Summary

Distribution of the top 100 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments



Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#)

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	PREDICTED: Populus trichocarpa chaperonin 60 subunit beta 2, chloroplastic (LOC7483228), mRNA	3369	3369	100%	0.0	100%	XM_002297581.3
<input type="checkbox"/>	Populus trichocarpa clone WS01225_K11 unknown mRNA	3369	3369	100%	0.0	100%	EF146973.1
<input type="checkbox"/>	PREDICTED: Populus euphratica chaperonin 60 subunit beta 2, chloroplastic-like (LOC105142297), mRNA	3179	3179	99%	0.0	98%	XM_011049861.1
<input type="checkbox"/>	PREDICTED: Populus trichocarpa chaperonin 60 subunit beta 2, chloroplastic (LOC7461344), mRNA	2736	2736	99%	0.0	94%	XM_002303947.3
<input type="checkbox"/>	PREDICTED: Populus euphratica chaperonin 60 subunit beta 2, chloroplastic (LOC105126079), mRNA	2736	2736	99%	0.0	94%	XM_011026824.1
<input type="checkbox"/>	PREDICTED: Hevea brasiliensis ruBisCO large subunit-binding protein subunit beta, chloroplastic (LOC110632315), transcript variant X3, mRNA	2056	2056	100%	0.0	87%	XM_021780503.1
<input type="checkbox"/>	PREDICTED: Hevea brasiliensis chaperonin 60 subunit beta 2, chloroplastic (LOC110632898), transcript variant X1, mRNA	2039	2039	99%	0.0	87%	XM_021781267.1
<input type="checkbox"/>	PREDICTED: Jatropha curcas chaperonin 60 subunit beta 2, chloroplastic (LOC105640997), transcript variant X2, mRNA	2023	2023	97%	0.0	87%	XM_020682048.1
<input type="checkbox"/>	PREDICTED: Jatropha curcas chaperonin 60 subunit beta 2, chloroplastic (LOC105640997), transcript variant X1, mRNA	2023	2023	97%	0.0	87%	XM_012225434.2

[Download](#) [GenBank](#) [Graphics](#)

PREDICTED: Hevea brasiliensis ruBisCO large subunit-binding protein subunit beta, chloroplastic (LOC110632315), transcript variant X3, mRNA

Sequence ID: [XM_021780503.1](#) Length: 2260 Number of Matches: 1

Range 1: 164 to 1987 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
2056 bits(1113)	0.0	1591/1828(87%)	8/1828(0%)	Plus/Plus
Query 1	ATGGCATCAACGTTTACTGCCATGTCCTCAGCTGGAACTTGGCTGCTCCTAATGCCCGT	60		
Sbjct 164	ATGGCGTACACCTTACTGCTAATGTCCTCGGTGGATCCTTGGCTGCAGCTAATGGCTGT	223		
Query 61	GCCCTGGATAAGAAGTTTGCATTTTCTCAAAACAAGTGTCACTTTTGGCTCCATTCT	120		
Sbjct 224	GTGATGGATAAGAAGCTTGCATTTTCTCAAAACAAGTGTCACTTTTGGCTCCATTCT	280		
Query 121	GCAAG-TCGAATTTGGTAGACCACAGAATGTAGTTCTACCAAGATCGCGTTCTCTCAAGGT	179		
Sbjct 281	GGAAAGCTC-ATTGGGTAGGAGACAGAATGTAGTTTACGTTAGGCCACCTTCACTAAGGT	339		
Query 180	TAATG-C--GGCCAAGGAGCTTCAATTTCAACAGGACGGGTGAGCAATAGGAAATGCA	236		
Sbjct 340	GTTTGGCATGGCCAAGGAGTTGCAATTTCAATAAGGATGGATCAGCAATTAAGAAATGCA	399		
Query 237	AACGGTGTGAACAAGCTTGCAGATCTAGTTGGGGTTACCCTTGGACCTAAAGGCCGGAA	296		
Sbjct 400	AACGGCGTGAATAAAGCTTGCAGATCTAGTTGGGGTTACTCTCGGACCTAAAGGCCGGAA	459		
Query 297	TGTTGTTCTTGAGAGCAAGTACGGTTCACCTAAAATTTGCAATGATGGTGTGACTGTTGC	356		
Sbjct 460	TGTTGTTCTTGAAAGCAAGTATGGCTCTCCAAAATAGTTAATGATGGTGTGACTGTTGC	519		
Query 357	TAAAGAGGTTGAATTTGGAGGATCCAGTTGAGAACAATTTGGTCTAAGCTAGTGAGACAAGC	416		
Sbjct 520	TAAAGAGGTTGAATTTGGAGGATCCAGTTGAGAATAATTTGGTGCACCAAGTTGGTGAAGCAGGC	579		
Query 417	TGCTGCCAAGACAATGACTTGGCTGGTATGGGACCACAACATCTGTTGTTCTTGCACA	476		
Sbjct 580	AGCTGCCAAGACAATGACTTGGCTGGTATGGGACCACAACATCTGTTGTTCTTGTCTCA	639		
Query 477	GGGCCTAATTCAGAAAGGTGTCAAGGTTGTGGCGGCTGGTGCACCAACCTGTTTTAATCAC	536		
Sbjct 640	AGGCTCTAATTCAGAAAGGTGTCAAGGTTGTAGCTGCTGGTGCACCAACCTGTTCTGATTAC	699		
Query 537	TAGAGGCATTGAAAAGACCACAAGAGCTCTTTGTAATGAACCTAATTTGATGTCAAAAGA	596		
Sbjct 700	TAGGGTAATGAGAAGACCACAAGAGCTCTTAGTGAATGAACCTAATTTGATGTCAAAAGA	759		

Attention aux annotations automatiques en cascade!

Pfam

= base de données de familles de protéines qui classe diverses propriétés des **domaines protéiques** sur la base de leurs alignements de séquences multiples

Published online 15 December 2015

Nucleic Acids Research, 2016, Vol. 44, Database issue D279–D285
doi: 10.1093/nar/gkv1344

The Pfam protein families database: towards a more sustainable future

Robert D. Finn^{1,*}, Penelope Coggill¹, Ruth Y. Eberhardt^{1,2}, Sean R. Eddy^{3,4,5}, Jaina Mistry¹, Alex L. Mitchell¹, Simon C. Potter¹, Marco Punta^{1,6}, Matloob Qureshi¹, Amaia Sangrador-Vegas¹, Gustavo A. Salazar¹, John Tate^{1,2} and Alex Bateman¹

¹European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, ²Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, ³Department of Molecular & Cellular Biology, Harvard University, Biological Laboratories 1008, 16 Divinity Avenue, Cambridge, MA 02138, USA, ⁴John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA, ⁵Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA and ⁶Sorbonne Universités, UPMC-Univ P6, CNRS, Laboratoire de Biologie Computationnelle et Quantitative - UMR 7238, 15 rue de l'Ecole de Médecine, 75006 Paris, France

Family: *Pkinase* (PF00069)

Loading page components (1 remaining)...

9546 architectures

349448 sequences

73 interactions

7104 species

4704 structures

Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Interactions

Structures

Jump to...

enter ID/acc

Go

Summary: Protein kinase domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Protein kinase domain Pfam InterPro

This is the Wikipedia entry entitled "[Protein kinase domain](#)". [More...](#)Protein kinase domain [Edit Wikipedia article](#)

The **protein kinase domain** is a structurally [conserved protein domain](#) containing the catalytic function of [protein kinases](#).^{[2][3][4]} Protein kinases are a group of [enzymes](#) that move a phosphate group onto proteins, in a process called phosphorylation. This functions as an on/off switch for many cellular processes, including metabolism, transcription, cell cycle progression, cytoskeletal rearrangement and cell movement, apoptosis, and differentiation. They also function in embryonic development, physiological responses, and in the nervous and immune system. Abnormal phosphorylation causes many human diseases, including cancer, and drugs that affect phosphorylation can treat those diseases.^[5]

Protein kinases possess a catalytic subunit which transfers the gamma phosphate from nucleoside triphosphates (often *ATP*) to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function. These enzymes fall into two broad classes, characterised with respect to substrate specificity: [serine/threonine specific](#) and [tyrosine specific](#).^[6]

Contents [\[hide\]](#)

- Function
- Structure
- Examples
- References
- External links

Function

[Protein kinase](#) function has been evolutionarily conserved from [Escherichia coli](#) to [Homo sapiens](#). Protein kinases play a role in a multitude of cellular processes, including division, proliferation, apoptosis, and differentiation.^[7] Phosphorylation usually results in a functional change of the target protein by changing enzyme activity, cellular location, or association with other proteins.

Structure

The catalytic subunits of protein kinases are highly conserved, and several structures have been solved,^[8] leading to large screens to develop kinase-specific inhibitors for the treatments of a number of diseases.^[9]

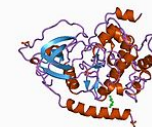
Eukaryotic protein kinases^{[2][3][10][11]} are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common with both serine/threonine and tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases. In the N-terminal extremity of the catalytic domain there is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP binding. In the central part of the catalytic domain there is a conserved aspartic acid residue which is important for the catalytic activity of the enzyme.^[12]

Examples

The following is a list of human proteins containing the protein kinase domain:^[13]

AAK1; ABL1; ABL2; ACVR1; ACVR1B; ACVR1C; ACVR2A; ACVR2B; ACVRL1; **ADCK1**; **ADCK2**; ADCK3; **ADCK4**; **ADCK5**; ADRBK1; ADRBK2; AKT1; AKT2; AKT3; ALPK1; **ALPK2**; **ALPK3**; STRADB; **CDK15**; AMHR2; ANKK1; ARAF; ATM; ATR; AURKA; AURKB; AURKC; AXL; BCKDK; BLK; BMP2K; BMPR1A; BMPR1B; BMPR2; BMX; BRAF; BRSK1; BRSK2; BTK; BUB1; C21orf7; CALM1; CALM2; CALM3; CAMK1; CAMK1D; **CAMK1G**; CAMK2A; CAMK2B; CAMK2D; CAMK2G; CAMK4; CAMKK1; CAMKK2; **CAMKV**; CASK; **CDK20**; CDK1; CDK11B; CDK11A; CDK13; **CDK19**; CDC42BPA; **CDC42BPB**; **CDC42BPG**; CDC7; CDK10; CDK2; CDK3; CDK4; CDK5; CDK6; CDK7; CDK8; CDK9; CDK12; CDK14; CDK16; CDK17; CDK18; **CDKL1**; CDKL2; **CDKL3**; **CDKL4**; CDKL5; CHEK1; CHEK2; CHUK; CIT; CKB; CKM; CLK1; CLK2; CLK3; **CLK4**; CSF1R; CSK; CSNK1A1; **CSNK1A1L**; CSNK1D; CSNK1E; **CSNK1G1**; **CSNK1G2**; **CSNK1G3**; CSNK2A1; CSNK2A2; DAPK1; DAPK2; DAPK3; DCLK1; **DCLK2**; **DCLK3**; DDR1; DDR2; DMPK; DYRK1A; DYRK1B; DYRK2; DYRK3; **DYRK4**; EGFR; EIF2AK1; EIF2AK2; EIF2AK3; EIF2AK4; ELK1; EPHA1; EPHA2; EPHA3; EPHA4; EPHA5; EPHA6; EPHA7; EPHA8; EPHB1; EPHB2; EPHB3; EPHB4; ERBB2; ERBB3; ERBB4; ERN1; **ERN2**; FER; FES; FGFR1; FGFR2; FGFR3; FGFR4; FGR; FLT1; FLT3; FLT4; FYN; GAK; GRK1; GRK4; GRK5; GRK6; GRK7; GSK3A; GSK3B; GUCY2C; GUCY2D; **GUCY2E**; GUCY2F; HCK; HIPK1; HIPK2; HIPK3; **HIPK4**; **HUNK**; ICK; IGF1R; IGF2R; IKBK; IKBE; ILK; INSR; IRAK1; IRAK2; IRAK3; IRAK4; ITR; JAK1; JAK2; JAK3; KALRN; KDR; SIK3; KSR2; LATS1; LATS2; LIMK1; LCK; LIMK2; LRRK1; LRRK2; LYN; MAK; MAP2K1; MAP2K2; MAP2K3; MAP2K4; MAP2K5; MAP2K6; MAP2K7; MAP3K1; MAP3K10; MAP3K11; MAP3K12; MAP3K13; MAP3K14;

Protein kinase domain

Structure of the catalytic subunit of cAMP-dependent protein kinase.^[14]

Identifiers

Symbol	Pkinase
Pfam	PF00069 ↗
InterPro	IPR000719 ↗
SMART	TyrKc ↗
PROSITE	PD0C00100 ↗
SCOP	1apm ↗
SUPERFAMILY	1apm ↗
OPM superfamily	417 ↗
OPM protein	2w5a ↗
CDD	cd00180 ↗

Available protein structures: [\[show\]](#)

KEGG = Kyoto Encyclopedia of Genes and Genomes

KEGG désigne un ensemble de bases de données relatives aux génomes, aux voies métaboliques et aux composés biochimiques.



KEGG [Help](#)
[» Japanese](#)

KEGG Home

[Release notes](#)
[Current statistics](#)
[Plea from KEGG](#)

KEGG Database

[KEGG overview](#)
[Searching KEGG](#)
[KEGG mapping](#)
[Color codes](#)

KEGG Objects

[Pathway maps](#)
[Brite hierarchies](#)
[KEGG DB links](#)

KEGG Software

[KEGG API](#)
[KGML](#)

KEGG FTP

[Subscription](#)

GenomeNet

[DBGET/LinkDB](#)

Feedback

[Copyright request](#)

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes](#) (October 15, 2018) for new and updated features.

Article in 2019 NAR DB Issue

[New approach for understanding genome variations in KEGG](#)

Main entry point to the KEGG web service

KEGG2 [KEGG Table of Contents](#) [[Update notes](#) | [Release history](#)]

Data-oriented entry points

KEGG PATHWAY [KEGG pathway maps](#)
KEGG BRITE [BRITE hierarchies and tables](#)
KEGG MODULE [KEGG modules](#)
KEGG ORTHOLOGY [KO functional orthologs](#) [[Annotation](#)]
KEGG GENOME [Genomes](#) [[Pathogen](#) | [Virus](#) | [Plant](#)]
KEGG GENES [Genes and proteins](#) [[SeqData](#)]
KEGG COMPOUND [Small molecules](#)
KEGG GLYCAN [Glycans](#)
KEGG REACTION [Biochemical reactions](#) [[RModule](#)]
KEGG ENZYME [Enzyme nomenclature](#)
KEGG NETWORK [Disease-related network elements](#)
KEGG DISEASE [Human diseases](#) [[Cancer](#)]
KEGG DRUG [Drugs](#) [[New drug approvals](#)]

KEGG MEDICUS [Health information resource](#) [[Drug labels search](#)]

Organism-specific entry points

KEGG Organisms Enter org code(s) [hsa](#) [hsa eco](#)

Analysis tools

KEGG Mapper [KEGG PATHWAY/BRITE/MODULE mapping tools](#)
BlastKOALA [Genome annotation and KEGG mapping](#)
GhostKOALA [Metagenome annotation and KEGG mapping](#)
BLAST/FASTA [Sequence similarity search](#)
SIMCOMP [Chemical structure similarity search](#)

Classification

[Pathway](#)
[Brite](#)
[Brite table](#)
[Module](#)
[KO \(Function\)](#)
[Organism](#)
[Compound](#)
[Network](#)
[Disease \(ICD\)](#)
[Drug \(ATC\)](#)
[Drug \(Target\)](#)

De novo transcriptome assembly based on RNA-seq and dynamic expression of key enzyme genes in loganin biosynthetic pathway of *Cornus officinalis*

Chengke Bai¹ · Yongmei Wu¹ · Bo Cao² · Jun Xu¹ · Guishuang Li¹



Attention aux classifications qui n'ont pas de sens biologique!

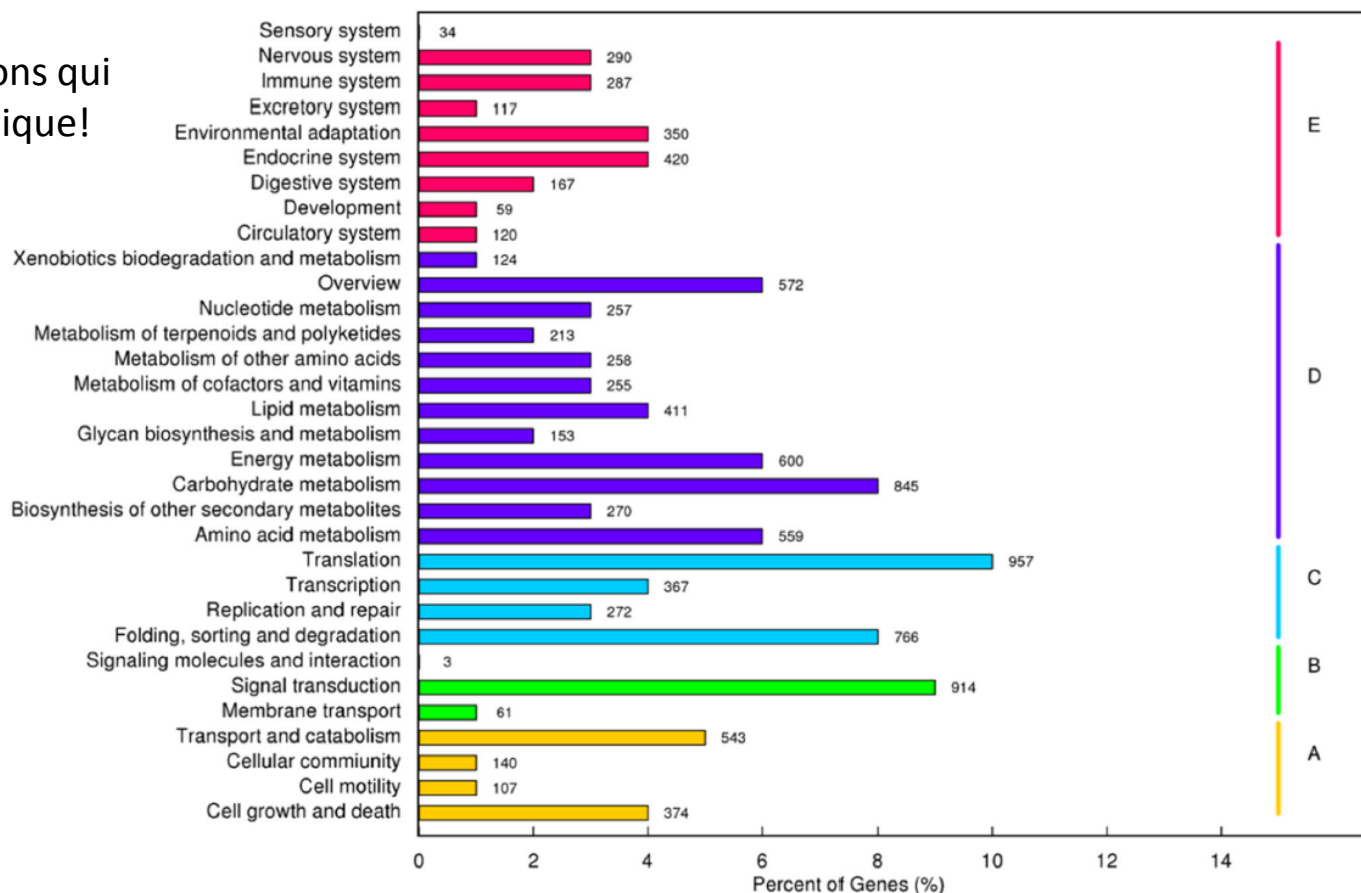


Fig. 4 KEGG classification of assembled unigenes. A, cellular processes; B, environmental information processing; C, genetic information processing; D, metabolism; E, organismal systems. 4517 unigenes with a high mapping ratio of 45.3% were involved in “metabolism” processes,

which includes “phenylpropanoid biosynthesis” (166 genes, ko00940), “flavonoid biosynthesis” (30 genes, ko00941), “tropane, piperidine and pyridine alkaloid biosynthesis” (30 genes, ko00960), and “isoquinoline alkaloid biosynthesis” (28 genes, ko00950)

Gene Ontology (GO)

Gene Ontology est un projet bio-informatique destiné à structurer la description des gènes et de leurs produits dans le cadre d'une **ontologie commune à toutes les espèces**.

Une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations

Le vocabulaire est contrôlé : le même terme pour parler de la même chose

Dans le cadre de GO, les propriétés des produits géniques sont décrites selon trois axes :

- **Cellular component** : les composants cellulaires auxquels ils s'appliquent, qu'il s'agisse du milieu intracellulaire ou de l'environnement extracellulaire,
- **Molecular function** : la fonction moléculaire réalisée, par exemple une activité catalytique pour une enzyme
- **Biological process** : le processus biologique dans lequel il est impliqué

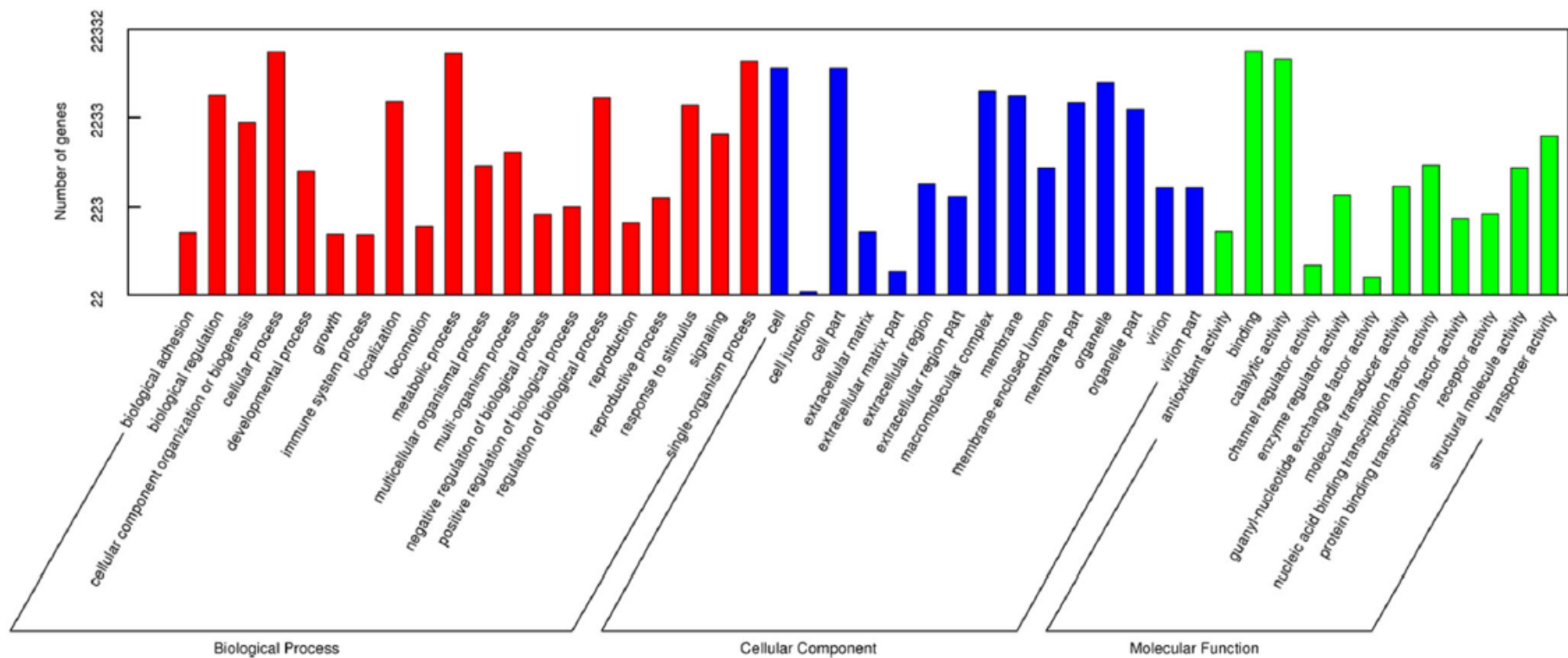


Fig. 3 Histogram of the GO classifications of annotated unigenes from *C. officinalis* transcriptome. 22,332 unigenes (40.73%) were assigned to at least one functional group in GO. The most abundant proteins were found

in cellular processes (12,993 unigenes) and metabolic processes (12,544 unigenes) of biological process group, in the cell (8587 unigenes) and catalytic activity (10,768 unigenes) of molecular function group

Gene : **Potri.001G002500** *P. trichocarpa*Define: (1 of 2) PTHR11353:SF8 - CHAPERONIN 60 SUBUNIT BETA 1, CHLOROPLASTIC-RELATED Name: ▾ Potri.001G002500Secondary Identifier: ▾ PAC:27044297proteins [Potri.001G002500.1](#), [Potri.001G002500.2](#)transcripts [Potri.001G002500.2](#), [Potri.001G002500.1](#)

Quick Links:

[Summary](#) [Genomics](#) [Expression](#) [Other](#)

Genome feature

Region:	gene	Length:	4676 FASTA...
Location:	Chr01:155815-160490		

Ontology Annotations

There are 19 ontology annotations.

Ontology	Term	Name	Namespace	Description
GO	GO:0005524	ATP binding	molecular_function	Interacting selectively and non-covalently with ATP, adenosine 5'-triphosphate, a universally important coenzyme and enzyme regulator.
GO	GO:0005737	cytoplasm	cellular_component	All of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures.
GO	GO:0042026	protein refolding	biological_process	The process carried out by a cell that restores the biological activity of an unfolded or misfolded protein, using helper proteins such as chaperones.
PFAM	PF00118	Cpn60_TCP1		TCP-1/cpn60 chaperonin family
PANTHER	PTHR11353			CHAPERONIN
PANTHER	PTHR11353:SF8			CHAPERONIN 60 SUBUNIT BETA 1, CHLOROPLASTIC-RELATED
KEGG	K04077	groEL, HSPD1		chaperonin GroEL
TIGRFAMs	TIGR02348	TIGR02348		GroEL: chaperonin GroL
PRINTS	PR00298	CHAPERONIN60		60kDa chaperonin signature
PRINTS	PR00304	TCOMPLEXTCP1		Tailless complex polypeptide 1 (chaperone) signature

[Show 9 more rows](#)

0 Pathways

Genomics

Gene Ontology

cellular component

[GO:0005737](#) ▾ [cytoplasm](#) ▾ [ISS](#) ▾

molecular function

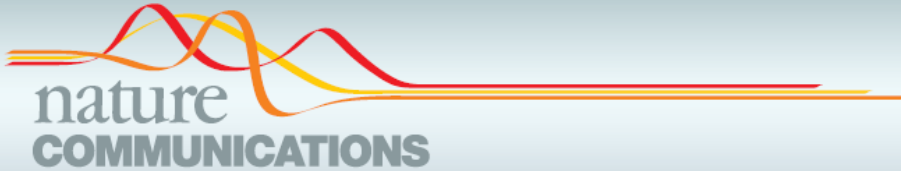
[GO:0005524](#) ▾ [ATP binding](#) ▾ [ISS](#) ▾

biological process

[GO:0042026](#) ▾ [protein refolding](#) ▾ [ISS](#) ▾

Gene models - Potri.001G002500

[Gene models](#)[Transcripts](#): 2 [Exons](#): 30 [Introns](#): 28 [CDSs](#): 2



ARTICLE

Received 23 Aug 2013 | Accepted 10 Mar 2014 | Published 7 Apr 2014

DOI: [10.1038/ncomms4606](https://doi.org/10.1038/ncomms4606)

OPEN

The seco-iridoid pathway from *Catharanthus roseus*

Karel Miettinen^{1,*}, Lemeng Dong^{2,*}, Nicolas Navrot^{3,*}, Thomas Schneider^{4,†}, Vincent Burlat⁵,
Jacob Pollier⁶, Lotte Woittiez^{4,†}, Sander van der Krol², Raphaël Lugan³, Tina Ilc³, Robert Verpoorte¹,
Kirsi-Marja Oksman-Caldentey⁷, Enrico Martinoia⁴, Harro Bouwmeester², Alain Goossens⁶,
Johan Memelink¹ & Danièle Werck-Reichhart³

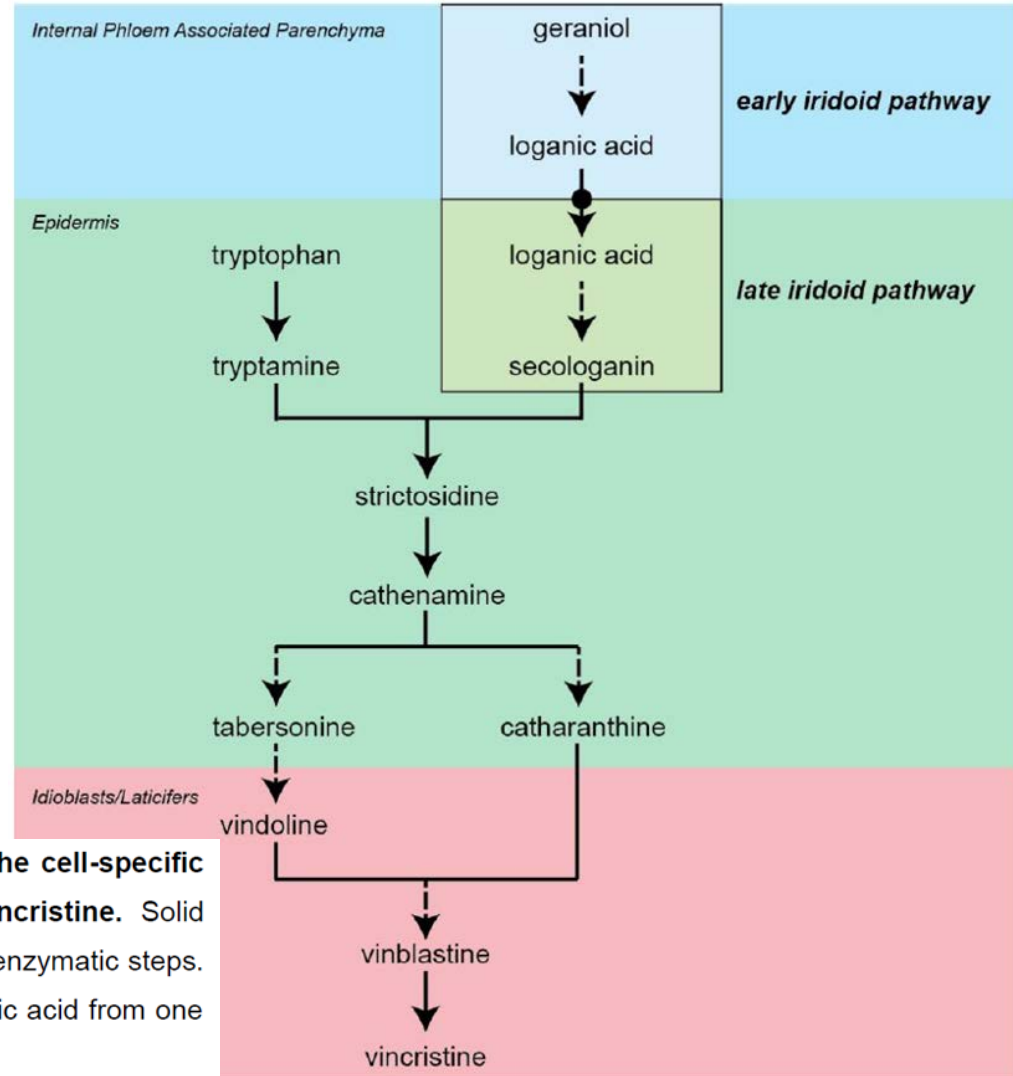
Alcaloïdes indolo-monoterpéniques =
composés naturels dérivés de certaines
plantes qui ont des propriétés
pharmacologiques

Traitement des cancers : camptothecine,
vincristine, vinblastine

Anti-malaria : quinine

Secologanin = iridoid ou securidoid

Pervenche de Madagascar



Supplementary Figure 1: Overview of the MIA pathway and the cell-specific localization of the branches leading to vinblastine and vincristine. Solid arrows represent single enzymatic steps, dashed arrows multiple enzymatic steps. The arrow bearing a black circle represents the transport of loganic acid from one type of cells to the other.

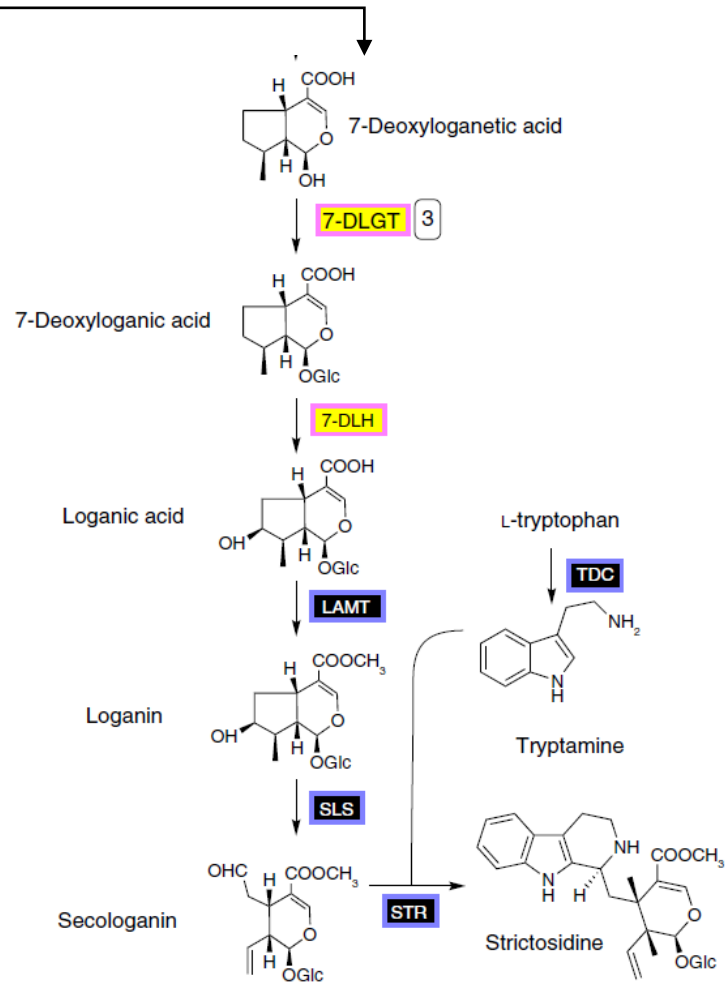
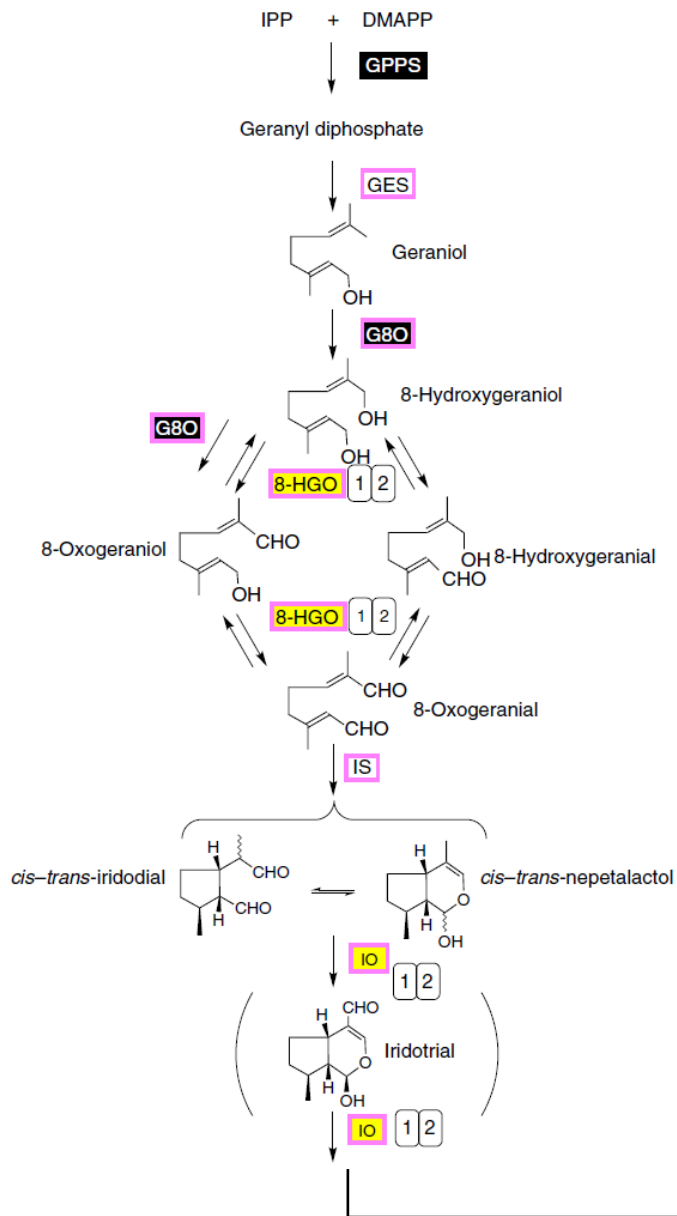


Figure 1 | The secologanin-strictosidine pathway. Genes indicated in boxes were published before (black background) or during (white background) the present study, or are reported here (yellow background). Frames indicate mRNA localization in the leaf IPAP (pink) or epidermis (blue). Numbers indicate predicted enzyme classes in the initial gene discovery strategy. 1: oxidoreductase, 2: cytochrome P450, 3: UGT. IPP, isopentenyl pyrophosphate; DMAPP, dimethylallyl pyrophosphate; Glc, glucose; GPPS, geranyl diphosphate synthase; GES, geraniol synthase; G8O, geraniol 8-oxidase; 8-HGO, 8-hydroxygeraniol oxidoreductase; IS, iridoid synthase; IO, iridoid oxidase; 7-DLGT, 7-deoxyloganic acid glucosyl transferase; 7-DLH, 7-deoxyloganic acid hydroxylase; LAMT, loganic acid O-methyltransferase; SLS, secologanin synthase; STR, strictosidine synthase; TDC, tryptophan decarboxylase. Iridotrial indicated in brackets was a previously proposed intermediate that we did not detect *in vitro* or *in vivo*.

Activité 2 : A partir de l'article Miettinen et al., 2013:

Quelle stratégie mise en œuvre pour identifier les gènes manquant dans la voie de biosynthèse de la strictosidine?

Comment les auteurs démontrent-ils que le gène *caros003452* est l'enzyme 8-hydroxygeraniol oxydoreductase recherchée?

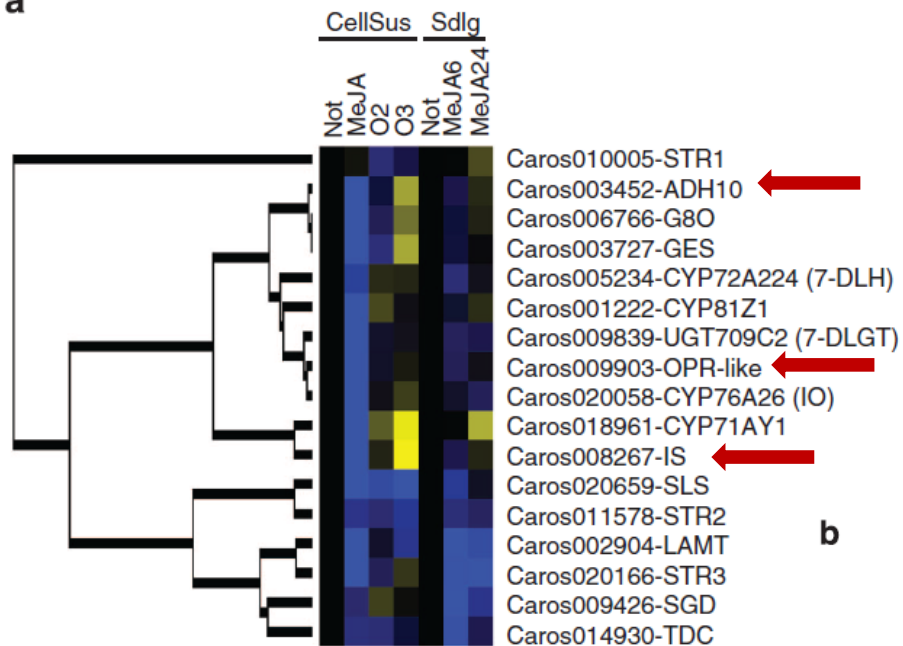
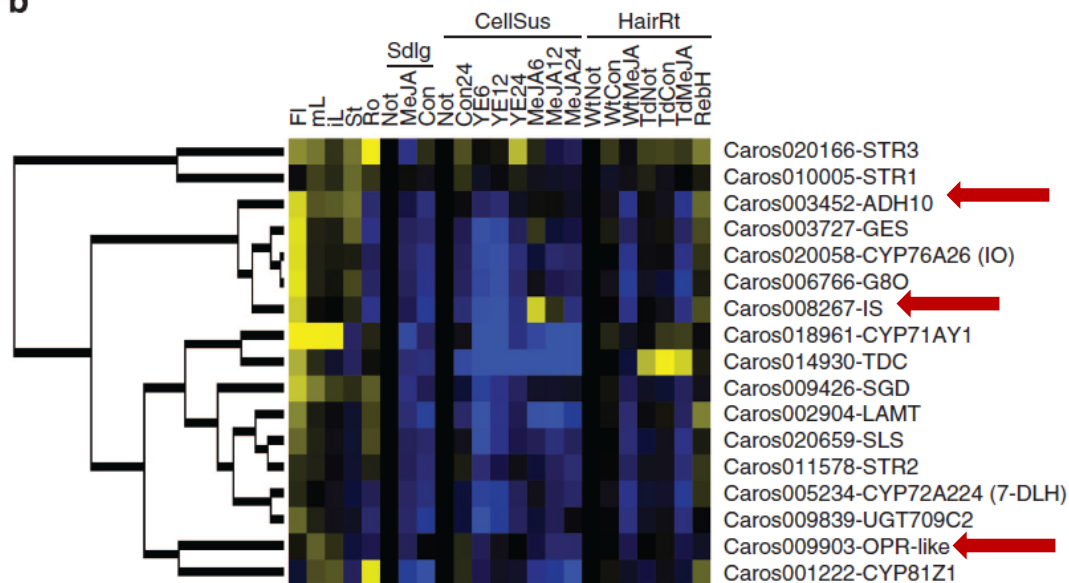
a**b**

Figure 2 | Gene discovery strategy. (a,b) Complete-linkage hierarchical clustering of early MIA pathway gene expression in *C. roseus* based on our data (a) or the Medicinal Plant Genomics Resource consortium (<http://medicinalplantgenomics.msu.edu>) (b). Colours indicate transcriptional activation (blue) or repression (yellow) relative to untreated samples. Tissues: Fl, flower; mL, mature leaves; iL, immature leaves; St, stem; Ro, root; Sdlg, seedling. Suspension cells (CellSus): Wt, wild-type; O2, ORCA2; O3, ORCA3. Hairy roots (HairRt): Wt, wild-type; Td, TDCi; RebH, RebH_F. Treatments: Not, no treatment; MeJA, methyl jasmonate (6, 12 or 24 h); Con, mock; YE, yeast extract. (c) Candidate P450 protein hits in the epidermis and mesophyll

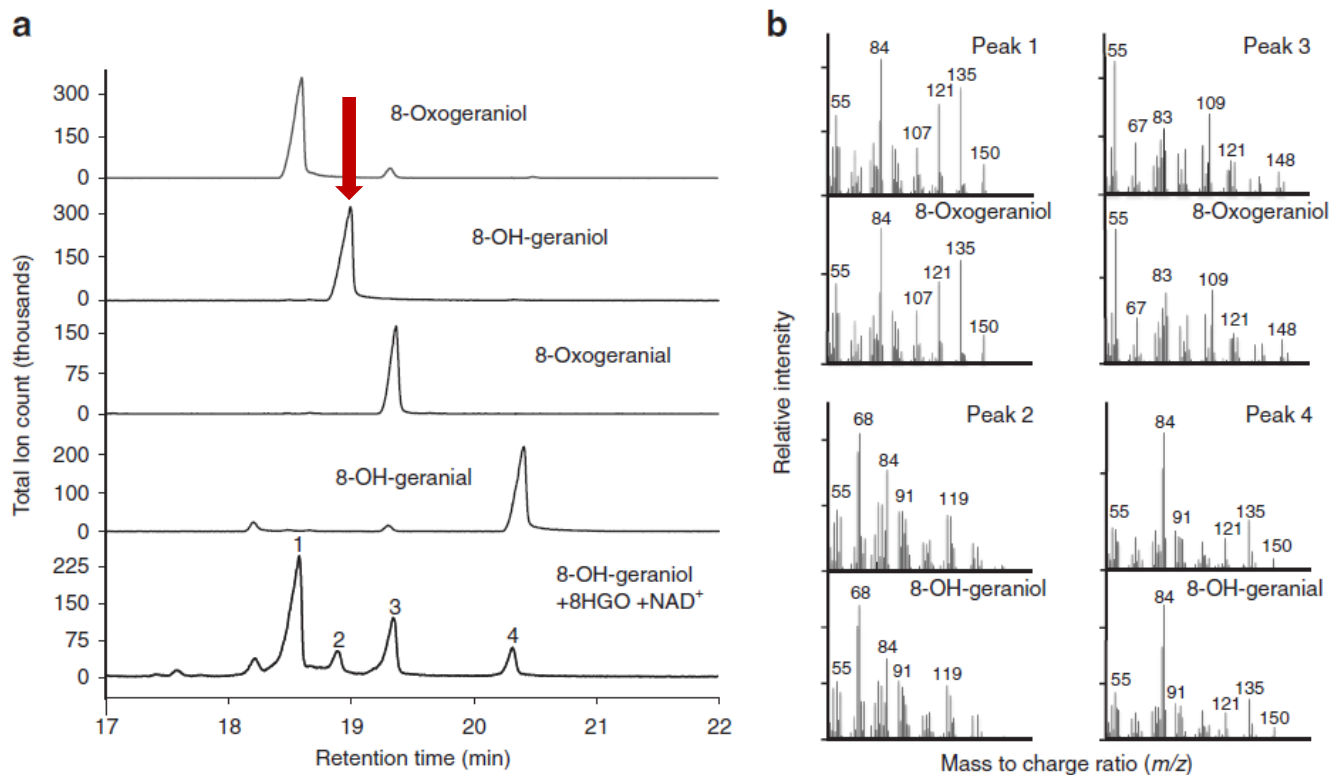


Figure 3 | Functional characterization of recombinant 8-HGO. Affinity-purified enzyme expressed in *E. coli* was incubated with 8-OH-geraniol and NAD⁺. **(a)** GC-MS profile of the reaction extract compared with authentic standards. **(b)** The identity of peaks 1-4 was confirmed by comparison of MS spectra with authentic standards. 8-HGO catalyses the stepwise conversion of 8-OH-geraniol into 8-oxogeraniol or 8-OH-geraniol and then into 8-oxogeraniol.

a

Gene combination	
1	PaGPPS+VoGES
2	PaGPPS+VoGES+G8O
3	PaGPPS+VoGES+G8O+8-HGO
4	PaGPPS+VoGES+G8O+8-HGO+IS
5	PaGPPS+VoGES+G8O+8-HGO+IS+IO
6	PaGPPS+VoGES+G8O+8-HGO+IS+IO+7-DLGT
7	PaGPPS+G8O+8-HGO+IS+IO+7-DLGT (no VoGES)
8	Empty vector
9	IO+7-DLGT+7DLH+LAMT+SLS
10	IO+7-DLGT+7-DLH+LAMT+SLS+TDC+STR

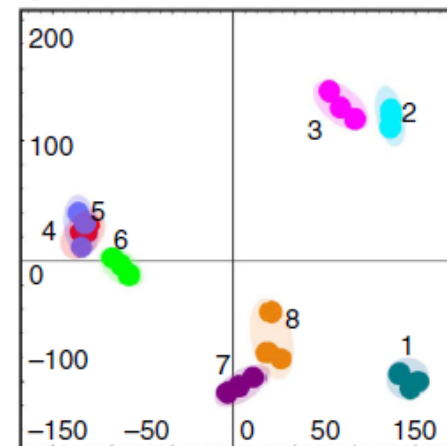
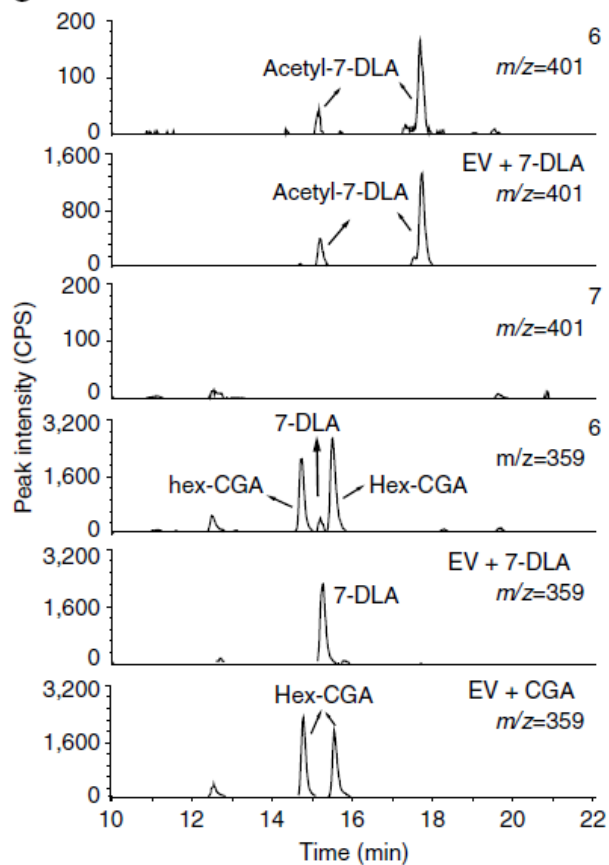
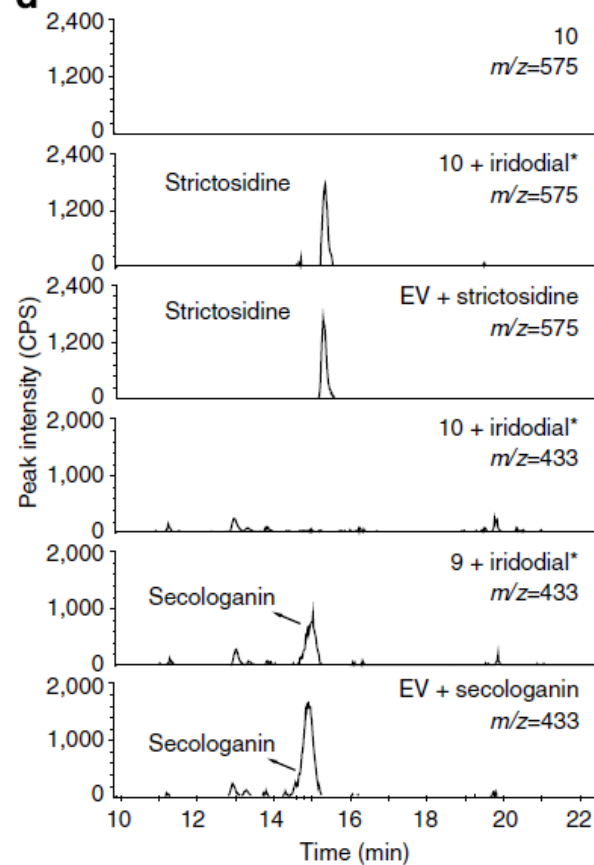
b**c****d**

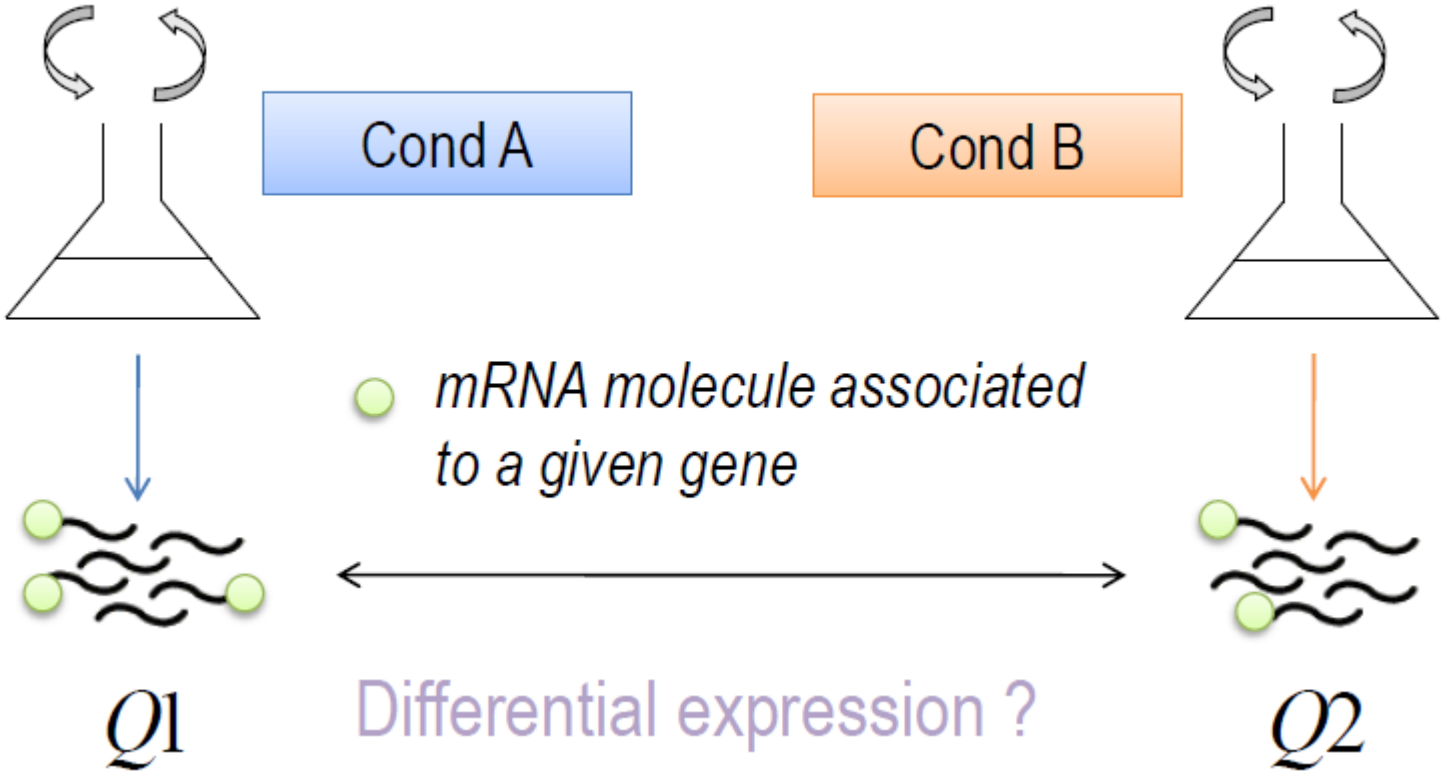
Figure 8 | Reconstitution of the strictosidine pathway in *N. benthamiana*. (a) Gene combinations infiltrated in leaves in triplicate. (b) Principal component analysis. PC1 and PC2 describe 36.2 and 31.1% of the total mass variation, respectively. (c) LC-MS analysis showing selected masses 401 and 359 representing (acetylated) 7-deoxyloganic acid (7-DLA) from infiltrations with 8-carboxygeranic acid (CGA), 7-DLA or gene combinations 6 or 7 (negative control). The two peaks likely represent 7-DLA acetylated at two different positions in the glucose moiety. (d) LC-MS analysis showing selected masses 433 (formic acid adduct of secologanin) and 575 (formic acid adduct of strictosidine) from infiltrations with secologanin or strictosidine, or with gene combinations 9 or 10, with or without iridodial. *Identical profiles with iridotrial or 7-DLH. Hex = hexosyl; CPS = counts per second.

Expression de gènes et implication dans certains processus biologiques

Gènes différentiellement exprimés

On compare deux conditions ou un traitement et une condition contrôle

Hypothèse sous-jacente : les gènes dont l'expression varie sont impliqués dans la réponse au traitement



Gènes co-exprimés

On compare plusieurs conditions, plusieurs traitements ou un processus de développement avec plusieurs étapes bien identifiées

Hypothèse sous-jacente : les gènes co-exprimés sont impliqués dans les mêmes grandes fonctions biologiques

The Plant Cell, Vol. 29: 1585–1604, July 2017, www.plantcell.org © 2017 ASPB.



LARGE-SCALE BIOLOGY ARTICLE

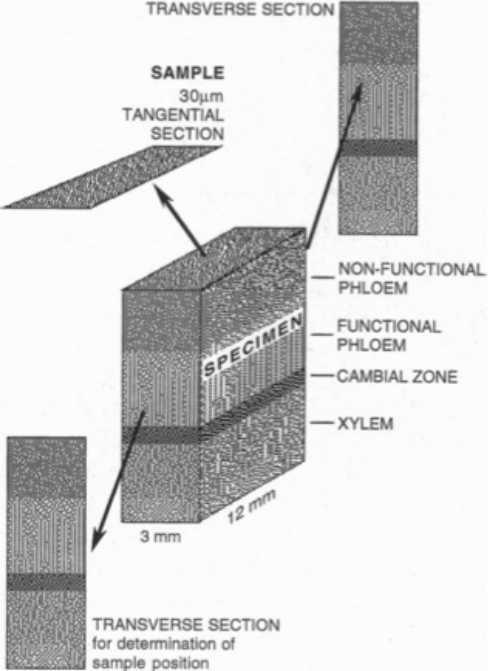
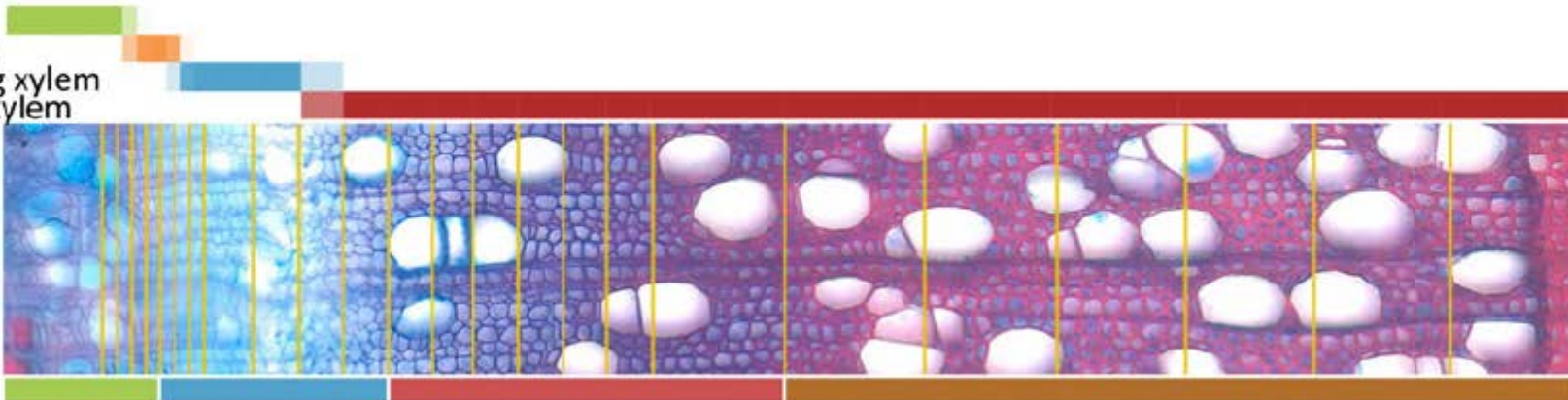
AspWood: High-Spatial-Resolution Transcriptome Profiles Reveal Uncharacterized Modularity of Wood Formation in *Populus tremula* OPEN

David Sundell,^{a,1} Nathaniel R. Street,^{a,1} Manoj Kumar,^{b,1} Ewa J. Mellerowicz,^b Melis Kucukoglu,^b
Christoffer Johnsson,^b Vikash Kumar,^b Chanaka Mannapperuma,^a Nicolas Delhomme,^a Ove Nilsson,^b
Hannele Tuominen,^a Edouard Pesquet,^{a,c} Urs Fischer,^b Totte Niittylä,^b Björn Sundberg,^b and Torgeir R. Hvidsten^{a,d,2}

Echantillonnage par préparation de coupes tangentielles

A

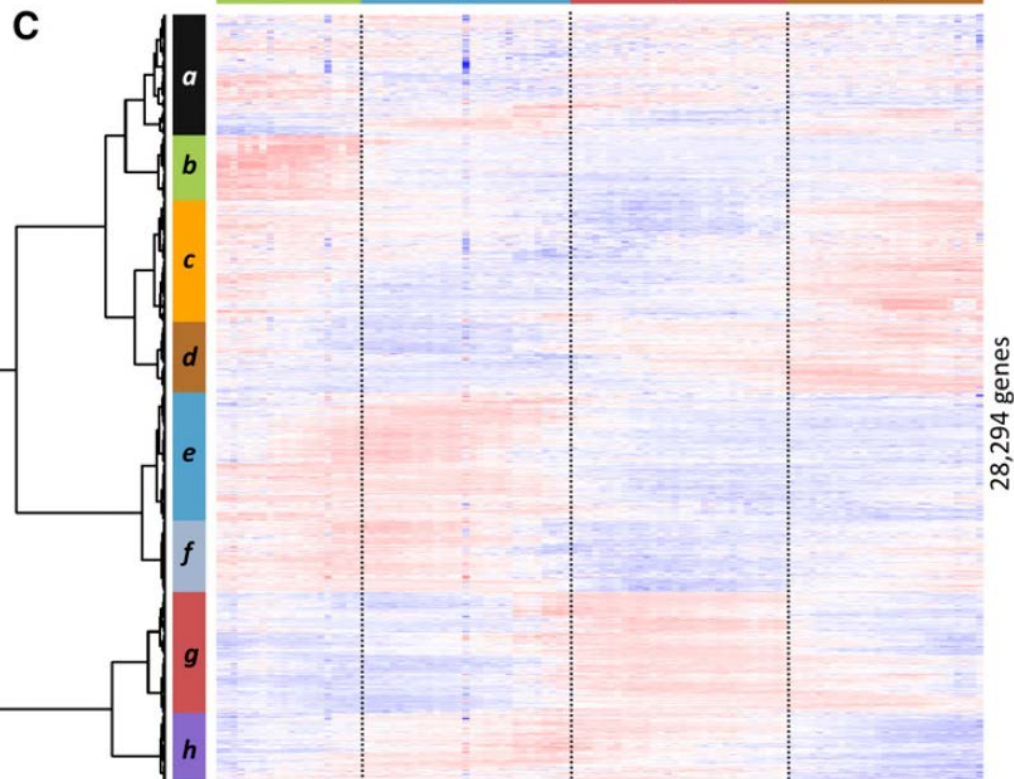
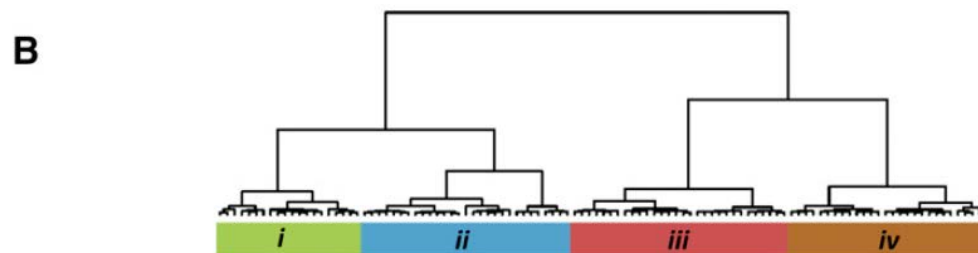
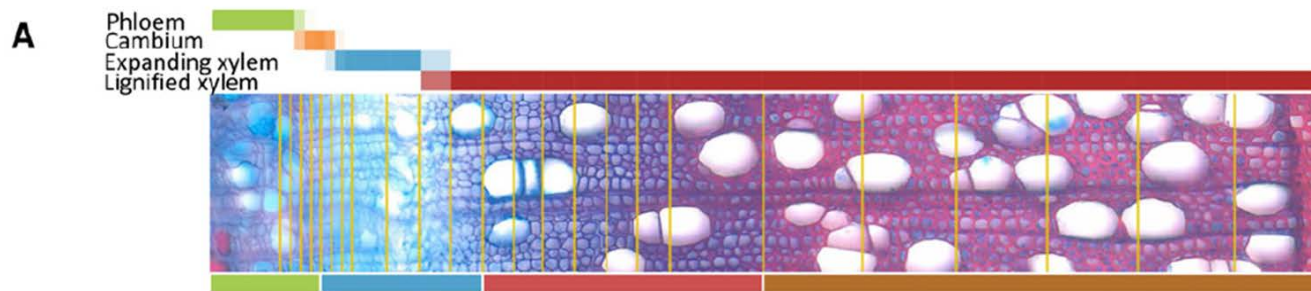
Phloem
Cambium
Expanding xylem
Lignified xylem



Activité 3 : A partir de l'article Sundell et al., 2017:

Etude de la figure 1

Qu'est-ce que présente la figure 1?



106 samples from four trees

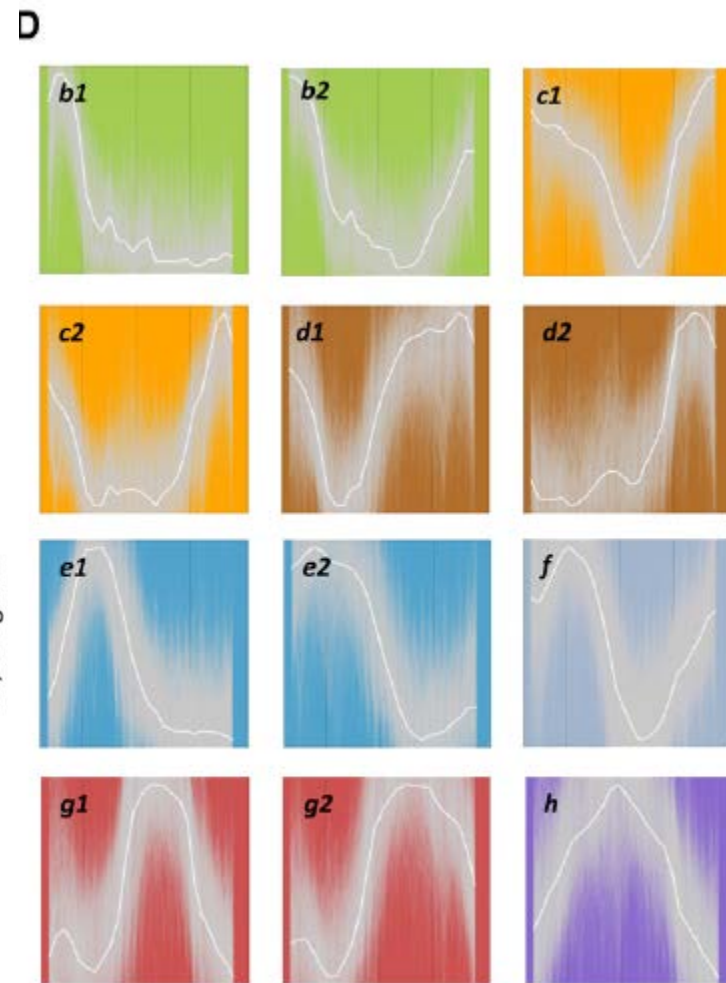


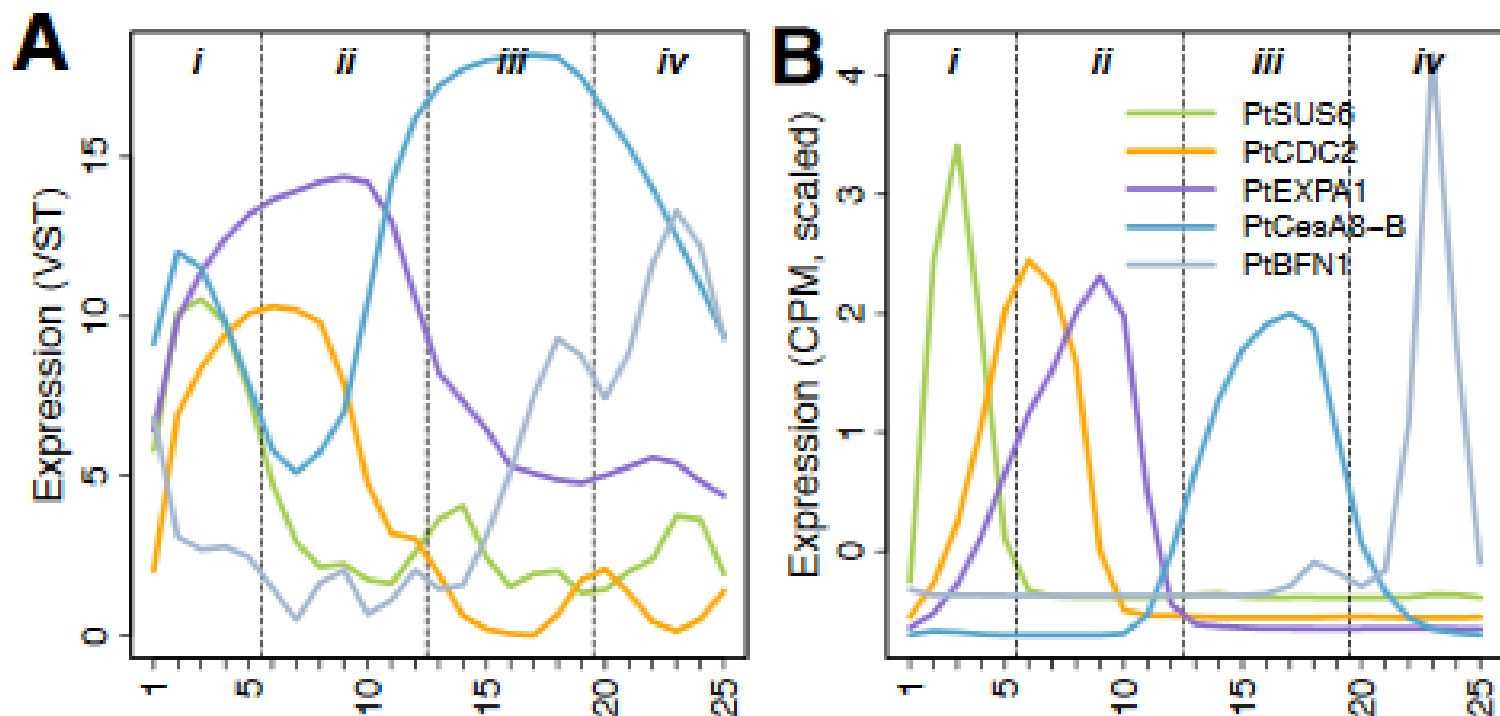
Figure 1. Hierarchical Clustering of Samples and Genes across Developing Xylem and Phloem Tissues.

(A) Transverse cross-section image from one of the sampled trees (tree T1). The pooled samples used for RNA-seq are visualized by overlaying them on the transverse cross section (positions of the pooled samples on the section are approximated). The color bar below the image shows four sample clusters identified by hierarchical clustering (see **[B]**). The color bar above the image shows the estimated tissue composition for each sample.

(B) Hierarchical clustering of all 106 samples from the four replicate trees using mRNA expression values for all expressed genes. The four main clusters are indicated with colors.

(C) Heat map describing hierarchical clustering of the 28,294 expressed annotated genes using mRNA expression values for all samples. Expression values are scaled per gene so that expression values above the gene average are shown in red and below average in blue. Eight main clusters have been assigned colors and are denoted *a* to *h*.

(D) Average expression profiles in tree T1 for each gene expression cluster and distinct subclusters (solid white lines). The expression profiles of all individual genes assigned to each cluster are shown as gray lines in the background.



Supplemental Figure 3.

The expression profiles of selected marker genes. Expression is shown with (A) the variance stabilized transformation (VST) and (B) scaled counts per million (CPM, calculated as 2^{VST} , scaled: mean centered and normalized by the standard deviation of each gene). *PtSUS6*/Potri.004G081300: *A. thaliana* sucrose synthase (*SUS*) genes *SUS5* and *SUS6* are known to be phloem localized (Barratt et al., 2009). The *Populus* homologs of *SUS5* and *SUS6* peaked in the two or three outermost sections of the single section series covering the cambial meristem, thus marking a phloem identity to these samples. The pooled phloem sample, consisting of six to seven section samples showed lower levels indicating that *SUS5* and *SUS6* expression decreased in the older phloem tissues. Expression was shown for one *SUS6* homolog (*PtSUS6*). *PtCDC2*/Potri.016G142800: The dividing meristem was marked by a set of cyclins typically known to be involved in the different stages of cell cycling. These markers indicated that sample cluster *i* and *ii* from the hierarchical clustering was split in the middle of the meristem. Expression was shown for *PtCDC2*. *PtEXPA1*/Potri.001G240900: Cell expansion was marked by the *Populus* alpha expansin (Gray-Mitsumune et al., 2004). The *PtEXPA1* expression marked the broader region covering both the phloem and xylem side, and decreased sharply at the border to sample cluster *iii*. *PtCesA8-B*/Potri.004G059600: *CesA4*, 7 and 8 are well-described hallmark genes for secondary wall formation in xylem cells (Kumar et al., 2009). The corresponding transcripts in aspen showed two distinct peaks, one minor on the phloem side marking secondary wall formation in phloem, and another major peak on the xylem side marking the bulk of secondary wall formation in xylem vessels, fibers and rays. Expression was shown for *PtCesA8-B*. *PtBFN1*/Potri.011G044500: Vessel cells undergo cell death before fiber cells. Xylem specific proteases and nucleases are responsible for the autolysis of xylem cells following cell death were selected to mark the timing of the cell death of fibers and vessel elements (reviewed by (Escamez and Tuominen, 2014)). In aspen, a homolog of the *A. thaliana* bifunctional nuclease 1 (BFN1) protein showed two peaks of expression; one appearing at the end of sample cluster *iii* and the other at the end of sample cluster *iv*. This corresponds well with the locations of cell death, as estimated on the basis of the viability assays of the xylem tissues (Supplemental Data Set 1). Thus, the sharp decrease in the expression of *PtBFN1* marks the border of living vessel elements and fibers, respectively.

Réseaux de coexpression

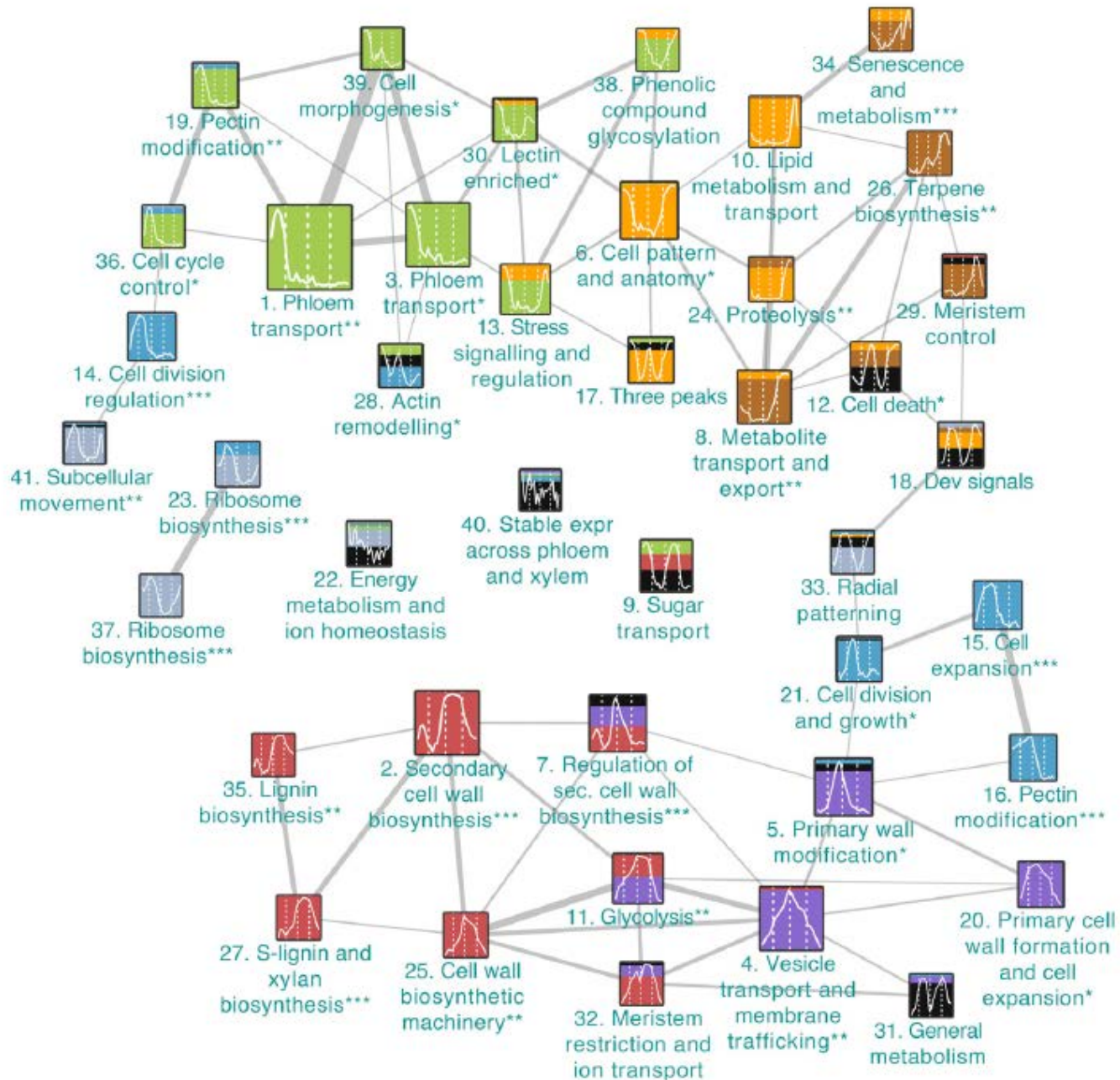


Figure 3. A Modular Version of the Coexpression Network.

Genes with representative expression profiles were identified in the coexpression network (at a Z-score threshold of 5) by iteratively selecting the gene with the highest centrality and a coexpression neighborhood not overlapping with any previously selected genes' neighborhood. Only annotated genes and positively correlated coexpression links were considered (i.e., Pearson correlation > 0). The selected genes and their coexpression neighborhoods (network modules) were represented as nodes in a module network. The modules were numbered according to the order in which they were selected (and hence according to size) and given descriptive names based on gene function enrichment analysis (Supplemental Data Set 8). The nodes are colored according to the hierarchical clusters in Figure 1 and reflect the proportion of genes in each network module belonging to the different hierarchical clusters. Nodes were linked if the neighborhoods overlapped at a lower coexpression threshold (Z-score threshold of 4). Link strengths are proportional to the number of common genes. Overlaps of fewer than five genes were not represented by links, and only the 41 network modules with at least 20 genes were displayed. Asterisks next to the module names indicate conservation in Norway spruce: *, 6–24% of the genes have conserved coexpression neighborhoods; **, 25–50%; and ***, >50% (Supplemental Data Set 8D).

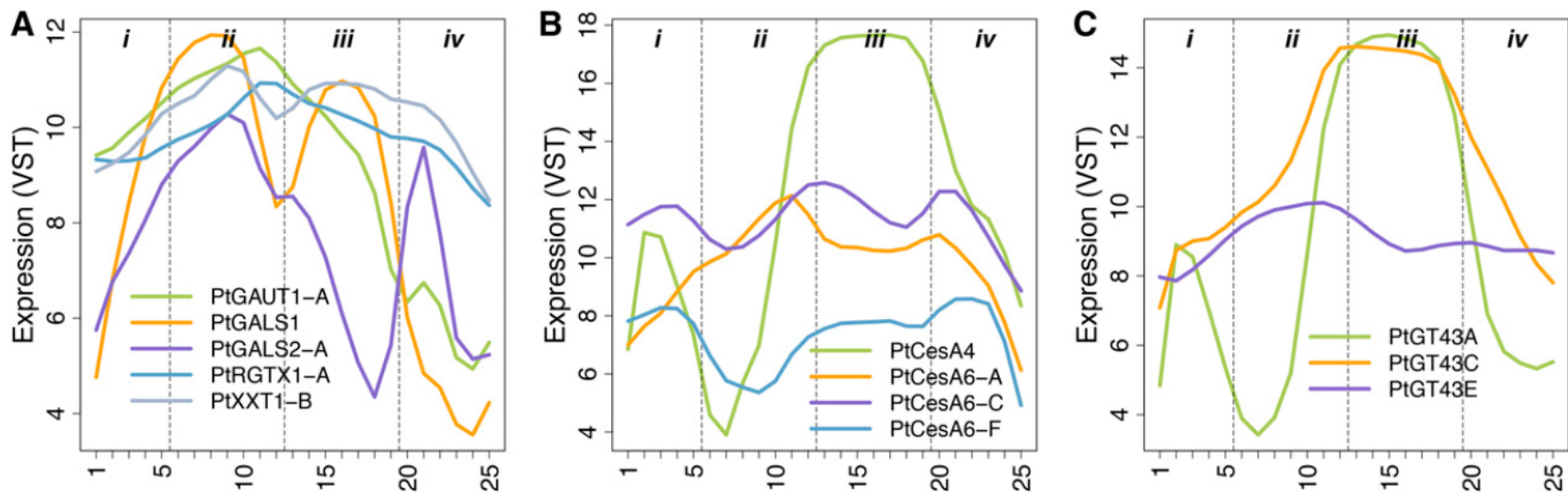


Figure 5. Primary and Secondary Cell Wall Biosynthesis Genes.

(A) Expression profiles for pectin and xyloglucan biosynthetic genes. All genes are highly expressed during primary wall biosynthesis, but also at later stages during xylem development. (i) Representative expression for homogalacturonan biosynthesis genes (illustrated by *PtGAUT1-A*); (ii) expression pattern of *PtGALS1*; (iii) representative pattern of other RG-I biosynthesis genes (represented by *PtGALS2-A*); and (iv) representative expression pattern for three of the putative xyloglucan biosynthesis genes (illustrated by *PtXXT1-B*). For a list of identified putative pectin and xyloglucan biosynthesis genes, see Supplemental Data Set 9.

(B) Expression patterns for *CesA* genes. (i) Members responsible for cellulose biosynthesis in the secondary wall layers are all induced in SCW biosynthesis zones in the xylem and phloem (illustrated by *PtCesA4*); (ii) members classified as primary wall *CesAs* typically peak in primary wall biosynthesis zone, but are also highly expressed during later stages of xylem differentiation (illustrated by *PtCesA6-A*); and (iii and iv) some members even peak during these later stages (illustrated by *PtCesA6-C* and *PtCesA6-F*). For a complete list of putative *CesA* genes, see Supplemental Data Set 9.

(C) The GT43 gene family responsible for xylan biosynthesis comprises three clades, A/B, C/D, and E, each having different expression here illustrated by (i) *PtGT43A*, (ii) *PtGT43C*, and (iii) *PtGT43E*. Expression profiles support the hypothesis that *PtGT43A/B* and *PtGT43C/D* are members of secondary wall xylan synthase complex, whereas *PtGT43E* and *PtGT43C/D* are members of the primary wall xylan synthase complex. For a complete list of genes coregulated with the three clades of *GT43* genes, see Supplemental Data Set 9.

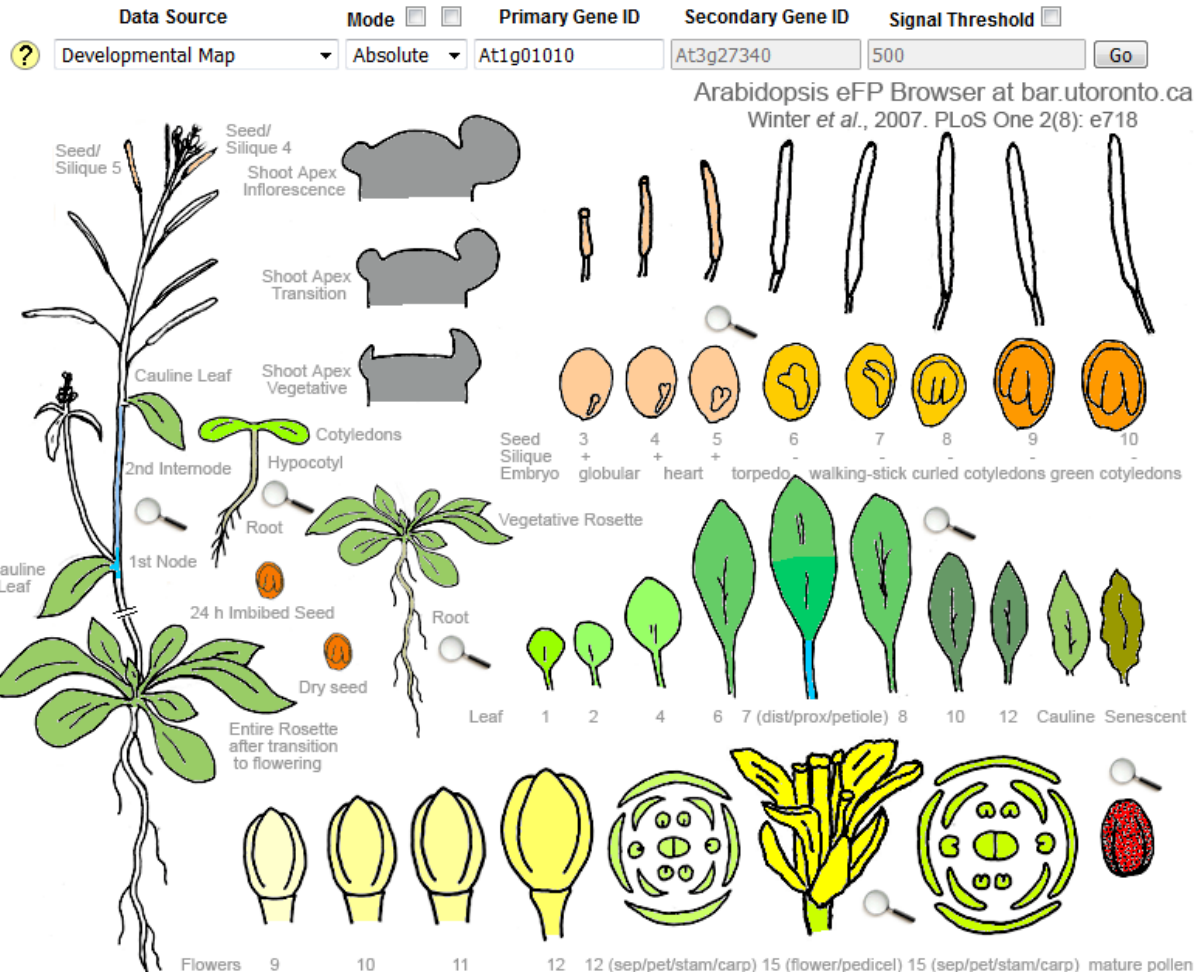
Transcript profiling ou gene atlas

Avoir une carte / un atlas de l'expression des gènes selon l'organe, le tissu, etc...

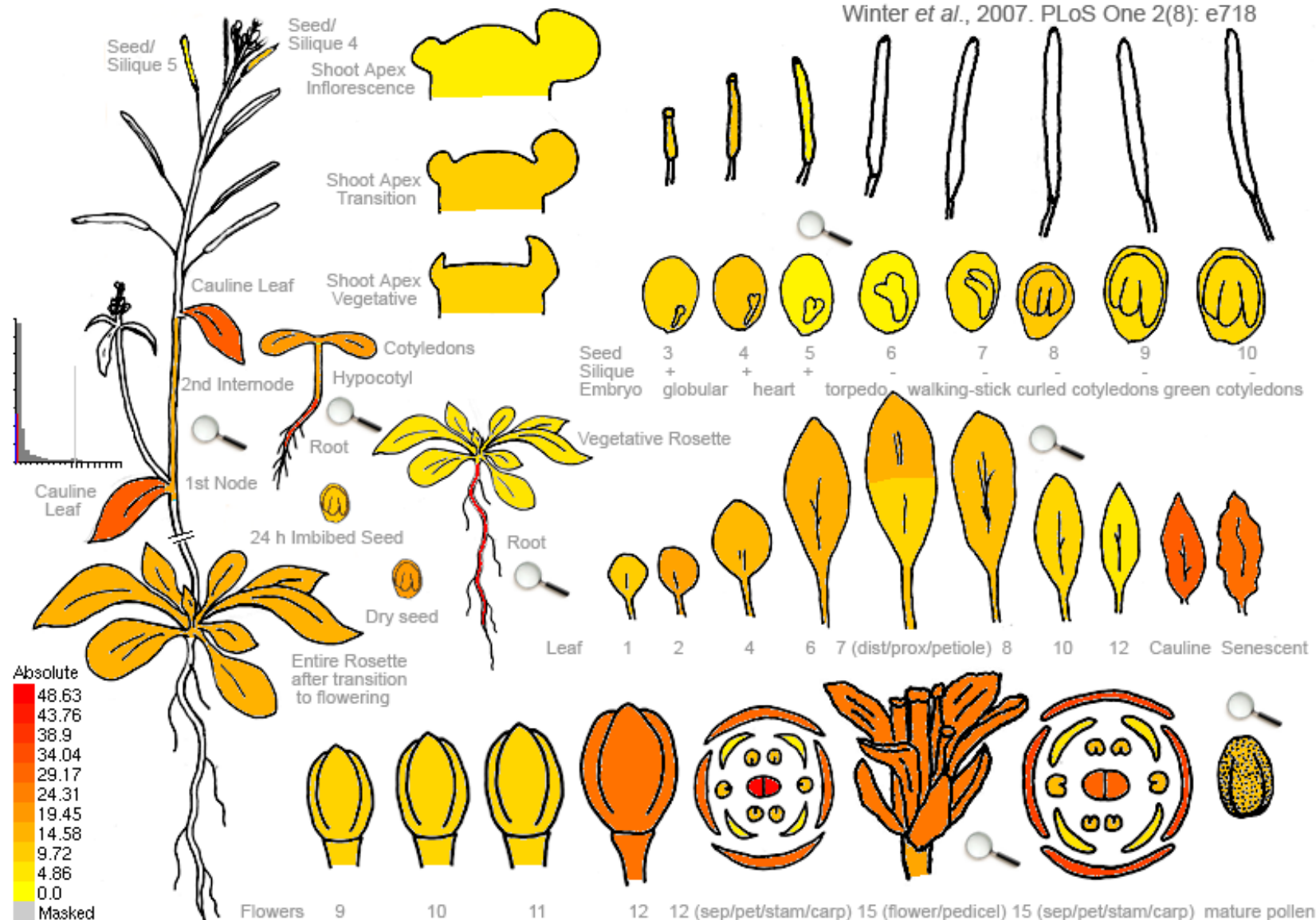


Arabidopsis eFP Browser

J'aime 1 K



eFP Browser by B. Vinegar, drawn by J. Alls and N. Provart. Data from Gene Expression Map of Arabidopsis Development: Schmid et al., 2005, Nat. Gen. 37:501, and the Nambara lab for the imbibed and dry seed stages. Data are normalized by the GCOS method, TGT value of 100. Most tissues were sampled in triplicate.

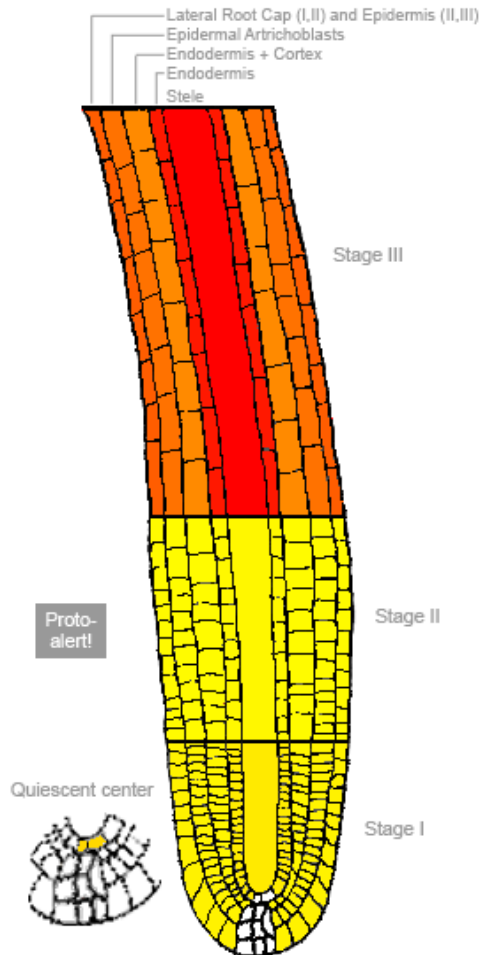


eFP Browser by B. Vinegar, drawn by J. Alls and N. Provar. Data from Gene Expression Map of Arabidopsis Development: Schmid et al., 2005, Nat. Gen. 37:501, and the Nambara lab for the imbibed and dry seed stages. Data are normalized by the GCOS method, TGT value of 100. Most tissues were sampled in triplicate.

[Click Here for Table of Expression Values](#)

[Click Here for Chart of Expression Values](#)

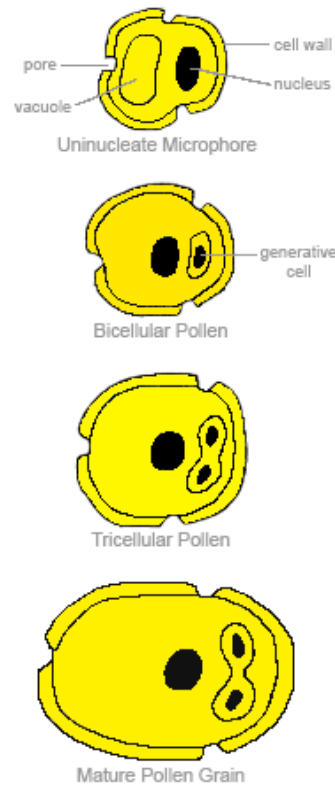
Root Cells Types



- Material from the roots of 6 day old wild-type Col-0 *Arabidopsis thaliana* plants was analyzed
- Plants grown under 16/8 hour light/dark conditions on MS + 4.5% sucrose media
- Root cell types were isolated by protoplasting and fluorescence-activated cell sorting. RNA was extracted from protoplasts and hybridized to ATH1 GeneChips
- The data were normalized by GCOS normalization, TGT 100. Triplicate measurements were made.

Results from Birnbaum et al. (2003) Science 302:1956 and Naway et al. (2005) Plant Cell 17:1908.

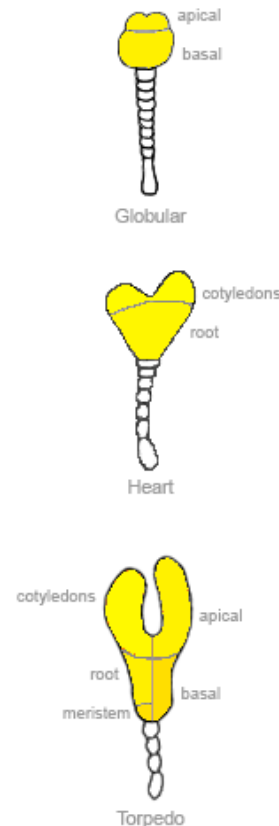
Microgametogenesis (Pollen Development)



- Plant material from the pollen of 5.10 growth stage wild-type *Arabidopsis thaliana* plants of Ler-0 ecotype was analyzed
- Plants grown under 16/8 hour light/dark conditions at 21°C
- All measurements were taken in duplicates - the average of which is shown
- RNA was isolated and hybridized to the ATH1 GeneChip
- The data were normalized by GCOS normalization, TGT 100

Honyes & Twell (2004) Gen. Biol. 5:R85

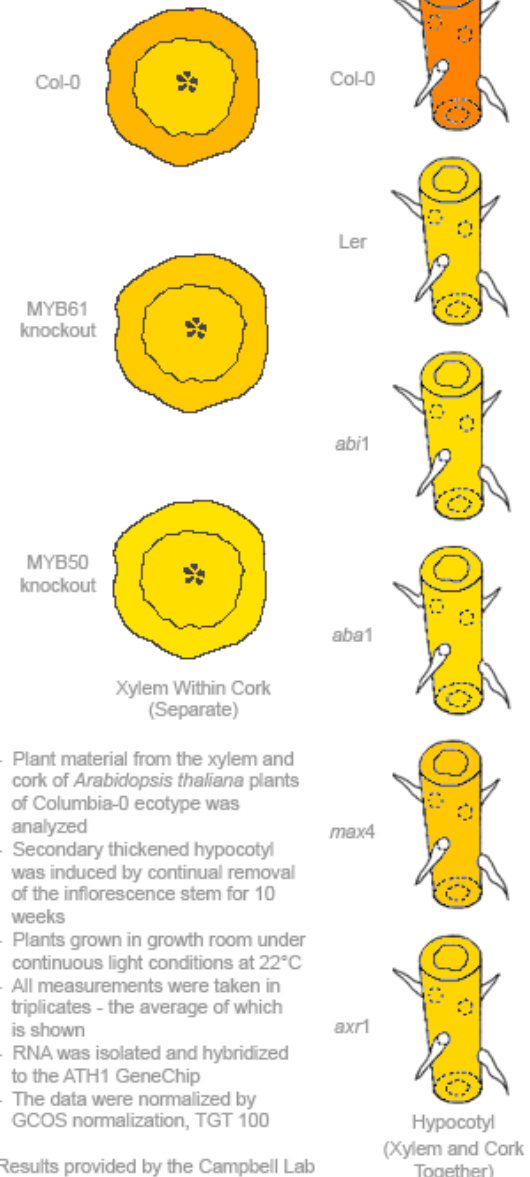
Embryo Development



- Plant material from embryos of wild-type Col-0 *Arabidopsis thaliana* plants of was isolated by laser capture microdissection
- Plants grown under 16/8 hour light/dark conditions
- RNA was amplified and hybridized to the ATH1 GeneChip. Note: 3' bias!
- All measurements were taken in triplicates - the average is shown.
- Results can be highly variable - standard deviation filtering advisable!
- The data were normalized by GCOS normalization, TGT 100

Casson et al. (2005) Plant J. 42:111

Xylem and Cork

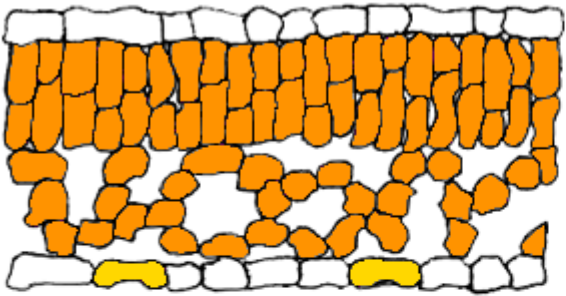


- Plant material from the xylem and cork of *Arabidopsis thaliana* plants of Columbia-0 ecotype was analyzed
- Secondary thickened hypocotyl was induced by continual removal of the inflorescence stem for 10 weeks
- Plants grown in growth room under continuous light conditions at 22°C
- All measurements were taken in triplicates - the average of which is shown
- RNA was isolated and hybridized to the ATH1 GeneChip
- The data were normalized by GCOS normalization, TGT 100

Results provided by the Campbell Lab

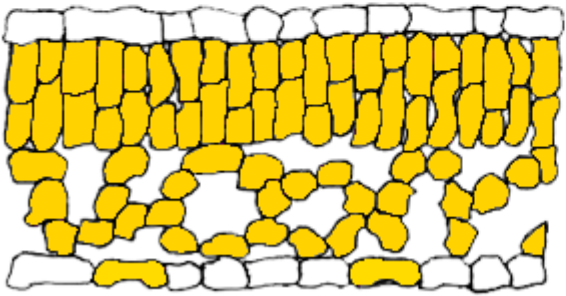
Guard and Mesophyll Cells

Water Spray for 4 Hours

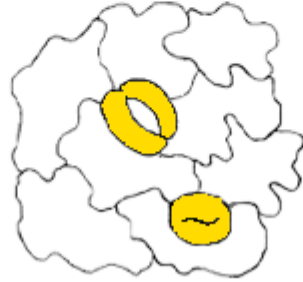
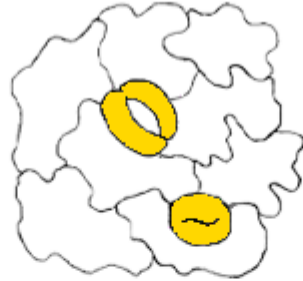


Proto-alert!

100 uM ABA for 4 Hours



Cross Section of Leaf

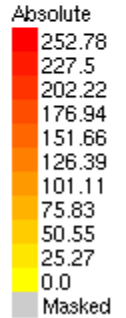
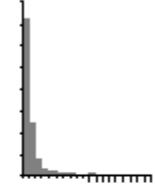


Surface View of Leaf

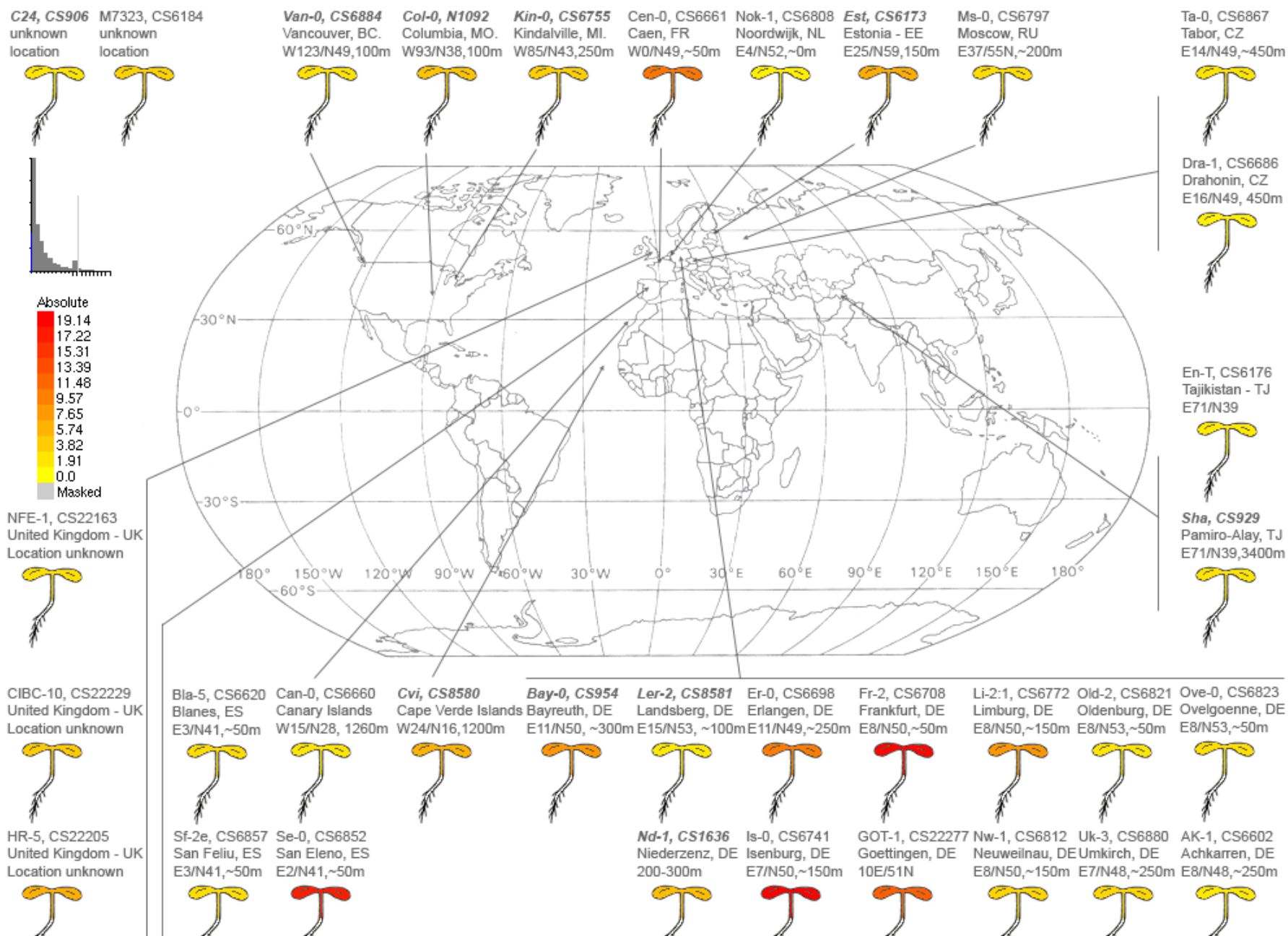
- Plant material from the leaves of 5 week old *Arabidopsis thaliana* plants of Columbia-0 ecotype was analyzed
- Mesophyll and guard cell protoplasts were used
- Plants grown in growth room under 16/8 hour light/dark conditions at 22°C
- Measurements were taken in duplicates - the average of which is shown. Actinomycin and cordycepin were added to one of the reps. The individual reps are displayed in the table below.
- RNA was isolated and hybridized to the ATH1 GeneChip. The data were normalized by GCOS normalization, TGT 100.

Results from Yang et al. (2008) Plant Methods 4:6

	-	+	
			actinomycin and cordycepin
			GCs without ABA treatment
			GCs with ABA treatment
			Mesophyll cells without ABA
			Mesophyll cells with ABA



Natural Variation eFP Browser. Data from the Weigel Lab (Lempe *et al.*, 2005, PLoS Genetics 1:e6). Aerial parts of 4 day old seedlings greenhouse-grown in soil at 23C under continuous light were sampled. Data normalized by the GCOS method, TGT value of 100. Plant material sampled in triplicate where *indicated by italics*, otherwise once.



Détection de polymorphisme de séquence et GWAS

GWAS : Genome-wide association study = étude d'association pangénomique

Sur des milliers d'individus, on cherche des associations entre variations génétiques (SNP) et variations de certains caractères étudiés

SNP = Single nucleotide polymorphism = polymorphisme d'un seul nucléotide (une base modifiée dans une séquence)

Wang *et al.* *Genome Biology* (2018) 19:72
<https://doi.org/10.1186/s13059-018-1444-y>

Genome Biology

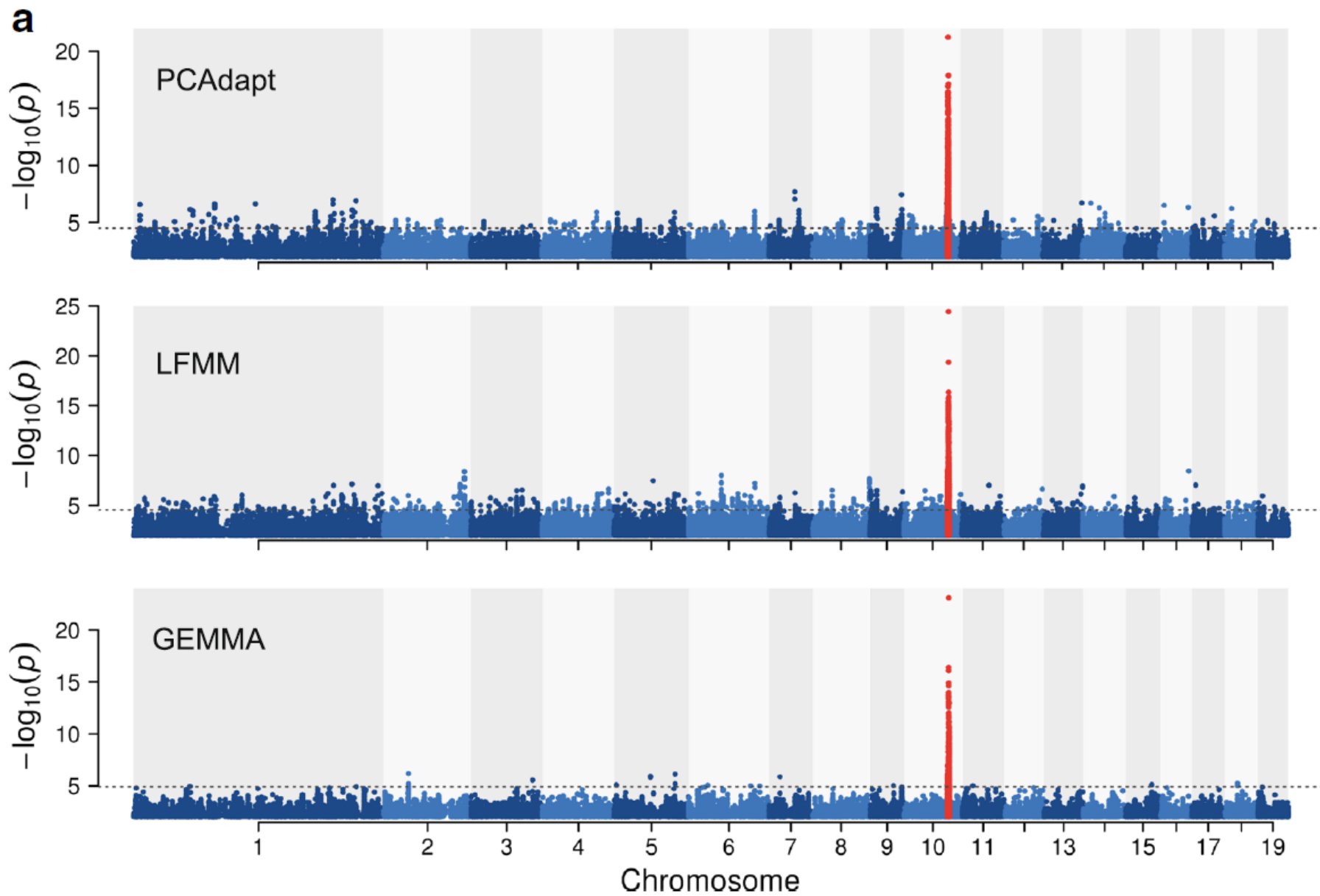
RESEARCH

Open Access

A major locus controls local adaptation and adaptive life history variation in a perennial plant



Jing Wang^{1,2*} , Jihua Ding³, Biyue Tan^{1,4}, Kathryn M. Robinson⁵, Ingrid H. Michelson⁵, Anna Johansson⁶, Björn Nystedt⁶, Douglas G. Scofield^{1,7,8}, Ove Nilsson³, Stefan Jansson⁵, Nathaniel R. Street⁵ and Pär K. Ingvarsson^{1,9*}



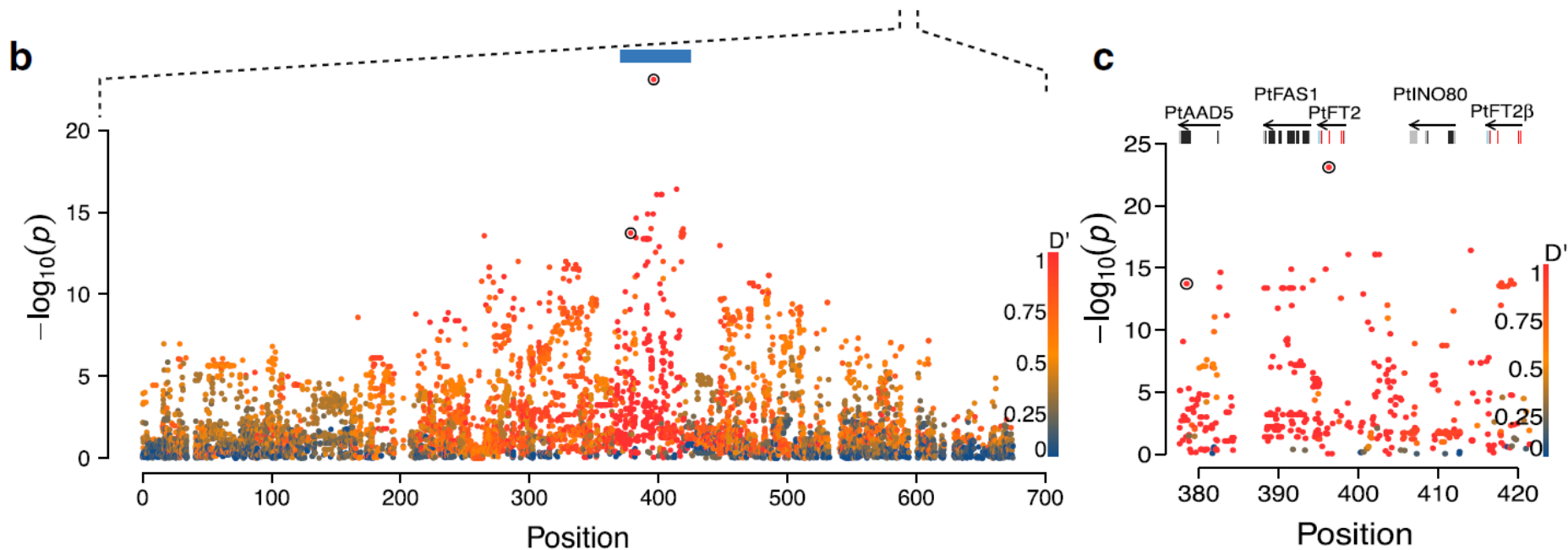


Fig. 2 Local adaptation signals across the genome. **a** Manhattan plots for SNPs associated with population structure (PCAdapt), climate variation (LFMM), and phenotype (GEMMA). The 700-kbp region surrounding *PtFT2* gene (marked in red) is identified by all methods. The dashed line represents the significance threshold for each method. Quantile-quantile plot is displayed in the right panel, with significant SNPs highlighted in red. **b** Magnification of the phenotype association results (from GEMMA) for the region surrounding *PtFT2* on Chr10. The coordinates correspond to the region 16.3 Mbp-17.0 Mbp on Chr10. Individual data points are colored according to LD with the most strongly associated SNP (Potra001246:25256). The two potential causal variants identified by CAVIAR within this region are marked by black circles. **c** Close-up view of the phenotype association results (from GEMMA) in a region corresponding to the blue bar in (b). This region contains the two *PtFT2* homologs (red - exons, blue - UTRs) and several other genes (dark gray - exons, light grey - UTRs)

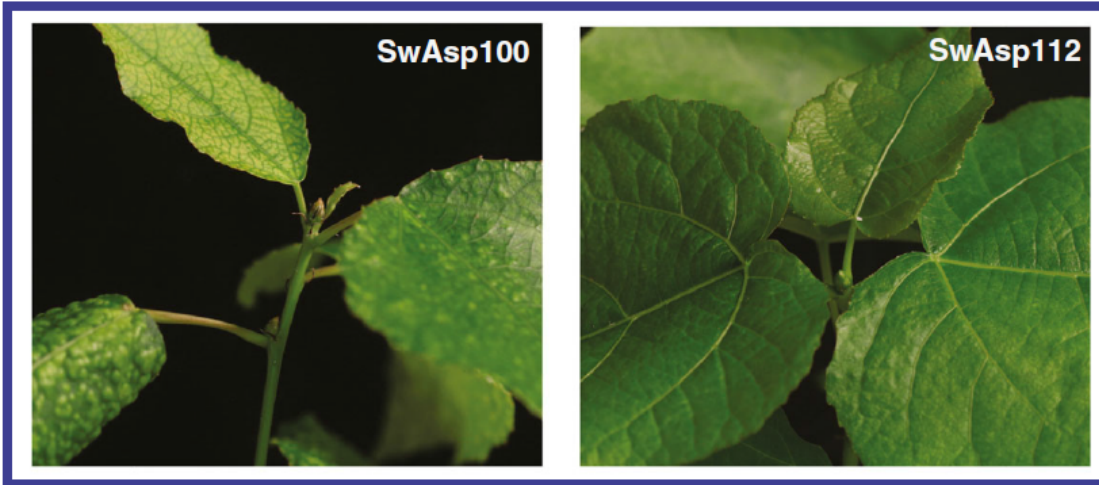
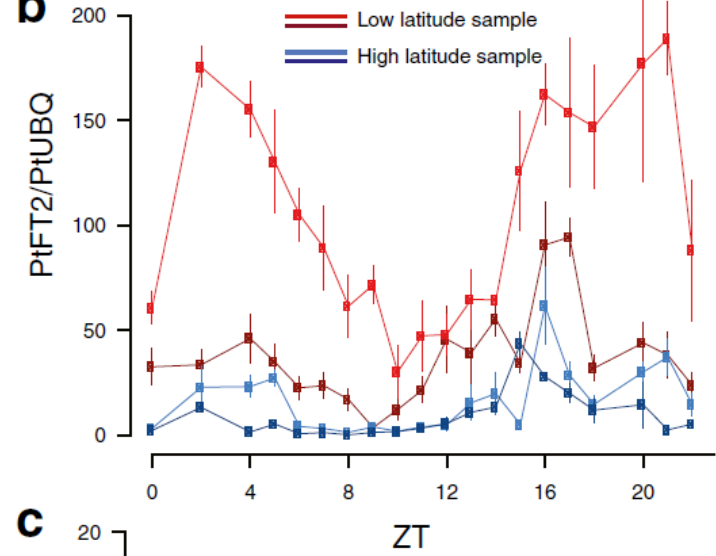
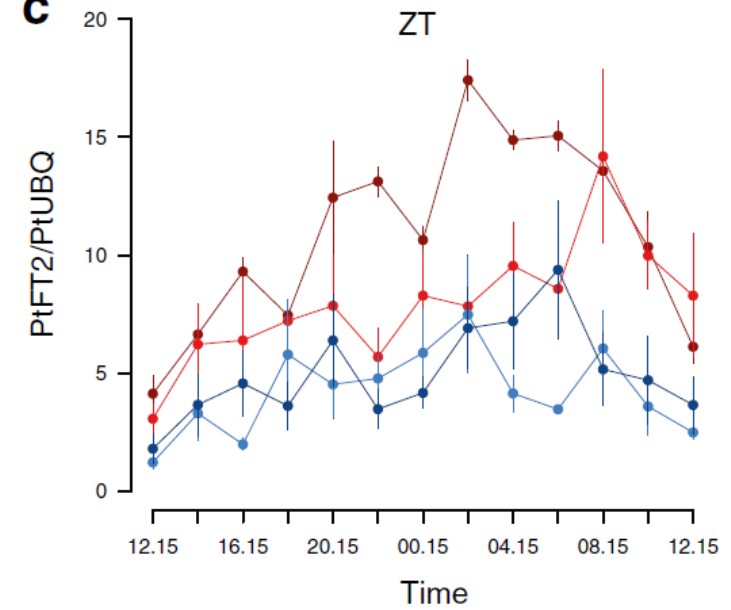
a**b****c**

Fig. 5 *PtFT2* expression affects short-day induced growth cessation and bud set in *P. tremula*. **a** Bud set phenotype under 19-h day-length conditions. Two southern clones (marked with a *red box*, SwAsp 018, Ronneby, latitude 56.2 °N; SwAsp 023, Vårgårda, latitudes 58 °N) and two northern clones (marked with a *blue box*, SwAsp 100, Umeå, latitude 63.9 °N; SwAsp 112, Luleå, latitudes 65.7 °N) were chosen to be analyzed. Trees were grown under 23-h day length for one month and then shifted to 19-h day length. Photos were taken one month after the shift to 19-h day length. **b** Dynamic expression analysis of *PtFT2* in two southern clones (*red*, SwAsp018 and SwAsp023) and two northern clones (*blue*, SwAsp100 and SwAsp112) from the greenhouse experiment. The genotypes of these trees at the most strongly associated *PtFT2* SNP are SwAsp018: T/T, SwAsp023: T/T, SwAsp100: not available and SwAsp 112: G/G. Samples for RT-PCR were taken two weeks after the trees were shifted to 19-h day length. *Error bars*, ±standard deviation. ZT zeitgeber time. **c** Dynamic expression analysis of *PtFT2* in two southern clones (*red*, SwAsp005 and SwAsp023) and two northern clones (*blue*, SwAsp100 and SwAsp116) from common garden experiment. The genotypes of these trees at the most strongly associated *PtFT2* SNP are SwAsp005: T/T, SwAsp023: T/T, SwAsp100: not available and SwAsp 112: G/G. Samples were collected in the Sävar common garden in early July 2014

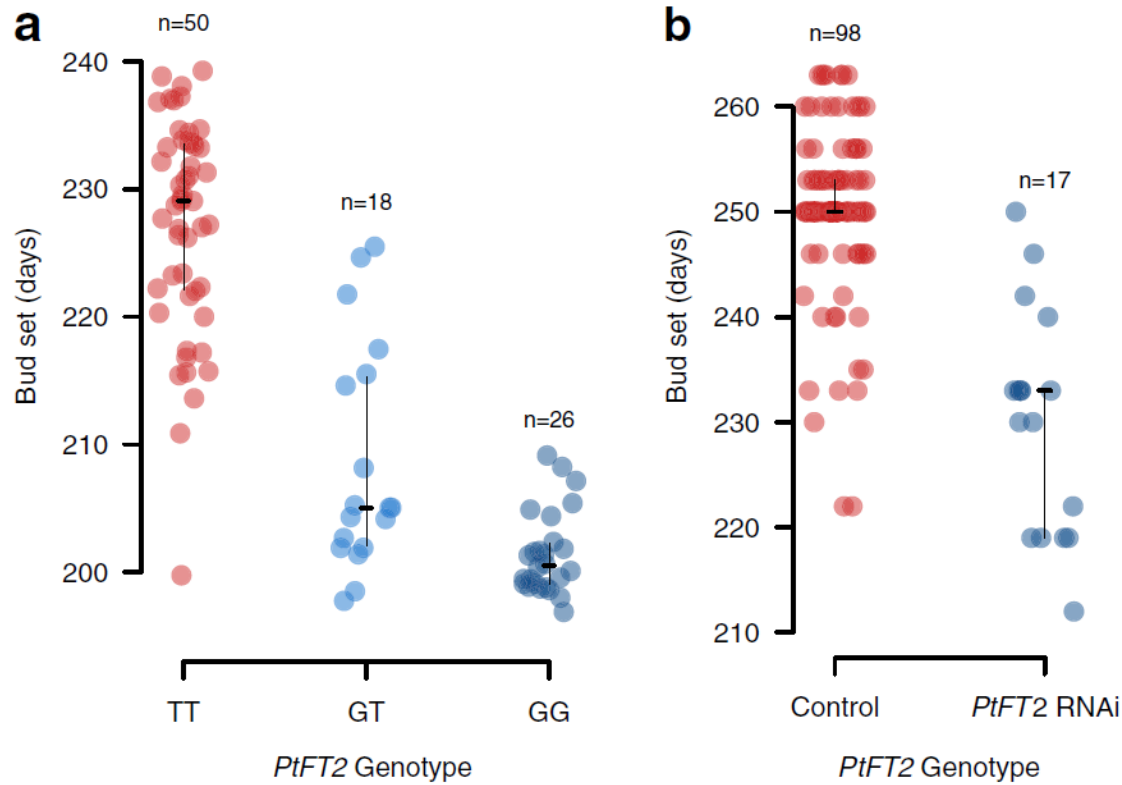


Fig. 6 Phenotypic effects of *PtFT2*. **a** The timing of bud set for the three genotypes classes at the *PtFT2* SNP (Potra001246:25256) that displays the strongest signal of local adaptation identified by all three methods as shown in Fig. 2a. The plot displays mean genotype bud set after correcting for common garden site, year, and block effects. The horizontal line indicates the median value and the vertical line marks the interquartile range. The number of genotypes in the respective classes is indicated above the figure. **b** The timing of bud set for wild type control lines and transgenic *PtFT2* lines in the field experiments at Våxtorp. The structure of the plots is the same as in (a)

IV. Ressources disponibles

Liste non exhaustive...

Bases de données Génomes

Phytozome

<https://phytozome.jgi.doe.gov/pz/portal.html>

10/2018 : 93 génomes de 82 espèces de plantes supérieures

Sites dédiés pour une espèce, généralement avec des données expertisées

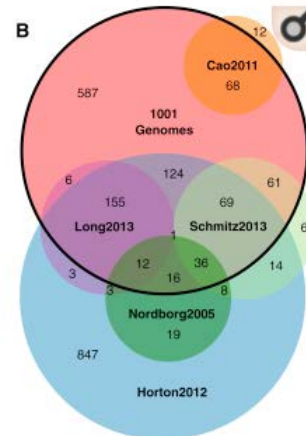
Ex : Arabidopsis

TAIR = The Arabidopsis Information Resource

<https://www.arabidopsis.org/>

<https://1001genomes.org/>

<i>Amaranthus hypochondriacus</i> v1.0	<i>Cucumis sativus</i> v1.0	<i>Panicum hallii</i> var. <i>hallii</i> v2.1
<i>Amaranthus hypochondriacus</i> v2.1	<i>Daucus carota</i> v2.0	<i>Panicum virgatum</i> v1.1
<i>Amborella trichopoda</i> v1.0	<i>Dunaliella salina</i> v1.0	<i>Panicum virgatum</i> v4.1
<i>Anacardium occidentale</i> v0.9	<i>Eucalyptus grandis</i> v2.0	<i>Phaseolus vulgaris</i> v2.1
<i>Ananas comosus</i> v3	<i>Eutrema salsugineum</i> v1.0	<i>Physcomitrella patens</i> v3.3
<i>Aquilegia coerulea</i> v3.1	<i>Fragaria vesca</i> v1.1	<i>Populus deltoides</i> WV94 v2.1
<i>Arabidopsis halleri</i> v1.1	<i>Glycine max</i> Wm82.a2.v1	<i>Populus trichocarpa</i> v3.0
<i>Arabidopsis lyrata</i> v2.1	<i>Gossypium hirsutum</i> v1.1	<i>Populus trichocarpa</i> v3.1
<i>Arabidopsis thaliana</i> Araport11	<i>Gossypium raimondii</i> v2.1	<i>Porphyra umbilicalis</i> v1.5
<i>Arabidopsis thaliana</i> TAIR10	<i>Helianthus annuus</i> r1.2	<i>Prunus persica</i> v2.1
<i>Asparagus officinalis</i> V1.1	<i>Hordeum vulgare</i> r1	<i>Ricinus communis</i> v0.1
<i>Boechera stricta</i> v1.2	<i>Kalanchoe fedtschenkoi</i> v1.1	<i>Salix purpurea</i> v1.0
<i>Botryococcus braunii</i> v2.1	<i>Kalanchoe laxiflora</i> v1.1	<i>Selaginella moellendorffii</i> v1.0
<i>Brachypodium distachyon</i> Bd21-3 v1.1	<i>Lactuca sativa</i> V8	<i>Setaria italica</i> v2.2
<i>Brachypodium distachyon</i> v3.1	<i>Linum usitatissimum</i> v1.0	<i>Setaria viridis</i> v1.1
<i>Brachypodium hybridum</i> v1.1	<i>Malus domestica</i> v1.0	<i>Setaria viridis</i> v2.1
<i>Brachypodium stacei</i> v1.1	<i>Manihot esculenta</i> v6.1	<i>Solanum lycopersicum</i> iTAG2.4
<i>Brachypodium sylvaticum</i> v1.1	<i>Marchantia polymorpha</i> v3.1	<i>Solanum tuberosum</i> v4.03
<i>Brassica oleracea capitata</i> v1.0	<i>Medicago truncatula</i> Mt4.0v1	<i>Sorghum bicolor</i> Rio v2.1
<i>Brassica rapa</i> FPsc v1.3	<i>Micromonas pusilla</i> CCMP1545 v3.0	<i>Sorghum bicolor</i> v3.1.1
<i>Capsella grandiflora</i> v1.1	<i>Micromonas</i> sp. RCC299 v3.0	<i>Sphagnum fallax</i> v0.5
<i>Capsella rubella</i> v1.0	<i>Mimulus guttatus</i> v2.0	<i>Spirodela polyrhiza</i> v2
<i>Carica papaya</i> ASGPBv0.4	<i>Miscanthus sinensis</i> v7.1	<i>Theobroma cacao</i> v1.1
<i>Chenopodium quinoa</i> v1.0	<i>Musa acuminata</i> v1	<i>Trifolium pratense</i> v2
<i>Chlamydomonas reinhardtii</i> v5.5	<i>Olea europaea</i> var. <i>sylvestris</i> v1.0	<i>Triticum aestivum</i> v2.2
<i>Chromochloris zofingiensis</i> v5.2.3.2	<i>Oropetium thomaeum</i> v1.0	<i>Vigna unguiculata</i> v1.1
<i>Cicer arietinum</i> v1.0	<i>Oryza sativa</i> Kitaake v3.1	<i>Vitis vinifera</i> Genoscope.12X
<i>Citrus clementina</i> v1.0	<i>Oryza sativa</i> v7_JGI	<i>Volvox carteri</i> v2.1
<i>Citrus sinensis</i> v1.1	<i>Ostreococcus lucimarinus</i> v2.0	<i>Zea mays</i> Ensembl-18
<i>Coccomyxa subellipsoidea</i> C-169 v2.0	<i>Panicum hallii</i> v2.0	<i>Zea mays</i> PH207 v1.1
<i>Coffea arabica</i> UCDv0.5	<i>Panicum hallii</i> v3.1	<i>Zostera marina</i> v2.2



Bases de données Stockage de données brutes Transcriptome

GEO

<https://www.ncbi.nlm.nih.gov/geo/>

Gene Expression Omnibus



GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.


Getting Started

- [Overview](#)
- [FAQ](#)
- [About GEO DataSets](#)
- [About GEO Profiles](#)
- [About GEO2R Analysis](#)
- [How to Construct a Query](#)
- [How to Download Data](#)

Tools

- [Search for Studies at GEO DataSets](#)
- [Search for Gene Expression at GEO Profiles](#)
- [Search GEO Documentation](#)
- [Analyze a Study with GEO2R](#)
- [Studies with Genome Data Viewer Tracks](#)
- [Programmatic Access](#)
- [FTP Site](#)

Browse Content

Repository Browser	
DataSets:	4348
Series: 	103808
Platforms:	18995
Samples:	2689466

GEO DataSets ▾ poplar AND wood |

[Create alert](#) [Advanced](#)

[Summary](#) ▾ [20 per page](#) ▾ [Sort by Default order](#) ▾

Search results

Items: 1 to 20 of 181

Series GSE49911

[Query DataSets for GSE49911](#)

Status	Public on Apr 01, 2014
Title	SND1 Transcription Factor-Directed Quantitative Functional Hierarchical Genetic Regulatory Network in Wood Formation in <i>Populus trichocarpa</i>
Organism	Populus trichocarpa
Experiment type	Expression profiling by high throughput sequencing
Summary	We focused on RNA-seq-based full transcriptome responses to PtrSND1-B1 overexpression at 7, 12, and 25 h in stem differentiating xylem (SDX) protoplasts
Overall design	We transfected PtrSND1-B1 and sGFP into stem differentiating xylem protoplasts and performed RNA-seq to reveal the whole transcriptome.
Contributor(s)	Chiang VL
Citation(s)	Lin YC, Li W, Sun YH, Kumari S et al. SND1 transcription factor-directed quantitative functional hierarchical genetic regulatory network in wood formation in <i>Populus trichocarpa</i> . <i>Plant Cell</i> 2013 Nov;25(11):4324-41. PMID: 24280390
Submission date	Aug 15, 2013
Last update date	Jun 17, 2016
Contact name	Vincent L Chiang
E-mail	vchiang@ncsu.edu
Organization name	North Carolina State University
Department	Forestry and Environmental Resources
Lab	Forest Biotechnology Group
Street address	840 Main Campus Drive
City	Raleigh
State/province	NC
ZIP/Postal code	27695
Country	USA
Platforms (1)	GPL17579 Illumina Genome Analyzer Iix (<i>Populus trichocarpa</i>)
Samples (18)	GSM1209600 GFP (7h)-1 GSM1209601 GFP (7h)-2 GSM1209602 GFP (7h)-3 More...
Relations	
BioProject	PRJNA215318
SRA	SRP028843

Bases de données Intégration de données transcriptome

Ex : PopGenIE = Populus Genome Integrative Explorer

<https://www.ncbi.nlm.nih.gov/geo/>

NEW popgenie.ORG

Potri.018G142700.5,POPTR_0001s03760, AT2G32080.2 or desc

Genome Tools Expression Tools Analysis Tools Community Annotation FTP Projects Help **Populus trichocarpa**

Welcome to v3 PopGenIE: The Populus Genome Integrative Explorer

Click here to [Start a tour](#) of PopGenIE basic features.

Genome Data

- ✓ *P. trichocarpa* v3.0 - Phytozome v10.1
- P. tremula* v1.1 - UPSC
- P. tremuloides* v1.1 - UPSC
- P. tremula X tremuloides*-'T89' v0.1 - UPSC [draft version]
- P. tremula X alba* - '717-1B4' v1.1 - AspenDB
- S. suchowensis* v4.1 - Nanjing Forestry University

Expression Data

exImage and exPlot

- ✓ *P. trichocarpa* Tissues microarray
- ✓ *P. tremula* exDiversity RNA-seq
- ✓ *P. tremula* exAtlas RNA-seq

exNet and exHeatmap

- ✓ *P. tremula* exAtlas RNA-seq
- ✓ *P. trichocarpa* microarray

ComPIEx

atgenie.ORG

plantgenie.ORG

congenie.ORG

- ✓ BLAST
- ✓ GeneList
- ✓ Gene Pages
- ✓ exImage
- ✓ exNet
- ✓ exPlot
- ✓ exHeatmap
- ✓ Enrichment
- ✓ Chromosome Diagram
- ✓ GBrowse
- ✓ Sequence Search
- ✓ WebApollo

NB : Chaque année, la revue NAR sort un numéro spécial sur les bases de données de biologie moléculaire



Volume 46, Issue D1
4 January 2018

Article Contents

Abstract

NEW AND UPDATED
DATABASES

NAR ONLINE MOLECULAR
BIOLOGY DATABASE
COLLECTION

ACKNOWLEDGEMENTS



FUNDING

REFERENCES

Comments (0)

Next >

The 2018 *Nucleic Acids Research* database issue and the online molecular biology database collection

Daniel J Rigden , Xosé M Fernández 

Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D1–D7, <https://doi.org/10.1093/nar/gkx1235>

Published: 18 December 2017 **Article history** ▼

Abstract

The 2018 *Nucleic Acids Research* Database Issue contains 181 papers spanning molecular biology. Among them, 82 are new and 84 are updates describing resources that appeared in the Issue previously. The remaining 15 cover databases most recently published elsewhere. Databases in the area of nucleic acids include 3DIV for visualisation of data on genome 3D structure and RNArchitecture, a hierarchical classification of RNA families. Protein databases include the established SMART, ELM and MEROPS while GPCRdb and the newcomer STCRDab cover families of biomedical interest. In the area of metabolism, HMDB and Reactome both report new features while PULDB appears in NAR for the first time. This issue also contains reports on genomics resources including Ensembl, the UCSC Genome Browser and ENCODE. Update papers from the IUPHAR/BPS Guide to Pharmacology and DrugBank are highlights of the drug and drug target section while a number of proteomics databases including proteomicsDB are also covered. The entire Database Issue is freely available online on the *Nucleic Acids Research* website (<https://academic.oup.com/nar>). The NAR online Molecular Biology Database Collection has been updated, reviewing 138 entries, adding 88 new resources and eliminating 47 discontinued URLs, bringing the current total to 1737 databases. It is available at <http://www.oxfordjournals.org/nar/database/c/>.

Issue Section: [Database Issue](#)

Conclusions

Analyse du transcriptome : génère des données de grandes dimensions, qu'il faut stocker et savoir traiter - Besoins en statistiques et en bioinformatique importants. Besoins de mettre en œuvre des méthodes pour les représenter.

Les cellules / le tissu de départ sont / est primordial pour l'interprétation biologique des résultats. Actuellement se développent des approches Single cell pour s'affranchir de l'effet « mélange de cellules » et permettant de prendre en compte les phénomènes de différenciation des cellules.

Ces données sont disponibles et partagées : Openscience / Opendata
On peut ré-analyser des données de séquence générées par des collègues avec une autre question de recherche.

L'analyse du transcriptome permet de faire un catalogue exhaustif des gènes exprimés dans un organe / un tissu / une cellule donnée, d'identifier de nouveaux gènes, de mieux comprendre certains processus biologiques et peut également être utilisée pour des études de polymorphismes génétiques.

Références

Articles de revue

Bunch, H. (2018). Gene regulation of mammalian long non-coding RNA. *Molecular Genetics and Genomics* 293, 1–15.

Cieślak, M., and Chinnaiyan, A.M. (2017). Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics* 19, 93–109.

Ijaq, J., Chandrasekharan, M., Poddar, R., Bethi, N., and Sundararajan, V.S. (2015). Annotation and curation of uncharacterized proteins-challenges. *Frontiers in Genetics* 6.

Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research*.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology* 13, e1005457.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10, 57–63.

Articles de recherche

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.

Bai, C., Wu, Y., Cao, B., Xu, J., and Li, G. (2018). De novo transcriptome assembly based on RNA-seq and dynamic expression of key enzyme genes in loganin biosynthetic pathway of *Cornus officinalis*. *Tree Genetics & Genomes* 14.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 44, D279–D285.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., Bhalerao, R., Uhlen, M., Teeri, T.T., Lundeberg, J., et al. (2001). A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences* 98, 14732–14737.

Li, Y., Wei, W., Feng, J., Luo, H., Pi, M., Liu, Z., and Kang, C. (2018). Genome re-annotation of the wild strawberry *Fragaria vesca* using extensive Illumina- and SMRT-based RNA-seq datasets. *DNA Research* 25, 61–70.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in *Arabidopsis*. *Cell* 133, 523–536.

Plass, M., Solana, J., Wolf, F.A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F.J., Kocks, C., and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 360, eaaq1723.

Sundell, D., Street, N.R., Kumar, M., Mellerowicz, E.J., Kucukoglu, M., Johnsson, C., Kumar, V., Mannapperuma, C., Delhomme, N., Nilsson, O., et al. (2017). AspWood: High-Spatial-Resolution Transcriptome Profiles Reveal Uncharacterized Modularity of Wood Formation in *Populus tremula*. *The Plant Cell* 29, 1585–1604.

Wang, J., Ding, J., Tan, B., Robinson, K.M., Michelson, I.H., Johansson, A., Nystedt, B., Scofield, D.G., Nilsson, O., Jansson, S., et al. (2018). A major locus controls local adaptation and adaptive life history variation in a perennial plant. *Genome Biology* 19.

Ouvrage

Tagu, D., and Moussard, C. (2003). *Principes des techniques de biologie moléculaire* (Paris: Institut National de la Recherche Agronomique).