Exploitation des ressources URGV pour l'inférence de réseaux chez *Arabidopsis thaliana*

Marie-Laure Martin-Magniette

Unité de Recherche en Génomique Végétale, Evry

UMR AgroParisTech/INRA MIA, Paris

Unité de Recherche en Génomique Végétale

UMR INRA 1165 - Université d'Evry Val d'Essonne
ERL CNRS 8196 'Bioinformatique et Génomique Végétale'

Directeur: H Hirt
Directeurs adjoints: C Lurin, A Bendahmane et B Sturbois
Directeur ERL CNRS: A Lecharny
Directeur administratif: A Charpentier

Vert = INRA
Rose = Université
Bleu = CNRS
Noir = non permanent

Génomique d'Arabidopsis

Génomique fonctionnelle d'Arabidopsis

C Lurin (DR2)
R Berthomé (CR1)
S Balzergue (IE2)
L Soubigou (AI)
A Avon (TR)
L Heurtevin (TR)
S Huguet (TR)

S Arribat (IE)
C Boussardon (PhD)
E Blondet (AI)
A Falcon (IE/PhD)
D Gey (IE)
D Monachello (IE)

Signalisation et Protéomique

H Hirt (DR1)
J Colcombet (CR2)
S Pateyron (TR)
J Bigeard (IE)

S Berriri (PhD)
E Busero (Post-doc)
A Charrier (PhD)
A Danquah (PhD)
A Evrard (Post-doc)
A Winger (Post-doc)
A Garcia (Post-doc)

Bioinformatique et génomique prédictive

S Aubourg (CR1)
A Lecharny (DR2)
ML Martin (CR1)
V Brunaud (IR2)
JP Tamby (IE2)
C Guichard (IE2)

S Dèrozier (IE)
O Rogier (IE)
A Cauchard (Appr.)

Génomique de la vigne et des arbres forestiers

A Adam-Blondon (DR2)
P Faivre-Rampant (CR1)
A Canaguier (IE2)
F Bitton (IE2)
I Le Clainche (TR)
R Bounon (TR)

C Houel (PhD)
CI Zah-Bi (PhD)
A Bresson (PhD)

Génomique des plantes cultivées

Génomique fonctionnelle des plantes cultivées

A Bendahmane (DR2)
B Sturbois (PR)
C Clépet (CR1)
A Boualem (IE2/PhD)
M Dalmais (IE2)
D Jublot (AI)
C Troadec (AI)

K Boudehri (Post-doc)
F Dahmani (IE)
J Eleblu (PhD)
S Elftieh (IE)
S Fleurier (Appr.)
C Foucart (Post-doc)
W Mouhaya (Post-doc)
A Leveau (PhD)
F Marcel (IE)
A Martin (Post-doc)
R Fernandez (PhD)
F Izhaq (PhD)
B Lasseur (ATER)

Organisation et évolution des génomes des plantes cultivées

B Chalhoub (DR2)
N Boudet (MdC)
H Belcram (AI)
C Huneau (AJT)

M Charles (Post-doc)
I Mestiri (PhD)

Administration

Directeur administratif — A Charpentier (IE2)
Secrétariat — S Genet (AT)
Finances — J Collinot-Maciel (TR)
Contrat et embauches — M Atlan (AI)

Services techniques

Chambres de culture et serres — P Audigier (TR), C Lepage (AT)
Laverie et milieux — L Bataille (AT)
Informatique — P Grevet (AI), JL Collomb (TR)
Maintenance des équipements — J Martinet (AI)

2, rue Gaston Crémieux, 91057 Evry Cedex, France - http://www.versailles.inra.fr/urgv/

20 juillet 2010

Contexte international

Arabidopsis thaliana séquencée en 2000, plante modèle chez les brassicacées
Mise à jour de son génome tous les ans

En 2007, Ma *et al.* (Genome Research) étudient le réseau de régulation.

- GGM avec estimation par la méthode de Schäffer et Strimmer.
- Données utilisées : 22 000 gènes à l'aide de 3000 expériences Affymétrix (environ 150 conditions expérimentales différentes).

Depuis 2008, quelques articles sur l'inférence de réseaux de co-expression.

- Exploitation des données Affymétrix déposés à GEO ou ArrayExpress
- Utilisation de la corrélation de Pearson ou Spearman
- Un article de review de Usadel *et al.* Plant, Cell & Environment en 2009

URGV: Transcriptomes et autres applications

(ChIP-chip, CGH….)

La plate-forme produit les puces, conseille pour les plans d'expériences, fait les hybridations et fournit les premières analyses statistiques . Elle dispose également de la base de données CATdb (Gagnot *et al*., 2008, *NAR*)

CATMA : puce Arabidopsis

2003-2010 : 200 collaborations : ANR/GnP Programmes, EU + nombreuses collabs avec INRA , CNRS, CEA, Univ. et labo étrangers (Bra, Sin, Cz, Ge, Be, UK, Sp, Swe) > **4300 swaps, 50 publications**

SAP: puce promoteurs Arabidopsis

2 Programmes GnP 53 hybridations, **Benhamed *et al.* 2008 *Plant J.***

Plate-forme Affymetrix (pour les plantes cultivées)

2 Programmes GnP, **55 collaborations > 1000 hybridizations, 9 publications**

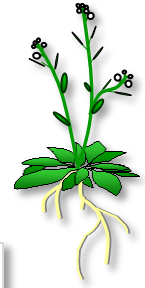
TAG : puce tiling-array Arabidopsis (Nimblegen)

1 Programme GnP

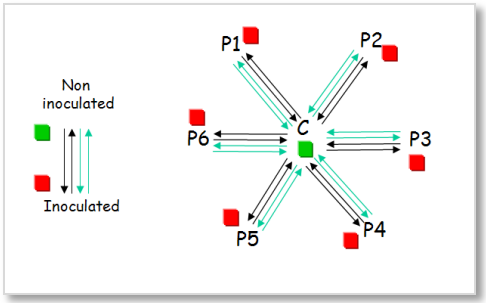
CATseq et BIOS : RNA-seq

AIP bioressources GAP et AIP bases de données

Démarche de la plate-forme



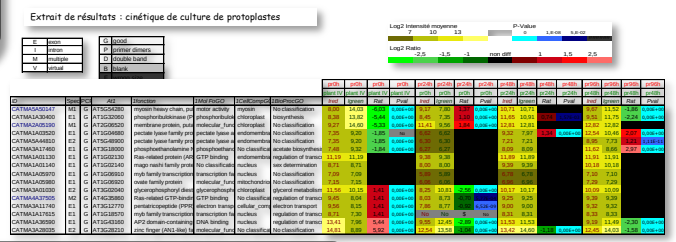
Choix de la puce (méthode)



Dessin des expériences

Hybridations

Analyses statistiques



Stockage dans CATdb

CATdb a Complete Arabidopsis Transcriptome database

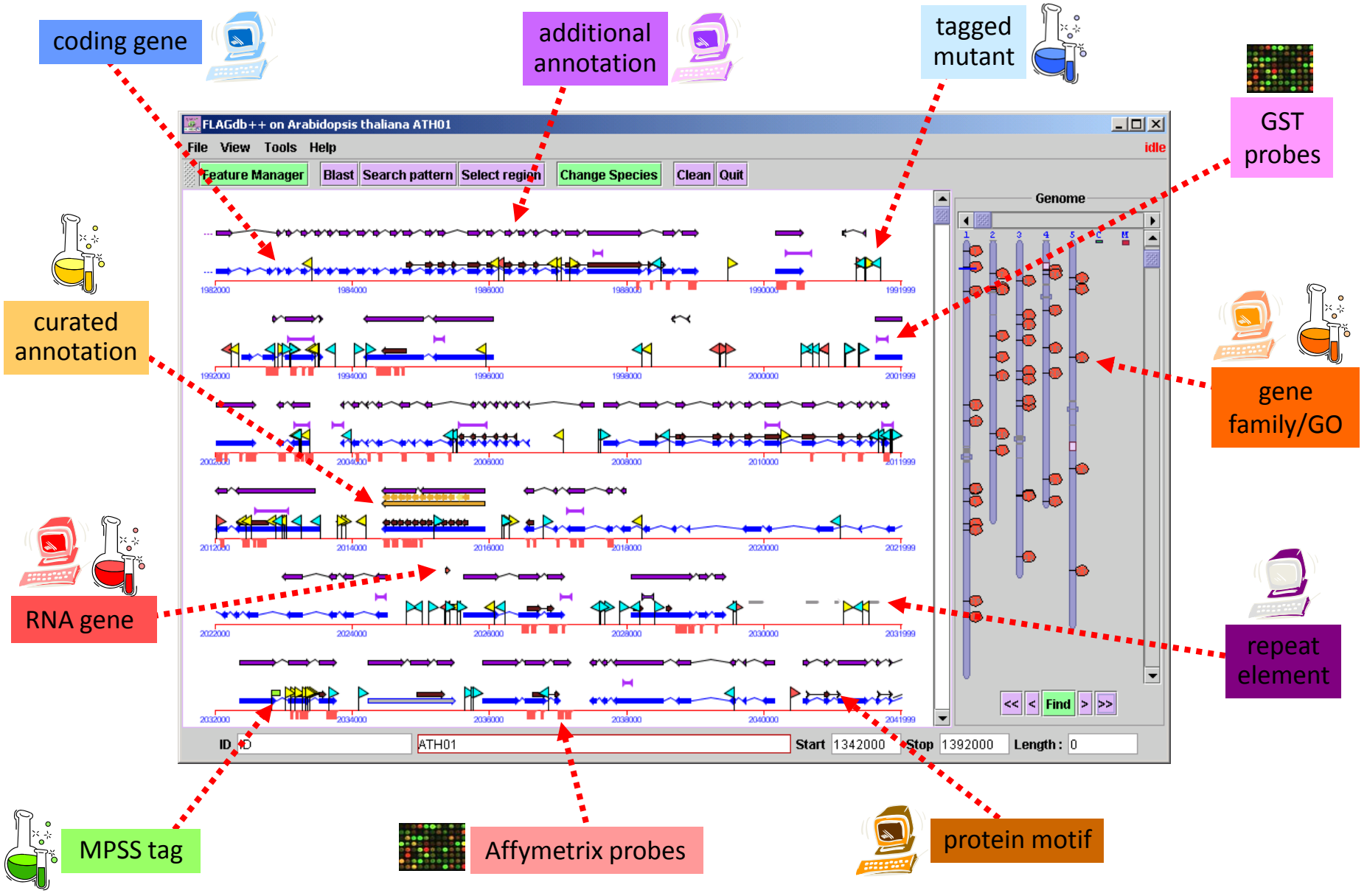
Aide à la lecture des résultats

FLAGdb++ CATdb

Au niveau statistiques

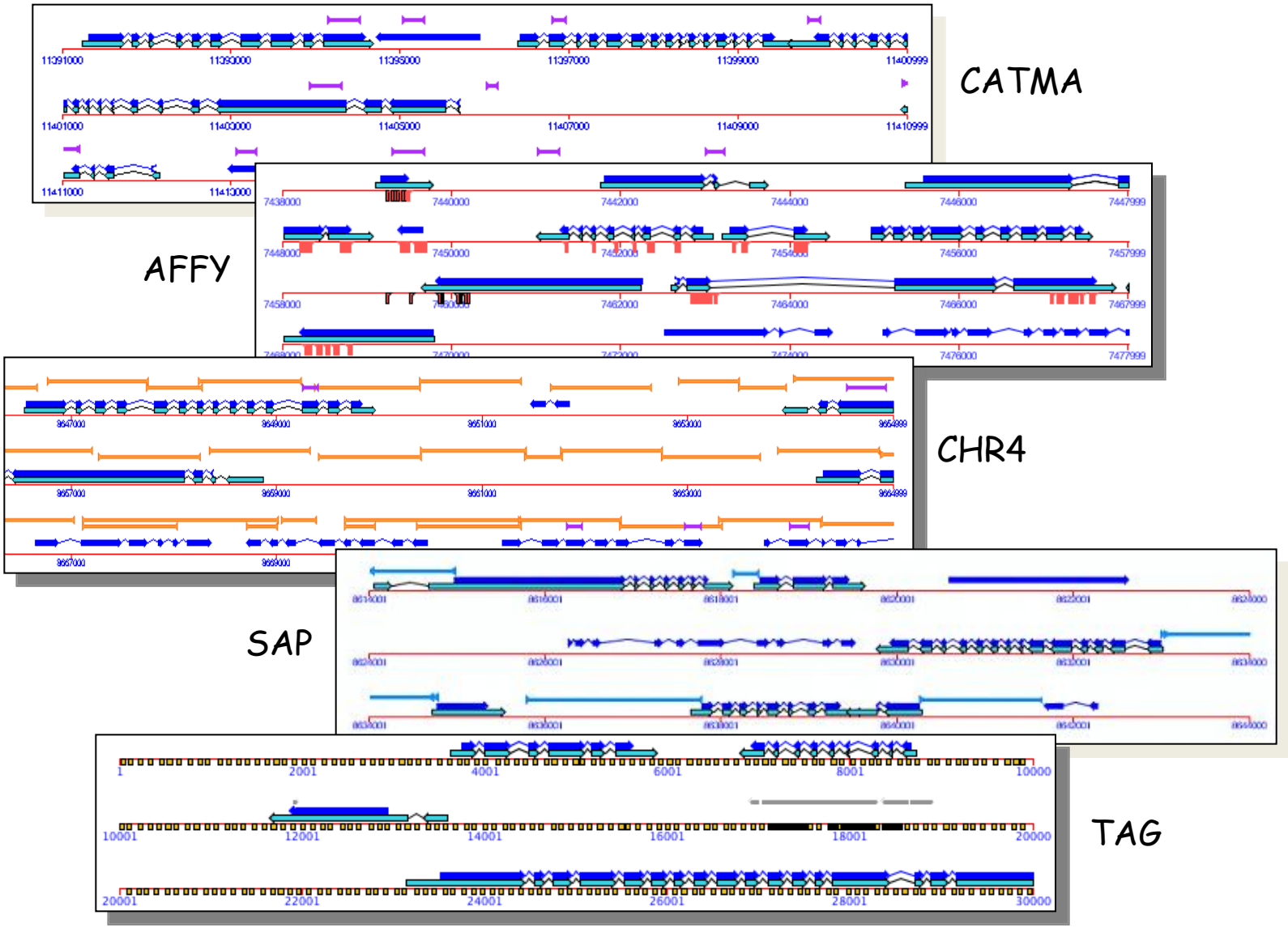
- **Planification expérimentale**
- **Normalisation mono, bi-couleurs et plus** (Bioinformatics 2005, BMC bioinformatics 2008)
- **Analyse différentielle** (Package R anapuce)
- Seuil d'hybridation (BMC Genomics 2007)
- Analyse de données de chIP-chip (Plos Genetics 2007, Plant Journal 2008, Bioinformatics 2008)
- Analyse de données tiling-array (thèse de Caroline Bérard)
- Classification non-supervisée avec sélection de variable (thèse de Cathy Maugis, Biometrics et CSDA 2009)

FLAGdb^{++} : intégration autour des génomes végétaux

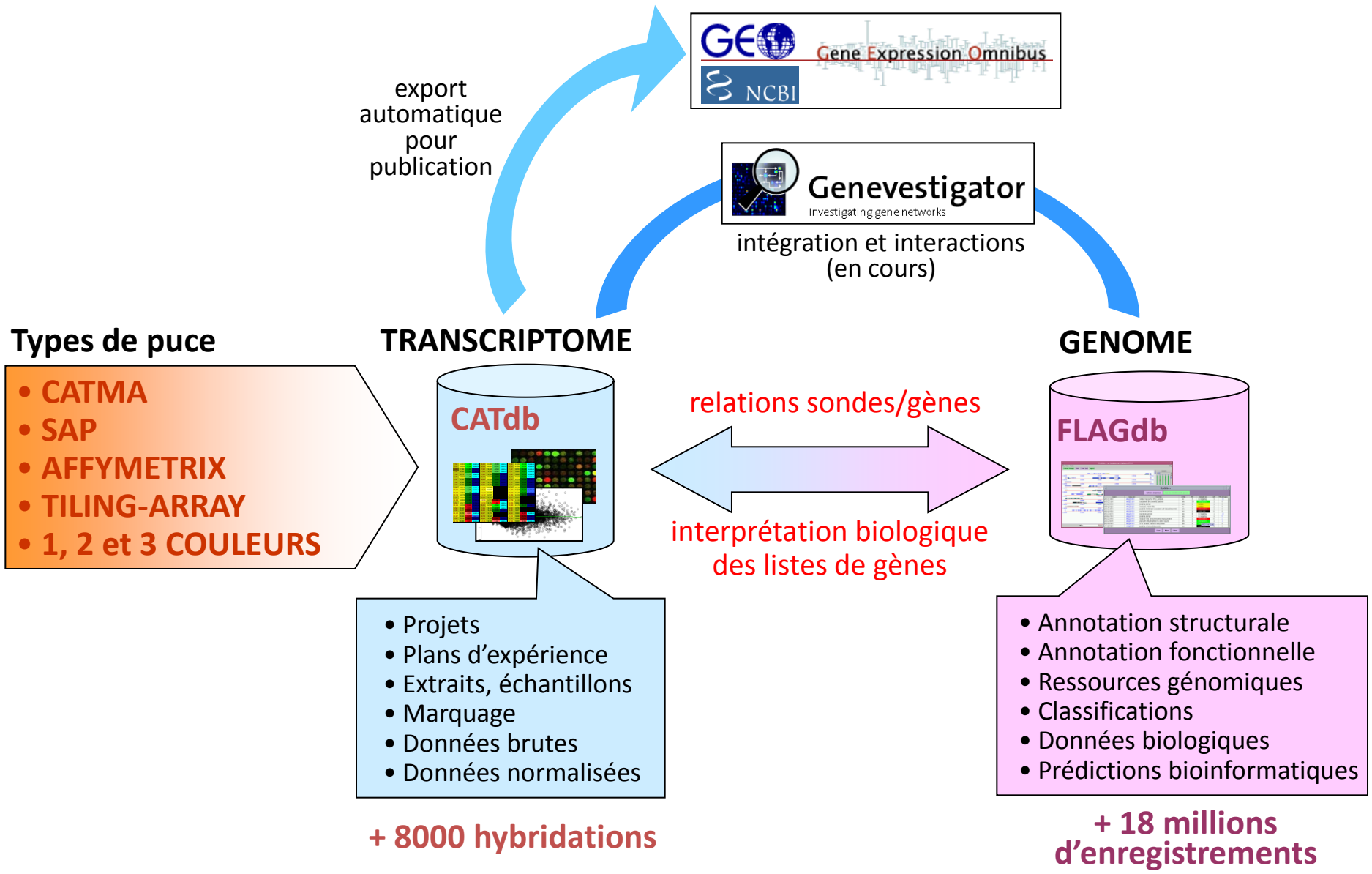


http://urgv.evry.inra.fr/FLAGdb

Outils de génomique



CATdb et FLAGdb^{++} : intégration pour le transcriptome



Les spécificités de la ressource transcriptome CATMA

- Sondes spécifiques et gestion dans FLAGdb^{++}
- Homogénéité en terme d'acquisition des données et de traitement statistique
- Puce 2 couleurs pour une étude de la dynamique de réponses des gènes
- Description des échantillons et plan d'expérience pour chaque projet

- 6000 gènes **exclusifs à CATMA**, absents de la puce Affymetrix:
  - 5000 gènes TAIR non représentés sur ATH1
  - 500 gènes EUGENE
  - 500 prédictions originales de miRNA

Bioinformatique pour la génomique prédictive à l'aide de l'analyse globale du transcriptome

BMC Genomics

Published: 2 November 2007

BMC Genomics 2007, 8:401

BioMed Central

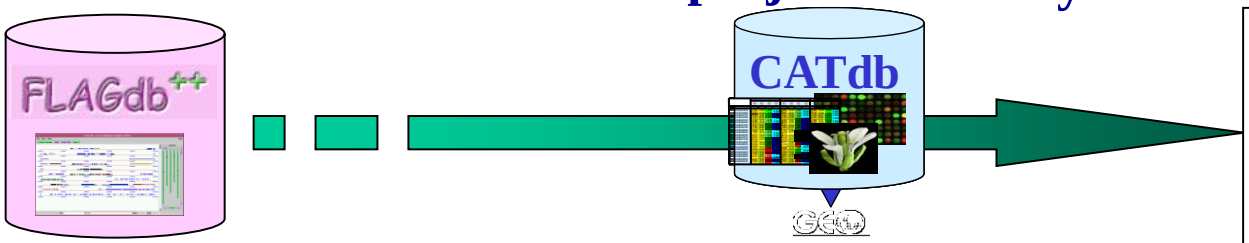
Research article

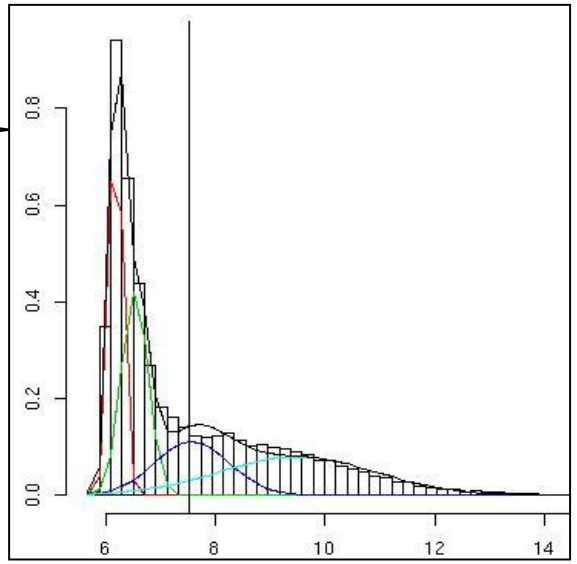
Open Access

Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome

Sébastien Aubourg*[1], Marie-Laure Martin-Magniette[1,2], Véronique Brunaud[1], Ludivine Taconnat[1], Frédérique Bitton[1], Sandrine Balzergue[1], Pauline E Jullien[3], Mathieu Ingouff[3], Vincent Thareau[4], Thomas Schiex[5], Alain Lecharny[1,4] and Jean-Pierre Renou*[1]

157 projets **6068 hybridations**

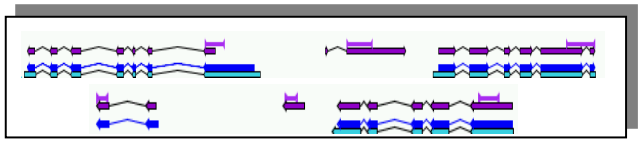




40 projets
522 hybridations

cell	61
protoplast	18
root	78
hypocotyl	28
stem	10
leaf	136
flower	10
pollen	2
silique	4
seed	16
aerial	40
whole plant	119

MixThres : mélange de Gaussiennes tronquées pour construire un seuil d'hybridation discriminant les sondes ayant un signal significatif

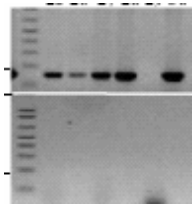


GST in EUGENE CDS

Les nouveaux gènes ont des spécificités

structurelles

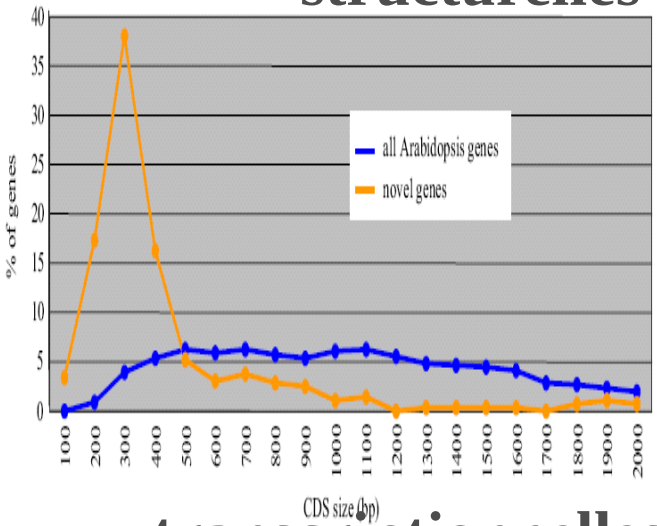
Validation par RT-PCR



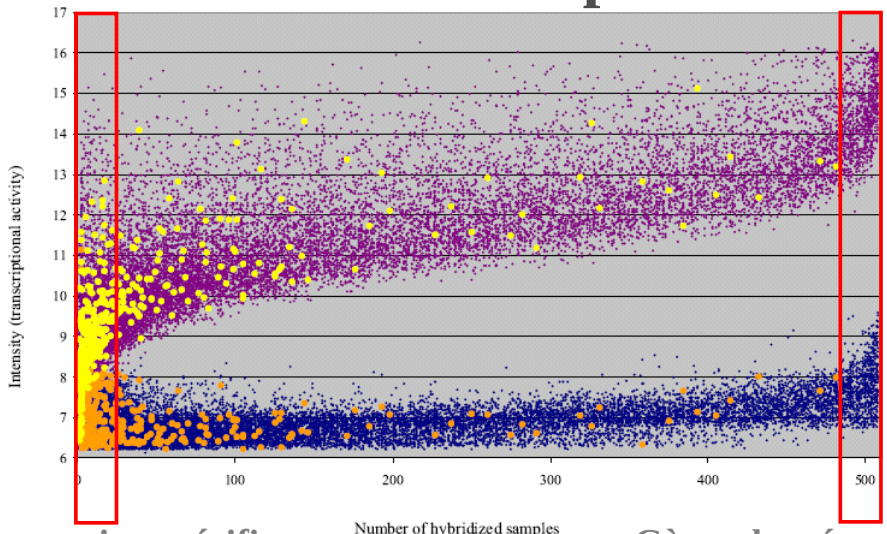
(88%)

CDS moyen
1247 bp
411 bp

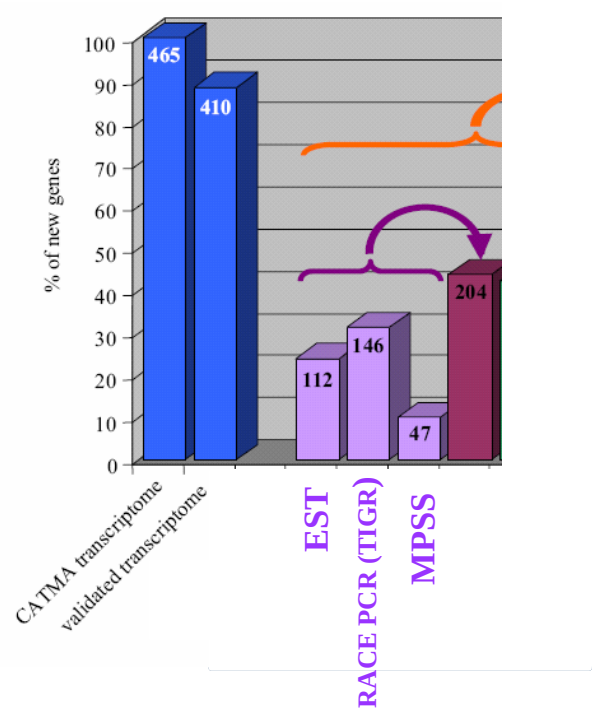
Nbre d'introns
4.2
0.67
191 sans intron



transcriptionnelles



Expression spécifique **Gènes de ménage**



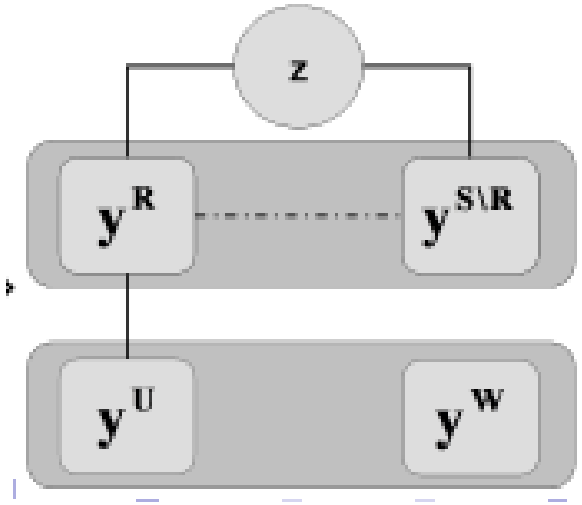
Amélioration de l'annotation avec CATMA
Intégration de ces résultats dans TAIR r9

Analyse globale du transcriptome

- 4616 gènes d'*Arabidopsis thaliana* dont 1430 gènes orphelins
- 33 swaps concernant des stress biotiques.
- Classification non-supervisée par mélange Gaussien avec sélection des swaps informatifs.

Maugis et al. 2009
La revue Modulad

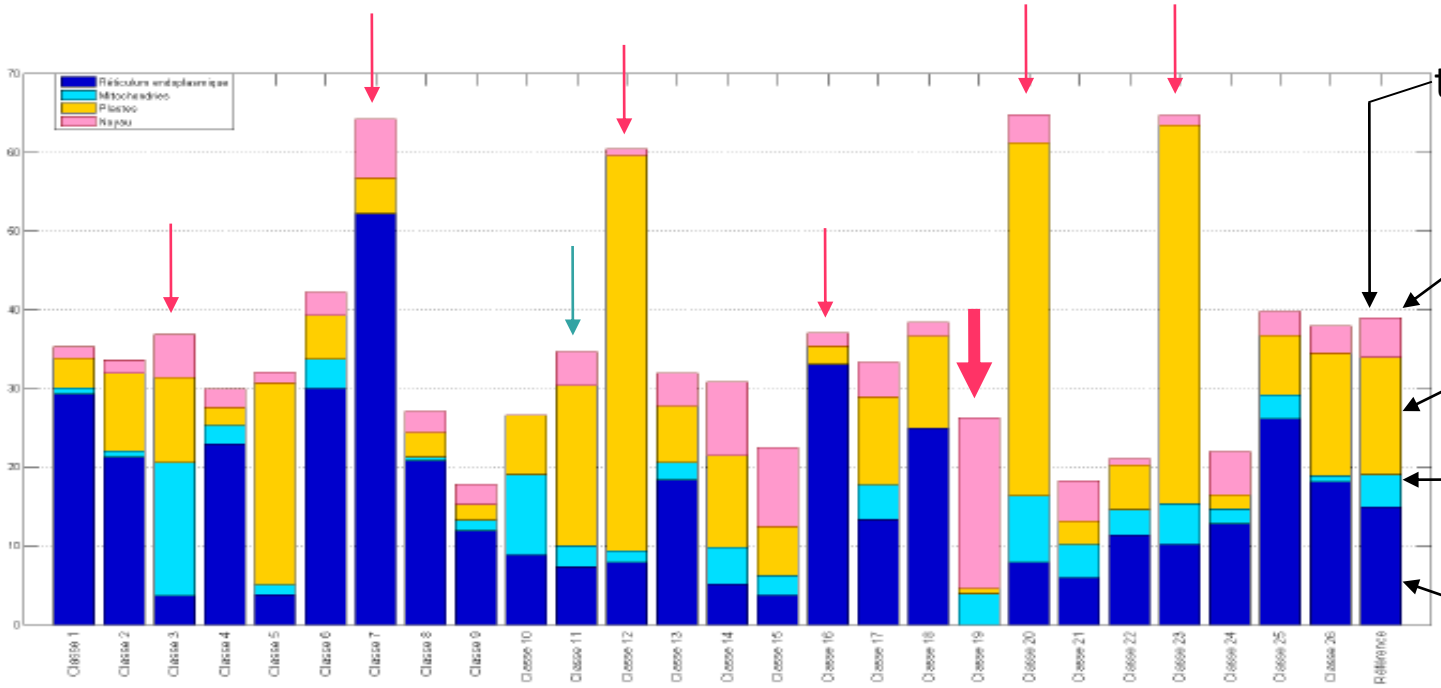
Mélange Gaussien avec sélection de variable
Modèle choisi avec un critère BIC
Estimation pour les mélanges avec 10 à 30 classes



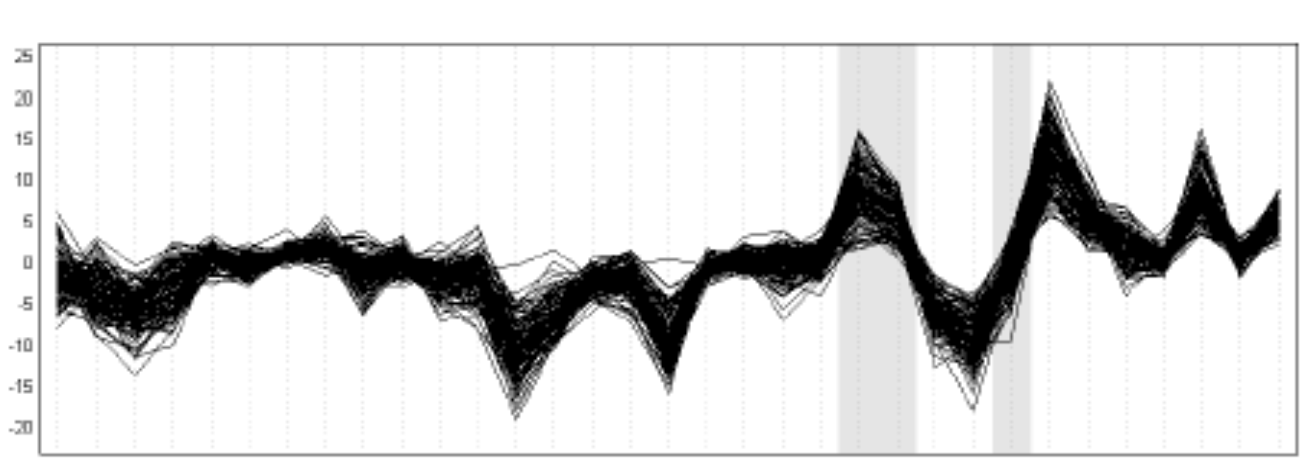
Description des classes

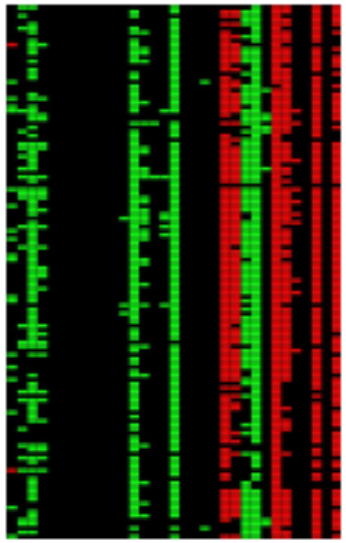
Taille variable des classes et les orphelins sont bien réparties

Numéro de classe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	Total
Nombre de gènes	133	113	277	87	78	303	67	221	151	79	329	255	602	214	80	127	45	60	129	165	333	123	156	109	264	116	4616
Nombre de gènes « sûrs »	112	79	221	60	60	200	50	140	103	57	219	187	313	158	72	105	37	42	108	120	238	88	98	97	178	85	3227
Nombre de gènes « mitigés »	21	34	56	27	18	103	17	81	48	22	110	68	289	56	8	22	8	18	21	45	95	35	58	12	86	31	1389
Nombre de gènes orphelins	34	35	78	24	26	94	14	68	41	25	129	103	192	60	27	35	14	16	8	50	124	26	56	36	76	39	1430


tendance générale
noyau
plastide
mitochondrie
Réticulum endoplasmique

Classe 19





Inférence de fonction par profilage transcriptomique

SONATA : **S**tress, **O**rphan, **N**etwork **A**nd **T**ranscriptome in **A**rabidopsis

Coordinateur : Sébastien Aubourg

Financement INRA–AllEnvi (2010-2013)
Réseau MIA-GAP en 2010 :SONATA-Stat

Contexte :

- Plus de 4000 gènes orphelins de fonction
- Environ 500 gènes nouvellement identifiés et 500 gènes à ARN prédits

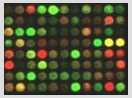
Modèle biologique : La réponse des plantes aux stress

Objectifs :

- liste complète des gènes régulés par les stress biotiques et abiotiques
- leur classification par homologie, profil phylogénétique, nature des stress et profils d'expression
- ébauche de réseaux de gènes impliqués dans les réponses aux stress
- transfert de connaissance vers les plantes d'intérêt agronomique

Collaborations : URGV, MIA, INRIA, IMT, X, Université Paris XI

I. Définition des gènes régulés par les stress

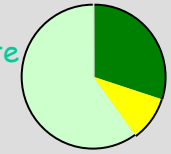


CATdb

~4000 swaps

stress

~450 swaps



Gènes de fonction déduite par homologie

Gènes de fonction inconnue

Gènes publiés impliqués dans la réponse aux stress

II. Clustering

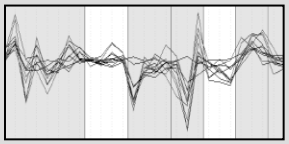
~450 swaps

~19 000 gènes diff exprimés

stress

SelvarClust
(modèle de mélange)

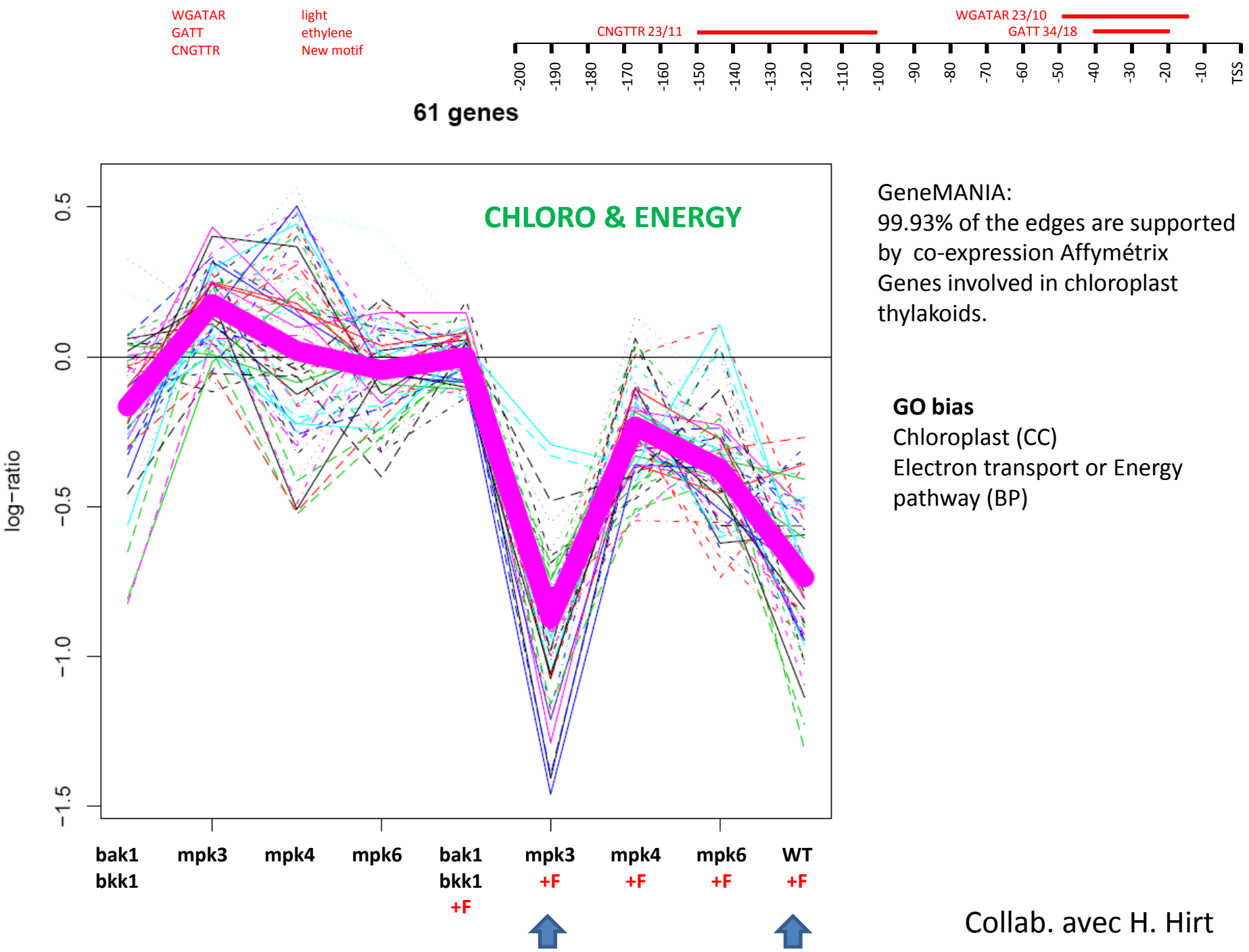
Clusters de gènes co-transcrits et caractère explicatif des expériences



III. Sélection de clusters d'intérêt par intégration

FLAGdb^{++}

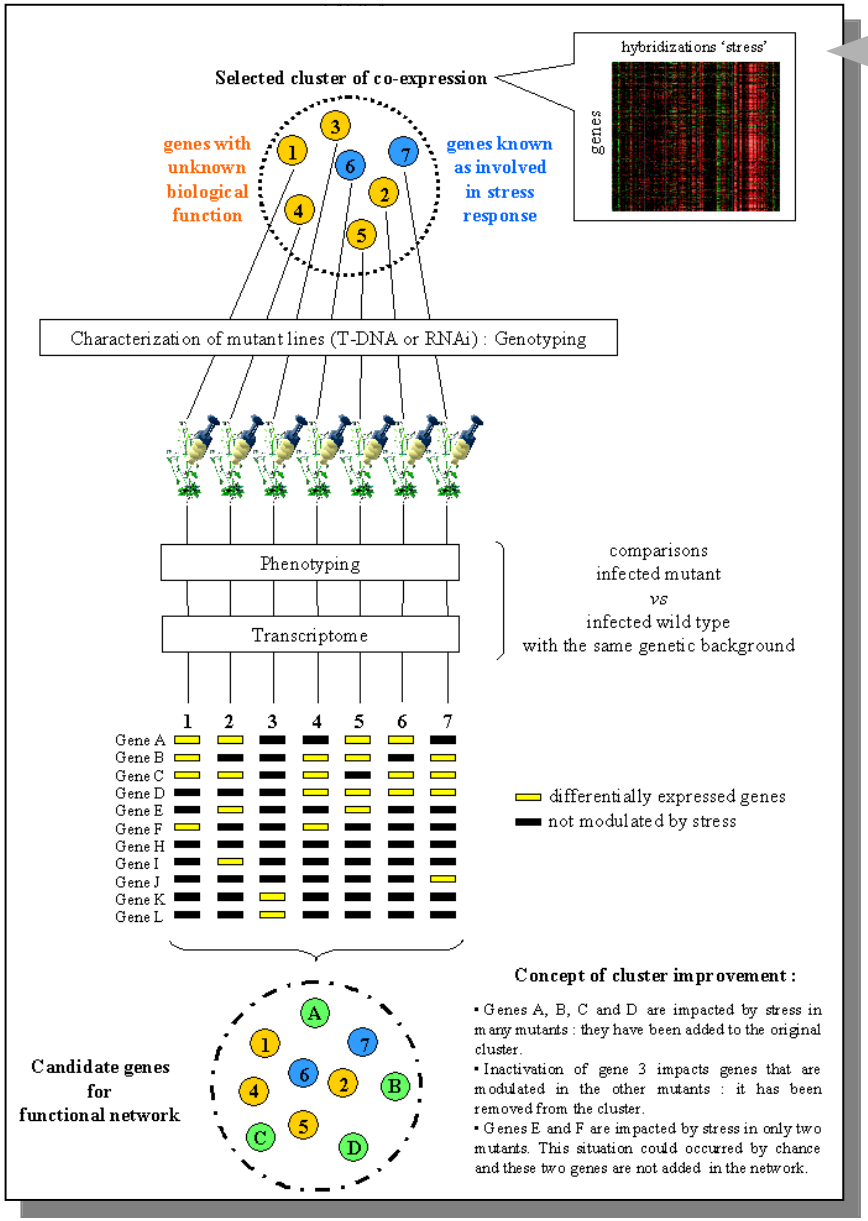
- Nature des gènes connus présents et des stress biotiques concernés
- Contenu des promoteurs en TFBS relatifs à la réponse aux stress (boîtes WRKY, whirly, GCC)
- Homogénéité du cluster en terme de profils phylogénétiques et de classification fonctionnelle (après une recherche d'homologues lointains via les structures secondaires prédites)
- Taille des clusters compatible avec l'approche expérimentale



GeneMANIA:
99.93% of the edges are supported by co-expression Affymétrix
Genes involved in chloroplast thylakoids.

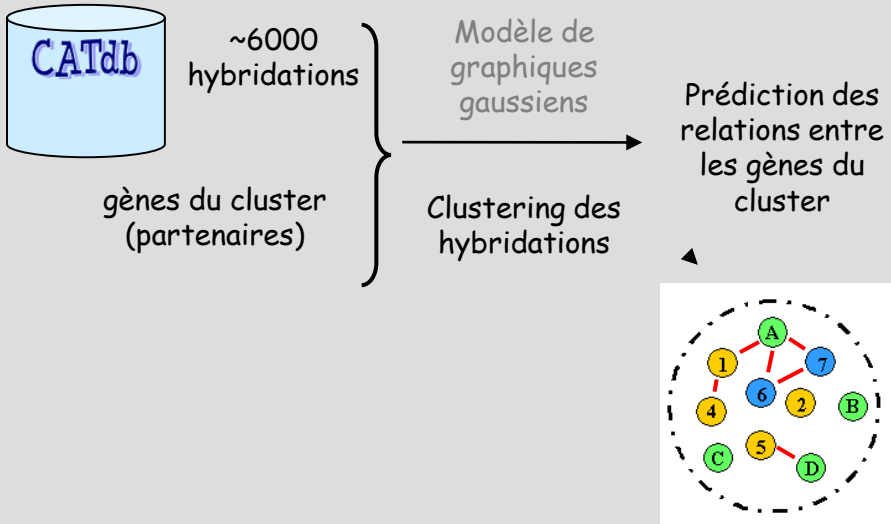
GO bias
Chloroplast (CC)
Electron transport or Energy pathway (BP)

Collab. avec H. Hirt



IV. Analyse fonctionnelle des mutants : Enrichissement des clusters

V. Du cluster au réseau de gènes



Projet similaire sur la graine sans la partie IV : 4D Virtual Seed (soumis, coord V. Fromion)

Inférence de réseaux : contexte favorable à l'URGV

Les ressources disponibles

- FLAGdb++ : annotation structurale et fonctionnelle
- CATdb : ressource transcriptome
- Interactome : matrice 9K testée en double hybride
- Données sur des lignées recombinantes
- A venir : 100 000 mutants TILLING d'Arabidopsis

Collaborateurs identifiés

- Les acteurs de la plate-forme
- L'équipe bioinformatique
- Des utilisateurs avertis de CATMA
- Des biologistes motivés pour la validation biologique des prédictions
- Des collègues statisticiens

… et évolution de la demande

- Sur la plate-forme, volonté d'accompagner encore plus les utilisateurs pour exploiter globalement les données transcriptomes

- Les biologistes souhaitent aller vers des analyses plus globales, rêvent de réseaux !

- Les besoins: développements de méthodologies et outils pour inférer des réseaux chez *Arabidopsis*

Questions de biologie

- Quels réseaux est-il possible de construire avec les données disponibles ?
  - Dans la littérature, beaucoup de co-expression pour identifier tous les acteurs du réseau avant de le construire
  - Peu de réseaux de régulation

- Comment intégrer les différents types de données

- Est-il possible de construire un réseau à partir de petits réseaux ?

Perspectives

- Encore du travail en classification non-supervisée
- **Inférer un réseau de gènes**
  - Gérer la grande dimension (en gènes), les expériences transcriptomes non i.i.d.
  - Intégrer plusieurs types de données (type et quantité différents)
  - Absence d'acteurs du réseau
- **Caractériser un réseau déjà disponible**
  - Interactome
  - Réseau de de co-expression à partir de la corrélation

Renforcer les discussions bio/bioinfo/méthodo

Echanger sur les données

Appel à collaborateurs pour constituer un projet méthodologique pour répondre aux demandes des biologistes

- Discussion pour préciser les questions biologiques
- Identifier les questions méthodologiques associées
- Vérifier l'existence de données adaptées et la motivation des partenaires