HUGO BÉRUBÉ

MISE EN PLACE D'UNE CHAÎNE D'ANALYSE ET DE TRAITEMENT DE BIOPUCES

Mémoire présentée
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de la maîtrise sur mesure en bioinformatique

FACULTÉ DES SCIENCES ET GÉNIES
UNIVERSITÉ LAVAL
QUÉBEC

2006

Résumé

Le but de ce mémoire était d'élaborer des outils informatiques et de les intégrer à une chaîne de traitement et d'analyse des données de biopuces. La chaîne d'analyse mise en place dans ce projet consiste d'abord en SLIMS, un logiciel conçu en PHP et MySQL utilisant des termes compatibles avec les standards MIAME. Il permet le suivi des expériences et des échantillons préalablement à l'extraction des ARN pour les expériences de biopuces. Les données sont prises en charge, à l'aide d'une procédure de transfert, par le logiciel BASE qui gère l'information relative aux biopuces. Finalement, les analyses de données sont réalisées avec différents outils disponibles dans Bioconductor et TM4. Un algorithme a été développé pour annoter tous les gènes de la biopuces. L'analyse d'une d'expérience comparant des épinettes transgéniques surexprimant le gène LIM2 a été faite à l'aide de la chaîne de traitement et d'analyse présentée dans ce mémoire.

Abstract

The goal of this dissertation was to design and implement a microarray analysis pipeline. The first tool of the microarray pipeline is a web-based LIMS: SLIMS. It allows the storage of all data related to experiments and samples from harvest to RNA extraction. This tool was designed in PHP and MySQL allowing easy access and manipulation of the data. A tranfer algorithm was designed to allow stored data to be automatically integrated into the BASE software, a tool that allows storage and analysis of microarray data. An annotation algorithm was also designed in order to annotate genes that are on the microarrays. A lignin/cellwall annotation was also included to enable the rapid indentification of all the genes related to the lignin biosynthesis pathway and cell wall assembly. This pipeline was used to analyze transgenic spruce overexpressing the pine LIM2 gene.

À Claude Déry et Ryszard Brzezinski de l'Université de Sherbrooke qui m'ont aidé à faire mes premiers pas dans le monde de la bioinformatique et sans qui je ne serais pas rendu où je le suis présentement. Merci pour votre confiance en moi

Avant-Propos

Ce mémoire contient un article qui n'a pas été soumis à ce jour. Le manuscrit « Management of biological experiments with SLIMS: example of use for a genomics project » a été réalisé en grande partie par Nathalie Pavy et moi. Nathalie Pavy a contribué à toutes les parties du manuscrit plus particulièrement à la section introduction, résultats et conclusion tandis que j'ai contribué principalement à la section implémentation et résultats. J'ai réalisé la plus grande partie (90%) de la conception et l'analyse de la base de données. Nathalie Pavy, François Larochelle et Nicolas Juge (10%) ont, par la suite, participé à l'amélioration du logiciel. J'ai aussi réalisé toute la programmation, l'implantation et le support nécessaire au bon fonctionnement du logiciel SLIMS.

Adresse email des auteurs:

Hugo Bérubé: hugo.berube@rsvs.ulaval.ca

François Larochelle: flarochelle@bioinfo.ulaval.ca

John Mackay: john.mackay@rsvs.ulaval.ca

Nathalie Pavy: nathalie.pavy@rsvs.ulaval.ca

Remerciement

Je voudrais tout d'abord remercier John Mackay pour m'avoir permis de faire ma maîtrise dans le projet Arborea ainsi que les membres du jury pour avoir évalué mon mémoire. Merci beaucoup à Nathalie Pavy pour son aide et son support tout au long de ma maîtrise et plus particulièrement pour le SLIMS qui fût un travail de longue haleine. Merci à Isabelle Giguère pour avoir fait les manipulations en laboratoire pour l'expérience ptLIM2. Merci à Jean Bousquet, François Larochelle, Jérôme Laroche et Nicolas Juge pour m'avoir pris sous leur aile à mon arrivé dans le projet Arborea et pour m'avoir initié au monde de la bioinformatique. Merci à Stéphane Larose pour m'avoir dépanné à l'occasion avec des problèmes informatiques. Merci aux gens de Bioneq qui m'ont permis de présenter SLIMS à un de leur colloque et pour leurs formations (BioPerl et TM4) qui furent très utiles. Merci aussi à Francis et Frank pour m'avoir permis de découvrir la ville de Québec et ses environs. Finalement, merci à Vicky pour son support moral tout au long de ce périple.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

4CL	4-coumarate : coenzyme A ligase
ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
ANOVA	Analysis of variance
ARN	Acide ribonucléique
ARNm	ARN messager
BLAST	Basic Local Alignment Search Tool
CCOAOMT	Caffeoyl-coenzyme A O-methyltransferase
CAD	Cinnamyl alcohol dehydrogenase
CCR	Cinamoyl: coenzyme A reductase
C4H	Cinnamate 4-hydroxylase
Cy3	CyDye fluors 3
Cy5	CyDye fluors 5
EST	Expressed sequence tag
FDR	Taux de faux positifs
HD-zip	homéodomaine leucine-zipper
KNOX	knotted-like homeobox
LIM	lily massages induced at meiosis
LIMS	Laboratory Information Management System
MAANONA	Microarray ANOVA
MIAME	Minimum Information About Microarray Experiments
MNID	Numéro d'identification MN
MYB	myeloblastosis
Perl	Practical Extraction and Report Language
PAL	phénylalanine amonnia lyase
PCR	polymerase chain reduction
QPCR	Quantitative PCQ
SAM	Significance Analysis of Microarrays
SLIMS	Sample LIMS
SQL	Structured Query Language
TIGR	The Institute for Genomic Research

Liste des définitions

A	$\log_2$ de la racine carré du produit des intensités : $\log_2 \sqrt{(R*G)}$
B	probabilité (en log) que le gène soit différentiellement exprimé
M	$\log_2$ du ratio des intensités : $\log_2 (R/G)$
MA-plots	Graphe de M en fonction de A
t	statistique t: le ratio de la valeur de M par rapport à l'erreur standard
P.Value	valeur de significativité
Q-value	valeur de significativité calculée à partir du taux de faux positif (FDR)

Introduction Générale

L'arrivée de nouvelles technologies en génomique demande des techniques et des outils d'analyses bien adaptés à celles-ci. Une des missions importantes de la bioinformatique et la statistique est l'analyse des quantités importantes de données provenant de la recherche en génomique. La grande taille des jeux de données, souvent couplée à une faible réplication impose des contraintes d'analyse, autant pour le statisticien que pour le biologiste. Il devient donc impératif de mettre en place un environnement permettant aux chercheurs d'accéder aux suites de programmes spécialisés et d'outils analyses appropriées pour gérer, entreposer et analyser cette quantité immense de données. Dans le cadre du projet Arborea, une plate-forme d'analyse de biopuces a été mise en place afin d'étudier et de caractériser les fonctions des gènes ayant un rôle dans la formation du bois. Les arbres consacrent une grande proportion de leur énergie et de leurs ressources à produire le bois et plusieurs gènes sont impliqués dans ce processus.

Le but de ce mémoire était d'élaborer des outils informatiques et de les intégrer à une chaîne de traitement et d'analyse des données de biopuces. À partir des profils d'expression, la chaîne d'analyse doit permettre l'accès à ces données ainsi qu'aux divers outils de cette chaîne d'analyse. Pour de nombreux organismes, dont les arbres forestiers, il est nécessaire de développer des biopuces sur mesure et des outils informatiques spécifiques. L'objectif était donc de bâtir une chaîne d'analyse adaptée pour permettre l'intégration de toutes les étapes nécessaires à l'analyse des biopuces. Le travail présenté dans ce mémoire visait aussi à répondre au besoin de suivi des données expérimentales, de la récolte des échantillons sur le terrain jusqu'à leur utilisation dans les hybridations.

Le mémoire illustre, par ailleurs, l'utilisation de cette chaîne d'analyse dans l'étude d'épinettes transgéniques surexprimant le facteur de transcription LIM2 dans le but de déterminer s'il joue un rôle dans la synthèse de la lignine lors de la formation du bois.

1.1 La Génomique

La génomique est une sous-discipline de la génétique qui s'intéresse à la structure, le fonctionnement et l'évolution des génomes entiers. Il s'agit d'un domaine d'étude en constante évolution qui vise à recueillir, à comprendre et à exploiter à grande échelle l'information biologique encodée par l'ADN. L'exemple le mieux connu est sans doute le projet du génome humain qui a mené au séquençage complet du génome de l'homme. Toutefois, la génomique s'étend également aux animaux, aux champignons, aux plantes, dont les arbres forestiers ainsi qu'à plusieurs microorganismes. Elle se divise en deux grand domaines : la génomique structurale, qui caractérise la nature physique du génome, et la génomique fonctionnelle qui cherche à caractériser et comprendre le fonctionnement des gènes. La génomique nous aide à comprendre le fonctionnement d'un organisme par la connaissance des gènes, de leur régulation et de leurs interactions avec l'environnement.

Les informations essentielles à la vie se retrouvent encryptées dans le génome d'un organisme, plus précisément dans son ADN. Celui-ci contient en quelque sorte les plans de constructions d'une espèce, l'information nécessaire pour bâtir et assurer le bon fonctionnement d'un organisme vivant. En connaissant la totalité des séquences d'un génome, il est possible de déterminer la différence intrinsèque entre deux espèces ou souches ce qui peut aider à répondre à plusieurs questions d'ordre biologique ou phylogénique. La connaissance de ces séquences peut aussi nous servir de point de départ d'une analyse fonctionnelle visant à déterminer les rôles des différents gènes dans l'organisme ainsi que l'interaction existant entre ceux-ci.

La génomique utilise plusieurs méthodes pour étudier la fonction des gènes. Par exemple, l'analyse des cadres de lecture ouverts (ORF) sur une séquence permet de déterminer les sections codantes d'un gène et auxquels l'ont peut attribuer une fonction. L'attribution d'une fonction débute avec la recherche d'homologie entre les ORFs et les séquences caractérisées chez d'autres organismes, entreposées dans des banques de données publiques (eg. NCBI, swissprot). La fonction d'un ORF peut aussi être déterminée à l'aide de la technique d'inactivation d'un gène par mutagenèse (Knock-out) dans laquelle les phénotypes engendrés par la mutation renseignent sur la fonction du gène. Il est ainsi

possible d'étudier la fonction d'un gène en caractérisant l'effet qu'une modulation de celui-ci aura sur un organisme ou sur tout un ensemble de gènes avec lesquels il interagit, une fois traduit en protéine.

1.2 Analyse de biopuces

Une technologie qui joue un rôle important dans le développement de la génomique est celle des biopuces. Il existe deux catégories de biopuces: les biopuces d'ADNc constituées à partir de longs fragments d'ADN complémentaires représentants des gènes complets ou partiellement complets, et les biopuces constituées d'oligonucléotides courts (e.g. produits Affymetrix) ou longs. Les biopuces d'ADNc sont souvent utilisées avec des hybridations simultanées de deux échantillons. Dans ce type d'expérience, une sonde est produite par transcription inverse des deux échantillons d'ARN extraits de sources différentes. Chaque sonde est marquée à l'aide d'un marqueur fluorescent spécifique. Les deux échantillons marqués sont hybridés avec la biopuce, permettant de déterminer le niveau d'hybridation avec chacune des sondes (gènes) sur la biopuce d'après l'intensité de la fluorescence.

Contrairement à l'approche à deux échantillons, les biopuces constituées d'oligonucléotides du type Affymetrix n'utilisent qu'un seul échantillon par hybridation. Des fragments d'ADN très courts sont synthétisés ou déposés sur la biopuce. La très forte densité de ces biopuces permet d'utiliser jusqu'à 40 sondes par transcrit. La moitié de ces sondes sont des répliques parfaites du gène tandis que l'autre moitié contient des imperfections qui serviront de contrôles pour tester la spécificité des signaux d'hybridations.

Les biopuces qui font l'objet de ce chapitre sont des biopuces à ADNc utlisés pour déterminer l'expression de deux différents sujets lors d'hybridations sur une même biopuces. Les biopuces sont utilisées pour des expériences visant à analyser la différence d'expression des gènes entre deux sujets, tests et contrôles (Figure 1.1). Elles consiste en une lame de verre sur laquelle sont fixés de manière structurée (en rangées ordonnées), des milliers de sondes qui sont fragments d'ADNc. En principe tous les gènes d'un

génome peuvent être représentés sur une seule ou quelques puces. L'intensité de la fluorescence ainsi obtenue pour chaque sonde placée sur la puce dépend de la quantité d'ARNm présente pour chacun des différents sujets étudiés. À l'aide d'un logiciel d'analyse d'image, le niveau de fluorescence de l'ARNm marqué se liant à chacun des sondes sur la biopuces est déterminé. Les échantillons marqués s'hybrideront différemment sur les sondes selon l'abondance du transcrit correspondant chez le sujet. Il est possible par exemple, de trouver des gènes qui sont plus fortement exprimés chez un individu malade que chez un individu en santé donnant une piste pour trouver les gènes ayant un rôle dans le système de défense.

Cette surexpression est détectée avec la biopuce par une plus importante hybridation et un signal plus fort pour les sondes relatives à ces gènes. Afin d'assurer que les différences d'intensité entre les sondes sont dûes aux ARNm hybridés et non à d'autres sources de variation liées à la méthode, les données brutes sont évaluées avec des outils d'analyse de la qualité de l'image. De plus, des algorithmes de normalisation et de standardisation devront être utilisés avant l'analyse et l'exploitation des données. Différents types d'analyses pourront ensuite être faites sur les données afin d'identifier des patrons d'expressions indiquant des gènes différentiellement exprimés.

1.2.1 Analyse d'image

Une fois les hybridations réalisées, la première étape de traitement est la numérisation de l'image et l'extraction des données à l'aide d'un logiciel. L'analyse d'image est une étape cruciale car les données à analyser sont extraites à partir de l'image analysée. Les données brutes sont fixées par cette analyse d'où l'importance d'assurer sa qualité. Deux facteurs principaux jouent dans l'obtention d'une image de qualité. Tout d'abord il y a la qualité des différentes manipulations dont l'extraction des ARN, le marquage et la qualité de l'hybridation et sans oublier la fabrication de la lame. S'il y a un problème avec un ou plusieurs de ces facteurs alors la qualité de l'image sera très affectée. Deuxièmement, la méthode de numérisation de l'hybridation est aussi un facteur très important. Une mauvaise numérisation aura pour effet de générer trop de points avec une intensité maximale saturée ou un bruit de fond trop important sur la puce. L'inspection visuelle de

l'image reste une façon simple d'évaluer la qualité d'une numérisation. Il est assez facile détecter un problème au niveau de l'intensité ou la saturation des différents points sur l'image.

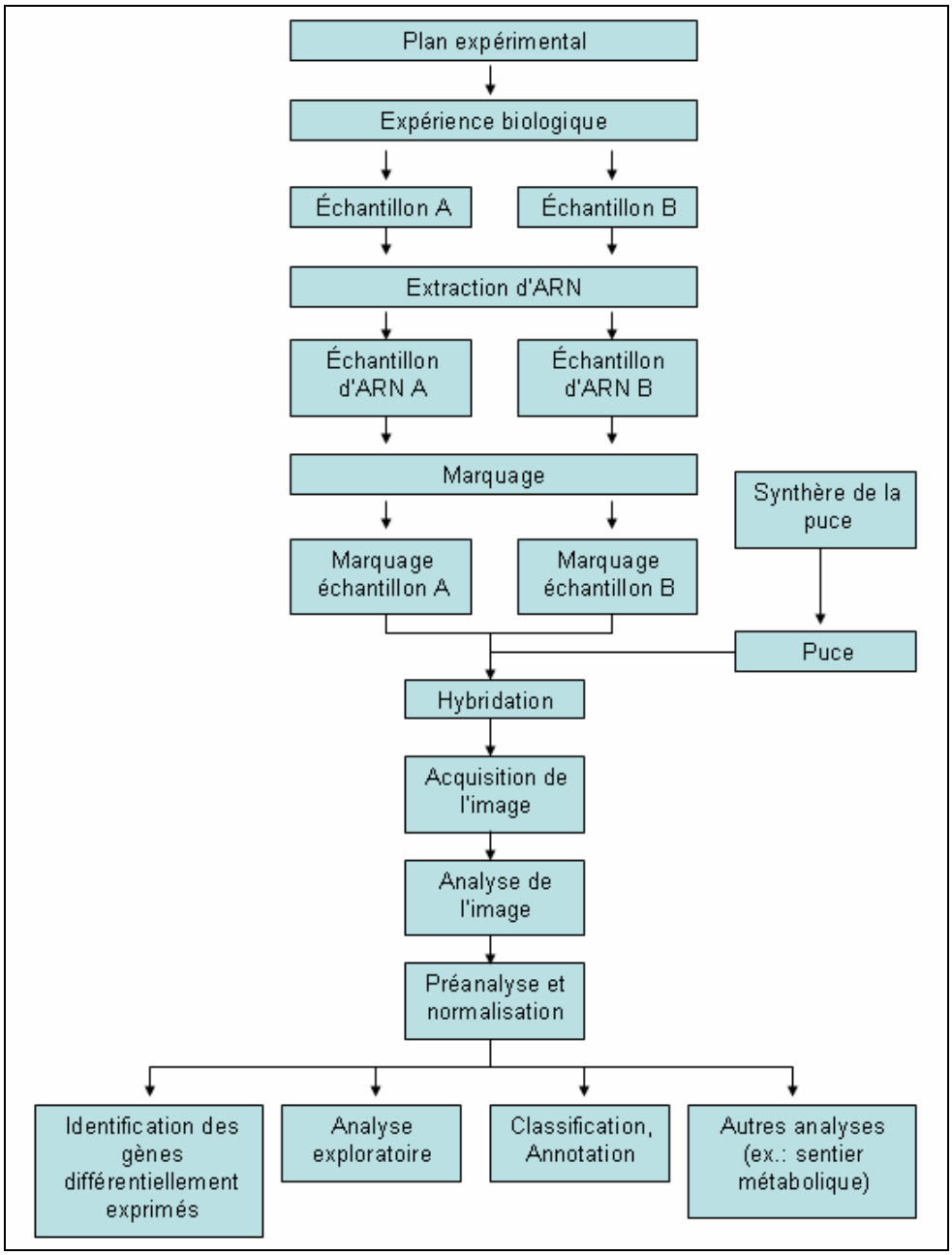


Figure 1.1 : Étapes fondamentales d'une analyse de biopuces d'ADNc *(Leung and Cavalieri, 2003)*

Il existe plusieurs logiciels permettant d'évaluer la qualité des biopuces d'après le diamètre, la circonférence des points, l'uniformité des réplications. Une fois les données lues et évaluées par le logiciel d'analyse d'image elles sont enregistrées dans un fichier texte où l'on y retrouve, entre autres, l'intensité de fond pour chaque 'point', l'intensité du point lui-même, l'intensité autour de celui-ci pour les longueurs d'onde se rapportant au type de marquage utilisé (e.g. cy3 et cy5). Ces données permettent de comparer différents échantillons en comparant les intensités entre les canaux.

L'analyse de l'image comporte quatre étapes: i) la numérisation de l'image, ii) l'identification des points, iii) la segmentation et iv) l'extraction des intensités et le calcul des ratios. L'identification des points est l'étape où l'utilisateur doit faire l'ajustement de la grille pour y contenir les points. Certains d'entre eux devront être fait manuellement et permettra d'identifier les points de faible qualité.

La segmentation consiste à différencier le bruit de fond de l'intensité des points. Cette partie est grandement affectée par la qualité de l'image car une image de faible qualité aura un bruit de fond très élevé. Une inspection visuelle permet aussi de voir les tâches et ou les imperfections qui pourraient nuire à notre analyse. Il existe plusieurs types d'algorithme de segmentation dont la segmentation par cercle fixe, la segmentation par cercle adaptée, la segmentation par forme adaptée et la segmentation par histogramme. Plus la qualité de l'image est bonne plus l'algorithme prédit les points et leur signal de façon précise (Leung and Cavalieri, 2003). Par contre, s'il y a contamination avec de la poussière ou un bruit de fond élevé alors l'algorithme rejettera les mauvaises données mais identifiera aussi des contamination comme des points.

Une fois l'extraction des intensités complétées les ratios sont calculés à partir des intensités des deux canaux.

1.2.2 Prétraitement et normalisation des données

L'étape de prétraitement consiste à éliminer les points de piètre qualité; par exemple ceux dont l'intensité totale est moindre que deux fois l'intensité du bruit de fond. Avant le prétraitement, il est recommandé de s'assurer que les données ne contiennent pas de biais ou d'erreurs systématiques dues à des erreurs de manipulations ou des imperfections sur la lame. S'ils ne sont pas corrigés, ils pourraient produire des données erronées et une mauvaise identification des gènes différentiellement exprimés. La plupart des expériences de biopuces recherche les relations entre les échantillons biologiques à partir des patrons d'expression en analysant les gènes différentiellement exprimés entre ceux-ci. Si une puce possède N éléments distincts et qu'on compare un contrôle avec un sujet test (que nous désignerons par R et V) alors le ratio T pour le gène i sera :

$$T_i = \frac{R_i}{V_i}$$

À partir du ratio il est possible de déterminer si un gène est plus exprimé dans le contrôle ou le sujet test. Les gènes surexprimés dans R par un facteur 2 auront un ratio d'expression de 2 alors que s'ils sont surexprimés par un facteur 2 dans V, le ratio d'expression T_i sera 0,5. Afin de rendre la calcul des ratios plus homogènes d'un côté par rapport à l'autre les ratios sont souvent écris en log de base 2. De cette manière les gènes sont traités de manière symétrique. La transformation logarithmique convertie les effets multiplicatifs en effets additifs selon une distribution normale plus facilement modélisable (Finkelstein *et al.*, 2002). Par exemple, un gène ayant un facteur 2 d'expression dans R aura une valeur de 1 tandis que ceux ayant un facteur d'expression de 2 dans V auront une valeur -1.

Le but de la normalisation est d'ajuster les données d'expression afin de réduire les biais dûs à d'autres sources que l'expression du gène entre les individus (Park *et al.*, 2003). Les biais peuvent être dûs des quantités différentes d'ARN pour différents échantillons, ou des différences dues à l'intensité de base dégagée par les différents fluorophore de marquage. Une normalisation doit permettre la comparaison des niveaux d'expressions

entre plusieurs hybridations. Il existe plusieurs méthodes de normalisation chacune basée sur des assomptions différentes.

Normalisation totale des intensités

La plus simple des méthodes de normalisation est probablement la normalisation totale des intensités. Elle prend pour acquis qu'il y a des quantités égales d'ARN pour les deux échantillons à comparer et que les gènes sur la puce constituent un ensemble représentatif des gènes dans l'organisme donc que la majorité des ratios se regroupent autour de 1. À partir de ce raisonnement un facteur de normalisation est calculé et servira à réajuster les données :

$$N = \frac{\sum_{i=1}^{Npuce} R_i}{\sum_{i=1}^{Npuce} V_i}$$

Ensuite, on ajuste une ou les deux intensités par le facteur calculé afin que le ratio moyen N soit égal à 1. Donc que

$$T_i = \frac{R_i}{V_i} = \frac{1}{N_{total}} \frac{R_i}{V_i}$$

Normalisation globale ou locale

Les normalisations peuvent être appliquées localement à un sous-ensemble de données ou globalement à toutes nos données. Les normalisations ont l'avantage de corriger les biais spatiaux comme les inconsistances d'impression avec les têtes d'impression et autres défauts dans la fabrication de la lame car l'algorithme traite chaque partie de la puce séparément des autres sections. La normalisation locale ajuste vers un ratio de 1 une section ayant une intensité anormalement élevée tandis que la normalisation globale visera à ce que la moyenne de tous les points de la biopuces soit de 1 et non d'une section

en particulier. La normalisation globale peut ainsi entraîner l'identification d'un nombre anormalement élevé de gènes différentiellement exprimés pour une région.

Normalisation Lowess

Plusieurs études ont démontrés qu'il existait un lien entre les ratios et les intensités. À plus faibles intensités les ratios ont tendance à fluctuer davantage et à souvent être plus intenses dans un canal que dans l'autre dépendamment du chromophore utilisé (Yang, Y.H. et al., 2002; Yang, I.V. et al., 2002). La méthode de normalisation 'Locally weighted linear regression (lowess)' a été proposée pour contrebalancer ce biais des données (Yang, Y.H. *et al.,* 2002; Yang, I.V. *et al.,* 2002). La normalisation lowess détecte les déviations systématiques visualisée par graphique le R - I (intensité) et les corrige en appliquant une régression linéaire locale en fonction du $\log_{10}$ des intensités et en soustrayant le best-fit de la moyenne $\log_2$ (ratio) des données expérimentales observées pour chaque point. Plus les points sont loin les uns des autres moins ils ont de poids dans le réajustement d'où le terme 'weighted' (Yang *et al.,* 2002).

OLIN

La méthode de normalisation OLIN (Futschik et Crompton, 2005) est une méthode optimisée de normalisation et de visualisation pour les biopuces à ADN. Bien que la méthode par régression locale soit une des plus populaire et efficace, elle peut engendrer des erreurs lorsque les paramètres par défaut sont utilisés sans tenir compte de leur impact dans l'analyse. À partir de cette problématique, OLIN offre deux méthodes de normalisation basées sur une version itérative de la régression locale. Ces deux normalisations visent à corriger les biais reliés aux intensités et aux biais locaux. Il a été prouvé à l'aide de comparaisons que ces normalisations réduisaient considérablement les erreurs systématiques dans les données de biopuces, amélioraient la qualité des données d'une expérience et la corrélation entre les données (Futschik et Crompton, 2004). OLIN (et OSLIN) sont implantés avec le langage R et ne peuvent être utilisé que dans l'environnement Bioconductor. Il existe d'autres fonctionnalités à l'intérieur du module OLIN qui permettent la visualisation de données de biopuces.

L'algorithme OLIN est basé sur une itération de la régression locale des intensités des points en fonction de leur intensité et de leur position suivi d'une correction dû au biais des ARN marqués. Pour les régressions locales, l'algorithme LOCFIT offre plus de flexibilité que la méthode lowess. Les paramètres du modèle sont optimisés à chaque étape de l'itération par validation croisée généralisée ce qui diminue grandement la complexité de l'algorithme par rapport à la validation croisée conventionnelle.

La deuxième normalisation utilisée est 'Optimized Scaled Local Intensity-dependant Normalisation (OSLIN)'. C'est une version optimisée de OLIN avec un ajustement des niveaux des intensités en tenant compte des dimensions spatiales de la puce. Il est très important de noter que OLIN et OSLIN assument que la majorité des gènes sur la puce ne sont pas différentiellement exprimés et que ceux qui le sont sont distribués également à l'échelle de la puce.

Normalisation composite

Les méthodes conventionnelles de normalisation prennent pour acquis deux assomptions majeures : 1) pour une même quantité de d'ARN provenant de différents organismes il y aura une quantité égale de cibles marquées, 2) aucune autre variable (position spatiale, intensité globale et la plaque) n'ont une influence sur le marquage sur l'ensemble de la lame. La normalisation composite (Yang et Dudoit, 2002) tient en compte les sources importantes de variation systémique comme le biais relié à l'intensité ou à la position spatiale. Aucune autre méthode, avant celle-ci, tenait compte de ces variations durant la normalisation.

Normalisation basée sur les réseaux neuronaux

Cette normalisation attaque aussi la problématique des biais reliés à l'intensité et la position spatiale. Cette méthode prend pour acquis que la majorité des gènes en présence sur la puce ne sont différentiellement exprimés et qu'il y aura en nombre semblable de gènes surexprimés que de gènes sous exprimés (Tarca *et al.*, 2005). Cette méthode produit des résultats indépendants qui peuvent varier légèrement à chaque fois qu'on l'exécute car l'algorithme basé sur les réseaux neuronaux ajuste ses paramètres de façon aléatoire et divisera de manière aléatoire les données entre groupe d'entraînement et

groupes ‘tests’. L’algorithme neuronal s’entraîne d’abord sur les données dites d’entraînement afin d’établir des critères d’évaluation et ensuite les appliquer sur les données tests. En faisant varier le jeu de données d’entraînement il faut s’attendre à voir ses critères d’évaluations changer. Ainsi, il est important que le jeu de données d’entraînement soit bien représentatif des données qui seront évaluées par la suite. La performance de cet algorithme est comparable à l’algorithme composite et se compare avantageusement aux autres méthodes.

Prétraitement : Filtrage et calcul des moyennes

La présentation graphique des données du ratio des instensités en fonction des intensités entre les échantillons (graphique RI) permet de constater que la variabilité augmente quand l’intensité diminue. Plus l’intensité est basse et plus l’erreur relative est élevée. L’application d’un filtre d’intensité est donc une méthode permettant d’assurer une certaine qualité du jeu de données. La méthode consiste à sélectionner les points qui sont statistiquement différents du bruit de fond, par exemple ceux qui ont une intensité égale à au moins deux déviations standards du bruit de fond.

Le filtrage à partir des réplications est basé sur la corrélation entre les ratios d’intensités existant entre les réplications d’une même sonde. Les sondes sont fréquemment déposées deux fois sur une même biopuces (réplicats), soit cote à cote ou à des adresses indépendantes. Il est possible d’éliminer les points pour lesquels les réplications ne donnent pas des ratios assez semblables. Théoriquement, le ratio d’un point multiplié par l’inverse de sa réplication devrait donné 1 ou un chiffre s’en approchant

Une autre méthode simplifie les données en faisant la moyenne entre les réplications. Elle peut être fait pour alléger la quantité de données à analyser. Cette approche entraîne une perte potentielle d’information car la corrélation entre les réplications ne sera plus disponible. La corrélation entre les réplications est une bonne indication de la qualité de la puce car la valeur obtenue avec cette corrélation existant entre un point et le point suivant contenant exactement la même sonde, donc théoriquement la valeur devrait être égale à 1.

1.2.3 Identification des gènes différentiellement exprimés

Le but de l'analyse de biopuces est d'identifier des gènes différentiellement exprimés entre les différents sujets, à partir des données normalisées et standardisées entre elles. Il existe différentes méthodes d'analyse pour identifier les gènes différentiellement exprimés par rapport à ceux qui ne le sont pas.

Méthodes de sélection basées sur la délimitation des ratios minimaux

Ces méthodes établissent une valeur d'expression qui délimite les gènes différentiellement exprimés des autres. Cette valeur représente le ratio requis pour qu'un gène soit considéré comme différentiel. En général, cette valeur est fixée arbitrairement à deux fois donc un gène qui est exprimé deux fois plus dans un sujet que dans l'autres est considéré comme différentiellement exprimé (Draghici, 2002; Schena *et al.*, 1996).

Une autre approche consiste à calculer la moyenne et la déviation standard de la distribution des ratios de log2 et de définir un seuil acceptable selon une valeur de confiance pour identifier les gènes différentiellement exprimés. Par contre cette démarche tout comme la précédente, applique un ratio fixe pour tous les points indépendamment de leur intensité alors qu'il est bien connu qu'il y a une plus grande variation à faible intensité. Ces méthodes identifient donc beaucoup de gènes différentiellement représentés par des sondes de faible intensité par rapport aux sondes donnant une forte intensité.

Analyse avec fenêtre coulissante

La méthode d'analyse avec fenêtre coulissante permet de calculer les ratios à l'intérieur d'une région entourant un certain nombre de points (sondes). Il est ainsi possible de calculer les ratios acceptables à partir de la déviation standard pour chaque intervalle à partir des points situés à l'intérieur de celui-ci. Donc on définit un ratio à partir duquel un gène est considéré comme différentiellement exprimé pour chacun des intervalles. On recherche habituellement les gènes différentiellement exprimés avec un niveau de confiance de 95%.

Test de t et tests de t modifiés

Un des tests les plus simples pour identifier les gènes différentiellement exprimés est basé sur la statistique student. Dans les expériences avec des réplications le test t peut être utilisé pour chaque gène en calculant la variance pour chacun des gènes à partir des ratios des intensités. Le test T gène par gène n'est pas influencé par l'hétérogénéité des valeurs (si le cas se présente) car il ne se concentre que sur un gène à la fois. Ce genre de test perd de la puissance s'il n'y a pas beaucoup de réplication par gène car la variance calculée peut être élevée. Il existe par contre un moyen d'appliquer un test t global, à l'ensemble des gènes, à condition, bien sur, que la variance soit homogène entre ceux-ci. Cette méthode sera l'équivalent d'un test t car il classera les gènes en ordre selon leur échelle de changement (fold change) et ne tiens pas compte de la variabilité génique donc souffre des mêmes problèmes que le test des ratios minimaux ('Fold Change Cut off').

Les modifications apportées au test t (test t modifié) tiennent compte du fait qu'il est difficile d'estimer la variance lorsque la taille des échantillons est petite. Il est possible d'obtenir un estimé plus stable en combinant les données de tous les gènes mais ceci implique que la variance est homogène entre les échantillons. Une autre version du test t, la 'significance analysis of microarrays' (SAM), ajoute ainsi une petite constante au dénominateur du test t gène par gène. Cette modification à pour effet de ne pas identifier comme significatif les gènes ayant un léger changement d'échelle éliminant ainsi la problématique discutée plus tôt (Tusher *et al.*, 2001).

Le 'test t régularisé' tient compte du test t gène par gène et du test t global en utilisant une moyenne pondérée entre les deux comme dénominateur du test t. La 'statistique B' proposé par Lonnsted et Speed (2002) calcule les probabilités des ratios différentiels de ceux non différentiels. Ce test permet une certaine variance spécifique à un gène tout en combinant l'information à travers tous les gènes donc devrait être plus stable que le test t.

ANOVA

L'analyse de variance (parfois appelé test F) est semblable au test T. La différence majeure est que, contrairement au test T, l'ANOVA cherche à tester l'hypothèse selon laquelle il existe une différence ou non entre deux ou plusieurs moyennes. C'est un

procédé arithmétique qui vise à répartir la variance entre les différentes sources de variations connues et de déterminer si la variation entre les traitements est significative en tenant compte de la variation entre les réplicats de chaque traitement. Un avantage important de l'ANOVA par rapport au test t est qu'il peut évaluer la différence pour plus de deux moyennes à la fois.

Il existe deux types de modèles au niveau de l'ANOVA pertinents à l'analyse des biopuces: le modèle simple et le modèle mixte. Le modèle simple teste la différence existant entre des groupes qui sont classifiés selon une ou plusieurs variables indépendantes tandis qu'un modèle mixte comprend des effets présumés fixes et des effets aléatoires. Dans les deux cas, il est possible de déterminer s'il existe des effets significatifs attribués aux variables indépendantes et s'il existe des interactions existantes entre celles-ci. Il existe une interaction entre deux variables indépendantes quand le résultat associé à une variable indépendante est influencé par le niveau d'une autre. (Kerr *et al.*, 2000).

MAANOVA

MAANOVA « A software package for the analysis of spotted cDNA Microarray experiments » a été développé pour l'analyse de variance aux expériences de biopuces. MAANOVA contient des fonctions, disponible dans R, Matlab et MIDAS, permettant de faire l'analyse l'expérience de biopuces (Wu *et al.,* 2002).

1.3 Considérations préalables et choix des outils d'une chaîne d'analyse de microarray

Une des problématiques actuelles en génomique est l'accessibilité à des outils d'analyse simples mais efficaces permettant d'interpréter des données produites par les expériences de biopuces. Ces expériences visent à déterminer le profil d'expression de milliers de gènes à la fois, dans différentes conditions. Au sein du projet Arborea, de telles

expériences ont été entreprises dans le cadre d'études de génomique fonctionnel de l'épinette et du peuplier.

La mise en place d'une chaîne d'analyse est essentielle pour assurer une bonne gestion et un bon suivi des données. Cette chaîne d'analyse doit être représentative des opérations en laboratoire et doit permettre l'entreposage de toutes les informations possibles concernant les données de laboratoire. Les logiciels choisis pour former la chaîne doivent bien répondre aux besoins et être compatibles entre eux. Dans le cadre du projet Arborea, les données à entreposer concernent toutes informations liées aux expériences et aux échantillons, provenant d'expériences majoritairement de comparaison de lignées transgéniques et sauvages, et factorielles qui génèrent un nombre très considérable d'échantillons. La chaîne d'analyse et de traitement devra servir non pas à simplement entreposer les échantillons et les expériences mais aussi à permettre un suivi de toutes les manipulations faites sur les échantillons de leur récolte jusqu'à l'hybridation des biopuces.

Nous présentons ici l'inventaire des programmes déjà disponibles qui pourront être utilisés pour la chaîne d'analyse. Plusieurs besoins doivent être comblés à diverses étapes du projet. Il faut assurer la gestion et le suivi des données de laboratoire; du début d'une expérience à la récolte des échantillons et de leur entreposage afin de pouvoir suivre étape par étape les manipulations faites sur les échantillons. Il doit aussi être possible de suivre ces échantillons jusqu'à leur utilisation dans les expériences de biopuces et dans l'analyse des données. Il est donc souhaitable d'utiliser un logiciel englobant toutes ces informations ou une suite de logiciels indépendants mais pouvant communiquer et s'échanger des données.

1.3.1 Standard MIAME

MIAME ('Minimum Information About Microarray Experiment') proposée par MGED (http://www.mged.org/) (Microarray Gene Expression Data Society), est un standard dans le développement de bases de données pour microarray et de système de gestion de celles-ci. Plus précisément, il propose une structure conceptuelle pour les descriptions

d'expériences de biopuces. Ils proposent l'utilisation d'un 'vocabulaire contrôlé' afin faciliter les requêtes et l'analyse automatique des données afin d'assurer l'utilisation des mêmes termes entre différentes équipes de recherches facilitant ainsi l'échange et le partage de données. Il est aussi recommandé d'y ajouter son propre vocabulaire étant donné la quantité réduite de terme existant. Ceci peut se faire en spécifiant son qualificatif et une valeur identifiant la source de la terminologie. MIAME comprend présentement cinq parties distinctes : 1) le plan expérimental, 2) les échantillons, 3) les hybridations, 4) méthode de mesures, 5) le plan des biopuces.

Le plan expérimental contient toute l'information concernant l'expérience tel le but de l'expérience, une description de celle-ci, les facteurs expérimentaux, le plan de l'expérience, les contrôles de qualité et les liens vers des publications ou autres site web pertinent à l'expérience.

La section sur les échantillons contient l'information sur l'origine de ceux-ci, les manipulations faites et les protocoles utilisé, les facteurs utilisés pour chaque échantillon, les protocoles techniques pour la préparation de l'extraction des échantillons (ARN ou ADN) à hybrider et les contrôles externes utilisés.

Les procédures d'hybridation décrivent le protocole et les conditions de l'hybridation, le blocage et le lavage incluant toutes les étapes pertinentes à ces étapes comme le marquage.

La partie sur les méthodes et les spécifications de mesure contient les données brutes comme les images des scans. Il y a aussi les données normalisées en plus des protocoles d'analyse et d'extraction des données comme les logiciels utilisés avec les procédures et paramètres en plus des méthodes de normalisation, de transformation et de sélection des données.

Le plan des biopuces (figure 1.2) doit contenir toute l'information générale concernant les biopuces tel le type de plate-forme utilisée, les spécifications des surfaces et de leur recouvrement ou le numéro d'identification commercial du produit ainsi que l'organisme utilisé. Il doit aussi y avoir les spécifications de la puce. L'information de chaque points

comme sa position, le rôle du rapporteur, sa séquence nucléotidique, son numéro d'accession, les amorces utilisées pour le PCR ainsi que toute annotation faite sur les gènes rapporteurs.

Table 1. Oligonucleotide array description file example:

Feature				Reporter						Biological annotation			
Coordinates on Array				Reporter ID (user defined) Oligo ID	Biosequence Type	Sequence	DDBJ/ EMBL/ Genbank	Reporter Usage	Control Type	ID	Designation	Related Gene Symbol, if appropriate	Database Entry
Meta Col	Meta Row	Col	Row										
1	1	1	1	Cy3Cy5	Oligo	AAAAAAAAAAAAAAAAAA	_	Control	Positive	C001_01	Labeled oligo	_	_
1	1	2	1	M00868_01	Oligo	ACCAGCAGATACCTCCTTG	D83002	Experimental	_	C002_01	Gene	ALK	LocusID 11682
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	6	10	8	M00264_01	Oligo	ATGTCCGTTGAATTGG	D83002	Experimental	_	C002_01	Gene	ALK	LocusID 11682
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	6	11	8	M02404_01	Oligo	AGTGGCGAGGAGGAGGAC	L11065	Experimental	_	C449_01	Gene	OPRK1	LocusID 18387
4	6	12	8	M03172_01	Oligo	CCACCACCAAGACCTACTCC	U34891	Experimental	_	C450_01	Gene	KLRA9	LocusID 16640

Table 2. cDNA array description file example:

Feature				Reporter						Biological annotation			
Coordinates on Array				Reporter ID (user defined) HGMP Ref	Biosequence Type	Clone ID	DDBJ/ EMBL/ Genbank	Reporter Usage	Control Type	ID	Designation	Related Gene Symbol	Database Entry
Meta Col	Meta Row	Col	Row										
1	1	1	1	370503	cDNA clone	IMAGE 32017	R17905	Experimental	_	C1	Gene	FNTA	LocusID2339
1	1	2	1	370504	cDNA clone	IMAGE 2962831	BC005866	Experimental	_	C2	Gene	MLH1	LocusID 4292
1	1	3	1	370505	Genomic clone	Cosmid 9H11	L40416	Control	Positive	_	_	_	_
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮			⋮	⋮
4	8	24	12	380696	cDNA clone	IMAGE 5214483	BC028215	Experimental	_	C285	Gene	PTEN	LocusID 5728

Figure 1.2 : Description selon les critères définis par MIAME d'une biopuces d'oligonucléotide et d'une biopuce d'ADNc

(http://www.mged.org/Workgroups/MIAME/miame.html)

MIAME propose donc un standard de conception pour les bases de données de biopuces que l'on doit intégrer dans les outils d'une chaîne d'analyse.

1.3.2 Les différentes architectures d'entreposage de données

Lorsqu'on développe (ou utilise) un système de gestion de données, le type de base de données sous-jacente a un impact considérable sur l'utilisation que l'on pourra faire des données. Les principales options disponibles dans l'entreposage sont Microsoft Access, MySQL, PostgreSQL, Oracle, Sybase et DB2. Chaque type de base de données est unique et possède ses avantages et désavantage.

Le langage SQL est utilisé pour manipuler les données dans chacun des logiciels d'entreposage de données connus. Certains logiciels offrent plus de fonctionnalités au niveau du langage SQL et plus de stabilités au niveau de la base de données que d'autres. SQL, un acronyme pour « Structured Query Language », permet d'interfacer directement avec une base de donnée. Il avait été développé par IBM dans les années 70 pour être utiliser avec les systèmes R et est maintenant un standard ISO et ANSI au niveau des bases de données.

Le langage SQL comprend trois niveaux d'utilisation : DDL, DML et DCL. DDL, qui signifie « Data Definition Language statements », inclus les fonctions qui permettent de définir, bâtir la base de données. Ces commandes servent à définir les attributs de chaque table et les relations existants entre celles-ci. Le niveau DML signifie « Data Manipulation Language statements » et correspond aux commandes servant à manipuler les données entreposées dans la base de données. C'est à partir de ces commandes que les accès aux informations sur la base de donnée seront faits. Le dernier niveau de langage est le DCL (« Data Control Language statements ») concerne toutes les commandes relatives à l'accès et la sauvegarde des données. Ce niveau permet d'assurer l'intégrité des donnés en sauvegardant l'état des données et de retourner en arrière en cas de corruptions de données. Il est important de bien identifier ses besoins car les logiciels de bases de données n'offrent pas les mêmes fonctionnalités (voir tableau 1.1).

Le choix entre Microsoft Access et MySQL se pose parfois étant donné qu'elles sont bien connues, sont disponibles gratuitement ou à un coup moindre et permettent un temps de développement relativement court. Par contre, leurs fonctionnalités sont très différentes l'une de l'autre. Access ne peut être utilisé que dans l'environnement Microsoft Windows tandis que MySQL est multi plateforme donc peut être utilisé autant sur Windows que sur UNIX ou Linux. De plus, le produit de Microsoft ne permet l'accès de ses données qu'à un seul utilisateur à la fois limitant ainsi grandement son utilisation dans une grosse équipe. MySQL est multi usagé donc permet à plusieurs utilisateurs d'accéder aux données en même temps sans danger de corruption des données. MySQL a été développé pour fonctionner dans un environnement réseau.

Il est important de s'assurer que les données sont entreposées de façon sécuritaire et ne peuvent être accédées que par les personnes autorisées. MS Access doit être entreposé sur une machine en local et peut être accédé par toute personne ayant accès à l'ordinateur tandis que MySQL nécessite une authentification. De plus, MySQL permet de gérer et de contrôler les actions faites par chacun des utilisateurs. Le nombre de données gérer par Access est significativement moins important que MySQL. À l'ère du 'open source' il est important de remarquer que MySQL est gratuit alors que MS Access est un logiciel commercial distribué par Microsoft.

Tableau 1.1 : Types de Bases de données disponibles

Nom	Description
Microsoft Access	Access est une base de données complètement intégrée dans la suite de logiciel Office de Microsoft. La force de cette base de données est sa disponibilité, son faible coût et sa facilité d'utilisation. Le développement d'une base de donnée Access peut se faire de façon intuitive avec un minimum de connaissance dans ce domaine. Access offre aussi la possibilité d'utiliser une interface graphique, et la création de rapport directement liée à la base de donnée qui nécessite très peu de développement. Microsoft Access est une base de donnée sous la forme d'un seul fichier qui peut être échangé, ou partagé sur un réseau. L'inconvénient d'Access est sa capacité de croissance car il permet à peu d'utilisateurs d'accéder aux mêmes données en même temps. De plus, la quantité de données entreposées et la sécurité entourant de la base de données sont significativement moins importantes que les autres types de bases de données disponibles. Access constitue néanmoins une option intéressante dans le cadre d'un projet nécessitant un développement rapide pour une gérer une quantité de données qui n'évoluera pas énormément au cours de son utilisation.
MySQL	MySQL offre deux options : une version disponible comme logiciel à code ouvert et une autre version appelée MaxDB. MySQL a beaucoup évolué ces dernières années et a passé d'un produit de qualité moindre à un produit qui se rapproche beaucoup des bases de données commerciales. MySQL peut être installé et utilisé sur plusieurs systèmes d'opérations incluant Windows, Unix et Linux. Les inconvénients sont le manque de réplication, de clés étrangères et de vues indexées. MySQL est gratuit à condition de développer un logiciel non commercial sinon il faudra payer une licence. MySQL est une très bonne solution, très bien documentée pour le développement

	d'application pour un projet de petite à moyenne envergure nécessitant une base de données relationnelle.
PostgreSQL	PostgreSQL est disponible librement comme MySQL. Il est très bien documenté et offre de nombreuses de fonctions intéressantes. PostgreSQL fonctionne sur tous les systèmes d'exploitation et gère de très grosses quantités de données. Il implémente les standard ANSI 99 et est conforme aux standards attendue dans les entreprises incluant les unions, vues, indexages, procédures enregistrés et beaucoup plus. Par contre PostgreSQL a deux inconvénients majeurs : il ne supporte pas des quantités de données aussi massive que DB2 ou Oracle mais mieux que MySQL et les exécutions de requêtes SQL sur PsotgreSQL peuvent prendre du temps à s'exécuter. PostgreSQL est mieux vue comme un standard d'entreprise que l'est MySQL mais moins que SQL Server ou DB2.
Oracle	Oracle est une des bases de données les plus anciennes encore utilisées à ce jour. Il offre des performances et des fonctionnalités très impressionnantes sur tous les systèmes d'exploitation connus. Oracle est une référence dans le domaine des bases de données. Par contre deux facteurs limitent quelque peu l'utilisation de Oracle : le coût et sa facilité d'utilisation et de développement. Une licence Oracle peut rapidement devenir très coûteux. Il faut aussi préciser que Oracle est plus complexe d'utilisation que les autres logiciels de bases de données mais offre une gamme impressionnante d'options. Oracle est le standard dans le développement de bases de données mais demande une licence payante.
Microsoft SQL Server	Microsoft SQL Server est une application permettant de créer une base de données sur les serveurs utilisant un système d'exploitation Windows. Cette application permet d'utiliser et de créer une base de données accessible par d'autre ordinateurs de tables, à travers Internet ou même à partir d'assistant digital (PDA). De plus, Microsoft SQL

	Server est accessible et bien documenté.
Sybase	Sybase est une compagnie informatique offrant des systèmes de bases de données. Très similaire au Server SQL, il possède néanmoins un avantage sur celui-ci : il performe sur plusieurs plateforme.
DB2	Développé par IBM, DB2 appelé communément « DB2 Universal database » est une suite de logiciel utilisé pour gérer un système de données. Il performe sur les principaux systèmes d'exploitation comme Windows. Le désavantage de DB2 est son manque de popularité en dehors de la compagnie IBM et aussi son coût dispendieux d'utilisation. DB2 peut être accédé de n'importe quelle logiciel en utilisant soit le module ODBC de Microsoft, JDBC de Java ou par l'interface CORBA.

Description des différents types de bases de données disponibles

1.3.3 Disponibilité des outils de la chaîne d'analyse

1.3.3.1 Logiciels de suivis d'expériences et d'échantillons

Les logiciels de suivis d'expériences et d'échantillons entreposent et gèrent les données de base de laboratoire. Ils visent à entreposer toute l'information nécessaire à propos d'une expérience, les différents tissus récoltés, les traitements appliqués aux échantillons, etc. Il existe une multitude de logiciels permettant l'entreposage de données d'échantillons de laboratoire. Ces logiciels sont majoritairement commerciaux ou ont été conçus pour affronter une problématique différente.

Dans le projet Arborea, les types d'informations sur les données expérimentales avaient été colligés à l'intérieur d'un fichier EXCEL préalablement au choix des outils d'entreposage et d'analyse. Le but fut donc de trouver un logiciel qui, en plus d'être compatible avec les données actuelles, permet de créer efficacement et rapidement des expériences factorielles contenant des dizaines, voir des centaines d'échantillons chacune. Aucun logiciel (http://ihome.cuhk.edu.hk/~b400559/arraysoft_database.html) n'offrait une compatibilité acceptable avec nos données actuelles en plus d'une efficacité à manipuler nos types d'expérience. La plupart des logiciels d'entreposage de données de biopuces gèrent bien les données relatives à celle-ci mais plutôt mal l'information en amont comme la description des expériences, les traitements appliqués et le prélèvement d'échantillons. Il donc a été décidé de créer nous-même un système de gestion d'expériences et d'échantillons répondant exactement aux besoins du projet. Ce système a été nommé SLIMS; son développement et sa structure sont décrit au chapitre 2.

1.3.3.2 Logiciels d'entreposage et d'analyse de biopuces

Il existe plusieurs types de logiciels d'entreposage et d'analyse de données de biopuces. La plupart ne font que l'entreposage ou l'analyse et non les deux à la fois, ce qui rend les choix plus complexes lorsqu'on cherche à rencontrer les deux objectifs.

Les logiciels d'analyse de biopuces performent plusieurs tâches fondamentales : l'analyse d'image des hybridations, le prétraitement des données qui comporte la transformation

des données afin de les rendre comparable pour les fins de l'analyse des données, l'analyse statistique pour l'identification des gènes différentiellement exprimés et la visualisation des données. Il existe un nombre impressionnant de logiciels disponibles pour chacune de ces étapes (http://ihome.cuhk.edu.hk/~b400559/arraysoft.html). Notre approche est de choisir un logiciel qui nous permettra de faire toutes ces étapes ou une suite de logiciels compatibles à travers les différentes manipulations.

Bien qu'il existe des logiciels commerciaux efficaces pour faire l'analyse et l'entreposage des biopuces nous avons préféré trouver l'équivalent au niveau de la communauté des logiciels distribué librement. Le tableau 1.2 donne une brève description des principaux logiciels d'entreposage et d'analyse de données de biopuces disponibles gratuitement.

Information supplémentaire concernant BASE

Le choix d'un logiciel pour l'entreposage des données de biopuces s'est arrêté sur le logiciel BASE (BioArray Software Environement). Ce logiciel permet de faire l'entreposage et l'analyse des données de biopuces. BASE nécessite un serveur Apache ainsi qu'une base de donnée MySQL et un interpréteur PHP, tous disponible gratuitement. L'installation du serveur BASE ne requiert aucune installation au niveau des utilisateurs et permet aux collaborateurs extérieurs de visualiser et accéder aux données du projet. Étant donné que les données numérisées sont directement entreposées sur BASE il est très facile et rapide de passer aux étapes suivantes comme l'assurance-qualité, la transformation et l'analyse des données. De plus, BASE permet l'ajout de scripts et de 'plug-ins' développé par la communauté donc reste à l'affût des développements dans le domaine. Avec ces nouveaux modules il est possible de faire le regroupement hiérarchique (hierarchical clustering), la normalisation entre les lames, etc.

Tableau 1.2 : Types Liste des logiciels de gestion de données de biopuces à ADN

Nom	Description
Nomad	Nomad (http://sourceforge.net/projects/ucsf-nomad/) est un système librement disponible et adaptable (« open source ») pour entreposer et interroger les résultats d'expérience. Développé par Michael Salazar en partenariat entre trois laboratoires soit l'Université de Californie, San Francisco et « Lawrence Berkeley National Laboratory ». Il est, par contre, à un stade très tôt de développement limitant ainsi son utilisation. Il ne semble pas y avoir eu de mise-à-jour autant au niveau des versions du logiciel que du site web depuis un an.
MADAM (Suite TM4)	Les outils TM4 consistent en quatre applications majeures, Microarray Data Manager (MADAM), TIGR_Spotfinder, Microarray Data Analysis System (MIDAS), et Multiexperiment Viewer (MeV), et une base de données conforme à MIAME. MADAM (http://www.tm4.org/madam.html) facilite l'entrée des données grâce à une interface graphique développée en C++. TIGR_Spotfinder permet de faire l'analyse des images d'hybridation et la quantification d'expression des gènes. MIDAS possède une interface Java qui permet aux utilisateurs de construire des protocoles d'analyses combinant un ou plusieurs étapes de normalisation et de filtrage. Finalement, MEV est capable de charger des fichiers « .tav », incluant ceux normalisés par MIDAS pour générer des résultats de données d'expression et d'annotation de une ou plusieurs expériences. La Suite TM4 développé par l'institut TIGR permet de faire le prétraitement et l'analyse des données de biopuces grâce à deux applications : MIDAS et MEV.

MeV (Suite TM4)	MeV (MultiExperiment Viewer ; http://www.tm4.org/mev.html) permet de faire l'analyse des données filtrées et normlisées. Il permet, entre autre, de faire la visualisation des hybridations et de leurs patrons d'expressions correspondantes. De nombreux algorithmes de 'clustering' (Bootstrapping, Jacknifing et K-means par exemple) sont disponibles pour identifier et travailler facilement avec des gènes d'intérêts. Il est aussi possible d'ajouter des annotations personnelles ou publiques aux données d'expression à l'aide des fichiers EASE. MeV permet la mise en place et l'échange de protocole d'analyse sous forme de fichier facilement échangeable et utilisable. Il intéressant de préciser que MeV acceept beaucoup de format de fichiers comme fichier d'entrée donc il n'est pas nécessaire, bien que souvent recommandé, de faire la normalisation et standardisation des données avec MIDAS avant de faire l'analyse avec MEV. Il accepte les fichier en format TIGR MeV (*.mev), les fichiers délimitées par des tabulation (*.TDMS), les fichiers TIGR Arrayviewer (*.tav), les formats Affymetrix, Genepix et Agilent.
2HAPI	HAPI (High-density Array Pattern Interpreter) a été implanté initialement avec Visual Foxpro par Dan Masys, Barney Welsh, et Jacques Corbeil du UCSD School of Medicine et est limité à ne pouvoir fonctionner que sur les plate-formes Windows. 2HAPI est l'incarnation web de HAPI. Permettant ainsi d'accéder à toutes les fonctionnalités à travers une interface web. Des algorithmes de « clustering » et des analyses en amonts seront disponibles dans l'avenir. Facteurs limitant l'utilisation de 2HAPI: 2HAPI est toujours en développement, ne contient encore pas tous les numéros d'accession NCBI en local. L'accès à l'application à l'air de se faire à partir de leur serveur et il ne semble pas y avoir de manière de télécharger 2HAPI et de l'utiliser en local.
RAD2	RAD (Stoeckert et al. (2001)) est une base de donnée publique sur l'expression des gènes qui entreposera des données provenant de (microarrays, high-density oligo arrays, microarrays) et nonarray-based (SAGE)

	experiments. Le but ultime de ce projet est de permettre l'analyse comparative d'expériences de laboratoires différents utilisant différentes plates-formes et étudiant différents systèmes biologiques. Pour atteindre ce but, RAD contient: des descriptions précises des expériences et des distinctions entre les données brutes et des résultats traités. De plus, un index de gènes est utilisé pour intégrer des éléments de microarrays et des "étiquettes" de gènes. La sélection des expériences incluses dans le RAD seront déterminés par les intérêts des chercheurs du projet et de leurs collaborateurs donc est un facteur très limitant dans notre cas.
Quicklims	Quicklims (http://www.dkfz-heidelberg.de/kompl_genome/Other/QuickLims/) a été programmé avec VBA donc avec Visual Basic utilisant Microsoft Access comme base de donnée. Le programme et le code sont disponibles gratuitement pour les utilisations non commerciales. Ils sont disponibles au DKFZ et les auteurs sont Felix Kokocinski et Gunnar Wrobel. Cette application n'utilise pas le standard MIAME et sa base de donnée Microsoft Access est un facteur très limitant lorsqu'on a plusieurs utilisateurs et qu'on s'attend à y entreposer beaucoup de données.
Arrayexpress	Arrayexpress (H. Parkinson et al. Nucleic Acids Research, 2005, Vol. 33, Database issue D553–D55) est un dépôt public développé par l'institut EBI (European Bioinformatics Institute) pour les biopuces, dont le but est d'entreposer les données d'annotation en accordance au standard MIAME. Il est possible de télécharger les fichiers et de l'utiliser en local. Il sert uniquement à entreposer les données de microarray et ne sert pas à faire l'analyse de biopuces. De plus il utilise une base de données Oracle, très dispendieuse d'utilisation.
MaxD	MaxD, a été développé sur les bases de Array Express par le « Microarray Bioinformatics Group » à l'Université de Manchester. Les modifications par rapport à Array Express sont sur la forme du modèle entités-

	relations. MaxD possède aussi une application appellée MaxView qui permet la visualisation et l'analyse des données d'expression (en Java).
BASE	BASE (Saal LH, Troein C, 2002) permet l'entreposage et l'analyse de données de microarray. Il permet de gérer l'information sur les biomatériels, les données brutes et les images. BASE permet aussi l'installation de 'plug-in' permettant de faire la normalisation, la visualisation et l'analyse de nos données. Il est disponible gratuitement sous la licence GNU GPL. BASE possède une interface web intuitive et s'installe sur un serveur central éliminant ainsi tout besoin d'installation sur les ordinateurs des utilisateurs. Il fait la distinction entre différents utilisateurs à l'aide de compte individuel. Il peut être installé sur UNIX ou Linux, nécessite un serveur Apache, un interpréteur PHP et C++ et est constitué d'une base de donnée MySQL (tous disponible gratuitement). L'utilité et la force de BASE au niveau de l'analyse de données d'expression sont la disponibilité immédiate des données brutes.
Bioconductor	Bioconductor (Genome Biol. 2004;5(10):R80. Epub 2004 Sep 15.) est un projet « open-source » fournie à la communauté scientifique des outils d'analyse permettant d'interpréter efficacement des données génomiques. L'équipe responsable de ce projet est celle de Jianhue Zang du Dana Farber Cancer Institute au Harvard Medical School de Boston. Le but de Bioconductor est de fournir un large éventail d'outils statistiques et graphiques pour l'analyse des données génomique ainsi que de faciliter l'intégration de « métadonnées » dans l'analyse de données expérimentales. Il est très facile d'imaginer l'utilité d'avoir accès à toute la littérature concernant un gène, ou toutes les données d'annotation reliées directement à nos données expérimentales. Le fait que Bioconductor est « open-source » est un immense atout pour toute la communauté scientifique. En rendant tout le

	code source disponible à la communauté il devient possible d'avoir accès aux algorithmes et de pouvoir ainsi apporter des mises à jour, des variantes de ceux-ci. Il est ainsi possible de prendre un algorithme ou un programme et de le modifier à notre guise pour qu'il soit mieux adapté à notre type d'expérience. Plusieurs modules ont été développés par la communauté scientifique pour traiter et analyser les données de biopuces. Les plus populaires sont 'Marray' et 'Limma'.
Marray	Le module Marray (Sandrine Dudoit, Yee Hwa Yang) de Biocondutor implémente les méthodes d'adaptation locale, les procédures de normalisation et de 'scaling' utilisées pour éliminer le biais dû au marquage et permet aussi l'utilisation de séquence 'spotté' comme contrôle.
Limma	Limma (Smyth, G. K., et Speed, T. P. (2003). Normalization of cDNA microarray data.) est un autre module de Bioconductor permettant de faire l'analyse des données d'expressions de biopuces. Limma contient plusieurs fonctionnalités déjà existantes dans marray mais possède une approche faisant plus la distinction entre la normalisation à l'intérieur d'une puce et entre les biopuces que celui-ci. Il utilise les modèles linéaires pour analyser des plans d'expérience et y évaluer les gènes différentiellement exprimés. Il propose une alternative intéressante et peut-être même plus intuitive que celle proposée par marray. De plus, il est possible d'utiliser Limma en conjonction avec marray afin d'utiliser le meilleur des deux modules.

BASE contient toute l'information concernant les biopuces et tous les résultats d'expérience d'hybridations. Les fichiers quantifiés avec un logiciel d'analyse d'image comme Quantarray sont entreposés sur BASE avec les images des hybridations. Les analyses préliminaires d'assurance qualité sont disponibles rapidement grâce aux outils intégrés dans le logiciel. BASE permet de créé des jeux de données appelés 'bioassay set' sur lesquels il est possible d'appliquer des algorithmes sans modifier les données originales. Chaque étape de l'analyse est affichée dans une organisation hiérarchique des étapes d'analyses permettant de voir chaque étape de l'analyse de façon chronologique. Il est possible d'accéder aux données à chaque étape de l'analyse. Les algorithmes d'analyses comprennent le calcul des corrélations de M et de A, le regroupement hiérarchique, le filtrage par rang, la normalisation global, lowess et basé sur la médiane des têtes d'impression (Figure 1.3). Un aspect très intéressant est d'avoir accès aux résultats des analyses qualités quelques minutes seulement après leur entreposage sur BASE. L'exécution des algorithmes sur BASE ne nécessite qu'un minimum de connaissance de leur fonctionnement. L'exportation des données est possible dans un format délimité par des tabulations et permet d'approfondir l'analyse statistique avec Bioconductor ou la suite TM4.

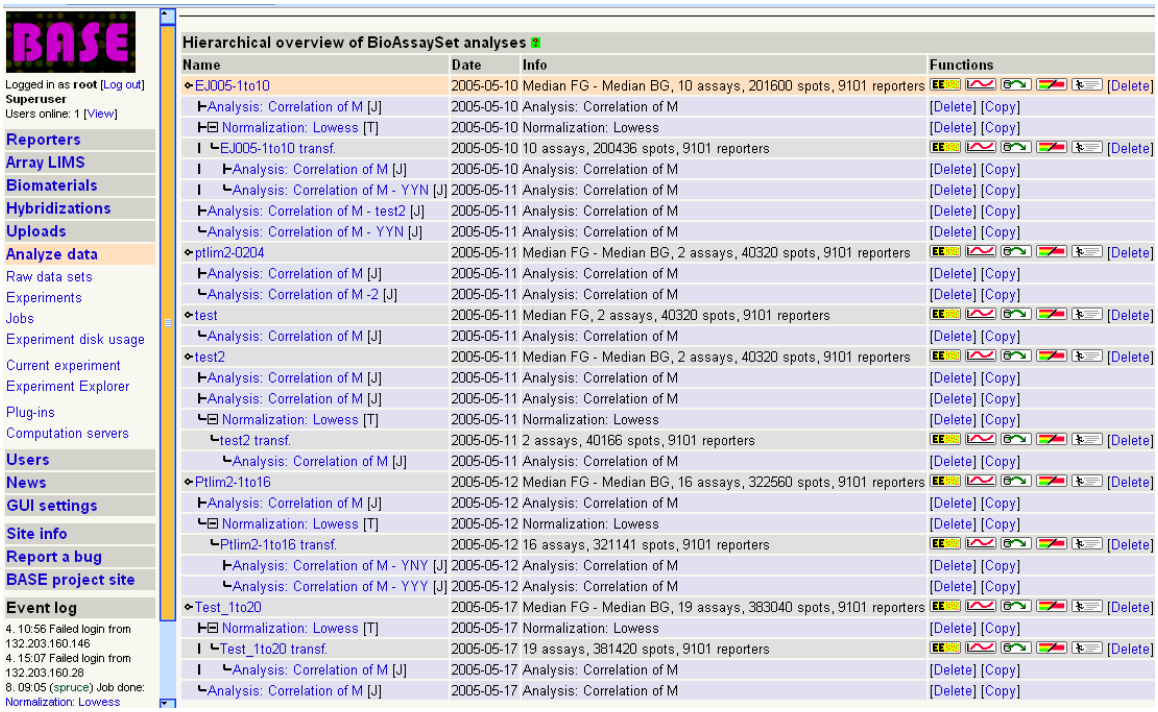


Figure 1.3 : Interface d'analyse de BASE

1.4 Description de la chaîne de traitement et d'analyse de biopuces

La chaîne de traitement et d'analyse des données de biopuces présentées dans ce mémoire utilise le système open-source BASE, des outils de la suite TM4 (de TIGR), Bioconductor et le système SLIMS que nous avons développé (figure 1.4). Le premier logiciel de la chaîne d'analyse permet l'entreposage et le suivit des données d'échantillons et d'expériences. C'est le point d'entrée des informations dans la chaîne d'analyse. Toutes les informations relatives à l'expérience et aux échantillons sont détaillées à cette étape et resteront accessible en tout temps à travers cette application. Le logiciel SLIMS, présenté au chapitre 2, a été conçu pour répondre au besoin précis du projet et permettre le transfert de ces données vers des logiciels d'entreposage et d'analyse de puce.

BASE gère toute l'information sur les échantillons à partir de l'extraction d'ARN jusqu'à l'hybridation sur les lames. Les données d'extractions sont disponibles dans BASE à l'aide d'une option de transfert existant entre SLIMS et celui-ci. Le logiciel gérant les données de biopuces, BASE, permet d'entreposer toute l'information concernant les biopuces et les données d'hybridations.

De plus, le logiciel SLIMS permet de transférer tous les échantillons appartenant à une expérience de la base de données de SLIMS vers BASE ce qui évite le processus fastidieux d'entrer manuellement tous les échantillons dans ce dernier. Le logiciel d'entreposage MADAM de TIGR ne fonctionne que sur le système d'exploitation Windows et ne peut être installé que sur un seul poste limitant beaucoup son utilisation. Les autres logiciels disponibles au moment de la sélection consistaient en majorité de site d'entreposage de données de biopuces extérieures, de logiciels commerciaux ou de logiciels n'offrant pas la simplicité d'utilisation de BASE.

Les analyses qualités de résultats d'hybridation sont faites majoritairement avec Bioconductor et MIDAS. Des scripts ont été développé en R pour calculer et faire des graphiques des intensités brutes et les corrélations à l'intérieur et entre les lames afin d'éliminer les lames problématiques.

La transformation et l'analyse des données ont été faite avec Bioconductor et MeV. Bioconductor a servi à faire les normalisations composites intra, inter lames et de faire l'analyse limma. L'annotation EASE et le 'clustering' des données a été fait par la SAM dans le logiciel MeV. Des annotations construites par un algorithme python sont présentées au chapitre 3.

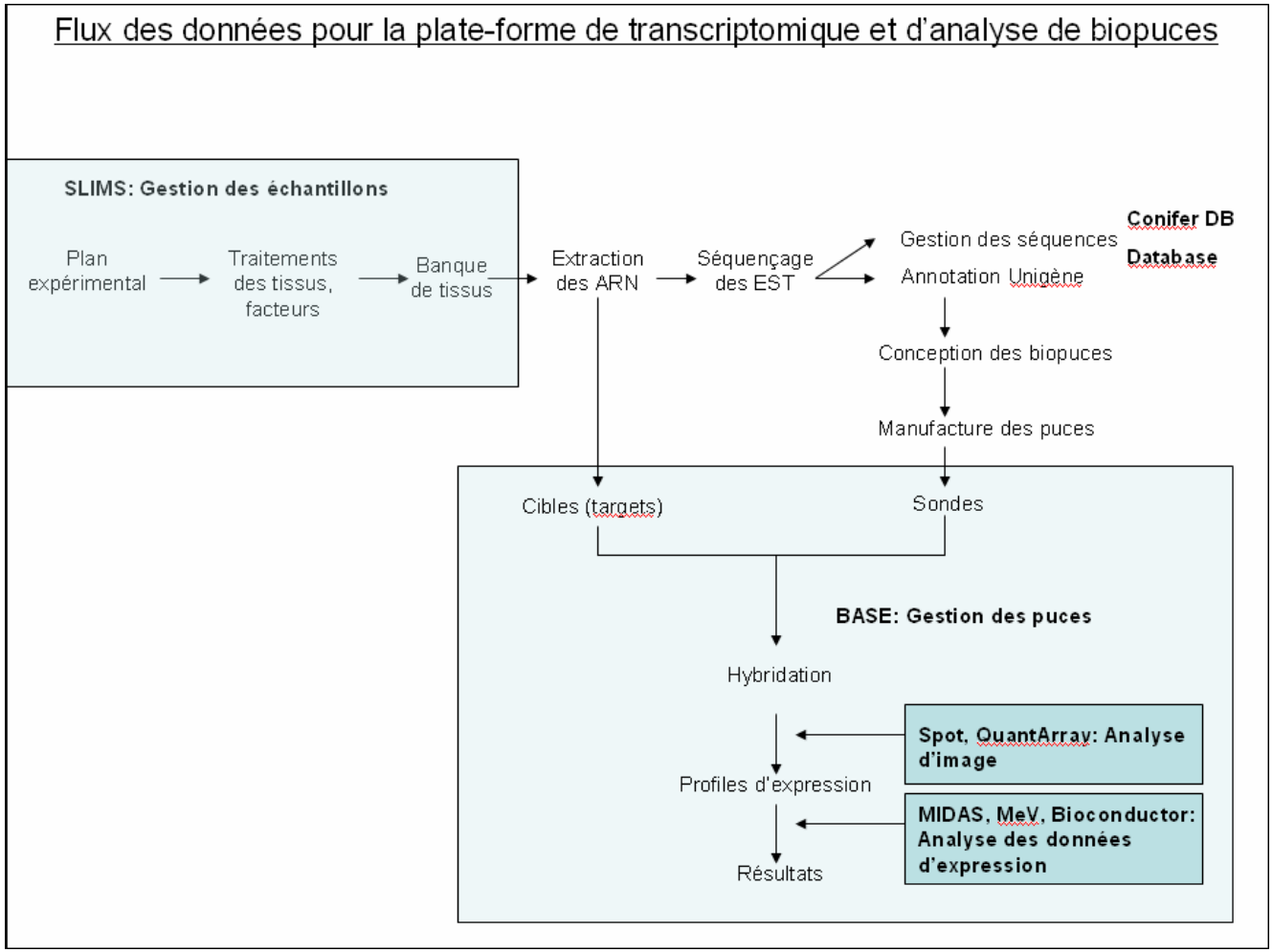


Figure 1.4 : Intégration des logiciels SLIMS et BASE à la plate-forme de transcriptomique et d'analyse de biopuces

1.5 Génomique fonctionnelle et formation du bois chez les arbres

L'ère de la génomique amène une nouvelle façon d'étudier les plantes. Plusieurs applications intéressantes découlent d'une analyse génomique approfondie de celles-ci.

Le 1er génome végétal complètement séquencé est celui d'Arabidopsis thaliana. Cette plante modèle a été choisie à cause de son génome relativement petit pour une plante. Par exemple, le riz contient trois fois plus d'ADN qu'Arabidopsis, la tomate sept fois, le maïs vingt fois et le blé cent vingt fois. Ce projet fut réalisé à l'aide d'un réseau de scientifique étalé à l'échelle de la planète. Le séquençage complet d'Arabidopsis a pour but d'aider l'étude de la structure et les fonctions vitales des plantes pour, éventuellement, appliquer ces connaissances au domaine de l'agriculture, de la santé et de l'environnement. Ce projet de séquençage a débuté en 1990 et fut terminé en 2000. Cet accomplissement a transformé la manière dont les biologistes étudient les plantes. Par ailleurs, plusieurs gènes identifiés dans le génome d'Arabidopsis ont été associés à des maladies chez l'humain. Par exemple, pour la maladie de Wilson, il a été possible de déterminer qu'un gène en particulier était nécessaire à la formation des récepteurs d'hormones chez Arabidopsis.

Dernièrement, les génomes complets du Peuplier et du riz ont été complétés permettant d'accéder au premier génome complet d'arbre. Le peuplier a été choisi étant donné son génome relativement compact (50 fois plus petit que celui du pin). L'intérêt de la biotechnologie pour les arbres et d'améliorer la gestion du milieu forestier. Aux niveaux des arbres forestiers l'étude des mécanismes de la formation du bois et la recherche des gènes impliqués sont d'une importance stratégique pour le Canada. Le Canada demeure l'un des principaux fabricants de produits de bois au monde en exportant vers une centaine de marchés, dont les plus importants ces dix dernières années ont été les États-Unis, le Japon et l'Europe. Une des problématiques actuelles en foresterie est l'impact par la lignine contenu dans le bois sur les procédés de transformation et sur l'environnement. Le procédé d'extraction de la lignine est polluant et coûteux. C'est ainsi que plusieurs études se concentrent sur des familles de gènes ayant un rôle dans la voie de biosynthèse de la lignine dans le but éventuel de pouvoir contrôler le taux de lignification.

1.5.1 Notions générales sur la formation du bois

Le bois est un tissu végétal (le xylème) qui joue un double rôle chez les plantes vasculaires : il est le conducteur de la sève brute et le tissu de soutien qui donne la résistance aux tiges. Il sert aussi de tissu de réserve. Le bois ou xylème contient différents types cellulaires, principalement des cellules vasculaires et parenchymeuses. Il est majoritairement composé de cellulose de 40 à 50%, de lignine de 20 à 30% et d'hemicellulose 25% à 35%.

1.5.1.1 La croissance primaire et la mise en place du système vasculaire primaire

Les mécanismes de formation du bois sont liés de près aux différentes étapes de croissance chez la plante soit la croissance primaire et la croissance secondaire. La phase primaire est responsable de l'élongation de la tige et du développement des tissues de base (Figure 1.5). Elle peut aussi être décrite comme la croissance en longueur à partir du méristème apical responsable de l'allongement de la tige. Elle comprend l'initiation de nouvelles feuilles, de bourgeons et de la création des trois méristèmes primaires. Les méristèmes sont des tissus formés par un ensemble de cellules jeunes et non différenciées, qui se multiplient activement et rapidement durant la saison végétative. Le mot méristème provient du grec meristos signifiant division. Vers 1750, les naturalistes s'intéressent à l'extrémité de la tige, partie où s'effectue le développement. C'est en effet une caractéristique du végétal que de croître par ses extrémités. Cette partie où se produit la croissance caulinaire a été appelée *punctum vegetationis* (De Wolf; 1759). Avec le développement des moyens optiques, les tissus puis les cellules on été délimités puis caractérisés, on parle alors de point végétatif ou zone apicale actuellement. (http://amap.cirad.fr/architecture/organo/organo1.html). Les trois méristèmes primaires sont le protoderme, le méristème fondamental et le procambium (Figure 1.6). On nomme protoderme la couche externe des cellules à l'extrémité de la tige (apex). Cette couche est appelée méristème principal ou apical car ces cellules sont en constante division.

Le méristème fondamental est situé au centre du sommet de la tige à l'intérieur du protoderme. Les principaux tissus formés par le méristème fondamental sont la moelle dans le centre de la tige et le cortex situé sous l'épiderme et entourant les tissus vasculaires. Le

procambium est formé de cellules longues et minces issues des méristèmes fondamentaux. Les cellules procambiales se divisent pour ensuite se différencier en xylème et phloème primaire. Chaque regroupement de cellules procambiales est à l'origine d'un regroupement vasculaire composé de xylème primaire à l'intérieur de la tige et de phloème primaire vers l'extérieur (Weier *et al.,* 1974).

Le rôle de ces tissues primaires est diversifié et essentiel au bon fonctionnement de la plante. L'épiderme protège les tissus situés à l'intérieur de la tige. Au niveau des tissus vasculaires, le phloème forme un vaste réseau de transport dans la plante, permettant aux organes producteurs de métabolites, les organes-sources, d'alimenter les organes-puits, qui utilisent les métabolites. Le cambium vasculaire produit le xylème et phloème secondaire. Le xylème permet le transfert de l'eau et des minéraux à travers la plante tout en renforcissant la tige. La moelle et les parenchymes de rayon servent à entreposer les réserves et leur transport.

Bref, la croissance primaire est responsable de l'élongation de la tige de la plante et des branches, du développement des organes et tissus de bases essentiels dont les feuilles et le système vasculaire primaire et autres. La phase secondaire est responsable du développement des nouveaux tissus vasculaire secondaire, augmentant ainsi le diamètre de la tige et permettant le maintien du réseau continue de cellules vivantes mis en place lors de la croissance primaire entre les tissus nouvellement formés des racines et des tiges en développement (Figure 1.7).

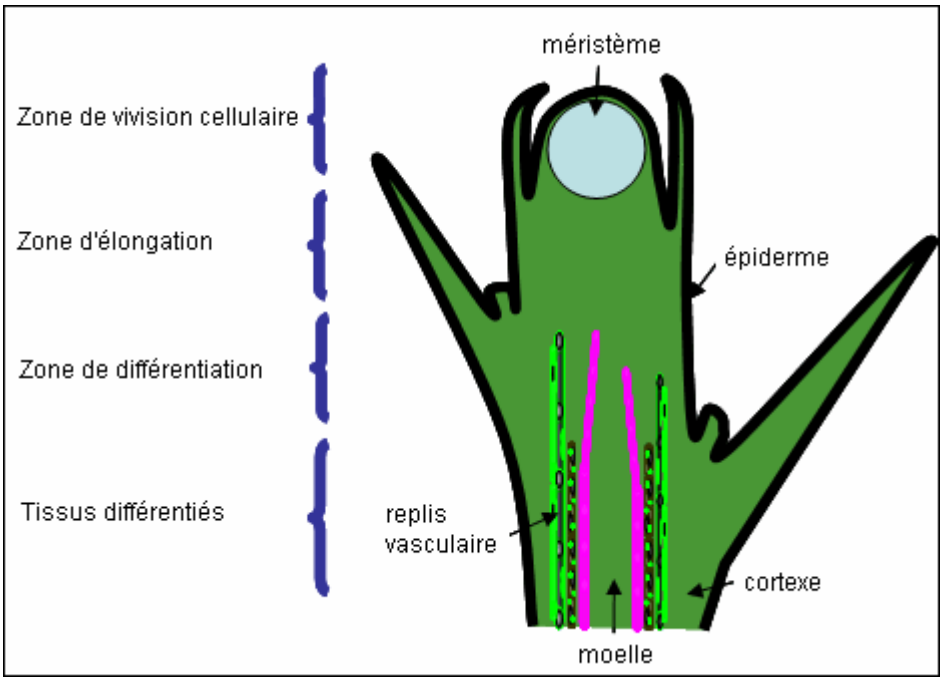


Figure 1.5 : Anatomie de la tige (provenant de http://www.steve.gb.com/images/science/apical_meristem.png)

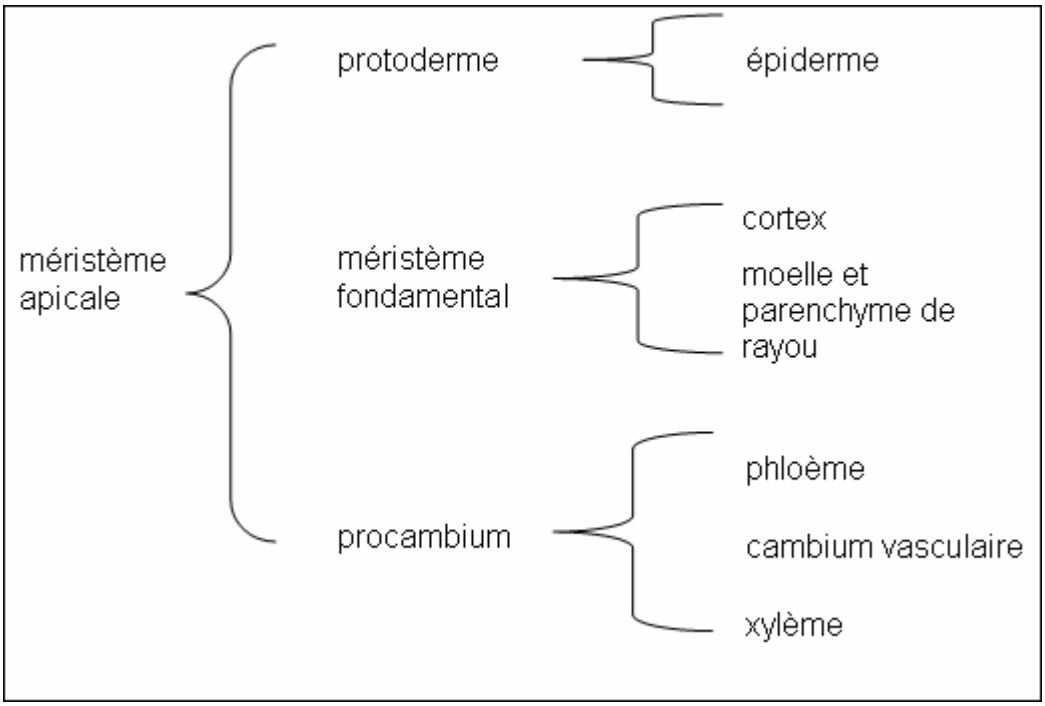


Figure 1.6 : Sommaire du développement primaire (Adatpé de Botany, An Introduction to Plant Biology, 5th Ed.)

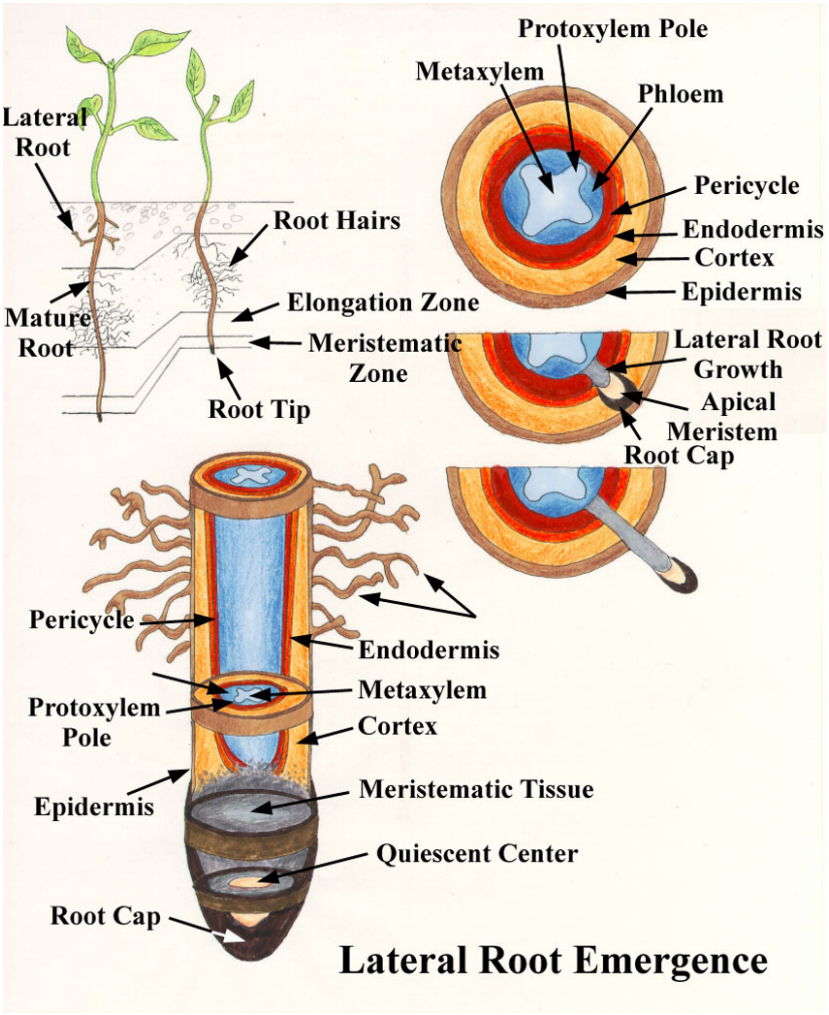


Figure 1.7 : Anatomie de la tige et la racine (provenant du site http://www.puc.edu/Faculty/Gilbert_Muth/art0060.jpg)

1.5.1.2 La croissance secondaire et l'élaboration du système vasculaire secondaire

La croissance secondaire est un phénomène qui n'est pas présent chez toutes les plantes. Par exemple, les plantes herbacées ne montrent aucune ou une très faible croissance secondaire car elles complètent leur cycle de vie en une saison. À l'inverse, plusieurs espèces dicotylédones et gymnospermes ont une croissance secondaire dès leur première année de croissance. Chez certaines plantes, appelées plantes ligneuses, ce processus peut continuer pendant plusieurs, voir des centaines d'années. Elle a lieu au niveau des cambiums ou de zones génératrices (histogènes). Les plantes ligneuses développent des tiges plus massives et épaisses à cause de la croissance du xylème secondaire à partir du méristème secondaire. La première étape nécessaire à la genèse du xylème et phloème secondaire est la formation du cambium vasculaire. Le cambium vasculaire provient de la division cellulaire, stimulée par des hormones végétales, coordonnée à l'intérieur du procambium.

Un cylindre de cellules cambiales se forme tout autour de la tige : il est appelé cambium vasculaire (Figure 1.7). Deux parties différentes contribue à la formation du cambium vasculaire : le cambium fasciculaire qui sont les cellules du méristème à l'intérieur du faisceau vasculaire et le cambium interfasciculaire qui sont les cellules du méristème entre les faisceau vasculaire. Quand ces deux cambiums se rejoignent il y a formation d'un cylindre au sein de la tige. Dès que ce cylindre est formé le cambium vasculaire devient actif et commence sa division à partir de sa surface intérieure et extérieure. Les nouvelles cellules formées sur la face interne du cambium se rattachent au xylème produit antérieurement. Le phloème nouvellement formé du côté externe du cambium s'attache au phloème déjà existant. Deux types de cellules cambiales existent : les cellules orientées dans l'axe radiales de transport latéral qui sont les cellules parenchymes présente dans le xylème et dans le phloème et les cellules fusiformes qui permettent le transport vertical. Chaque année, le cambium vasculaire produit une nouvelle couche de xylème secondaire (plus communément appelé bois). Le cambium à une activité mitotique à partir du printemps jusqu'à l'automne et est inactif durant l'hiver (Weier *et al.,* 1974).

1.5.2 La lignine : un constituant majeur du bois et une cible pour la biotechnologie

La lignine est un hétéropolymère aromatique présent majoritairement dans les parois cellulaire des cellules vasculaires primaires et secondaires. Elle joue un rôle important dans le support mécanique, le transport de l'eau et la résistance aux pathogènes. La composition de la lignine peut varier en fonction des espèces mais consiste de différents phénylpropanoides, plus précisément de trois alcools cinnamyliques : l'alcool p-coumarylique, l'alcool coniférylique et alcool sinapylique.

La lignine forme l'un des principaux constituants du bois et le deuxième polymère naturel dans la biomasse terrestre après la cellulose. Les lignines sont évaluées à 300 milliards de tonnes et le taux annuel de la biosynthèse est de l'ordre de 20 milliards de tonnes et constitue de 20% à 30% de la masse totale du bois lui donnant ainsi sa rigidité (Pr P. Tisnès http://spcmib.ups-tlse.fr/themes/theme1/detail/trav1a.html). Au niveau pratique, trois secteurs sont concernés par la lignine: les papeteries, le domaine de la nutrition animale et l'utilisation du bois pour le chauffage. Dans la fabrication des pâtes à papier, les lignines sont indésirables, à cause de la coloration qu'elles donnent au papier. Pour obtenir un papier blanc, il faut les extraire de la pâte à papier. Les industriels réalisent cette opération par traitement chimique à l'aide de produits chlorés qui sont polluants. Des travaux de recherche visent l'obtention de bois offrant une teneur en lignine plus faible ou desquels la lignine serait plus facile à extraire. Au niveau de la nutrition animale la présence de lignine dans les plantes fourragères réduit la disgestabilité par les animaux. Tandis qu'au niveau du bois de chauffages, les lignines contribuent au potentiel calorifique et ont donc un apport bénéfique.

Certaines études se sont penchées sur la quantité et la composition de la lignine (Anterola and Lewis, 2002 ; Humphreys and Chapple, 2002), et sur le rôle qu'y joue les enzymes de la voie de biosynthèse. Les avancées technologiques au niveau de la génomique ont permis d'agir directement dans le sentier métabolique de la lignine. Il est maintenant possible d'étudier les effets de la modification d'un gène sur l'ensemble du sentier. Le sentier métabolique est relativement bien connu et la plupart des gènes agissant dans ce sentier ont

été identifiés chez plusieurs plantes (Whetten *et al.*, 1998). La modification génétique de la lignine au niveau de son contenu et sa composition est une des applications potentielle de la biotechnologie forestière. Il y a des exemples de manipulations génétiques de certains des gènes de la voie de biosynthèse de la lignine qui ont permis de modifier la quantité et la composition de lignine produite par les plantes. (Whetten *et al.*, 1998).

La lignine est un polymère des alcools cinnamyliques (monolignols). Les gymnospermes contiennent majoritairement l'alcool conifférylique, les angiospermes contiennent l'alcool conifférylique et l'alcool sinapylique et les trois types sont trouvés dans les lignines des graminées. Les gènes impliqués dans le transport des monolignoles à travers la paroi et leur polymérisation à la lignine sont peu connus. La polymerisation des monolignoles, la lignification proprement dite serait catalysée par les peroxydases et les laccases (Eckardt**,** 2002). Les principales étapes et enzymes nécessaires à la synthèse de chacun des précurseurs de la lignine sont présentés à la figure 1.8.

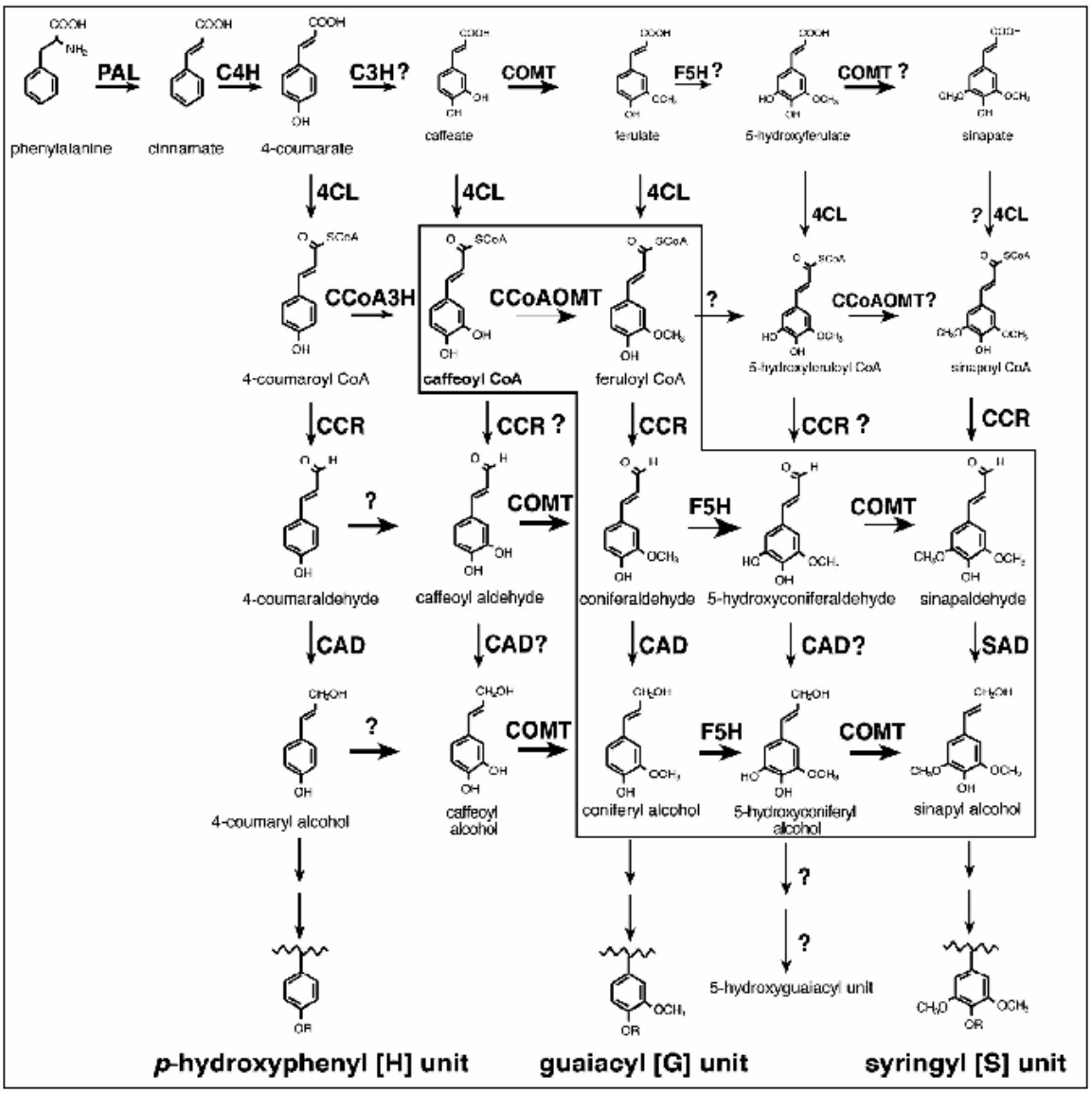


Figure 1.8 : Voie de biosynthèse de la lignine

1.5.3 Les facteurs de transcription LIM

La famille des LIM est formée de gènes ayant un rôle important dans les processus cellulaire chez les eucaryotes, dont la transcription et l'organisation du cytosquelette d'actine. Ils sont connus principalement chez les animaux mais on les retrouve aussi chez les plantes. Ils sont caractérisés par la présence de un ou plusieurs motifs doigts de zinc, qu'on appelle les domaines LIM, qui possèdent une fonction dans l'interaction protéine-protéine ou ADN-protéine. Basé sur la comparaison de leur séquence et sur l'association avec d'autres domaines fonctionnels, il a été proposé de catégoriser les LIM selon trois groupes majeurs. Le premier est formé des facteurs de transcription LIM-homéodomaine qui comportent deux domaines LIM à leur position N-terminale en plus d'un homéodomaine. Les LIM du deuxième groupe comportent un domaine kinase, en plus des deux domaines LIM. Les gènes provenant du troisième groupe contiennent uniquement deux domaines LIM.

Le premier LIM décrit chez les plantes est PLIM1 (originalement appelé SF3). Il fut isolé du pollen de tournesol (Baltz et al., 1992) et est structurellement parent avec les séquences animales LIM CRP et MLP. PLIM1 a été détecté dans la structure cytoplasmique à l'intérieur des microspores et dans la région corticale des grains de pollen mature où elle était présente en plus forte concentration dans les cônes de germination (Eliasson *et al.,* 2000). PLIM1 lie des acides nucléiques in vitro suggérant un rôle dans la transcription des gènes ou dans le transport des ARN messagers. Des séquences homologues ont été identifiées à partir de sa séquence protéique. WLIM1, identifié comme ayant une haute ressemble avec le domaine LIM du tournesol, participerait comme son homologue animal, à deux fonctions distinctes : une dans le cytoplasme et une dans le noyau (Mundel, 2000).

La protéine Ntlim1 isolée chez le tabac se lie spécifiquement à la séquence 'Pal-box' présente dans le promoteur de plusieurs gènes de la voie de biosynthèse de la lignine (Kawaoka *et al.*, (2000). Ntlim1 possède deux domaines LIM riches en cystéine et formant deux doigts de zinc. Il a été démontré que Ntlim1 se lie à la séquence 'Pal-box' et contrôle l'expression de certains gènes de la voie de biosynthèse de la lignine. La suppression de

Ntlim1 dans des plants de tabac transgéniques cause une réduction simultanée du niveau de transcrit de certains gènes de la voie de biosynthèse de la lignine et du niveau du contenu ligneux. Il a donc été proposé que des gènes de la famille des LIM constituent des régulateurs potentiels de la lignine chez les arbres toutefois ils n'ont pas encore été caractérisés chez ces derniers. Des travaux ont été entrepris afin de préciser ce rôle potentiel.

1.6 Cadre du projet et objectifs

Dans le cadre du projet Arborea, une plate-forme d'analyse de biopuces à été mise en place afin d'étudier la régulation des gènes à l'échelle du transcriptome. Ces expériences visent à déterminer le profil d'expression de milliers de gènes à la fois, dans différentes conditions. Au sein du projet Arborea, des expériences ont été entreprises sur l'épinette et le peuplier, notamment en lien avec la caractérisation de facteurs de transcription ayant un rôle dans la formation du bois.

Suites aux hybridations, on souhaite que le processus d'analyse permette un accès dynamique aux données ainsi qu'aux outils de la chaîne d'analyse. Pour de nombreux organismes, il est nécessaire de créer des biopuces sur mesure et de développer des outils informatiques spécifiques. Il est donc important de bâtir une chaîne d'analyse qui tienne compte de toutes les étapes nécessaires à l'analyse des biopuces. Il doit y avoir un suivi des données expérimentales, de la récolte des échantillons sur le terrain jusqu'à leur utilisation au laboratoire. Ce projet de mémoire visait à développer les outils informatiques et les algorithmes nécessaires à la mise en place d'une chaîne complète d'entreposage des informations du début à la fin des expériences, du champ à la plate-forme de transcriptomique. Les objectifs spécifiques de ce projet étaient :

1 – Faire l'inventaire des besoins et des outils. Identifier les composantes et la chaîne de traitement et d'analyse (discuté au chapitre 1). BASE a été sélectionné pour l'entreposage des donnes de biopuces, TM4 et Bioconductor pour le traitement et l'analyse des données de biopuces.

2 – Développement d'un LIMS permettant la gestion des échantillons en amont des analyses de biopuces. Le résultat de ce travail a permis l'implantation la base de données d'échantillons SLIMS (chapitre 2).

3 – Développer des algorithmes permettant de bâtir des annotations spécifiques aux sondes présentes sur nos biopuces (chapitre 3).

4 – Analyser une expérience de biopuces (chapitre 3). La chaîne d'analyse sera utilisée pour interpréter l'expérience sur les transgéniques surexprimant un facteur de transcription LIM afin d'identifier des gènes différentiellement exprimés.

2.0 Développement d'outils informatiques et leur intégration dans la chaîne d'analyse : SLIMS

Le choix de développer SLIMS au lieu de l'utilisation d'un logiciel déjà existant est discuté dans l'article que j'ai rédigé avec l'aide de Nathalie Pavy, John Mackay et François Larochelle. SLIMS fut la réponse à un manque de systèmes de gestion d'échantillons adaptés aux expériences réalisées en biologie expérimentale. Aucun outil disponible ne permettait de générer et de gérer facilement les échantillons reliés aux expériences réalisées dans le projet Arborea. SLIMS est un outil permettant de définir son plan expérimental et de créer automatiquement tous les échantillons possibles découlant de ce plan. Il est par la suite possible de préciser des traitements supplémentaires qu'on applique aux échantillons. Pour chaque étape de manipulation des données, SLIMS mémorise quand et qui a fait les manipulations permettant ainsi de faire un suivi des manipulations et des origines de celles-ci.

J'ai réalisé la plus grande partie (90%) de la conception et l'analyse de la base de données. Nathalie Pavy, François Larochelle et Nicolas Juge (10%) ont, par la suite, participé à l'amélioration du logiciel. J'ai aussi réalisé toute la programmation, l'implantation et le support nécessaire au bon fonctionnement du logiciel SLIMS. La mise en place et le support du serveur sur lequel SLIMS est installé a été faite par Stéphane Larose.

J'ai rédigé toute la documentation nécessaire à l'utilisation, l'administration et l'installation de SLIMS. Par la suite, des corrections ont été apportées à celles-ci par les co-auteurs. Le manuscrit soumis a été réalisé en grande partie par Nathalie Pavy et moi. Nathalie Pavy a contribué à toutes les parties du manuscrit plus particulièrement à la section introduction, résultats et conclusion tandis que j'ai contribué principalement à la section implémentation et résultats.

Management of biological experiments with SLIMS: example of use for a genomics project

Hugo BERUBE[1], François LAROCHELLE[2], John MACKAY[1] and Nathalie PAVY[1§]

1 Project Arborea, Laval University, Québec, Canada, G1K 7P4. www.arborea.ulaval.ca/en

2 Bioinformatics Center, Laval University, Québec, Canada G1K 7P4.

[§]Corresponding author

Email addresses:

HB: hugo.berube@rsvs.ulaval.ca

FL: flarochelle@bioinfo.ulaval.ca

JM : john.mackay@rsvs.ulaval.ca

NP: nathalie.pavy@rsvs.ulaval.ca

Phone : 418- 656-2131

Fax : 418-656-7493

Abstract

Background: In human genomics, research projects often rely on existing tissue banks linked to clinical databases containing extensive, patient data, disease status and tissue

samples. At this time, there is a lack of publicly available databasing tools that meet the sample and data management needs for experimentation with laboratory animals, plants, or cell cultures, for gene expression, proteomic and metabolomic profiling in response to diverse biotic or abiotic factors under controlled conditions. We have developed a user friendly database system to share data that should effectively facilitate the management, tracking and mining of information generated by such experiments in multi-laboratory collaborations.

Results: SLIMS (Sample Laboratory Information Management System) is a web-based relational database facilitating the management of experiments leading to the massive production of samples. It was designed to manage both complex multifactorial experiments and downstream processes applied to the samples. To keep track of experimental procedures, it stores protocols, biological metadata, samples derived from a tissue collection and it automatically generates sample identifiers. Thus, data pertaining to hundreds of samples can be simultaneously uploaded in the database with a minimum manual intervention. Furthermore, each annotation related to the samples can be directly transferred in the BioArray Software Environment (BASE), an open source tool handling microarray data in an SQL database.

Conclusion: SLIMS is a multi-user system designed to store information about experiments and samples. It has been intensively used to handle samples generated for gene expression analyses. Its simple design will make it attractive for those looking for an easy to customize LIMS. SLIMS is available for testing on the SLIMS project web page and for download at http://www.arborea.ulaval.ca/en/slims/.

Background

In biology, dose response and time course experiments, as well as factorial designs are frequently used to investigate the effect of multiple factors simultaneously. Large sets of samples are often generated in experiments testing physiological or cellular hypotheses, where experimental subjects are dissected into several tissues or when several cell types are harvested. Experimenters may also wish to apply diverse biochemical assays or molecular analyses, like gene expression, thus further increasing the complexity of sample and data management. Finally, due to the magnitude and complexity of sample information, data management is paramount in large genomics programs where multiple users, sometimes from several institutions, wish to work in a coordinated manner. In our project, controlled experiments are conducted to analyse the effect of transgene expression, as well as environmental and nutritional factors on tree growth and development. On a yearly basis, we isolate thousands of tissue samples that enter into an analysis pipeline involving different technology platforms for transcript profiling and a variety of phenotypic analyses. A software was developed to handle the data pertaining the production, processing and analysis of the samples, and became essential to organize the overall project.

SLIMS, a Sample Laboratory Information Management System includes a database appropriate for storing experiments and samples through a easy to understand web interface. The stored data include the methods used to produce the tissue samples, to harvest and process them. The design and functions of SLIMS are generic and adaptable, making SLIMS easy to customize to nearly any biological analyses. Two types of information describing experiments are stored in SLIMS: designs ranging from basic single factor experiments to complex multifactorial designs conducted to produce the tissue samples, and simple processes used in molecular biology protocols like RNA extraction.

This report presents the use of SLIMS to handle samples for gene expression analyses. To facilitate the flow of sample information into a microarray analysis platform, we developed a link between SLIMS and the Bio Array Software Environment [BASE; 1]. BASE is a system for the management of data generated in microarray analyses, which is reaching widespread use in the scientific community. Linking SLIMS and BASE is a solution to easily manage the upstream biological metadata associated with the samples used as targets in microarray hybridizations. Sample data already available in SLIMS can be directly uploaded into BASE. Complete and standardized information necessary to reproduce and describe the experiments are thus made available all along the sample production and analysis process.

Implementation

SLIMS is a relational database implemented using MySQL [2] version 3.23.49 or greater. The web interfaces for database query and data upload are coded in PHP [3] version 4.3.8 and embedded in HTML pages. Forms are made available through an Apache [4] web server version 1.3.26. SLIMS supports multiple users and permissions; it handles large databases and runs on UNIX and Windows systems. Control of user access to the database utilizes the MySQL authentication system to increase the protection level and to ensure control over data entered into SLIMS.

SLIMS database

To design a tool able to handle data from numerous experiments such as those encountered in experimental biology, we developed a database that models the flow of the data during the experiments. The SLIMS database has eleven tables (Figure 1). The "SAMPLE" table includes the following data: origin of the biological sample, description, treatments, storage location and quantity of collected tissue. Each item stored in the tables is identified with a unique auto-incrementing ID that is used to create links between the tables. Therefore, the logical links between the tables (i.e. the database integrity) are maintained regardless of the user's modifications of the data. All physical

activities or objects associated with the experiment are stored in their appropriate tables while maintaining a logical link to their corresponding experiments through foreign keys. For each account created through SLIMS by the superuser, an account with the same login and password is created in MySQL. Control of user access to the database by MySQL instead of PHP increases the protection level and the control over data entered into SLIMS. It also allows for easier and more accurate methods of data recovery.

SLIMS interface

Entering and modifying data in the database is achieved through a web interface that is easy to use and to understand. For the administrator, one menu specifically addresses the configuration requirements. For the users, seven menus enable to manipulate experiments, biomaterials, samples, processed samples, protocols, species, and tissue types, respectively. Each menu has the same structure, enabling to add, browse, edit, search, and modify the data. It is also possible to save the data in a comma-separated format in order to export them to spreadsheet software. The standardization of all menus is made to keep the interface simplicity. For the experiment type and the biological material, the descriptive terms showed in forms are conform to the MIAME standards [6] with the vocabulary extracted from the MGED ontology [5].

For each table of the database, at least one dynamically created form displays the fields required for data entry. Each field from a given database table is represented on the corresponding form according to its data type (numerical, string or text field) and relation with other tables (primary or foreign key). For example, foreign keys consist of drop-menus with all the items to ensure the selection of an already existing item. Consequently, modifications that are made to the tables are followed by the corresponding change in the forms. Therefore, modifications to the PHP code are only required to reflect table relation changes.

Installation

SLIMS can be installed either on the user machine or on a server. Access is controlled by two kinds of logins: the superuser login (i.e. administrator) and the user logins. The

administrator enters and curates the information common to all the experiments that are needed to operate SLIMS locally. These data include the user's accounts, the laboratory personnel, locations where experiments are conducted and samples are stored, the species and genotypes, and the tissues available for the users. Once all members of a group of experimenters from a laboratory or project have been entered through the AdminOptions/PeopleMenu, operators may easily be tracked for process or experiment. Tracking of experimenters is crucial in large programs where one experiment may be conducted by several people and data may be managed by someone else.

Results and discussion

SLIMS was designed to store background information pertaining to the experimental design, procedures, biological materials and sample production that is as comprehensive as possible. Experimental data that are stored in SLIMS include: experiment type, protocols, experiment design and conditions (e.g. controlled environment conditions during the experiment) observations made on experimental subjects during samples production, and quality control procedures. An example of a typical factorial experiment managed in SLIMS is illustrated on Figure 2; it is comprised of 9 treatments applied to tree seedlings from which three tissues were collected. It generated a total of 216 tissue samples including technical and biological replicates

Management of experimental procedures

Whenever the samples are used for microarray studies, experiments need to be classified in one of the experiment type from the MIAME list. In SLIMS, this classification is handled through the so called "Experiment type" field consisting of a drop-down list of controlled terms issued by the MGED society [5]. This list lies in the "EXPERIMENT_TYPE" table and may be updated by the administrator. The description of procedures required to accurately reproduce the experiments are stored in the "PROTOCOL" table. Tracking of each step during the sample production process and recording of relevant information are easily achieved through data entries into optional fields labelled "comments" and "details". The basic features of SLIMS thus help to

centralize and standardize protocols within a group of users, in addition to providing a complete and uniform information registry of experimental procedures.

Stepwise data entry

The data flow in SLIMS reflects the way biological samples are produced and used in diverse analyses. We have implemented a general procedure to enter data into the SLIMS database. First, the experiment is defined to enable a comprehensive definition and identification of the biological units that are handled. Second, the biological material used in the experiment is defined and all related information is entered. Third, based upon experimental design information provided by the user, the system automatically generates sample identifiers that are uploaded in the database. Finally, these tissue samples can be sub-divided into as many "processed samples" as desired. An option is provided to transfer the samples and their associated annotations into BASE for later RNA extraction and microarray hybridization.

Automatic assignment of sample identifiers

The core information needed to define an experiment is comprised of the biomaterial description and the treatments. The biomaterial is identified by the species name and a genotype descriptor. The treatments are defined based upon the number of factors tested in a given experiment, the number of levels of each factor. The number of technical and biological replicates and the types of collected tissues must also be entered. For each collected sample, a unique identifier is automatically assigned by the software, incorporating the experiment name, treatment code, replicate and tissue information (Figure 3). The user is prompted to enter experimental design information into SLIMS in a stepwise preset process (Figure 4a), ultimately resulting in the automatic display of a table that reflects the design (Figure 4c). All the combinations of the treatments, the tissues and the biological replicates are presented in the table thus generated. From this point, the entry of sample data is completed in only two steps: the user must specify the identifier of each experimental replicate and validate the identity of each sample that is defined by default by the system. The program then displays the data on a confirmation page before it is uploaded into the database. Each experimental detail is only entered

once, no matter how many samples it relates to (Figure 4b). This feature helps to avoid errors, and enables the upload of hundreds of samples with minimum manual intervention. After their upload in the database, data can be edited, modified, or deleted if the user has the appropriate access rights. The process enables some curation of the database without the intervention of its administrator. Thus, an advantage of SLIMS is its ability to rapidly handle complex experiments involving several factors and large numbers of samples.

Traceability of tissues and processed samples

For experimental purposes, collected tissues may be divided into several sub-samples or aliquots for use in multiple analyses, involving several techniques. For example, tissues sub-samples may be subjected to protein extraction, RNA extractions or simply stored in different facilities. Thus, a processed sample is derived from a sample previously entered into the database through processing with a defined protocol. The source sample can either be an unprocessed original sample, or processed sample to which a new procedure was applied. For example, to study gene expression, tissues may be ground and RNA extracted, to ultimately be analyzed using RT-PCR, northern hybridization or microarrays. Details related to each process that is applied to the samples, and the remaining quantities of the biomaterial are transferred to the "PROCESSED_SAMPLE" table. SLIMS keeps track of every process applied to the samples, along with the date and the manipulator, thus allowing to trace each sample along the analytical pipeline.

Simple transfer of sample data and annotations into BASE

SLIMS provides an option to transfer data into the BioArray Software Environment (BASE), which is a system for microarray data management and analysis. In SLIMS, terms were chosen to comply with the BASE nomenclature, thus facilitating the coordinated use of both tools. Sample data and annotations are transferred directly into the BASE database, which must also be installed locally. If the SLIMS user is not registered as a BASE user, his SLIMS login and password are used to create a new user account in BASE. The following data are written into BASE database tables: species data (genus, species, and genotype), annotation types (including factors, tissues, and

biological replicates), samples and sample annotations (for each sample, the level of each factor, the tissue name, and the biological unit name are uploaded in the "sample annotation" table from BASE). The sample data are uploaded from the SLIMS "SAMPLE" table into three BASE tables called "Sample", "SampleAnnotation" and "SampleAnnotationType" tables. The relational structure of the BASE tables is respected during this transfer process. Sample data are transferred into BASE only when the appropriate option is chosen; the link between SLIMS and BASE is temporary. Modifications entered into SLIMS do not automatically affect the same sample in BASE although it is possible to overwrite it by executing another transfer. All existing BASE samples with the same name will be overwritten during the upload, thus updates made to sample data in SLIMS may be carried forward to BASE through the same transfer mechanism.

Scalability and Testing

SLIMS currently operates on a pentiumIII 233 MHz on SQL and Linux, and can handle up to thirty thousand samples and then successfully transferred into BASE. It has been installed and successfully tested on windows-based (XP and 98) computers using EasyPHP software, thus fully using the power and flexibility of the dynamic language PHP for efficient use of the databases. Although the execution speed of SLIMS is data-dependant, it is safe to say that SLIMS can run on a low end personal computer.

Conclusions

The goal of this project was to produce a tool to facilitate the management of biological experiments producing large sets of samples. To this end, a new software package called SLIMS has been developed; it uses a graphical interface to handle the data and relies on a MySQL database. Sample IDs are automatically generated to facilitate data entry and reduce the entry of erroneous data. When combined with the BASE microarray database, SLIMS enables to keep track of all the processes applied to a sample from its production to its use in microarray hybridizations. In this report, we illustrated the use of SLIMS in a genomics project. SLIMS should also be useful in other contexts like in agronomy or genetics. Its simple, light design (3 MB installation) and dynamic interface make it very

straightforward. SLIMS is written in PHP and has been tested on Unix and Windows. It is freely available for download under the terms of the GNU General Public License (GPL) at http://www.arborea.ulaval.ca/en/slims/ .

Availability and requirements

Project name : SLIMS (Sample Laboratory Information Management System)

Project home page : http://www.arborea.ulaval.ca/en/slims/

Operating systems : UNIX ,WindowsXP, Windows98

Programming language : MySQL and PHP

Other requirements : Apache server

License : GNU General Public License

Restrictions to use by non-academics : none

List of abbreviations

BASE. BioArray Software Environment

SLIMS. Sample Laboratory Information Management System

SQL. Structured Query Language

Author's contributions

HB developed and tested the system as well as the SLIMS project web site. FL and NP contributed to the design of the database and interface, to the writing of the documentations. FL contributed to the connectivity between SLIMS and BASE. All authors contributed to the manuscript.

Acknowledgements

This work was supported by funding from GénomeQuébec and GénomeCanada to JM for the Arborea project.

References

1. Saal, L.H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A., Peterson, C. (2002) **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** Genome Biol. , 3(8):SOFTWARE0003.
2. http://www.mysql.com
3. http://www.php.net
4. http://www.apache.org
5. http://www.mged.org/
6. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M. (2001) **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** Nat Genet., 29(4):365-71.

Figure legends

Figure 2.1 : Database structure. The fingerprint variable contains the ID of the user who created the data in the database whereas the ID_PERSON represents the person who conducted the experiment in the laboratory. Each field starting with 'ID_' represents a foreign key.

Figure 2.2 : Example of a factorial experiment managed in SLIMS. The experiment was run with tree seedlings and compared 2 factors (dose of a nutrient applied to the plants, time after applications began), and used three levels for each factor, for a total of 9 treatments. For each treatment, there were 4 technical replicates each comprised of one seedling (the experimental unit) and 2 biological replications, achieved by replicating the entire experiment in two growth cabinets. Finally, 3 tissues (leaf, shoot, root). were

collected from each plant in the experiment. In this example, 216 samples are collected (= 3 Dose Levels x 3 Time Levels x 3 Tissues x 4 Technical Replicates x 2 Biological Replicates).

Figure 2.3: Sample naming. Experimental design information is used to create the sample ID that contains the experiment code, the treatment code, the technical replicate number, the tissue code and the biological replicate number. In this example, N stands for Needles.

Figure 2.4: Data upload in SLIMS and BASE from the experiment illustrated on Figure 2. Colors have been added to visualize the flow of the data at the different steps of their upload. blue : Experiment_code, red : biomaterial definition, green: level 1 of factor 1, yellow : level 1 of factor 2, pink: origin of the tissue sample **(a)** SLIMS interface to enter the experimental design **(b)** SLIMS interface to enter the core information about the experiment **(c)** SLIMS display generated based on data provided by the user **(d)** BASE screenshot illustrating the annotations of one sample from the treatment 1(Dose level =1 and Time level =1), from replicate number 1-1, from tissue = N. The nomenclature used for the sampleID is illustrated on Figure 3.

Figure 2.1.

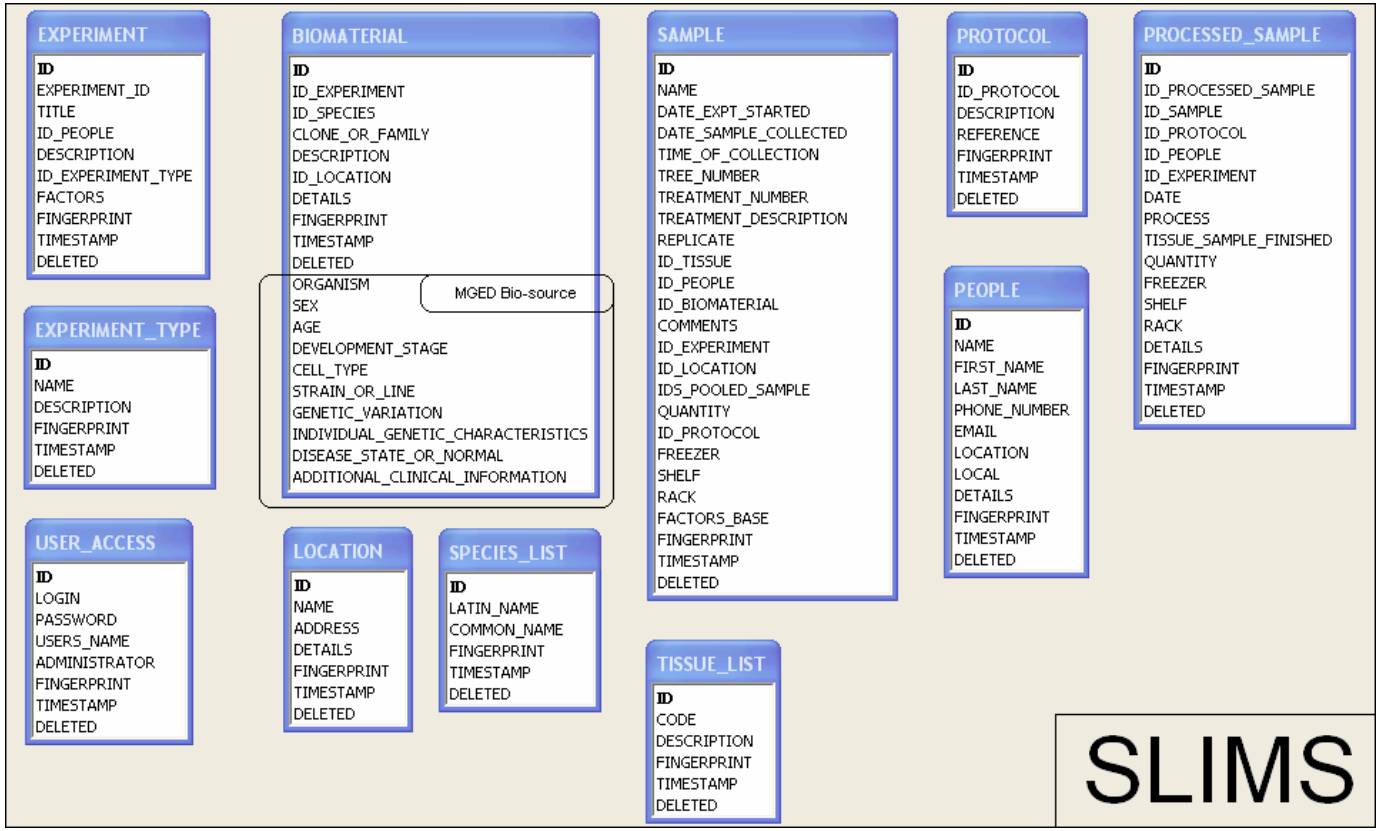

EXPERIMENT
ID
EXPERIMENT_ID
TITLE
ID_PEOPLE
DESCRIPTION
ID_EXPERIMENT_TYPE
FACTORS
FINGERPRINT
TIMESTAMP
DELETED
EXPERIMENT_TYPE
ID
NAME
DESCRIPTION
FINGERPRINT
TIMESTAMP
DELETED
USER_ACCESS
ID
LOGIN
PASSWORD
USERS_NAME
ADMINISTRATOR
FINGERPRINT
TIMESTAMP
DELETED
BIOMATERIAL
ID
ID_EXPERIMENT
ID_SPECIES
CLONE_OR_FAMILY
DESCRIPTION
ID_LOCATION
DETAILS
FINGERPRINT
TIMESTAMP
DELETED
ORGANISM
SEX
AGE
DEVELOPMENT_STAGE
CELL_TYPE
STRAIN_OR_LINE
GENETIC_VARIATION
INDIVIDUAL_GENETIC_CHARACTERISTICS
DISEASE_STATE_OR_NORMAL
ADDITIONAL_CLINICAL_INFORMATION
MGED Bio-source
LOCATION
ID
NAME
ADDRESS
DETAILS
FINGERPRINT
TIMESTAMP
DELETED
SPECIES_LIST
ID
LATIN_NAME
COMMON_NAME
FINGERPRINT
TIMESTAMP
DELETED
SAMPLE
ID
NAME
DATE_EXPT_STARTED
DATE_SAMPLE_COLLECTED
TIME_OF_COLLECTION
TREE_NUMBER
TREATMENT_NUMBER
TREATMENT_DESCRIPTION
REPLICATE
ID_TISSUE
ID_PEOPLE
ID_BIOMATERIAL
COMMENTS
ID_EXPERIMENT
ID_LOCATION
IDS_POOLED_SAMPLE
QUANTITY
ID_PROTOCOL
FREEZER
SHELF
RACK
FACTORS_BASE
FINGERPRINT
TIMESTAMP
DELETED
TISSUE_LIST
ID
CODE
DESCRIPTION
FINGERPRINT
TIMESTAMP
DELETED
PROTOCOL
ID
ID_PROTOCOL
DESCRIPTION
REFERENCE
FINGERPRINT
TIMESTAMP
DELETED
PEOPLE
ID
NAME
FIRST_NAME
LAST_NAME
PHONE_NUMBER
EMAIL
LOCATION
LOCAL
DETAILS
FINGERPRINT
TIMESTAMP
DELETED
PROCESSED_SAMPLE
ID
ID_PROCESSED_SAMPLE
ID_SAMPLE
ID_PROTOCOL
ID_PEOPLE
ID_EXPERIMENT
DATE
PROCESS
TISSUE_SAMPLE_FINISHED
QUANTITY
FREEZER
SHELF
RACK
DETAILS
FINGERPRINT
TIMESTAMP
DELETED
SLIMS

Figure 2.2.

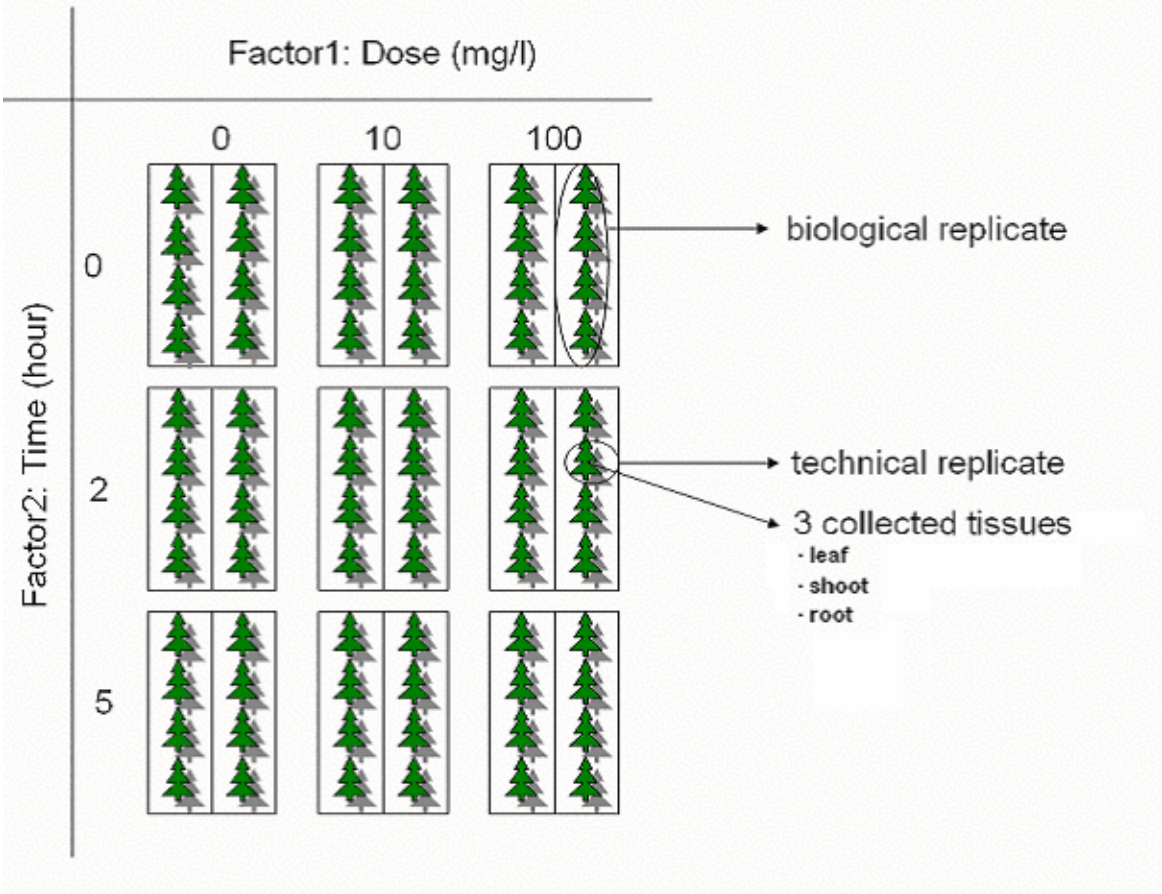


Figure 2.3.

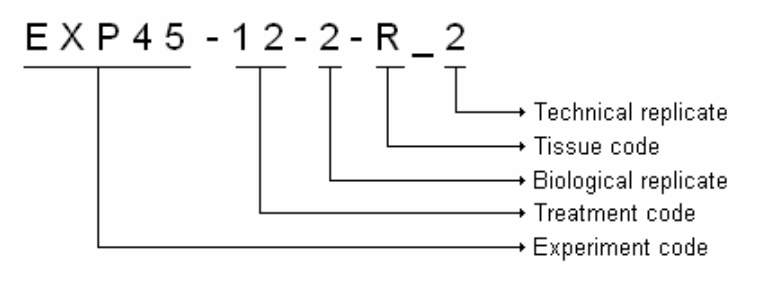


Figure 2.4a.

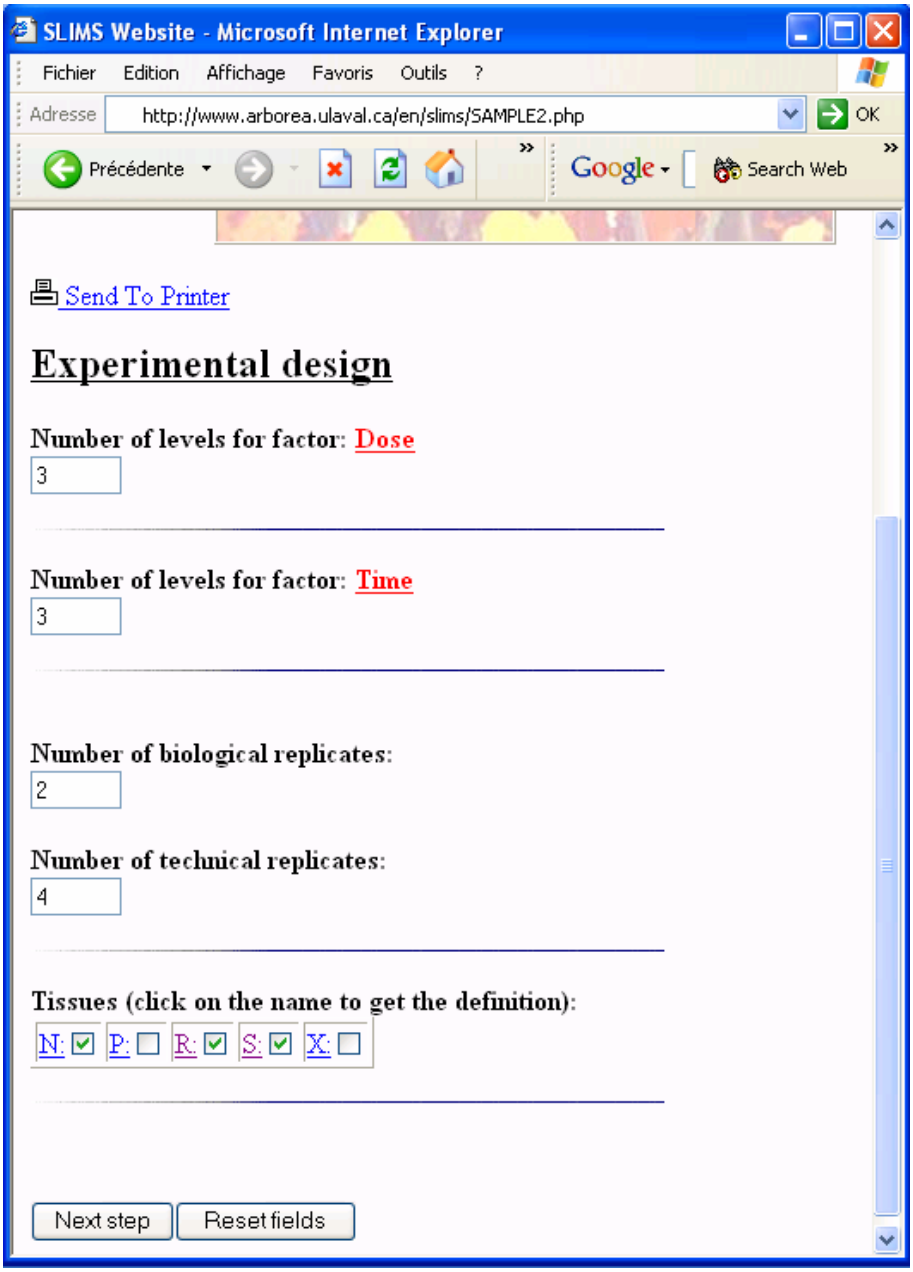

SLIMS Website - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Adresse http://www.arborea.ulaval.ca/en/slims/SAMPLE2.php
OK
Précédente
Google
Search Web
Send To Printer
Experimental design
Number of levels for factor: Dose
3
Number of levels for factor: Time
3
Number of biological replicates:
2
Number of technical replicates:
4
Tissues (click on the name to get the definition):
N: P: R: S: X:
Next step
Reset fields

Figure 2.4b.

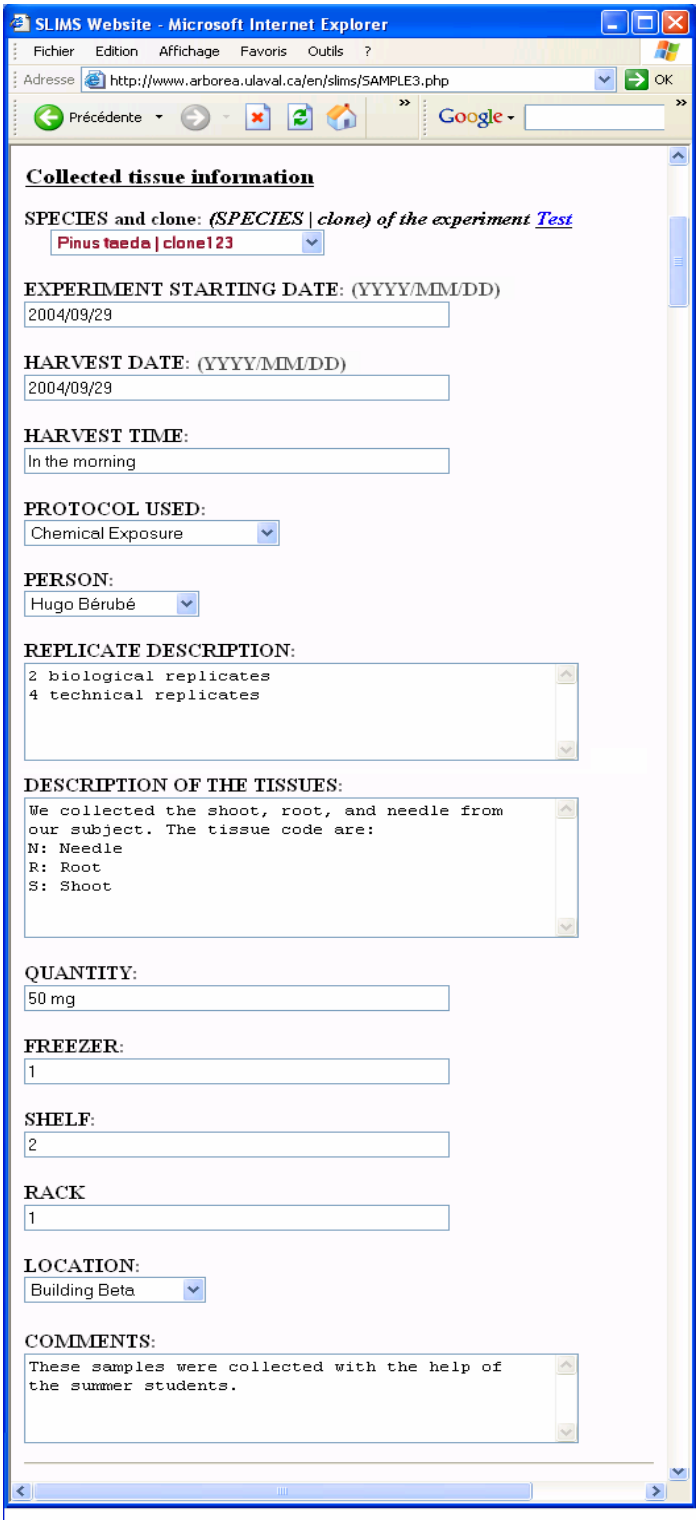

SLIMS Website - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Adresse http://www.arborea.ulaval.ca/en/slims/SAMPLE3.php
OK
Précédente
Google
Collected tissue information
SPECIES and clone: (SPECIES | clone) of the experiment Test
Pinus taeda | clone123
EXPERIMENT STARTING DATE: (YYYY/MM/DD)
2004/09/29
HARVEST DATE: (YYYY/MM/DD)
2004/09/29
HARVEST TIME:
In the morning
PROTOCOL USED:
Chemical Exposure
PERSON:
Hugo Bérubé
REPLICATE DESCRIPTION:
2 biological replicates
4 technical replicates
DESCRIPTION OF THE TISSUES:
We collected the shoot, root, and needle from our subject. The tissue code are:
N: Needle
R: Root
S: Shoot
QUANTITY:
50 mg
FREEZER:
1
SHELF:
2
RACK
1
LOCATION:
Building Beta
COMMENTS:
These samples were collected with the help of the summer students.

Figure 2.4c.

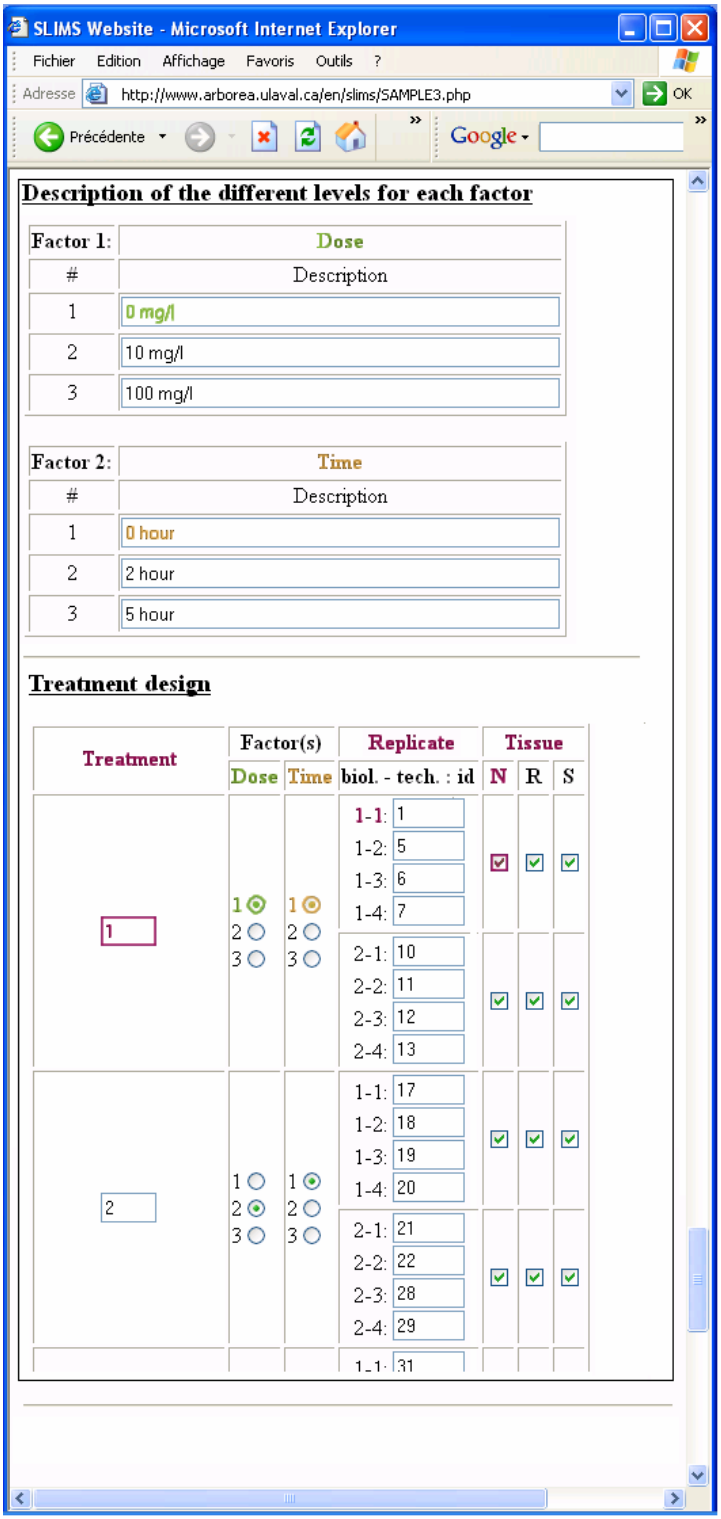
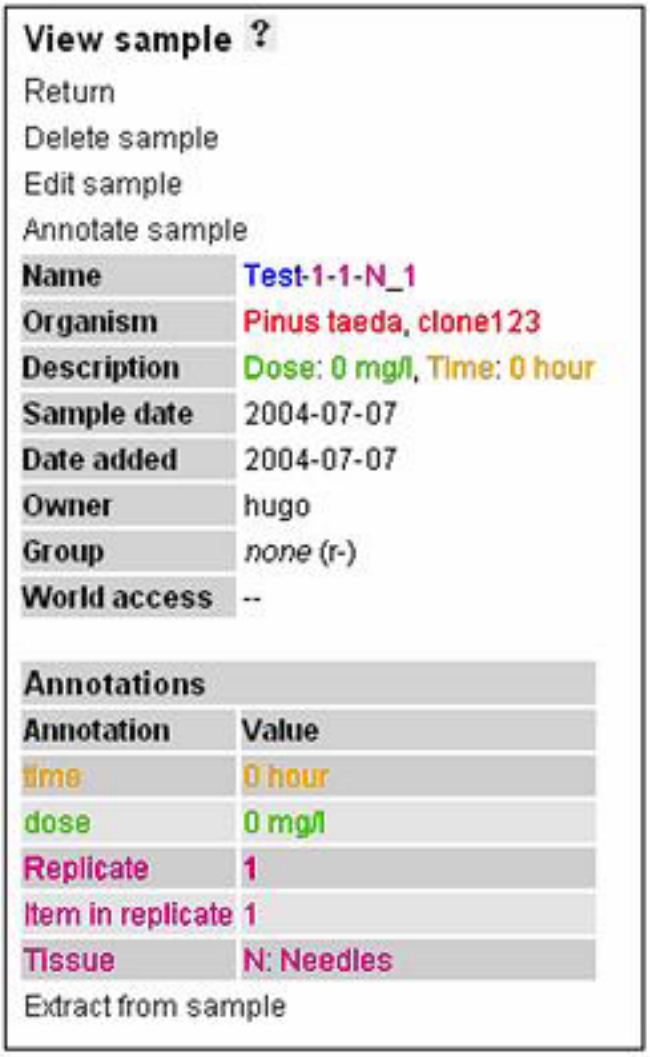

SLIMS Website - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Adresse http://www.arborea.ulaval.ca/en/slims/SAMPLE3.php
OK
Précédente
Google
Description of the different levels for each factor
Factor 1: Dose
Description
1 0 mg/l
2 10 mg/l
3 100 mg/l
Factor 2: Time
Description
1 0 hour
2 2 hour
3 5 hour
Treatment design
Treatment
Factor(s)
Dose Time
Replicate
biol. - tech. : id
Tissue
N R S
1
1-1: 1
1-2: 5
1-3: 6
1-4: 7
2-1: 10
2-2: 11
2-3: 12
2-4: 13
2
1-1: 17
1-2: 18
1-3: 19
1-4: 20
2-1: 21
2-2: 22
2-3: 28
2-4: 29
1-1: 31

Figure 2.4d.



View sample ?

Return

Delete sample

Edit sample

Annotate sample

Name	Test-1-1-N_1
Organism	Pinus taeda, clone123
Description	Dose: 0 mg/l, Time: 0 hour
Sample date	2004-07-07
Date added	2004-07-07
Owner	hugo
Group	*none* (r-)
World access	--

Annotations	
Annotation	**Value**
time	0 hour
dose	0 mg/l
Replicate	1
Item in replicate	1
Tissue	N: Needles

Extract from sample

3.0 Analyse par biopuces d'épinettes transgéniques surexprimant un facteur de transcription LIM

3.1 Introduction

Plusieurs études de l'expression des gènes ont été réalisées en lien avec la formation bois, soit dans le but mieux caractériser les mécanismes cellulaires sous-jacents ou encore afin d'identifier des marqueurs potentiellement reliés à certaines caractéristiques désirables du bois (Whetten *et al*., 2001). Certaines de ces études ont été réalisées avec des approches génomique à l'aide d'hybridation contre des puces à ADN ou des membranes contenant quelques centaines de gènes, toutefois les études menées jusqu'à maintenant ont été strictement descriptives (Hertzberg *et al.,* 2003; Egertsdottter *et al*., 2004). Une des utilisations prometteuses des biopuces permet d'identifier les loci qui conditionnent le niveau d'expression pour un grand nombre de gènes par l'analyse d'expression au sein d'une famille en ségrégation faisant l'objet de cartographie génétique (Kirst *et al.,* 2004). Une approche qui vise à mettre en évidence les mécanismes sous jacents à l'expression des gènes est d'entreprendre l'étude fonctionnelle des facteurs de transcription, comme les LIM. Les études de gain de fonction ou de perte de fonction permettent d'une part de circonscrire le rôle des facteurs de transcription, mais aussi d'identifier leurs cibles potentielles et mettre en lumière des groupes de gènes co-régulés.

La famille des LIM est formée de gènes ayant un rôle important dans les processus cellulaires chez les eucaryotes, dont la transcription et l'organisation du cytosquelette d'actine (Dawid *et al.,* 1995) . Il a été démontré que la suppression d'un des membres de cette famille multigénique, soit Ntlim1 cause une réduction simultanée du niveau de transcrit de certains gènes de la voie de biosynthèse de la lignine et du contenu ligneux dans des plants de tabac transgéniques (Kawaoka *et al.,* 2000). Deux autres séquences, possédant un niveau de similarité élevé au gène NtLIM1 ont aussi été identifiés mais semblent posséder des rôles fort différents. PLIM1 a été détecté dans la structure cytoplasmique à l'intérieur des microspores et dans la région corticale des grains de

pollen mature où elle était présente en plus forte concentration dans les cônes de germination (Baltz *et al.,* 1999). WLIM1 à une expression plus ubiquitaire et comme les plus proches homologues animales, participerait à deux fonctions distinctes, soient au niveau du cytoplasme et du noyau (Mundel *et al.*, 2000). Nous avons donc émis l'hypothèse qu'un des gènes de la famille des LIM constitue un régulateur potentiel de la lignine chez les conifères, un groupe taxonomique fort distant des plantes comme Arabidopsis et le tabac chez lesquelles les LIMs ont été étudiés précédemment.

Plusieurs études ont mis en évidence de nombreuses séquences de facteurs de transcription chez les conifères (Kirst *et al.,* 2001 ; Pavy *et al.,* 2004 et 2005) caractérisés certaines familles comme les Knox (Guillet-Claude *et al.*, 2004), analysé leur expression (Ingouff *et al.*; 2003), toutefois seulement quelques facteurs de transcription ont été étudier au niveau fonctionnel chez les chez ces espèces. Des études de réalisées à l'aide protéines MYBs recombinantes ont mis en évidence leur activité de régulation transcriptionnnelle (Xue *et al., 2003*). La surexpression du gène ptMYB4 du pin, réalisée dans des plantes de tabac transgénique, a montré qu'il agit comme régulateur positif de la synthèse de la lignine (Patzlaff *et al.*, 2003). Aucune étude n'est disponible démontrant la perte ou le gain de fonction dans un conifère, donc toutes les études réalisées jusqu'à maintenant utilise des systèmes d'expresssions hétérologues.

L'étude de deux LIM (LIM1 et LIM2) a été entreprise chez l'épinette et le pin. La première étape a été d'isoler et de caractériser les séquences des gènes à partir des banques d'ESTs de pins (*Pinus taeda*). Des séquences orthologues ont aussi été isolées chez l'épinette (*Picea glauca*). Les gènes LIM1 et LIM2 sont les séquences du pin et de l'épinette qui montrent le plus haut niveau de similarité avec le gène NtLIM1, mais sont aussi fort similaires aux gènes PLIM et WLIM. Chez l'épinette, ces gènes sont exprimés dans plusieurs tissus dont le xylème en différentiation. Des lignées transgéniques surexprimant ces gènes ont été produites et des plantules d'épinette ont été régénérées, puis cultivés en serre. L'analyse transcriptomique des transgéniques LIM2 est présentée dans ce chapitre. Un des objectifs centraux de l'analyse était de déterminer si ce gène agit comme régulateur des gènes de la synthèse de la lignine. Nous visions aussi

à évaluer si la surexpression de ce gène chez l'épinette constituerait une approche favorable qui permettrait de mettre en évidence son rôle.

3.2 Matériel et Méthode

3.2.1 Matériel végétal

L'expérience ayant comme sujet les lignées transgéniques surexprimant ptLIM2 vise à identifier les gènes différentiellement exprimés chez des plants transgéniques surexprimant ce gène. Quatre lignées transgéniques ont été analysées et des prélèvements ont été effectués à trois dates pour chacune des lignées : 02, 04, 08, 21. Chaque prélèvement a été considéré comme un réplicat biologique pour les fins de cette étude. La tige entière, les aiguilles et les bouts de pousses sont récoltés (Figure 3.1). Pour chacun des prélèvements, dix arbres en santé sont choisis au hasard et mélangés. Lors des prélèvements les plants avaient une hauteur variant entre 10 et 18 cm; ils approchaient la fin de leur premier cycle annuel de croissance mais n'étaient pas encore en dormance.

Nous considérons que les plants sont à des stades de développement semblables. Les témoins sont des plants transformés un vecteur vide (pCambia) Les hybridations ont toutes été constituées d'un échantillon de transgéniques et de témoins, récoltés à la même date.

3.2.2 Extraction des ARN, biopuces et hybridation

Le protocole d'extraction d'ARN choisi est le protocole « Protocole maxi extractions » (Chang *et al.,* 1993). Le protocole de marquage est basé sur le protocole d'Invitrongen « Superscript Plus Indirect cDNA Labeling System kit » (cat. # L1014-06). La biopuce d'épinette est composée de 9053 clones différents, en plus des 22 gènes candidats provenant de pinus taeda qui ont été utilisés pour générer les transgéniques. De plus, 27 contrôles différents font partie de la biopuce. Les protocoles d'hybridation et de lavage sont fournis en annexe (annexe 3A et 3B).

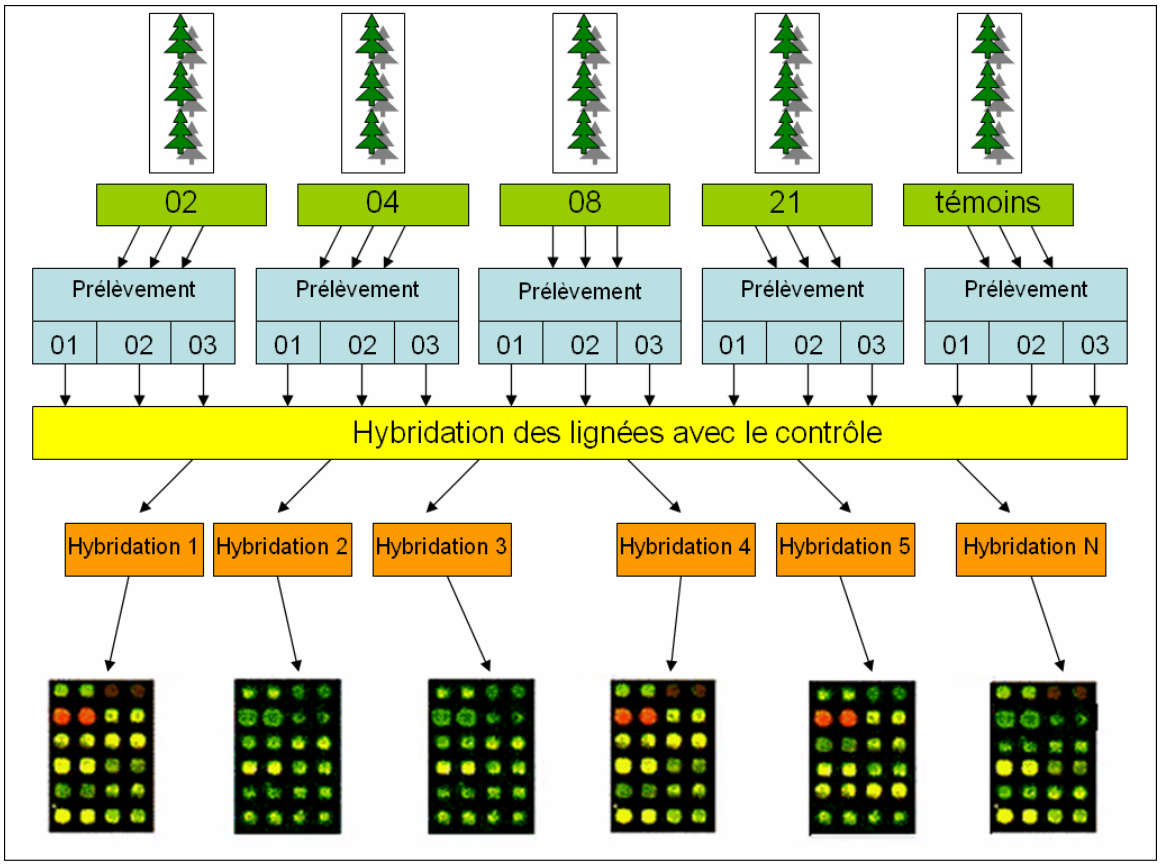


Figure 3.1 : Plan expérimental de l'expérience ptLIM2

3.2.3 Analyse qualité des hybridations, normalisation et analyse

L'analyse qualité des hybridations a été faite à partir de fichiers textes produits par le logiciel QuantArray qui permet de quantifier les intensités de chaque point sur la biopuce. C'est à partir de ces fichiers de données que nous avons analysé la qualité de l'hybridation pour déterminer s'il est nécessaire de reprendre les hybridations. Le but de l'analyse qualité d'image d'hybridation a été à d'identifier les jeux de données à retenir pour l'analyse d'après des imperfections au niveau de la lame ou anomalies au niveau de l'expression des gènes sur la lame.

Plusieurs méthodes de normalisation (printtiploess, Neural Net, Composite) ont été testées afin d'utiliser celle qui donnait la plus grande corrélation entre les réplicats à l'intérieur d'une lame et entre les réplicats d'hybridation. La normalisation des données est faite à l'aide de Bioconductor et MIDAS. MIDAS a été utilisé lors des analyses préliminaires pour normaliser les données, valider entre les réplicats et entre les hybridations inversées. Les analyses plus approfondies ont été menées avec limma et marray. Ces modules de Bioconductor ont permis de créer tous les différents graphiques d'analyse dont MA-plot, les graphiques des intensités et les calculs de corrélation. De plus, Bioconductor a permis de faire les normalisations plus complexes comme la normalisation composite (Yang et Dutoit) et la normalisation par les réseaux neuronaux (Tarca et al, 2005.) qui ne sont pas disponible avec MIDAS. La méthode de normalisation des données retenue en vue de l'analyse statistique s'est fait avec Bioconductor afin d'utiliser la méthode de normalisation composite. L'analyse statistique des données à été faite avec limma et MeV en utilisant la méthode SAM.

3.2.4 Annotation et analyses des résultats : Python et MeV

Nous avons créé des annotations propres aux gènes placés sur la biopuces d'épinette avec le script Python 'CreateAnnotaion.py'. L'annotation des séquences sur la puce a été faite à l'aide de scripts programmés en Perl et Python pour analyser et extraire l'information

contenue dans des fichiers de résultats BLAST. Ces scripts ont comme rôle de rechercher le meilleur résultat parmi les différents résultats blasts pour chaque clone situé sur la puce. Un fichier est aussi créé associant chaque identifiant du clone (MNID) aux meilleurs résultats des analyses BLAST. Trois bases de données ont été choisies pour la recherche d'annotations : (1)la base de données 'nr'; (2)la banque de données PGI5 (Pinus Gene Indice), maintenue par le TIGR Genome Institute, et contenant 35053 séquences consensus dont 16666 contigs et 18250 singletons; (3) la banque de donnée Pfam avec ses modèles de 7868 familles de protéines basées sur la banque de données Swissprot et SP-TrEMBL.

À l'aide de la commande BLASTALL, toutes les séquences présentes sur la puce épinette ont été comparées à chacune de ces banques de données. BLASTALL, exécuté avec les arguments '-v1 –b1', ne retourne que le meilleur résultats pour chaque comparaison. À partir des résultats un script externe (parseblast.pl) a été utilisé pour simplifier les résultats à une ligne par résultat blast. Un script, programmé avec l'approche orientée objet, a été développé en Python pour analyser ces résultats. Les objets du script sont réutilisables par d'autres personnes sans avoir à comprendre le fonctionnement interne de l'objet

J'ai développé l'algorithme 'CreateAnnotaion.py' qui consiste de deux objets distincts. Le premier objet représente le contenu d'une ligne de fichier retourné par le script Perl 'parseblast.py'. Cet objet contient l'information relative au résultat BLAST : le nom de la séquence de requête, le nom de la séquence cible, la position et la longueur, le score , le E-Value, etc. Il est possible de questionner l'objet pour chacune de ces informations. Il existe un objet pour chaque ligne du fichier lu. Le deuxième objet gère les objets 'ParsedBlastResult'. Il est possible d'interroger cet objet à propos des résultats BLAST. Cet objet contient une liste de tous les objets 'ParsedBlastResult' et les questionne au besoin. Cet objet peut retourner tous les résultats BLAST ayant comme espèce cible Arabidopsis par exemple, les résultats ayant un une probabilité (E-Value) inférieur à une telle valeur. Les valeurs retournées par une requête sont toujours sous la forme de l'objet d'origine. Il est donc possible d'interroger de nouveau le nouvel objet correspondant à

notre requête précédente. Ce processus itératif permet d'approfondir continuellement un jeu de donnée qu'on peut par la suite faire afficher à l'écran ou écrire dans un fichier.

En exécutant le script python, un objet 'BlastResult' est créé où pour chaque ligne lue du fichier blast un nouvel objet 'ParsedBlastResult' lui est attribué. Par la suite, des méthodes sont appelées sur l'objet 'BlastResult' afin de ne garder que les résultats ayant une probabilité (E-Value) plus petit que 1^{e}-08. Un fichier est ensuite créé où les résultats blast sont annotés à chacun des clones (MNID).

Une annotation spécifique à lignine et à la paroi cellulaire a aussi été créée pour les gènes associés à leur biogénèse. La paroi cellulaire est le site de l'activité métabolique majeure dans la formation du bois. C'est dans la paroi que sont déposés la cellulose, la lignine et les hemiculloses (Whetten *et al.*, (2001) ; Peter et Neale. 2004). Des séquences codant pour l'ensemble des protéines jouant un rôle dans la formation de la paroi cellulaire ont été obtenue à partir de la base de données « Cell Wall Navigator » (Girke *et al.*,2004; Pavy *et al.*, 2005). À partir de ces séquences, des homologues ont été identifiés sur la puce épinette et annoter selon le nom de famille de l'enzyme. Nous avons ainsi identifié sur le biopuce une centaine de clones homologue à des enzymes participant dans la voie de biosynthèse de la lignine. Il existe donc deux fichiers d'annotations complémentaires pour l'analyse de cette expérience : l'annotation de la puce en entier avec les bases de données publiques nr, TIGR Gene Index et Pfam et une seconde annotation ne contenant que les clones reliés à l'élaboration de la paroi cellulaire décrit par le projet « Cell Wall Navigator » (Girke *et al.,*2004).

3.3 Résultats

L'Assurance qualité et choix des hybridations à analyser

L'assurance qualité a été faite à partir des graphiques des intensités, des corrélations à l'intérieur et entre les hybridations et de MA-plots, dans le but d'identifier les hybridations de qualité satisfaisante pour l'analyse. La première étape de l'assurance qualité est l'analyse du graphique des intensités pour identifier les hybridations dont l'intensité est trop faible (Figure 3.2). Un nuage de points loin de la diagonale indique

qu'un canal est plus intense que l'autre tandis qu'un nuage de point peu étendue indique une intensité globale faible. La majorité des nuages de points des hybridations se situent autour de la diagonale zéro donc les données semblent être bien balancées entre les deux canaux (figure 3.2a). Pour certaines hybridations nous avons identifiés des intensités trop faibles (figure 3.2b), et avons repris les analyses d'images en ajustant la saturation sur le logiciel d'analyse. Les hybridations dont l'analyse d'images a été recommencée sont H10, H14, H13, H18, H20, H7, H16, H19, H15, H5 et H9.

Les MA-plots des vingt-quatre hybridations ont été analysés afin de déterminer la distribution des ratios d'intensité entre les échantillons. Cette analyse a été faite sur les données non normalisées et normalisées. L'hybridation quatre (H4) possède un nuage de points qui s'étend vers le bas plutôt que de se situer majoritairement autour du ratio zéro (Figure 3.3a). Pour un très grand nombre de points il n'y a de l'hybridation ou signal que dans un seul canal indiquant un problème au niveau du marquage. Les points ne sont pas distribués de manière homogène sur l'axe des A. Un nuage de points trop décallé d'un côté de l'axe des M ou un nuage autour de l'axe de 0 signifie un problème de marquage ou un problème lors de l'analyse d'image. Cette distribution hétérogène persiste malgré la normalisation (Figure 3.3b). Les MA-plots de données normalisés donnent généralement une meilleure indication de la qualité des données. Il a donc été décidé de reprendre l'hybridation H4.

La corrélation intra lame est calculée entre un point et son réplicat à l'intérieur d'une hybridation et fournie un indice de répétitivité. Les hybridations ayant des corrélations intra lame trop basse pour les données normalisées (<0.2) ont été analysées de nouveau. La majorité des hybridations dont l'analyse d'image a été recommencée (i.e. « rescans ») affichent des meilleures corrélations intra lame que l'analyse originale (voir Tableau 3.1). Malgré les reprises et les « rescans », certaines lames ont quand même des corrélations intra lame faibles donc ont été ignoré lors des étapes subséquentes d'analyse. La corrélation intra lame servi à comparer l'efficacité de plusieurs méthodes de normalisation (Tableau 3.1).

L'analyse des corrélations inter lame compare chaque point d'une hybridation avec le point correspondant d'une autre hybridation et calcule la corrélation moyenne entre tous les points de ces deux hybridations. Elle a permis d'identifier les hybridations qui sont les plus similaires de chacune des lignées. Nous avons retenu pour l'analyse les quatre meilleures hybridations pour chacune des lignées (voir Tableau 3.2) en se basant, en grande partie, sur les hybridations ayant les meilleures corrélations inter lame avec les autres hybridations de sa lignée. Bien que les meilleures hybridations n'ont pas toujours les signes attendus lors de l'analyse inter lame, ce sont les meilleures hybridations sur l'ensemble des critères d'évaluations (MA-plots, graphique des intensités, corrélation intra et inter lame).

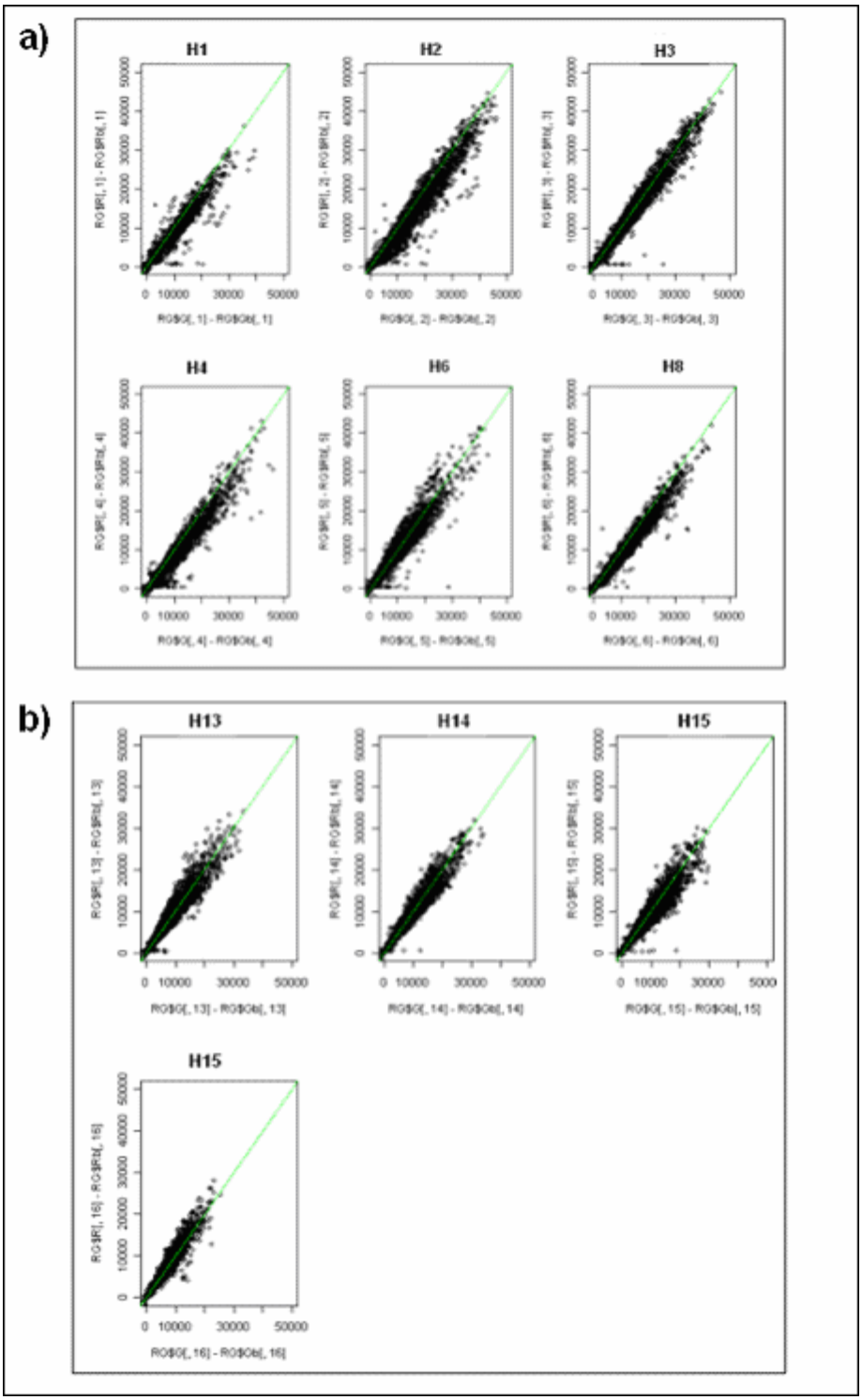


Figure 3.2 : Graphiques des intensités calculés sur les donnés brutes avec limma. A) Hybridations 1, 2, 3, 4, 6, 8 b) Hybridations 13, 14, 15

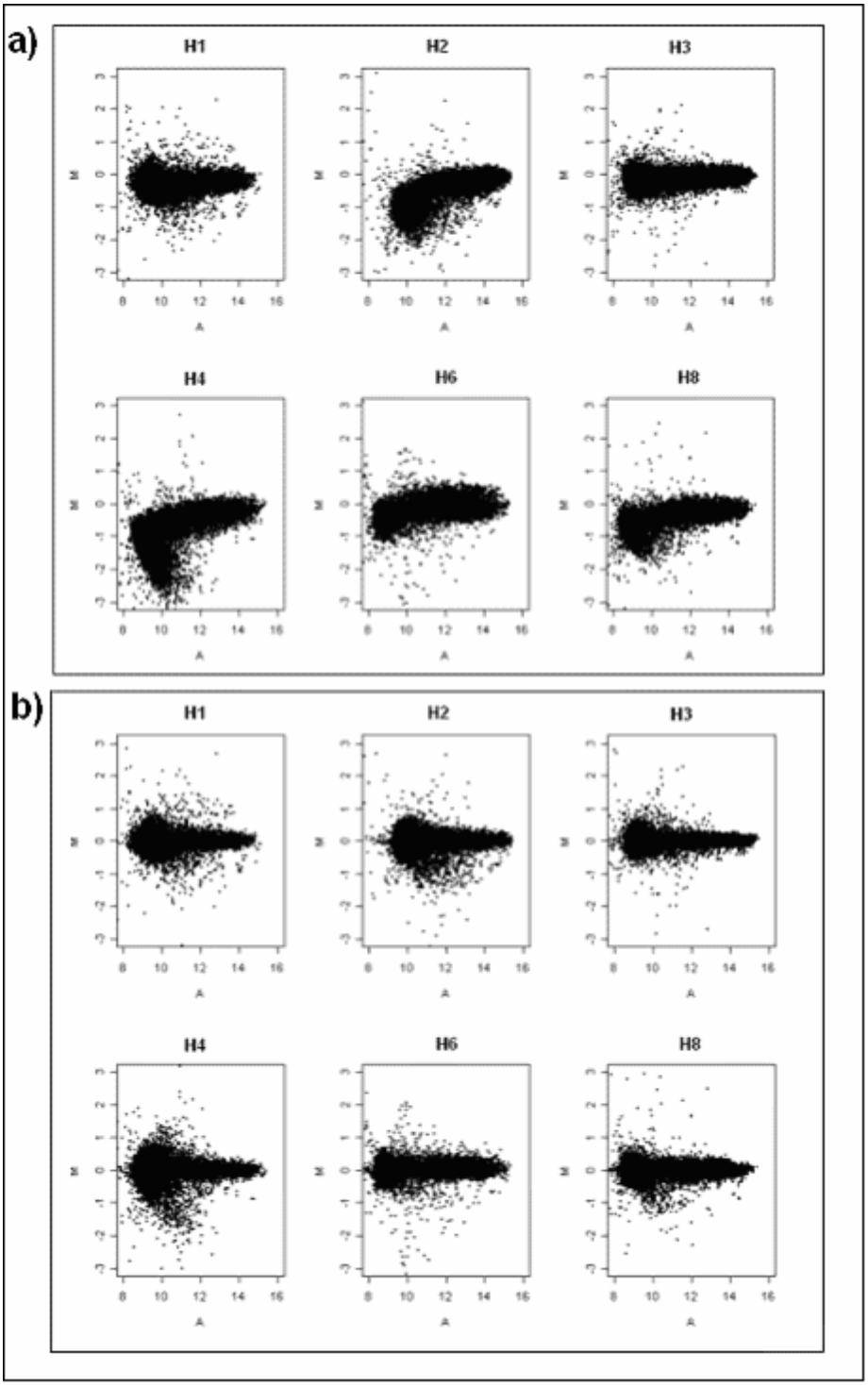


Figure 3.3 : MA-plots des hybridations 1 à 6: a) avec donnée brutes b) normalisés par la méthode « printiploess » réalisés avec le module limma de bioconductor

Tableau 3.1 : Corrélation intra-lame des hybridations selon différentes méthodes de normalisation

Hybridation	None	PTLOess	Composite
H1_2103-TZ3	0,319	0,223	0,221
H2_0801-TZ1*	0,759	0,232	0,474
H3_TZ1-040	0,280	0,121	0,127
H4_TZ2-0202-REP*	0,730	0,173	0,417
H4_TZ2-0202	0,768	0,250	0,460
H5_2102-TZ2-R*	0,274	0,140	0,141
H5_2102-TZ2	0,332	0,045	0,090
H6_0803-TZ3*	0,529	0,106	0,230
H7_0402-TZ2-R	0,618	0,215	0,335
H7_0402-TZ2*	0,790	0,209	0,521
H8_TZ2-2102*	0,704	0,270	0,429
H9_TZ3-2103-R	0,614	0,121	0,312
H9_TZ3-2103	0,346	0,109	0,108
H10_0201-TZ1-R*	0,614	0,116	0,318
H10_0201-TZ1	0,470	0,111	0,166
H11_0403-TZ3-R*	0,645	0,139	0,343
H11_0403-TZ3	0,362	0,157	0,163
H12_2101-TZ1*	0,482	0,135	0,204
H13_TZ1-0201-R	0,773	0,194	0,475
H13_TZ1-0201	0,393	0,121	0,100
H14_0202-TZ2-R*	0,728	0,139	0,441
H14_0202-TZ2	0,529	0,110	0,214
H15_TZ2-0802-R*	0,565	0,133	0,288
H15_TZ2-0802	0,482	0,110	0,187
H16_TZ2-0402-R*	0,802	0,225	0,563
H16_TZ2-0402	0,704	0,272	0,440
H17_TZ3-0803*	0,668	0,374	0,379
H18_TZ3-0203-R*	0,688	0,437	0,505
H18_TZ3-0203	0,470	0,382	0,376
H19_TZ3-0403-R	0,725	0,445	0,522
H19_TZ3-0403*	0,628	0,394	0,424
H20_0401-TZ1-R	0,460	0,258	0,247
H20_0401-TZ1	0,460	0,258	0,247
H21_TZ1-0801	0,675	0,603	0,605
H22_TZ1-2101*	0,647	0,376	0,411
H23_0203-TZ3	0,683	0,294	0,400
H24_0802-TZ2	0,595	0,296	0,298
Moyenne	0,576	0,224	0,329

Les hybridations marquées d'un astérisque (*) sont celle retenue pour la phase d'analyse, celle marqué de « -R » représente les « rescans » alors que « -REP » représente les reprises.

Tableau 3.2 : Corrélation inter-lame des hybridations[1]

Lignée 02

	H10-R*	H10	H14-R*	H14	H23	H13-R	H13	H4-REP*	H4	H18-R*	H18
H10-R*	1,00	0,54	0,47	0,29	-0,28	0,50	0,07	-0,43	-0,34	0,21	-0,11
H10	0,54	1,00	0,30	0,22	-0,16	0,31	0,08	-0,25	-0,18	0,13	-0,06
H14-R*	0,47	0,30	1,00	0,54	-0,31	0,60	0,09	-0,53	-0,46	0,38	0,00
H14	0,29	0,22	0,54	1,00	-0,19	0,40	0,13	-0,34	-0,31	0,29	0,06
H23	-0,28	-0,16	-0,31	-0,19	1,00	-0,35	-0,05	0,28	0,21	-0,20	0,00
H13-R	0,50	0,31	0,60	0,40	-0,35	1,00	0,31	-0,50	-0,39	0,33	-0,09
H13	0,07	0,08	0,09	0,13	-0,05	0,31	1,00	0,01	0,03	-0,04	-0,13
H4-REP*	-0,43	-0,25	-0,53	-0,34	0,28	-0,50	0,01	1,00	0,46	-0,26	0,10
H4	-0,34	-0,18	-0,46	-0,31	0,21	-0,39	0,03	0,46	1,00	-0,33	-0,08
H18-R*	0,21	0,13	0,38	0,29	-0,20	0,33	-0,04	-0,26	-0,33	1,00	0,52
H18	-0,11	-0,06	0,00	0,06	0,00	-0,09	-0,13	0,10	-0,08	0,52	1,00

Lignée 04

	H20-R	H20	H7-R	H7*	H11-R*	H11	H3	H16-R*	H16	H19-R	H19*
H20-R	1,00	1,00	0,20	0,13	-0,04	-0,05	0,05	0,04	0,02	0,28	0,16
H20	1,00	1,00	0,20	0,13	-0,04	-0,05	0,05	0,04	0,02	0,28	0,16
H7-R	0,20	0,20	1,00	0,56	-0,21	0,14	0,18	-0,25	-0,13	-0,23	-0,25
H7*	0,13	0,13	0,56	1,00	-0,35	0,18	0,19	-0,49	-0,33	-0,42	-0,36
H11-R*	-0,04	-0,04	-0,21	-0,35	1,00	0,32	0,04	0,51	0,47	0,16	0,02
H11	-0,05	-0,05	0,14	0,18	0,32	1,00	0,14	-0,03	0,07	-0,26	-0,28
H3	0,05	0,05	0,18	0,19	0,04	0,14	1,00	-0,01	0,07	-0,14	-0,19
H16-R*	0,04	0,04	-0,25	-0,49	0,51	-0,03	-0,01	1,00	0,70	0,37	0,20
H16	0,02	0,02	-0,13	-0,33	0,47	0,07	0,07	0,70	1,00	0,21	0,03
H19-R	0,28	0,28	-0,23	-0,42	0,16	-0,26	-0,14	0,37	0,21	1,00	0,67
H19*	0,16	0,16	-0,25	-0,36	0,02	-0,28	-0,19	0,20	0,03	0,67	1,00

Lignée 08

	H2*	H24	H6*	H21	H15-R*	H15	H17*
H2*	1,00	0,12	0,37	0,19	-0,32	-0,21	-0,28
H24	0,12	1,00	0,13	0,19	-0,16	-0,09	-0,08
H6*	0,37	0,13	1,00	0,20	-0,22	-0,10	-0,19
H21	0,19	0,19	0,20	1,00	-0,04	0,03	-0,32
H15-R*	-0,32	-0,16	-0,22	-0,04	1,00	0,48	0,10
H15	-0,21	-0,09	-0,10	0,03	0,48	1,00	0,03
H17*	-0,28	-0,08	-0,19	-0,32	0,10	0,03	1,00

Lignée 21

	H12*	H5-R*	H5	H1	H22*	H8*	H9-R	H9
H12*	1,00	0,10	-0,11	-0,06	0,00	-0,37	0,22	0,02
H5-R*	0,10	1,00	0,36	0,00	-0,05	-0,06	0,14	-0,01
H5	-0,11	0,36	1,00	-0,05	-0,10	0,17	-0,20	-0,05
H1	-0,06	0,00	-0,05	1,00	0,17	0,10	0,16	0,12
H22*	0,00	-0,05	-0,10	0,17	1,00	-0,07	0,14	0,09
H8*	-0,37	-0,06	0,17	0,10	-0,07	1,00	-0,22	0,02
H9-R	0,22	0,14	-0,20	0,16	0,14	-0,22	1,00	0,33
H9	0,02	-0,01	-0,05	0,12	0,09	0,02	0,33	1,00

[1]Les hybridations marquées d'un astérisque(*) sont celle retenue pour la phase d'analyse.

Les champs ombragés sont les hybridations où des corrélations positives sont attendues.

3.3.2 Normalisation des données

Plusieurs méthodes de normalisation ont été comparées afin d'évaluer laquelle est la plus appropriée pour nos données. La corrélation intra et inter lame ont été calculées avec plusieurs méthodes de normalisation avec et sans correction du bruit de fond (Tableau 3.3). La correction du bruit de fond vise à éliminer l'intensité de la fluorescence qui serait causée par autre chose que l'hybridation de la cible à la sonde (Dudoit *et al.*, 2002*)*. L'analyse subséquente a été faite sur les données normalisées avec la méthode composite étant donné que c'est la méthode de normalisation qui donne les meilleures corrélations intra et inter lame (Tableau 3.3).

Tableau 3.3 : Corrélation intra et inter lame avec différentes méhodes de normalisation

	Corrélation moyenne	
Normalisation	Intra	Inter
None_B	**0,58**	**0,32**
PTLOess_B	**0,23**	**0,09**
Composite_B	**0,34**	**0,21**
nNets_B	**0,27**	**0,05**
None_NB	0,69	0,35
PTLOess_NB	0,28	0,10
Composite_NB	0,42	0,24
nNets_NB	0,32	0,06

Corrélation moyenne intra et inter lame pour différentes méthodes de normalisation calculée à partir de R avec limma.

3.3.3 Identification et annotation des gènes différentiellement exprimés

3.3.3.1 Analyse limma

Les données retenues suite à l'analyse qualité et qui ont été normalisées avec l'algorithme composite ont été soumises à l'analyse limma, avec la commande 'lmfit' qui estime les

variations d'échelle d'expression en les appliquant à un modèle linéaire, avec un adoucissement (« smoothing ») Bayesien aux erreurs standard. Les résultats de cette analyse (les 10 meilleurs résultats par lignée) sont présentés ordonnée selon leur niveau de significativité (3.4). Aucun gène différentiellement exprimé n'a pu être identifié pour les analyses individuelles ou combinées des lignées. Ce résultat est probablement dû à un faible taux de corrélation intra lame observés pour certaines hybridations ou le faible nombre de lame de qualité retenue pour l'analyse. Le faible taux de corrélation intra lames réduit la sensibilité du test et nuit à l'identification de gènes qui seraient réellement différentiés. L'un des paramètres déterminant pour l'analyse 'lmfit' est un taux de corrélation intra lame moyen très élevé entre toutes les lames afin de conclure que les gènes d'expression élevés sont différentiels. Alternativement, il n'y a aucun gène différentiellement exprimé entre les transgéniques et les témoins. Nous avons donc entrepris une nouvelle analyse avec une méthode différente.

Tableau 3.4 : Résultat de l'analyse limma

Lignée 02 :	ID	MNID	M	A	t	P.Value
	1812	MN5195855	-0.8103834	0.2648679769	-6.419033	0.3752139
	7856	MN5161830	-0.5740780	0.2504171965	-4.938850	0.6996672
	6248	MN5176391	-0.4148196	0.0732548855	-4.930442	0.6996672
	2469	MN5234527	-0.3964071	0.0179983001	-4.843494	0.6996672
	1064	MN5177260	-0.9710889	0.4208950008	-4.731963	0.6996672
	4810	MN5161824	-0.5357821	0.2227238627	-4.682098	0.6996672
	3699	MN5236092	-0.7485500	0.3928893653	-4.501264	0.6996672
	3959	MN5253642	0.4012829	0.0073406706	4.396058	0.6996672
	1735	MN5253163	-0.4396931	-0.0297748794	-4.381286	0.6996672
	396	MN5195432	-0.8914863	0.3588420676	-4.210680	0.6996672
Lignée 04 :	ID	MNID	M	A	t	P.Value
	761	MN5254450	0.3628840	0.009231792	5.774358	0.2434856
	76	MN5258455	0.3404068	-0.037521625	5.226663	0.2434856
	8386	MN5254032	0.3498666	0.068087590	5.067104	0.2434856
	1729	MN5255853	0.3251910	0.009380222	5.053880	0.2434856
	5896	MN5169821	0.3171061	0.019925949	5.018950	0.2434856
	1924	MN5246151	0.3438076	-0.053022775	5.017545	0.2434856
	186	MN5257573	0.3213512	0.025206358	4.971276	0.2434856
	589	MN5176058	0.3410179	0.079445048	4.932109	0.2434856
	8349	MN5244313	0.3164651	-0.012882005	4.899844	0.2434856
	682	MN5235024	-0.3453574	-0.050629988	-4.817252	0.2434856
Lignée 08 :	ID	MNID	M	A	t	P.Value
	5599	MN5191492	-0.3644873	-6.359681e-02	-5.882787	0.1983295
	4822	MN5237499	0.3344332	6.717139e-02	5.304584	0.1983295
	7742	MN5159358	-0.2986141	-1.540258e-02	-5.260695	0.1983295
	8187	MN5193445	-0.3153957	2.831768e-03	-5.250479	0.1983295
	5682	MN5169533	0.3019440	1.716383e-02	5.249001	0.1983295
	5937	MN5191991	-0.3810001	1.801099e-02	-5.212771	0.1983295
	4831	MN5165019	0.3128374	-1.946503e-02	5.145921	0.1983295
	741	MN5171164	-0.3853422	5.417153e-02	-4.939367	0.1983295
	7008	MN5255489	-0.4280725	7.804404e-02	-4.932722	0.1983295
	3340	MN5196218	0.3436621	-8.671141e-02	4.918895	0.1983295
Lignée 21:	ID	MNID	M	A	t	P.Value
	5868	MN5248924	0.5778973	0.0673612985	8.135370	0.1402427
	7021	MN5235638	0.7288052	-0.0548306087	7.915542	0.1402427
	1011	MN5182713	0.4490339	-0.0387222506	7.028344	0.2298165
	6122	MN5249533	0.5993303	0.0714593135	6.643123	0.2357817
	463	MN5249914	0.6551607	-0.0763097915	6.410265	0.2357817
	6629	MN5256787	0.4129626	0.0209383793	6.373782	0.2357817
	5946	MN5245227	0.4177364	-0.0307043895	6.233709	0.2372741
	1663	MN5249369	0.6288760	-0.1511660121	6.027753	0.2395679
	1997	MN5191560	0.6685655	-0.1600376968	6.010377	0.2395679
	3699	MN5236092	0.4598221	-0.0531628250	5.895564	0.2471795

3.3.3.2 Analyse SAM et annotation

L'analyse SAM est une des méthodes d'analyse de biopuces les plus utilisée (635 citations depuis Octobre 2004; http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1173086). L'analyse SAM a été faite sur chaque lignée séparément et par la suite en combinant les différentes lignées par paires et toutes ensembles. Pour les fins de l'analyse, les quatre meilleures hybridations ont été choisies pour chacune des lignées (Tableau 3.1, 3.2). Pour chaque analyse le seuil de FDR (false discovery rate) a été fixé le plus bas possible tout en permettant d'identifier le plus des gènes différentiellement exprimés possible. Le FDR contrôle le pourcentage de faux positifs attendus. C'est pourquoi une valeur beaucoup plus élevée est permise pour le FDR que pour la p-value. Généralement le FDR se situe en dessous de 20% (Tusher *et al.*, *2001;* Xiao *et al.*, 2002*)*. Le type d'analyse SAM sélectionnée est une analyse du type «une classe» donc la comparaison se fait par rapport à un ratio de 0. Les gènes différentiellement exprimés sont ceux qui le sont par rapport à un ratio d'expression de 0 plutôt qu'une comparaison entre deux groupes définis.

L'analyse SAM a permis d'identifier des gènes dans trois lignées différentes et dans une analyse de deux lignées combinées (Tableau 3.5). Nous avons retenu les gènes différentiellement exprimés identifiés par la SAM qui ont une valeur q (« q-value »), le niveau de confiance, plus petit que 5%. Le nombre de gènes identifiés pour chacune des lignées est très variable (Tableau 3.5). De plus, on retrouve très peu de recoupement entre les différentes lignées (Tableau 3.8). Le sens de l'expression des gènes (surexpression ou suppression chez le transgénique) varie beaucoup entre les lignées indiquant une grande différence entre celles-ci

L'analyse de lignée 02 a révélé vingt-et-un gènes différentiellement exprimés, tous sous exprimés dans le transgénique, tandis que dans la lignée 08, vingt-sept gènes sont sous exprimés. La lignée 21 affiche quarante-quatre gènes surexprimés dans le transgénique (Tableau 3.11). Les analyses de lignées combinées trouvent, en général, moins de gènes que les analyses de lignées individuelles. Ceci peut être expliqué par le fait que les

lignées sont tellement différentes entre elles que lorsqu'on les combine on trouve peu de points en commun. L'analyse combinée de la lignée 04 avec la lignée 08 est la seule exception où l'effet du transgène est le plus semblable. Nous avons trouvé beaucoup plus de gènes que lorsqu'on les analyse individuellement.

Tableau 3.5 : Liste des gènes différentiellement exprimés trouvés à l'aide de l'analyse SAM

Combinaison	+[1].	-[1].	All Sign.	Non.Sign.	FDR
L02	0	21	21	9032	21,28%
L08	0	27	27	9026	16,87%
L21	44	0	44	9009	16,14%
L04+L08	115	3	118	8935	15,30%

[1] + : gènes surexprimés chez le transgénique

\- : gènes sousexprimés chez le transgénique

Annotation des gènes identifiés par l'analyse SAM

Tous les gènes de la biopuces ont été annotés avec trois banques de données différentes afin de maximiser la puissance de l'annotation. Cette annotation a été faite à l'aide d'un script python pour réunir et filtrer tous les résultats BLAST en un seul fichier EASE facilement utilisable par MeV (voir Tableau 3.6).

Tableau 3.6 : Résultats du script d'annotation python

Résultat de l'annotation	
Nombre d'annotation par gène	
Une seule annotation	1375
Deux annotations	1480
Trois annotations	2804
Total des séquences annotées	5659
Annotation par banque	
Séquences annotées avec nr	4199
Séquences annotées avec PGI5	5003
Séquences annotées avec Pfam	3545

Le nombre d'annotations a été déterminé avec un filtrage à un seuil de probabilité 1^{e}-08.

Les annotations dont le contig est suivi du caractère '*' indique que le contig est représenté par un amplicon de mauvaise qualité (double bandes ou faible concentration). Les annotations ne contenant que le nom du contig correspondent à des séquences qui n'ont pas été associées à des séquences connues des banques de données nr, PGI5 et Pfam. L'annotation a identifié cinq classes fonctionnelles parmi les des gènes différentiellement exprimés: les gènes liés au pollen, à la photosynthèse, au stress, à la transcription et à la paroi cellulaire (Tableau 3.7). Il y a cinq gènes liés à la photosynthèse (« chlorophyle a/b », la « plastocyanine » et le « oxygen evolving complex »), qui sont soit surexprimés dans la lignée 21 ou sous exprimés dans la ligné 8. Un phénomène semblable s'applique aux gènes liés au stress qui sont sous-exprimés dans la lignée 2 et 8, alors que ceux identifiés dans la lignée 21 sont surexprimés. Une classe qui nous intéresse particulièrement est celle des gènes liés à la paroi cellulaire où on y retrouve 7 gènes différentiellement exprimés. Parmi ces gènes on trouve la « cellulose synthase » surexprimée dans la lignée 21 et dans l'analyse combinée de 4 & 8 avec un ratio relativement faible de 1.19 et 1.30. Seule une séquence semblable à la cinnamoyl-CoA reductase, surexprimée dans les lignées 4 et 8 a un lien avec la synthèse de la lignine.

Le patron d'expression varie beaucoup entre les lignées ce qui indique une variabilité entre celles-ci plus particulièrement entre la lignée 21 et les trois autres. Ces données ne permettent pas de déterminer très clairement quels gènes ou groupes de gènes pourraient être contrôlés par le gène LIM2. Les enzymes de la synthèse de la lignine semblent peu affectées par sa surexpression. Les quelques gènes affichants une expression différentielle sont liés au stress, à la photosynthèse ou à la paroi cellulaire suggère aussi un rôle autre que la synthèse de la lignine.

Tableau 3.7 : Gènes différentiellement exprimés classés selon leur fonction

Lignée	MNID	surepxression (+) suppression (-) transgénique	Annotation	q-val,	M (log 2)	Ratio (ch1/ch2)
Liés au pollen						
8	MN5249369	-	Contig8955 [1:] [2:putative Pollen specific protein C13 precursor [Oryza sativa (japonica cultivar-group)] ref|NP_921099.1| putative Pollen specific protein C13 precursor [Oryza sativa (japonica cultivar-group)] gb|AAN31783.1| Putataive pollen specific p	4,07	-0,59	1,51
21		+		0,00	0,63	1,55
Liés à la Photosynthèses						
8	MN5191560	-	Contig10528 [1:homologue to UP|CB2A_PINSY (P15193) Chlorophyll a-b binding protein type II 1A] [2:unnamed protein product [2:Pinus sylvestris] sp|P15193|CB2A_PINSY Chlorophyll a-b binding protein type II 1A] [3:PF00504.11;Chloroa_b-bind]	0,00	-0,64	1,56
21		+		0,00	0,67	1,59
8	MN5235638	-	Contig7847 [1:-] [2:chlorophyll a/b-binding protein [2:Pinus thunbergii] pir||S22522 chlorophyll a/b-binding protein (cab-6) precursor - Japanese black pine] [3:PF00504.11;Chloroa_b-bind]	4,07	-0,63	1,55
21		+		0,00	0,73	1,66
21	MN5182713	+	Contig10754 [1:similar to UP|PSBP_CUCSA (Q9SLQ8) Oxygen-evolving enhancer protein 2] [2:23 kDa oxygen evolving protein of photosystem II [2:Solanum tuberosum] sp|P93566|PSBP_SOLTU Oxygen-evolving enhancer protein 2] [3:PF01789.6;PsbP]	0,00	0,45	1,37
21	MN5163062	+	Contig7897 [1:similar to UP|Q9XQB2 (Q9XQB2) Chlorophyll a/b binding protein CP29] [2:chlorophyll a/b binding protein CP29 [2:Vigna radiata]] [3:PF00504.11;Chloroa_b-bind]	0,00	0,44	1,36
21	MN5164094	+	Contig7254 [1:similar to UP|PLAS_VICFA (P00288) Plastocyanin] [2:unnamed protein product [2:Spinacia oleracea] pir||CUSP plastocyanin precursor - spinach sp|P00289|PLAS_SPIOL Plastocyanin] [3:PF00127.9;Copper-bind]	2,17	0,32	1,25
Liés au Stress						
2	MN5236092	-	Contig7487 [1:similar to UP|Q9ZP84 (Q9ZP84) Heat shock protein 17.4] [2:small heat-shock protein [Pseudotsuga menziesii]] [3:PF00011.8;HSP20]	0,00	-0,75	-1,68
2	MN5253163	-	Contig8835 [1:UP|Q9M4R0 (Q9M4R0) Ubiquitin conjugating protein] [2:ubiquitin conjugating protein [Avicennia marina]] [3:PF00179.14;UQ_con]	0,00	-0,44	-1,36
8	MN5236092	-	Contig7487 [1:similar to UP|Q9ZP84 (Q9ZP84) Heat shock protein 17.4] [2:small heat-shock protein [2:Pseudotsuga menziesii]] [3:PF00011.8;HSP20]	4,22	-0,41	1,33
21	MN5190745	+	Contig8973 [1:-] [2:ubiquitin-like protein [2:Phaseolus vulgaris] pir||T12035 polyubiquitin 4.4 - kidney bean] [3:PF00240.11;ubiquitin]	0,00	0,59	1,50
21	MN5248924	+	Contig8511 [1:-] [2:water-stress-inducible protein LP3 - loblolly pine gb|AAB07493.1| water deficit inducible protein LP3] [3:-]	0,00	0,58	1,49

21	MN5195778	+	Contig7709 [1:UP|Q9AR09 (Q9AR09) Ubiquitin fused to ribosomal protein L40] [2:ubiquitin / ribosomal protein CEP52 - wood tobacco gb|AAA34064.1| ubiquitin fusion protein] [3:PF00240.11;ubiquitin]	0,00	0,48	1,40
21	MN5232943	+	Contig8945 [1:weakly similar to UP|Q8VWQ1 (Q8VWQ1) Dehydration-induced protein RD22-like protein] [2:dehydration-induced protein RD22-like protein [2:Gossypium hirsutum]] [3:PF03181.5;BURP]	2,96	0,42	1,34
21	MN5164094	+	Contig7254 [1:similar to UP|PLAS_VICFA (P00288) Plastocyanin] [2:unnamed protein product [2:Spinacia oleracea] pir||CUSP plastocyanin precursor - spinach sp|P00289|PLAS_SPIOL Plastocyanin] [3:PF00127.9;Copper-bind]	2,17	0,32	1,25
Liés au facteur de transcription						
21	MN5239863	+	Contig12838 [1:similar to UP|Q9FHG8 (Q9FHG8) Similarity to C3HC4-type RING zinc finger protein] [2:-] [3:-]	0,00	0,43	1,34
21	MN5164545	+	Contig13919 [1:similar to UP|Q86EX7 (Q86EX7) Clone ZZD1326 mRNA sequence] [2:unknown [2:Arabidopsis thaliana] gb|AAM51593.1| At1g19310/F18O14_14 [2:Arabidopsis thaliana] ref|NP_564078.1| zinc finger (C3HC4-type RING finger) family protein [2:Arabidopsis	2,96	0,39	1,31
4 & 8	MN5253245	+	Contig11104 [1:weakly similar to UP|Q06979 (Q06979) Ocs-element binding factor 3.2] [2:bZIP transcription factor [2:Nicotiana tabacum]] [3:-]	1,31	0,20	1,15
4 & 8	MN5233591	+	Contig9571 [1:similar to UP|Q9LUR1 (Q9LUR1) RING zinc finger protein-like] [2:putative zinc finger protein [Arabidopsis thaliana] pir||T52079 probable zinc finger protein [2:imported] - Arabidopsis thaliana] [3:-]	0,84	0,19	1,14
4 & 8	MN5252349	+	Contig10963 [1:weakly similar to GB|AAL11556.1|15983376|AF424562 AT3g59080/F17J16_130 {Arabidopsis thaliana;}] [2:putative chloroplast nucleoid DNA binding protein [Oryza sativa (japonica cultivar-group)] dbj|BAD15987.1| putative chloroplast nucleoid D	0,84	0,18	1,13
Liés au parois cellulaire et à la lignine						
2	MN5195544	-	Contig9170 [1:similar to UP|Q9FIU5 (Q9FIU5) Serine/threonine-specific protein kinase-like protein] [2:putative serine/threonine-specific protein kinase [Oryza sativa (japonica cultivar-group)] dbj|BAD03013.1| putative serine/threonine-specific protein k	4,02	-0,68	-1,60
21	MN5196269	+	Contig8766 [1:similar to UP|P93156 (P93156) Cellulose synthase (Fragment)] [2:cellulose synthase (EC 2.4.1.-) catalytic chain celA2 - upland cotton (fragment) gb|AAB37767.1| cellulose synthase] [3:PF03552.3;Cellulose_synt]	0,00	0,54	1,46
21	MN5160221	+	Contig8815 [1:-] [2:Calmodulin (CaM) emb|CAA09302.1| calmodulin 3 protein [2:Capsicum annuum] dbj|BAB61908.1| calmodulin NtCaM2 [2:Nicotiana tabacum] dbj|BAB61907.1| calmodulin NtCaM1 [2:Nicotiana tabacum] gb|AAA34144.1| calmodulin emb|CAC84563.1| putati	0,00	0,38	1,30
4 & 8	MN5234677	+	Contig8082 [1:UP|Q6GUG6 (Q6GUG6) Cellulose synthase catalytic subunit] [2:cellulose synthase 6 [2:Populus tremuloides]] [3:PF03552.3;Cellulose_synt]	0,00	0,25	1,19
4 & 8	MN5256816	+	Contig7542 [1:] [2:cinnamoyl-CoA reductase-like protein [Arabidopsis thaliana] ref|NP_194776.1| cinnamoyl-CoA reductase-related [Arabidopsis thaliana] gb|AAK68826.1| cinnamoyl-CoA reductase-like protein [Arabidopsis thaliana] pir||D85356 cinnamoyl-CoA re	0,00	0,21	1,15

4 & 8	MN5251271	+	Contig11336 [1:similar to UP|BRI1_LYCES (Q8GUQ5) Brassinosteroid LRR receptor kinase precursor (tBRI1) (Altered brassinolide sensitivity 1) (Systemin receptor SR160)] [2:putative receptor protein kinase [Arabidopsis thaliana] gb|AAD20088.1| putative r	1,20	0,20	1,15
4 & 8	MN5236817	+	Contig3549 [1:weakly similar to UP|Q9FKC2 (Q9FKC2) Receptor protein kinase-like protein] [2:leucine-rich repeat family protein / protein kinase family protein [Arabidopsis thaliana]] [3:PF00069.12;Pkinase]	1,19	0,20	1,15

Tableau 3.8 : Tableau des gènes différentiellement exprimés retrouvés dans plusieurs lignées analysées séparément avec SAM

groupement	Lignée	MNID	Annotation	q-value(%)	M	Ratio
1	2	MN5171347	Contig6724 [1:homologue to UP|Q6WSR8 (Q6WSR8) Class IV chitinase Chia4-Pa2] [2:putative class IV chitanase [Picea abies]] [3:PF00182.8;Glyco_hydro_19]	3,82	-0,51712771	-1,4311032
	8	MN5171347		0,00	-0,63457656	-1,55248202
2	2	MN5177212	Contig8858 [1:similar to UP|TPIC_FRAAN (Q9M4S8) Triosephosphate isomerase] [2:triosephosphate isomerase [Fragaria x ananassa] sp|Q9M4S8|TPIC_FRAAN Triosephosphate isomerase] [3:PF00121.7;TIM]	0,00	-0,81054679	-1,75387604
	21	MN5177212		2,77	0,31554037	1,24447769
3	8	MN5191560	Contig10528 [1:homologue to UP|CB2A_PINSY (P15193) Chlorophyll a-b binding protein type II 1A] [2:unnamed protein product [2:Pinus sylvestris] sp|P15193|CB2A_PINSY Chlorophyll a-b binding protein type II 1A] [3:PF00504.11;Chloroa_b-bind]	0,00	-0,64458497	-1,5632895
	21	MN5191560		0,00	0,66856548	1,5894917
4	8	MN5194939	Contig9058 [1:-] [2:glutamyl-tRNA(Gln) amidotransferase B family protein [2:Arabidopsis thaliana] gb|AAL67097.1| At1g48520/T1N15_12 [2:Arabidopsis thaliana] gb|AAL06883.1| At1g48520/T1N15_12 [2:Arabidopsis thaliana] gb|AAG29096.1| Glu-tRNA(Gln) amidotransferase subunit B [2:Arabidopsis thaliana]] [3:PF02637.6;GatB_Yqey]	4,22	-0,49758813	-1,41185128
	21	MN5194939		2,06	0,26796434	1,20410761
5	8	MN5196602	Contig4166 [1:-] [2:unknown [2:Arabidopsis thaliana] dbj|BAB11109.1| unnamed protein product [2:Arabidopsis thaliana] ref|NP_196885.1| glutaredoxin family protein [2:Arabidopsis thaliana] gb|AAL31176.1| AT5g13810/MAC12_24 [2:Arabidopsis thaliana] gb|AAK63958.1| AT5g13810/MAC12_24 [2:Arabidopsis thaliana]] [3:PF00462.9;Glutaredoxin]	4,22	-0,45103207	-1,36701784
	21	MN5196602		0,00	0,40167745	1,32104302
6	8	MN5235638	Contig7847 [1:-] [2:chlorophyll a/b-binding protein [2:Pinus thunbergii] pir||S22522 chlorophyll a/b-binding protein (cab-6) precursor - Japanese black pine] [3:PF00504.11;Chloroa_b-bind]	4,07	-0,62798867	-1,54540896
	21	MN5235638		0,00	0,72880521	1,65726603
7	2	MN5236092	Contig7487 [1:similar to UP|Q9ZP84 (Q9ZP84) Heat shock protein 17.4] [2:small heat-shock protein [2:Pseudotsuga menziesii]] [3:PF00011.8;HSP20]	0,00	-0,74855002	-1,68010339
	8	MN5236092		4,22	-0,41265307	-1,33113147
	21	MN5236092		0,00	0,45982208	1,37537219
8	8	MN5249369	Contig8955 [1:] [2:putative Pollen specific protein C13 precursor [Oryza sativa (japonica cultivar-group)] ref|NP_921099.1| putative Pollen specific protein C13 precursor [Oryza sativa (japonica cultivar-group)] gb|AAN31783.1| Putataive pollen specific protein C13 precursor [Oryza sativa (japonica cultivar-group)] gb|AAM08621.1| Putative Pollen specific protein C13 precursor [Oryza sativa (japonica cultivar-group)]] [3:PF01190.7;Pollen_Ole_e_I]	4,07	-0,59455436	-1,51000608
	21	MN5249369		0,00	0,62887599	1,54635975

Validation des données par QPCR

L'analyse des microarrays indique qu'un seul gène de la biosynthèse de la lignine est différentiellement exprimé chez deux lignées à un niveau relativement faible. Ce résultat suggère que la surexpression de LIM2 a eu peu d'effet sur cette voie de synthèse. Des analyses QPCR ont été faites (par Vicky Roy) en parallèle afin d'étudier l'expression de trois autres gènes de la biosynthèses de la lignine en lien avec la surexpression du gène ptLIM2.

L'expression relative de la CAD, CCOAOMT et de la 4CL ont été analysées dans différentes lignées de transgéniques ptLIM2. L'abondance des transcrits a été mesuré sur des ADNc synthétisés à partir d'ARN extrait de la tige de pousse de 21 semaine des lignées transgéniques 2, 8 , 16 et 21 et du témoin T (Figure 3.4). Les analyses QPCR indiquent que la surexpression de ptlim2 n'aurait pas un effet direct sur la voie de biosynthèse de la lignine. Ces résultats concordent avec les analyses de biopuce, c'est-à-dire qu'il n'y a pas d'indication claire de surexpression des gènes associés à la lignine dans ces lignées.

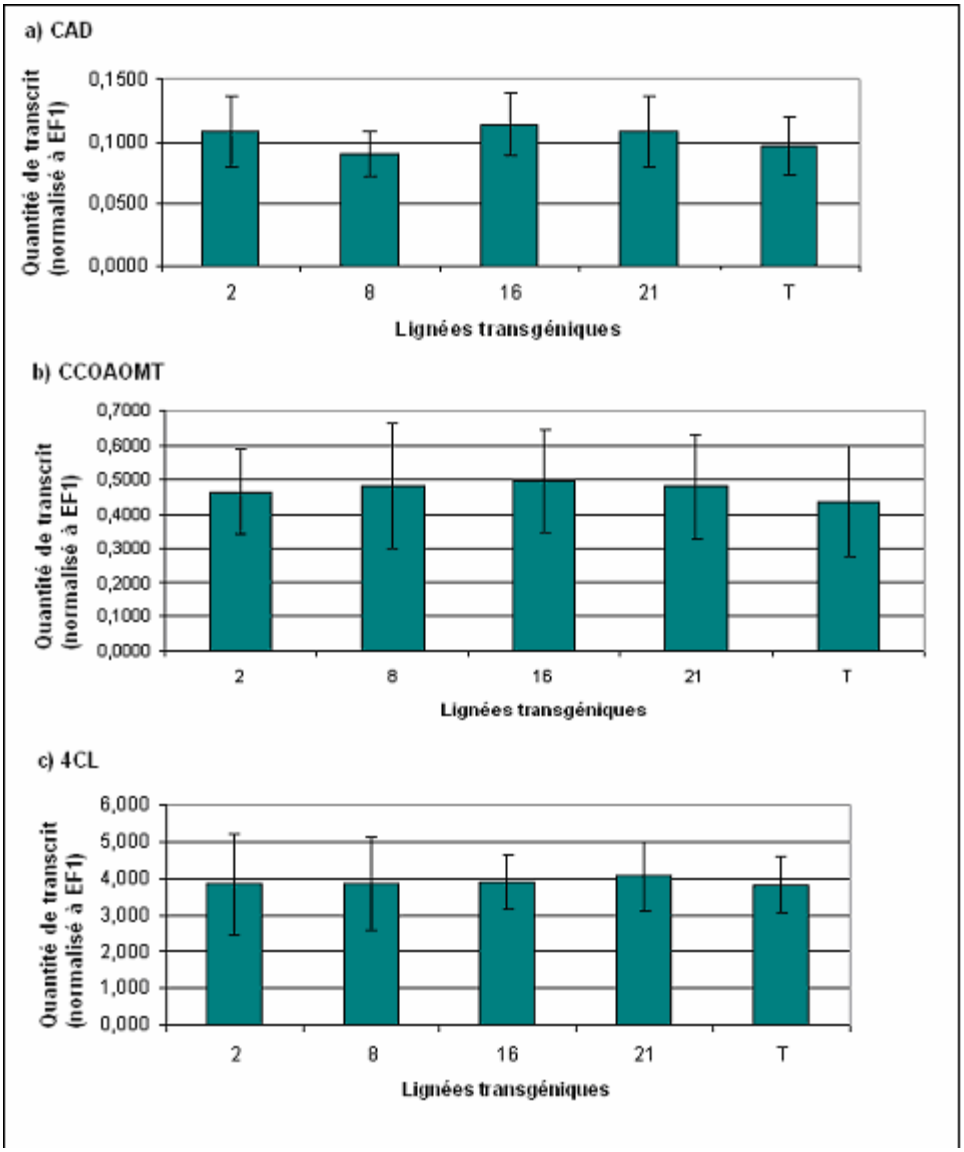


Figure 3.4 : Résultats des analyses QPCR

3.4 Discussions

Des lignées transgéniques d'épinette blanche surexprimant le gène LIM2 isolé du pin ont été analysées à l'aide de biopuces contenant plus de 9000 séquences de cDNA différentes de l'épinette. Le gène LIM2 étant considéré comme un régulateur potentiel de la synthèse de la lignine, cette étude visait à déterminer si sa surexpression entraînerait une modification de l'expression des enzymes nécessaire à la synthèse de ce composé essentiel à la formation du bois. Plusieurs méthodes de pré-traitement des données ont été utilisées pour assurer la qualité des images et afin de sélectionner les meilleures hybridations pour la recherche de gènes différentiellement exprimés. Dans ce but, les données normalisées avec la méthode composite (avec soustraction du bruit de fonds) ont été soumises à deux méthodes d'analyse statistique, LIMMA et SAM. Seule la méthode SAM a permis d'identifier des gènes soient surexprimés ou encore sous-exprimés chez les transgéniques. Les gènes ainsi mis en évidence varient toutefois considérablement en nombres entre les différentes lignées et seulement quelques gènes ont été trouvés en commun chez plus d'une lignée.

L'analyse des résultats de biopuces et l'identification de gènes différentiellement exprimés avec différentes méthodes statistiques a montré que le résultat dépendait grandement de la méthode utilisée. Les différentes analyses SAM ont identifié un nombre considérable de gènes différentiellement exprimés pour chacune des lignées, mais ces gènes ne semblent pas être les mêmes entre les différentes lignées. Par surcroît, l'expression relative des gènes retrouvés dans plusieurs lignées par rapport aux témoins n'est pas le même à travers toutes les lignées. Seul l'analyse regroupée des lignées 4 et 8 montre certains gènes en commun. On remarque par ailleurs que l'expression des gènes de la voie de biosynthèse de la lignine est très peu affectée par la surexpression de ptLIM2, sauf pour le gène de la CCR, qui est faiblement surexprimé dans deux des trois lignées analysées. L'expression de quelques autres enzymes associées à la paroi cellulaire comme la cellulose synthetase et la calmodulin est aussi légèrement augmentée. Les faibles différences entre les transgéniques et les témoins, ainsi que la variabilité observée entre les lignées portent aussi à questionner

l'importance biologique de ces observations. Enfin, l'analyse indépendante des niveaux de transcrits de quelques gènes de la synthèse de la lignine par RT-QPCR confirme l'absence de changement au niveau de ce métabolisme chez les lignées analysées.

Étant donnée l'absence d'effet clair au niveau de la synthèse de la lignine, nous avons examiné les groupes de gènes liés à d'autres processus physiologiques. Des gènes reliés au stress et à la photosynthèse sont surexprimés dans certaines lignées mais sousexprimés chez d'autres. Donc, il est peu probable qu'ils soient régulés par LIM2. Par contre, au moins un facteur de transcription du type RING, une protéine à doigt de zinc, est surexprimé dans chacune des lignées. Des gènes de celluloses synthetases sont aussi surexprimés dans chacune des lignées. D'après l'ensemble de ces observations, il faut conclure que les différentes lignées d'épinette transgénique étudiées apportent quelques pistes sur le rôle de LIM2, toutefois ces résultats sont préliminaires et devront être validés.

Certaines plantes angiospermes comme le tabac et *Arabidopsis thaliana*, dont le génome a été complètement séquencé sont reconnus comme des systèmes modèles pour l'étude moléculaire des mécanismes qui régissent la croissance et le développement chez l'ensemble des plantes. Ainsi la découverte du gène ntLIM1 chez le tabac ((Kawaoka, 2000) a permis de développer l'hypothèse que des séquences similaires pourraient agir comme régulateur de la synthèse de la lignine chez les conifères (des gymnospermes) et donc avoir un rôle dans la différentiation du tissu vasculaire et la formation du bois chez ces espèces à valeur commerciale. L'épinette est toutefois fort différente des plantes modèles comme Arabidopsis, du point vue anatomique, physiologique et génétique. L'étude des facteurs de transcription KNOX (Guillet-Claude, *et al.,* 2004), entre autres, a clairement démontré que la structure des familles de gènes peut évoluer fort différemment chez les conifères par rapport aux angiospermes. Ainsi, nous avons entrepris l'étude des facteurs de transcription LIM de conifère par la surexpression de séquences de pin dans un système homologue (l'épinette) afin d'étudier leur rôle physiologique et évaluer leur rôle potentiel dans la synthèse de la lignine.

L'étude de la formation du xylème secondaire chez des conifères transgéniques est un processus qui demande beaucoup de temps contrairement à d'autres espèces même ligneuses tel le peuplier qui se développe beaucoup plus rapidement permettant ainsi

d'avoir accès à beaucoup plus de matériel. Ainsi, les analyses présentées dans cette étude représentent une première partie du travail à accomplir pour l'étude des LIMs de conifères. D'une part, des arbres plus développés devront être étudiées et d'autres part, des transgéniques surexprimants d'autres gènes LIM restent à analysés. Au moins trois séquences (LIM1, 2, 3) sont présentes chez les conifères et leur rôle restent à déterminer.

4.0 Conclusion générale

Le développement d'outils informatiques et la mise en place d'une chaîne de traitement et d'analyse de données de biopuces sont décrits dans ce mémoire. Un inventaire complet des LIMS a été fait soulevant leur forces et faiblesses. Cette analyse a permis de démontrer qu'aucun LIMS ne faisait une gestion efficace d'échantillons dans un contexte d'expérience factorielle. Un outil a donc été conçu, en PHP et MySQL, pour combler ce besoin. SLIMS permet le suivi des échantillons de leur récolte jusqu'à l'extraction des ARN pour les expériences de biopuces. À cette étape les données peuvent être transférées automatiquement à un logiciel spécialisé dans la gestion de données de biopuces. L'utilisation de cette chaîne a été illutstrée par l'analyse d'une expérience visant à identifier des gènes différentiellement exprimés chez les transgéniques surexprimant le gène ptLIM2. Elle consiste principalement de trois outils : SLIMS, BASE et Bioconductor (ou MeV). L'outil SLIMS, développé spécifiquement pour le projet Arborea, a assuré le suivi de tous les échantillons liés à cette expérience. SLIMS est présentement installé sur Meije, le serveur central du centre de bioinformatique de l'Université Laval. Utilisé quotidiennement par une dizaine de personnes différentes, il gère présentement plus de 200 expériences et environ 16000 échantillons. Ici, SLIMS a permis de faire le suivi d'une expérience avec des transgéniques ptLIM2 de la récolte des échantillons jusqu'à l'extraction des ARN pour les expériences de biopuces. La procédure de transfert des données de SLIMS vers le logiciel BASE permet de faire le suivi des échantillons à partir de l'extraction de l'ARN jusqu'aux résultats d'expérience de biopuces.

BASE sert à l'entreposage des données de biopuces et des résultats d'expériences. Il a donc été possible de faire l'analyse qualité des hybridations directement à partir de BASE accélérant significativement le procédé d'analyse. L'analyse des données numérisées de ces hybridations réalisées avec Bioconductor et la suite TM4 nous a permis d'accéder aux méthodes d'analyse les plus récentes et les plus performantes dans le domaine. Nous avons entre autres comparé plusieurs méthodes de normalisation et sélectionné la méthode la plus efficace pour nos données. Plusieurs types d'analyses ont été réalisés en parallèle. La méthode SAM a permis d'identifier le plus grand nombre de gènes différentiellement exprimés. Les gènes identifiés et leur expression sont toutefois variables entre les lignées

transgéniques suggérant des différences significatives entre les celles-ci. Par contre, deux lignées, 04 et 08, ont démontré plus de similarité, 118 gènes ont ainsi pu être identifiés lors de cette analyse.

Une annotation exhaustive a été faite permettant d'associer à chaque gène de la biopuce l'information recueillie à partir de trois banques de données publiques soit nr, PGI5 et PFam. Les gènes identifiés lors de l'analyse des différentes lignées sont majoritairement associés au stress, à la photosynthèse, à la régulation transcriptionelle et à la paroi cellulaire.

Un seul gène de la voie de biosynthèse de la lignine a été identifié comme différentiellement exprimé. On remarque par ailleurs que l'expression des gènes de la voie de biosynthèse de la lignine est peu affectée par la surexpression de ptLIM2. Toutefois, au moins un facteur de transcription du type RING est surexprimé dans chacune des lignées étudiées. Il est encore trop tôt pour cerner le rôle de ptLIM2 au niveau de la transcription puisque ces résultats sont préliminaires et devront être validés. À la lumière de cette analyse il semble que le rôle biologique de ptLIM2 ne soit pas principalement au niveau de la biosynthèse de la lignine.

Bibliographie

Anterola, A.M. and Lewis, N.G. (2002). Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochemistry 61 221-294.

Baltz R, Evrard JL, Domon C, Steinmetz A. (1992). *A LIM motif is present in a pollen-specific protein.* Plant Cell. 4:1465-6.

Boerjan W, Ralph J, Baucher M. (2003). *Lignin biosynthesis*. Annu Rev Plant Biol. 54:519-46.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. (2001) *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet., 29:365-71.

Chang, S, Puryear, J, and Cairney J (1993). *A simple and Efficient Method for Isolating RNA from Pine Trees.* Plant Molecular Biology Reporter 11:113-116

Dawid IB, Toyama R, Taira M. (1995*) LIM domain proteins*. C R Acad Sci III 318: 295-306

Draghici S. (2002). *Statistical intelligence: effective analysis of high-density microarray data.* Drug Discov Today. 1:7

Dudoit S, Yang YH, Callow MJ, Speed TP. (2002). *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Statistica Sinica 12: 111-140.

Eckardt N. A. (2002). *Probing the Mysteries of Lignin Biosynthesis: The Crystal Structure of Caffeic Acid/5-Hydroxyferulic Acid 3/5-O-Methyltransferase Provides New Insights.* PLANT CELL, 14: 1185 - 1189.

Egertsdotter U, van Zyl LM, MacKay J, Peter G, Kirst M, Clark C, Whetten R, Sederoff R.. (2004). *Gene Expression during Formation of Earlywood and Latewood in Loblolly Pine: Expression Profiles of 350 Genes.* Plant biol 6: 654-663

Eliasson A, Gass N, Mundel C, Baltz R, Krauter R, Evrard JL, Steinmetz A. (2000). *Molecular and expression analysis of a LIM protein gene family from flowering plants.* Mol Gen Genet. 264:257-67.

Finkelstein DB, Gollub J, Cherry JM. (2002) *Normalization and systematic measurement error in cDNA microarray data.* Journal of Computational and Graphical Statistics, 5

Franke R, Hemm MR, Denault JW, Ruegger MO, Humphreys JM, Chapple C. (2002). *Changes in secondary metabolism and deposition of an unusual lignin in the ref8 mutant of Arabidopsis.* The Plant Journal 30:47-59

Futschik ME, Crompton T. (2005). *OLIN: Optimized normalization, visualization and quality testing for two-channel microarray data.* Bioinformatics, 21:1724-6

Futschik ME, Crompton T. (2004). *Model selection and efficiency testing for normalization of cDNA microarray data.* Genome Biology,5:R60

Guillet-Claude C, Isabel N, Pelgas B, Bousquet J. (2004). *The evolutionary implications of knox-I gene duplications in conifers: correlated evidence from phylogeny, gene mapping, and analysis of functional divergence.* Mol. Biol. Evol. 21: 2232-2245.

Hertzberg M, Aspeborg H, Schrader J, Blomqvist K, Andersson A, Bhalerao R, Marchant A, Bennett M, Uhlen M, Teeri TT, Lundeberg J, Sundberg B, Nilsson P, Sandberg G. 2001. *A transcriptional roadmap to xylogenesis*. Proc. Natl. Acad. Sci. USA 98: 14732-14737.

Humphreys JM, Chapple C. (2002). *Rewriting the lignin roadmap*. Plant Biol. 5, 224–229.

Ingouff M, Farbos I, Wiweger M, von Arnold S. (2003). *The molecular characterization of PaHB2, a homeobox gene of the HD-GL2 family expressed during embryo development in Norway spruce.* J Exp Bot. 54:1343-50.

Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada K, Ebinuma H. (2000). *Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis.* Plant Journal 22; 289-301

Kerr M, Martin M, Churchill G. (2001) *Analysis of variance for gene expression microarray data.* J Comp Biol 7: 819-837.

Kirst M, Myburg AA, De Leon JP, Kirst ME, Scott J, Sederoff R. (2004). Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiol. 135: 2368-2378

Leung and Cavalieri. (2003) *Fundamentals of cDNA microarray data analysis*. Trends Genet. 19: 649-59.

Mundel C, Baltz R, Eliasson A, Bronner R, Grass N, Krauter R, Evrard JL, Steinmetz A. (2000). *A LIM-domain protein from sunflower is localized to the cytoplasm and/or nucleus in a wide variety of tissues and is associated with the phragmoplast in dividing cells.* Plant Molecular Biology 42: 291-302,

Park T., Yi S-G, Kang S-H, Lee SY Lee Y-S and Simon S. (2003). *Evaluation of normalization methods for microarray data.* BMC Bioinformatics 4:33

Pavy N, Paule C, Parsons L, Crow JA, Morency MJ , Cooke J, Johnson JE, Noumen E, Guillet-Claude C, Butterfield Y, Barber S, Yang G, Liu J, Stott J, Kirkpatrick R, Siddiqui A, Holt R, Marra M, Seguin A, Retzel E, Bousquet J and MacKay J. (2004). *Large-scale statistical analysis of pine xylem ESTs lead to the discovery of regulatory genes expressed in root xylem.* Plant Molecular Biology. 57: 203-224

Pavy N, Paule C, Parsons L, Crow JA, Morency MJ, Cooke J, Johnson JE, Noumen E, Guillet-Claude C, Butterfield Y, Barber S, Yang G, Liu J, Stott J, Kirkpatrick R, Siddiqui A, Holt R, Marra M, Seguin A, Retzel E, Bousquet J and MacKay J. (2005). *Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters*. BMC Genomics. 6(1):144.

Peter G, Neale D. (2004) Molecular *basis for the evolution of xylem lignification*. Curr Opin Plant Biol. 7:737-42.

Plomion C, Richardson T, MacKay J. (2005). *Advances in forest tree genomics*. New Phytologist 166 : 713 -

Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. (2002) *BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.* Genome Biol. 3 (8)

Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. (1996). *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes*. Proc Natl Acad Sci U S A. 93:10614-9.

Tarca AL, Cooke JE, Mackay J. (2005). A robust neural networks approach for spatial and intensity dependent normalization of cDNA microarray data. Bioinformatics 21:2674-83

Tusher VG, Tibshirani R, Chu G. (2001) *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences USA 98: 5116-5121.

Whetten R, Sun YH, Zhang Y, Sederoff R. (2001). *Functional genomics and cell wall biosynthesis in loblolly pine.* Plant Mol Biol 47: 275-291

Whetten RW, MacKay JJ, Sederoff RR. (1998). *Recent advances in understanding ligning biosynthesis.* Annu Rev Plant Physiol Plant Mol Biol. 49:585-609.

Weier T Stocking CR, Barbour MG, Rost TL, (1974). *Botany: an introduction to plant biology.* John Wiley and Sons, New York, N.Y. 580

Wu H, Kerr MK, Cui X, and Churchill GA. (2002). *MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments.* An Overview of Methods and Software, Parmigiani G, Garret ES, Irizarry RA, Zeger SL, Springer, N.Y

Xue B, Charest PJ, Devantier Y, Rutledge RG. (2003). Characterization of a MYBR2R3 gene from black spruce (Picea mariana) that shares functional conservation with maize C1. Molecular Genetics and Genomics. 270: 78 - 86

Xiao Y, Segal MR, Rabert D, Ahn AH, Anand P, Sangameswaran L, Hu D, Hunt CA. (2002) *Assessment of differential gene expression in human peripheral nerve injury.* BMC Genomics 3:28

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. (2002). *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation*. Nucleic Acids Res. 30: 15.

van Zyl, L., von Arnold, S., Bozhkov. P., Chen, Y., Egertsdotter, U., MacKay, J., Sederoff, R., Shen, J., Zelena, L., Clapham, D. (2002). *Heterologous array analysis in Pinaceae: hybridization of Pinus taeda cDNA arrays with cDNA from needles and embryogenic cultures of P. taeda, P. sylvestris or Picea abies.* Functional and Comparative Genomics, 3: 306-318

Sites Internet cités

1. http://www.mysql.com

2. http://www.php.net

3. http://www.apache.org

4. http://www.mged.org/

5. http://ihome.cuhk.edu.hk/~b400559/arraysoft.html

6. http://amap.cirad.fr/architecture/organo/organo1.html

7. http://spcmib.ups-tlse.fr/themes/theme1/detail/trav1a.html

8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1173086

Annexe 1 – Documentation SLIMS

Documentations du SLIMS : http://132.203.160.236/demo/Documentation.php

Annexe 2 - Gènes différentiellement exprimés identifiés à l'aide de l'analyse SAM.

A. Lignée 2

MNID	+/-	Annotation	q-val.	M	Ratio
MN5177289	-	Contig10220 [1:-] [2:-] [3:-]	0,00	-1,01	-2,02
MN5177260	-	Contig7110 [1:weakly similar to UP|Q8C5F9 (Q8C5F9) Mus musculus adult male adrenal gland cDNA] [2:putative membrane protein [Oryza sativa (japonica cultivar-group)] gb|AAO73245.1| putative membrane protein [Oryza sativa (japonica cultivar-group)]] [3:P	0,00	-0,97	-1,96
MN5195432	-	Contig9006 [1:-] [2:-] [3:-]	0,00	-0,89	-1,86
MN5238365	-	Contig5810* [1:-] [2:-] [3:-]	0,00	-0,84	-1,79
MN5177212	-	Contig8858 [1:similar to UP|TPIC_FRAAN (Q9M4S8) Triosephosphate isomerase] [2:triosephosphate isomerase [Fragaria x ananassa] sp|Q9M4S8|TPIC_FRAAN Triosephosphate isomerase] [3:PF00121.7;TIM]	0,00	-0,81	-1,75
MN5195855	-	Contig5731* [1:-] [2:-] [3:-]	0,00	-0,81	-1,75
MN5196858	-	Contig9824 [1:-] [2:-] [3:-]	0,00	-0,81	-1,75
MN5236092	-	Contig7487 [1:similar to UP|Q9ZP84 (Q9ZP84) Heat shock protein 17.4] [2:small heat-shock protein [Pseudotsuga menziesii]] [3:PF00011.8;HSP20]	0,00	-0,75	-1,68
MN5193809	-	Contig4300 [1:-] [2:unknown protein [Arabidopsis thaliana] gb|AAM20141.1| unknown protein [Arabidopsis thaliana] ref|NP_973868.1| expressed protein [Arabidopsis thaliana] ref|NP_173410.1| expressed protein [Arabidopsis thaliana] ref|NP_973869.1| expresse	4,02	-0,69	-1,62
MN5195544	-	Contig9170 [1:similar to UP|Q9FIU5 (Q9FIU5) Serine/threonine-specific protein kinase-like protein] [2:putative serine/threonine-specific protein kinase [Oryza sativa (japonica cultivar-group)] dbj|BAD03013.1| putative serine/threonine-specific protein k	4,02	-0,68	-1,60

MN5255771	-	Contig6714 [1:similar to UP|Q6RSS6 (Q6RSS6) Defensin] [2:putative gamma-thionin protein [Picea abies] pir||T14866 probable gamma-thionin precursor SPI1 - Norway spruce] [3:PF00304.8;Gamma-thionin]	0,00	-0,62	-1,54
MN5161830	-	Contig1390 [1:-] [2:-] [3:-]	0,00	-0,57	-1,49
MN5161824	-	Contig13729* [1:-] [2:-] [3:-]	0,00	-0,54	-1,45
MN5171347	-	Contig6724 [1:homologue to UP|Q6WSR8 (Q6WSR8) Class IV chitinase Chia4-Pa2] [2:putative class IV chitanase [Picea abies]] [3:PF00182.8;Glyco_hydro_19]	3,82	-0,52	-1,43
MN5169388	-	Contig13895 [1:ribosomal protein S12 [1:Pinus koraiensis]] [2:ribosomal protein S7 [2:Pinus thunbergii] pir||T07582 ribosomal protein S7 - Japanese black pine chloroplast sp|P41652|RR7_PINTH Chloroplast 30S ribosomal protein S7 dbj|BAA04458.1| ribosoma	4,02	-0,49	-1,40
MN5238447	-	Contig12131 [1:-] [2:-] [3:-]	0,00	-0,47	-1,38
MN5239809	-	Contig12924 [1:-] [2:-] [3:-]	0,00	-0,47	-1,38
MN5253163	-	Contig8835 [1:UP|Q9M4R0 (Q9M4R0) Ubiquitin conjugating protein] [2:ubiquitin conjugating protein [Avicennia marina]] [3:PF00179.14;UQ_con]	0,00	-0,44	-1,36
MN5176391	-	Contig22 [1:weakly similar to UP|Q6J163 (Q6J163) Nodulin-like protein 5NG4] [2:putative nodulin MtN21 [Oryza sativa (japonica cultivar-group)] dbj|BAD07647.1| putative nodulin MtN21 [Oryza sativa (japonica cultivar-group)] dbj|BAD07925.1| putative nodul	0,00	-0,41	-1,33
MN5234527	-	Contig797 [1:-] [2:-] [3:-]	0,00	-0,40	-1,32
MN5157728	-	Contig13996 [1:-] [2:-] [3:-]	5,46	-0,39	-1,31

Les gènes sélectionnés proviennent de l'analyse SAM à partir d'un FDR 21,28% et ayant un q-value plus petit que 5%. Ils sont ordonnés selon le ratio de M.

B. Lignée 8

MNID	+/-	Annotation	q-val.	M	Ratio
MN5250989	-	Contig11342 [1:similar to UP|Q6PXE1 (Q6PXE1) Auxin-repressed protein] [2:-] [3:-]	4,07	-1,02	2,02
MN5177289	-	Contig10220 [1:-] [2:-] [3:-]	0,00	-0,74	1,68
MN5191560	-	Contig10528 [1:homologue to UP|CB2A_PINSY (P15193) Chlorophyll a-b binding protein type II 1A] [2:unnamed protein product [2:Pinus sylvestris] sp|P15193|CB2A_PINSY Chlorophyll a-b binding protein type II 1A] [3:PF00504.11;Chloroa_b-bind]	0,00	-0,64	1,56
MN5195855	-	Contig5731* [1:-] [2:-] [3:-]	0,00	-0,64	1,56

MN5171347	-	Contig6724 [1:homologue to UP|Q6WSR8 (Q6WSR8) Class IV chitinase Chia4-Pa2] [2:putative class IV chitanase [2:Picea abies]] [3:PF00182.8;Glyco_hydro_19]	0,00	-0,63	1,55
MN5235638	-	Contig7847 [1:-] [2:chlorophyll a/b-binding protein [2:Pinus thunbergii] pir||S22522 chlorophyll a/b-binding protein (cab-6) precursor - Japanese black pine] [3:PF00504.11;Chloroa_b-bind]	4,07	-0,63	1,55
MN5249369	-	Contig8955 [1:] [2:putative Pollen specific protein C13 precursor [Oryza sativa (japonica cultivar-group)] ref|NP_921099.1| putative Pollen specific protein C13 precursor [Oryza sativa (japonica cultivar-group)] gb|AAN31783.1| Putataive pollen specific p	4,07	-0,59	1,51
MN5196866	-	Contig3979 [1:] [2:OSJNBa0070C17.23 [2:Oryza sativa (japonica cultivar-group)] emb|CAE05216.3| OSJNBa0070C17.23 [2:Oryza sativa (japonica cultivar-group)]] [3:PF02705.5;K_trans]	4,07	-0,58	1,50
MN5240262	-	Contig12527 [1:-] [2:-] [3:-]	4,07	-0,56	1,47
MN5235024	-	Contig4812* [1:-] [2:-] [3:-]	0,00	-0,54	1,46
MN5233804	-	Contig9482 [1:] [2:-] [3:-]	0,00	-0,54	1,46
MN5242210	-	Contig12445 [1:similar to UP|O48767 (O48767) Expressed protein (At2g32980/T21L14.8)] [2:unknown [2:Arabidopsis thaliana]] [3:-]	3,88	-0,52	1,43
MN5194939	-	Contig9058 [1:-] [2:glutamyl-tRNA(Gln) amidotransferase B family protein [2:Arabidopsis thaliana] gb|AAL67097.1| At1g48520/T1N15_12 [2:Arabidopsis thaliana] gb|AAL06883.1| At1g48520/T1N15_12 [2:Arabidopsis thaliana] gb|AAG29096.1| Glu-tRNA(Gln) amidotran	4,22	-0,50	1,41
MN5193330	-	Contig9207* [1:-] [2:-] [3:-]	4,22	-0,46	1,38
MN5193444	-	Contig3630* [1:-] [2:-] [3:-]	4,07	-0,46	1,37
MN5196602	-	Contig4166 [1:-] [2:unknown [2:Arabidopsis thaliana] dbj|BAB11109.1| unnamed protein product [2:Arabidopsis thaliana] ref|NP_196885.1| glutaredoxin family protein [2:Arabidopsis thaliana] gb|AAL31176.1| AT5g13810/MAC12_24 [2:Arabidopsis thaliana] gb|AAK6	4,22	-0,45	1,37
MN5238136	-	Contig8764 [1:similar to UP|O65844 (O65844) Protein phosphatase 1] [2:protein phosphatase 1] [3:PF00149.14;Metallophos]	0,00	-0,43	1,35
MN5255489	-	Contig11908 [1:] [2:-] [3:-]	0,00	-0,43	1,35
MN5236092	-	Contig7487 [1:similar to UP|Q9ZP84 (Q9ZP84) Heat shock protein 17.4] [2:small heat-shock protein [2:Pseudotsuga menziesii]] [3:PF00011.8;HSP20]	4,22	-0,41	1,33
MN5162480	-	Contig7879 [1:similar to UP|RBS_LARLA (P16031) Ribulose bisphosphate carboxylase small chain] [2:ribulose-1] [3:PF00101.8;RuBisCO_small]	4,07	-0,39	1,31
MN5171164	-	Contig13375 [1:-] [2:-] [3:-]	4,07	-0,39	1,31
MN5191991	-	Contig3725 [1:-] [2:MYB-related transcription factor [2:Antirrhinum majus] pir||T17027 MYB-related transcription factor - garden snapdragon] [3:-]	2,85	-0,38	1,30
MN5242050	-	Contig6992* [1:-] [2:-] [3:-]	4,07	-0,37	1,29

MN5191492	-	Contig10522 [1:similar to UP|Q9FM07 (Q9FM07) Permease 1] [2:putative permease [2:Oryza sativa (japonica cultivar-group)] dbj|BAC99450.1| putative permease [2:Oryza sativa (japonica cultivar-group)]] [3:-]	0,00	-0,36	1,29
MN5190543	-	Contig4931* [1:-] [2:-] [3:-]	4,95	-0,35	1,27
MN5193445	-	Contig9239* [1:-] [2:-] [3:-]	4,07	-0,32	1,24
MN5159358	-	Contig898 [1:similar to GP|4099833|gb| bifunctional nuclease {Zinnia elegans}] [2:bifunctional nuclease [2:Zinnia elegans]] [3:PF02265.6;S1-P1_nuclease]	4,07	-0,30	1,23

Les gènes sélectionnés proviennent de l'analyse SAM de la lignée 8 avec un FDR de 16,87% et ayant un q-value plus petit que 5%. Ils sont ordonnés selon le ratio de M.

C. Lignée 21

MNID	+/-	Annotation	q-val.	M	Ratio
MN5182493	+	Contig10776 [1:similar to UP|Q9FVF0 (Q9FVF0) Pyruvate decarboxylase] [2:BTH-induced ERF transcriptional factor 3 [2:Oryza sativa (indica cultivar-group)] ref|XP_467107.1| putative AP2-related transcription factor [2:Oryza sativa (japonica cultivar-group)	2.86	1,00	2,01
MN5235638	+	Contig7847 [1:-] [2:chlorophyll a/b-binding protein [2:Pinus thunbergii] pir||S22522 chlorophyll a/b-binding protein (cab-6) precursor - Japanese black pine] [3:PF00504.11;Chloroa_b-bind]	0.0	0,73	1,66
MN5191560	+	Contig10528 [1:homologue to UP|CB2A_PINSY (P15193) Chlorophyll a-b binding protein type II 1A] [2:unnamed protein product [2:Pinus sylvestris] sp|P15193|CB2A_PINSY Chlorophyll a-b binding protein type II 1A] [3:PF00504.11;Chloroa_b-bind]	0.0	0,67	1,59
MN5249914	+	Contig8952 [1:] [2:disease resistance gene [2:Pinus sylvestris]] [3:PF01846.9;FF]	0.0	0,66	1,57
MN5249369	+	Contig8955 [1:] [2:putative Pollen specific protein C13 precursor [2:Oryza sativa (japonica cultivar-group)] ref|NP_921099.1| putative Pollen specific protein C13 precursor [2:Oryza sativa (japonica cultivar-group)] gb|AAN31783.1| Putataive pollen specif	0.0	0,63	1,55
MN5237237	+	Contig4813 [1:] [2:unknown [2:Xerophyta humilis]] [3:-]	0.0	0,60	1,52
MN5249533	+	Contig11298 [1:weakly similar to UP|Q8GSL4 (Q8GSL4) Origin recognition complex subunit 6-like protein] [2:origin recognition complex subunit 6 -like protein [2:Oryza sativa (japonica cultivar-group)] dbj|BAC22351.1| origin recognition complex subunit 6	0.0	0,60	1,52

MN5190745	+	Contig8973 [1:-] [2:ubiquitin-like protein [2:Phaseolus vulgaris] pir||T12035 polyubiquitin 4.4 - kidney bean] [3:PF00240.11;ubiquitin]	0.0	0,59	1,50
MN5248924	+	Contig8511 [1:-] [2:water-stress-inducible protein LP3 - loblolly pine gb|AAB07493.1| water deficit inducible protein LP3] [3:-]	0.0	0,58	1,49
MN5196269	+	Contig8766 [1:similar to UP|P93156 (P93156) Cellulose synthase (Fragment)] [2:cellulose synthase (EC 2.4.1.-) catalytic chain celA2 - upland cotton (fragment) gb|AAB37767.1| cellulose synthase] [3:PF03552.3;Cellulose_synt]	0.0	0,54	1,46
MN5241749	+	Contig8379 [1:GB|AAA34124.1|170354|TOBUBI4A pentameric polyubiquitin {Nicotiana sylvestris;}] [2:unnamed protein product [2:Pisum sativum] gb|AAK96602.1| AT4g05320/C17L7_240 [2:Arabidopsis thaliana] gb|AAD03344.1| ubiquitin [2:Pisum sativum] pir||UQPM	0.0	0,53	1,44
MN5242005	+	Contig8832 [1:weakly similar to UP|Q6PY83 (Q6PY83) Major ampullate spidroin 2-1 (Fragment)] [2:-] [3:-]	0.0	0,51	1,42
MN5249053	+	Contig11285 [1:] [2:-] [3:-]	0.0	0,50	1,41
MN5182745	+	Contig10734* [1:-] [2:-] [3:-]	2.28	0,48	1,40
MN5195778	+	Contig7709 [1:UP|Q9AR09 (Q9AR09) Ubiquitin fused to ribosomal protein L40] [2:ubiquitin / ribosomal protein CEP52 - wood tobacco gb|AAA34064.1| ubiquitin fusion protein] [3:PF00240.11;ubiquitin]	0.0	0,48	1,40
MN5181797	+	Contig6685 [1:] [2:-] [3:-]	0.0	0,47	1,39
MN5236092	+	Contig7487 [1:similar to UP|Q9ZP84 (Q9ZP84) Heat shock protein 17.4] [2:small heat-shock protein [2:Pseudotsuga menziesii]] [3:PF00011.8;HSP20]	0.0	0,46	1,38
MN5182713	+	Contig10754 [1:similar to UP|PSBP_CUCSA (Q9SLQ8) Oxygen-evolving enhancer protein 2] [2:23 kDa oxygen evolving protein of photosystem II [2:Solanum tuberosum] sp|P93566|PSBP_SOLTU Oxygen-evolving enhancer protein 2] [3:PF01789.6;PsbP]	0.0	0,45	1,37
MN5243035	+	Contig12503* [1:-] [2:-] [3:-]	2.06	0,45	1,36
MN5163062	+	Contig7897 [1:similar to UP|Q9XQB2 (Q9XQB2) Chlorophyll a/b binding protein CP29] [2:chlorophyll a/b binding protein CP29 [2:Vigna radiata]] [3:PF00504.11;Chloroa_b-bind]	0.0	0,44	1,36
MN5239177	+	Contig8643 [1:-] [2:At3g53980 [2:Arabidopsis thaliana] emb|CAB88360.1| putative protein [2:Arabidopsis thaliana] gb|AAO00805.1| putative protein [2:Arabidopsis thaliana] pir||T45938 hypothetical protein F5K20.280 - Arabidopsis thaliana ref|NP_850700.1| p	1.59	0,43	1,35
MN5239863	+	Contig12838 [1:similar to UP|Q9FHG8 (Q9FHG8) Similarity to C3HC4-type RING zinc finger protein] [2:-] [3:-]	0.0	0,43	1,34
MN5232943	+	Contig8945 [1:weakly similar to UP|Q8VWQ1 (Q8VWQ1) Dehydration-induced protein RD22-like protein] [2:dehydration-induced protein RD22-like protein [2:Gossypium hirsutum]] [3:PF03181.5;BURP]	2.96	0,42	1,34
MN5245227	+	Contig8660 [1:] [2:unknown [2:Arabidopsis thaliana]] [3:-]	0.0	0,42	1,34

MN5256787	+	Contig8768 [1:similar to UP|Q6PXE1 (Q6PXE1) Auxin-repressed protein] [2:dormancy-associated protein [2:Codonopsis lanceolata]] [3:PF05564.2;Auxin_repressed]	0.0	0,41	1,33
MN5196602	+	Contig4166 [1:-] [2:unknown [2:Arabidopsis thaliana] dbj|BAB11109.1| unnamed protein product [2:Arabidopsis thaliana] ref|NP_196885.1| glutaredoxin family protein [2:Arabidopsis thaliana] gb|AAL31176.1| AT5g13810/MAC12_24 [2:Arabidopsis thaliana] gb|AAK6	0.0	0,40	1,32
MN5160448	+	Contig8174 [1:similar to UP|Q9LKW3 (Q9LKW3) Dehydration-induced protein ERD15] [2:early response to dehydration 15-like protein [2:Pseudotsuga menziesii var. menziesii] gb|AAV92291.1| early response to dehydration 15-like protein [2:Pseudotsuga menziesi	2.47	0,40	1,32
MN5256941	+	Contig11869* [1:-] [2:-] [3:-]	0.0	0,39	1,31
MN5164545	+	Contig13919 [1:similar to UP|Q86EX7 (Q86EX7) Clone ZZD1326 mRNA sequence] [2:unknown [2:Arabidopsis thaliana] gb|AAM51593.1| At1g19310/F18O14_14 [2:Arabidopsis thaliana] ref|NP_564078.1| zinc finger (C3HC4-type RING finger) family protein [2:Arabidopsis	2.96	0,39	1,31
MN5160221	+	Contig8815 [1:-] [2:Calmodulin (CaM) emb|CAA09302.1| calmodulin 3 protein [2:Capsicum annuum] dbj|BAB61908.1| calmodulin NtCaM2 [2:Nicotiana tabacum] dbj|BAB61907.1| calmodulin NtCaM1 [2:Nicotiana tabacum] gb|AAA34144.1| calmodulin emb|CAC84563.1| putati	0.0	0,38	1,30
MN5237079	+	Contig5030 [1:] [2:putative formamidase [2:Arabidopsis thaliana] ref|NP_568029.1| formamidase] [3:PF03069.4;FmdA_AmdA]	2.22	0,37	1,30
MN5238381	+	Contig638* [1:-] [2:-] [3:-]	0.0	0,37	1,29
MN5249014	+	Contig7121 [1:weakly similar to UP|Q6NN03 (Q6NN03) At5g04080] [2:-] [3:-]	0.0	0,36	1,29
MN5182224	+	Contig10097* [1:-] [2:-] [3:-]	2.47	0,35	1,28
MN5181972	+	Contig7759 [1:similar to UP|Q08671 (Q08671) Peroxidase precursor] [2:peroxidase pir||T10790 peroxidase (EC 1.11.1.7) - upland cotton] [3:PF00141.10;peroxidase]	0.0	0,35	1,28
MN5196571	+	Contig9788 [1:] [2:putative DNA-binding protein [2:Oryza sativa (japonica cultivar-group)] dbj|BAD81093.1| putative DNA-binding protein [2:Oryza sativa (japonica cultivar-group)]] [3:-]	2.28	0,35	1,27
MN5258553	+	Contig8758 [1:UP|Q70XK1 (Q70XK1) ADP-ribosylation factor 1-like protein] [2:ADP-ribosylation factor [2:Oryza sativa (japonica cultivar-group)] emb|CAD48129.2| ADP-ribosylation factor 1-like protein [2:Hordeum vulgare subsp. vulgare] sp|P51823|ARF_ORYSA A	0.0	0,33	1,26
MN5164094	+	Contig7254 [1:similar to UP|PLAS_VICFA (P00288) Plastocyanin] [2:unnamed protein product [2:Spinacia oleracea] pir||CUSP plastocyanin precursor - spinach sp|P00289|PLAS_SPIOL Plastocyanin] [3:PF00127.9;Copper-bind]	2.17	0,32	1,25

MN5177212	+	Contig8858 [1:similar to UP|TPIC_FRAAN (Q9M4S8) Triosephosphate isomerase] [2:triosephosphate isomerase [2:Fragaria x ananassa] sp|Q9M4S8|TPIC_FRAAN Triosephosphate isomerase] [3:PF00121.7;TIM]	2.78	0,32	1,24
MN5164514	+	Contig8962 [1:homologue to UP|Q71F44 (Q71F44) Eukaryotic translation initiation factor 5A isoform VIII (Fragment)] [2:eukaryotic translation initiation factor 5A isoform VIII [2:Hevea brasiliensis]] [3:PF01287.8;eIF-5a]	2.61	0,31	1,24
MN5174710	+	Contig8084 [1:similar to UP|Q8L6A8 (Q8L6A8) Aspartic proteinase] [2:aspartic proteinase [2:Vigna unguiculata] gb|AAQ14346.1| aspartic proteinase [2:Vigna unguiculata]] [3:PF00026.12;Asp]	2.28	0,31	1,24
MN5192528	+	Contig10594 [1:-] [2:calmodulin-binding ion transporter-like protein [2:Physcomitrella patens]] [3:-]	3.03	0,28	1,21
MN5194939	+	Contig9058 [1:-] [2:glutamyl-tRNA(Gln) amidotransferase B family protein [2:Arabidopsis thaliana] gb|AAL67097.1| At1g48520/T1N15_12 [2:Arabidopsis thaliana] gb|AAL06883.1| At1g48520/T1N15_12 [2:Arabidopsis thaliana] gb|AAG29096.1| Glu-tRNA(Gln) amidotran	2.06	0,27	1,20
MN5192021	+	Contig4114 [1:] [2:unknown protein [2:Arabidopsis thaliana] gb|AAO42075.1| unknown protein [2:Arabidopsis thaliana] ref|NP_178129.1| expressed protein [2:Arabidopsis thaliana] gb|AAD55470.1| Unknown protein [2:Arabidopsis thaliana] pir||G96832 hypothetic	2.69	0,25	1,19

Les gènes sélectionnés proviennent de l'analyse SAM de la lignée21 à partir d'un FDR 16,14% et ayant un q-value plus petit que 5%. Ils sont ordonnés selon le ratio de M.

D. Lignées 4 et 8 combinées

MNID	+/-	Annotation	q-val,	M	Ratio
MN5253946	-	Contig11809* [1:-] [2:-] [3:-]	0,00	-0,35	1,27
MN5169439	+	Contig14024 [1:weakly similar to UP|Q9LKP0 (Q9LKP0) RNA-dependent RNA polymerase] [2:RNA-directed RNA polymerase-like protein [Arabidopsis thaliana] gb|AAF73959.1| SGS2 [Arabidopsis thaliana] gb|AAG52184.1| putative RNA-directed RNA polymerase; 73997-69	0,00	0,32	1,25
MN5254450	+	Contig2172 [1:weakly similar to UP|Q9LCZ9 (Q9LCZ9) Photoassimilate-responsive protein PAR-1b-like protein] [2:PAR-1b [Nicotiana tabacum] pir||S62699 photoassimilate-responsive protein PAR-1b precursor - common tobacco] [3:PF06521.1;PAR1]	0,84	0,30	1,23
MN5237431	+	Contig11953* [1:-] [2:-] [3:-]	0,84	0,29	1,23
MN5233623	+	Contig9577* [1:-] [2:-] [3:-]	1,20	0,28	1,22
MN5244505	+	Contig2794* [1:-] [2:-] [3:-]	1,31	0,27	1,21
MN5169756	+	Contig8068* [1:-] [2:-] [3:-]	1,20	0,27	1,20
MN5258455	+	Contig2663* [1:-] [2:-] [3:-]	0,84	0,26	1,20

MN5254889	+	Contig1960 [1:] [2:-] [3:-]	0,00	0,26	1,20
MN5237248	+	Contig11962* [1:-] [2:-] [3:-]	0,84	0,26	1,20
MN5235876	+	Contig5861 [1:weakly similar to UP|Q9MA24 (Q9MA24) T5E21.9] [2:expressed protein [Arabidopsis thaliana] pir||E86280 protein T5E21.9 [2:imported] - Arabidopsis thaliana gb|AAF63177.1| T5E21.9 [Arabidopsis thaliana]] [3:-]	1,24	0,26	1,20
MN5194697	+	Contig9031* [1:-] [2:-] [3:-]	0,72	0,26	1,20
MN5253155	+	Contig11097 [1:weakly similar to UP|ENL1_ARATH (Q9SK27) Early nodulin-like protein 1 precursor (Phytocyanin-like protein)] [2:putative blue copper binding protein [Oryza sativa (japonica cultivar-group)] dbj|BAD03003.1| putative blue copper binding prot	1,20	0,26	1,19
MN5255519	+	Contig11910 [1:-] [2:-] [3:-]	0,84	0,25	1,19
MN5169821	+	Contig14133* [1:-] [2:-] [3:-]	0,84	0,25	1,19
MN5253032	+	Contig11177* [1:-] [2:-] [3:-]	0,84	0,25	1,19
MN5254313	+	Contig11759* [1:-] [2:-] [3:-]	1,24	0,25	1,19
MN5234677	+	Contig8082 [1:UP|Q6GUG6 (Q6GUG6) Cellulose synthase catalytic subunit] [2:cellulose synthase 6 [2:Populus tremuloides]] [3:PF03552.3;Cellulose_synt]	0,00	0,25	1,19
MN5254032	+	Contig2069 [1:] [2:-] [3:-]	0,84	0,25	1,19
MN5171411	+	Contig28 [1:] [2:Putative ubiquitin protein [Oryza sativa (japonica cultivar-group)] gb|AAM22708.1| Putative ubiquitin protein [Oryza sativa (japonica cultivar-group)]] [3:PF00627.16;UBA]	1,31	0,25	1,19
MN5252328	+	Contig10959 [1:-] [2:-] [3:-]	0,72	0,24	1,18
MN5237895	+	Contig3924 [1:] [2:-] [3:-]	0,00	0,24	1,18
MN5164191	+	Contig13831* [1:-] [2:-] [3:-]	1,24	0,24	1,18
MN5237421	+	Contig5228 [1:] [2:putative Noc3p [Oryza sativa (japonica cultivar-group)]] [3:PF03914.4;CBF]	0,84	0,24	1,18
MN5192633	+	Contig10607* [1:-] [2:-] [3:-]	1,20	0,24	1,18
MN5256238	+	Contig2020 [1:-] [2:-] [3:-]	0,84	0,24	1,18
MN5196616	+	Contig9793* [1:-] [2:-] [3:-]	0,84	0,24	1,18
MN5176058	+	Contig13752 [1:weakly similar to UP|UFO5_MANES (Q40287) Flavonol 3-O-glucosyltransferase 5 (UDP-glucose flavonoid 3-O-glucosyltransferase 5)] [2:putative flavonol 3-O-glucosyltransferase [Oryza sativa (japonica cultivar-group)] dbj|BAC83989.1| putativ	0,84	0,24	1,18
MN5253745	+	Contig6968* [1:-] [2:-] [3:-]	0,00	0,24	1,18
MN5194633	+	Contig5471* [1:-] [2:-] [3:-]	0,00	0,24	1,18
MN5163637	+	Contig1841* [1:-] [2:-] [3:-]	0,00	0,24	1,18
MN5238519	+	Contig12112 [1:similar to GB|AAK93087.1|15291637|AY051663 LD21247p {Drosophila melanogaster;}] [2:Hypothetical protein CBG09177 [2:Caenorhabditis briggsae]] [3:PF00270.15;DEAD]	1,20	0,24	1,18
MN5246166	+	Contig2860* [1:-] [2:-] [3:-]	1,31	0,24	1,18
MN5251589	+	Contig11029* [1:-] [2:-] [3:-]	0,00	0,24	1,18
MN5255329	+	Contig6267 [1:weakly similar to GB|AAM70554.1|21700861|AY124845 At1g75220/F22H5_6 {Arabidopsis thaliana;}] [2:integral membrane protein [2:Beta vulgaris] pir||T14545 probable sugar transporter protein - beet] [3:PF00083.11;Sugar_tr]	0,72	0,23	1,18
MN5175280	+	Contig2491 [1:similar to UP|Q7QF64 (Q7QF64) AgCP13728 (Fragment)] [2:unknown protein [Arabidopsis thaliana] gb|AAK68815.1| Unknown protein [Arabidopsis thaliana]] [3:PF07933.1;DUF1681]	0,72	0,23	1,18
MN5257573	+	Contig11541* [1:-] [2:-] [3:-]	0,84	0,23	1,18
MN5241947	+	Contig2970* [1:-] [2:-] [3:-]	0,84	0,23	1,18
MN5234725	+	Contig1454* [1:-] [2:-] [3:-]	1,20	0,23	1,17
MN5254531	+	Contig2071 [1:similar to UP|TPK1_MOUSE (Q9R0M5) Thiamin pyrophosphokinase 1 (Thiamine pyrophosphokinase 1) (mTPK1)] [2:putative thiamin pyrophosphokinase 1 [Oryza sativa (japonica cultivar-group)] dbj|BAD87327.1| putative thiamin pyrophosphokinase 1 [O	0,84	0,23	1,17

MN5194031	+	Contig5385* [1:-] [2:-] [3:-]	0,72	0,23	1,17
MN5164495	+	Contig13930 [1:-] [2:-] [3:-]	0,00	0,23	1,17
MN5233895	+	Contig6108* [1:-] [2:-] [3:-]	1,31	0,23	1,17
MN5255334	+	Contig1918 [1:] [2:putative TIR/NBS/LRR disease resistance protein [2:Pinus taeda]] [3:-]	0,72	0,23	1,17
MN5234239	+	Contig4565* [1:-] [2:-] [3:-]	0,84	0,23	1,17
MN5255963	+	Contig5362* [1:-] [2:-] [3:-]	0,84	0,23	1,17
MN5243611	+	Contig5143* [1:-] [2:-] [3:-]	0,84	0,23	1,17
MN5237565	+	Contig6156 [1:similar to UP|PP12_ACECL (P48481) Serine/threonine protein phosphatase PP1 isozyme 2] [2:protein phosphatase 1] [3:PF00149.14;Metallophos]	1,31	0,23	1,17
MN5236301	+	Contig4270 [1:] [2:-] [3:-]	0,84	0,22	1,17
MN5193503	+	Contig9411* [1:-] [2:-] [3:-]	1,31	0,22	1,17
MN5158513	+	Contig2541 [1:-] [2:-] [3:-]	1,31	0,22	1,17
MN5237916	+	Contig12143* [1:-] [2:-] [3:-]	1,31	0,22	1,17
MN5258944	+	Contig7746 [1:] [2:pentatricopeptide (PPR) repeat-containing protein-like [2:Oryza sativa (japonica cultivar-group)] dbj|BAC99540.1| pentatricopeptide (PPR) repeat-containing protein-like [Oryza sativa (japonica cultivar-group)]] [3:PF03140.4;DUF247]	1,31	0,22	1,17
MN5195733	+	Contig6481 [1:] [2:putative vacuolar assembling protein [2:Ipomoea trifida]] [3:-]	1,31	0,22	1,17
MN5233196	+	Contig9891 [1:-] [2:putative NADPH oxidoreductase homolog [Oryza sativa (japonica cultivar-group)] dbj|BAD38525.1| putative NADPH oxidoreductase homolog [Oryza sativa (japonica cultivar-group)]] [3:PF00107.13;ADH_zinc_N]	0,84	0,22	1,16
MN5170326	+	Contig5355* [1:-] [2:-] [3:-]	1,20	0,22	1,16
MN5250378	+	Contig3720 [1:weakly similar to UP|Q93ZD6 (Q93ZD6) AT4g24880/F13M23_20] [2:P0672D08.12 [Oryza sativa (japonica cultivar-group)] dbj|BAB92128.1| hypothetical protein~similar to Arabidopsis thaliana chromosome 4] [3:-]	1,31	0,22	1,16
MN5235627	+	Contig12210 [1:weakly similar to GB|AAL15368.1|16323268|AY057738 At1g15670/F7H2_1 {Arabidopsis thaliana;}] [2:kelch repeat-containing F-box-like [Oryza sativa (japonica cultivar-group)] dbj|BAD25009.1| kelch repeat-containing F-box-like [Oryza sativa (	1,31	0,22	1,16
MN5236451	+	Contig12344 [1:similar to GP|10177634|dbj glycosyl transferase-like {Arabidopsis thaliana}] [2:glycosyl transferase family 1 protein-like [Oryza sativa (japonica cultivar-group)] dbj|BAD68686.1| glycosyl transferase family 1 protein-like [Oryza sativa (	0,00	0,22	1,16
MN5168994	+	Contig461 [1:similar to UP|Q93VE2 (Q93VE2) Peptide transporter] [2:oligopeptide transporter-like protein [Arabidopsis thaliana] pir||T47604 oligopeptide transporter-like protein - Arabidopsis thaliana] [3:PF00854.10;PTR2]	0,72	0,22	1,16
MN5257484	+	Contig11514* [1:-] [2:-] [3:-]	0,84	0,22	1,16
MN5252513	+	Contig8459* [1:-] [2:-] [3:-]	1,24	0,22	1,16
MN5239120	+	Contig1539 [1:] [2:-] [3:-]	1,31	0,22	1,16
MN5239508	+	Contig5034 [1:weakly similar to UP|Q7PC86 (Q7PC86) PDR7 ABC transporter] [2:putative PDR-like ABC transporter [Oryza sativa (japonica cultivar-group)] dbj|BAD05827.1| putative PDR-like ABC transporter [Oryza sativa (japonica cultivar-group)]] [3:PF01061	0,00	0,21	1,16
MN5250513	+	Contig11412 [1:similar to GB|AAO24565.1|27808570|BT003133 At3g53410 {Arabidopsis thaliana;}] [2:putative hydroxyproline-rich glycoprotein [Oryza sativa (japonica cultivar-group)] ref|NP_920425.1| putative hydroxyproline-rich glycoprotein [Oryza sativa	1,20	0,21	1,16
MN5195493	+	Contig9161* [1:-] [2:-] [3:-]	0,84	0,21	1,16
MN5259466	+	Contig11664 [1:] [2:-] [3:-]	0,72	0,21	1,16
MN5256816	+	Contig7542 [1:] [2:cinnamoyl-CoA reductase-like protein [Arabidopsis thaliana] ref|NP_194776.1| cinnamoyl-CoA reductase-related [Arabidopsis thaliana] gb|AAK68826.1| cinnamoyl-CoA reductase-like protein [Arabidopsis thaliana] pir||D85356 cinnamoyl-CoA re	0,00	0,21	1,15
MN5195321	+	Contig4164 [1:-] [2:-] [3:-]	0,84	0,21	1,15

MN5162292	+	Contig8238 [1:weakly similar to GB|AAL49941.1|17979249|AY070475 AT3g04610/F7O18_9 {Arabidopsis thaliana;}] [2:putative RNA binding protein [Oryza sativa] gb|AAL31692.1| putative RNA binding protein [Oryza sativa]] [3:PF00013.15;KH_1]	1,31	0,21	1,15
MN5259666	+	Contig4913 [1:similar to UP|Q7XVJ5 (Q7XVJ5) OJ000126_13.10 protein] [2:-] [3:-]	0,84	0,21	1,15
MN5239653	-	Contig12941 [1:-] [2:At3g05090/T12H1_5 [Arabidopsis thaliana] gb|AAN13212.1| unknown protein [Arabidopsis thaliana] gb|AAL07141.1| unknown protein [Arabidopsis thaliana] gb|AAM83246.1| AT3g05090/T12H1_5 [Arabidopsis thaliana] ref|NP_566246.1| transducin	0,00	-0,21	1,15
MN5251271	+	Contig11336 [1:similar to UP|BRI1_LYCES (Q8GUQ5) Brassinosteroid LRR receptor kinase precursor (tBRI1) (Altered brassinolide sensitivity 1) (Systemin receptor SR160)] [2:putative receptor protein kinase [Arabidopsis thaliana] gb|AAD20088.1| putative r	1,20	0,20	1,15
MN5243884	+	Contig12655* [1:-] [2:-] [3:-]	0,84	0,20	1,15
MN5161164	+	Contig3325 [1:-] [2:-] [3:-]	0,00	0,20	1,15
MN5237478	+	Contig4977* [1:-] [2:-] [3:-]	0,84	0,20	1,15
MN5253245	+	Contig11104 [1:weakly similar to UP|Q06979 (Q06979) Ocs-element binding factor 3.2] [2:bZIP transcription factor [2:Nicotiana tabacum]] [3:-]	1,31	0,20	1,15
MN5257107	+	Contig11855 [1:similar to UP|Q9LFP7 (Q9LFP7) Serine/threonine specific protein kinase-like] [2:-] [3:-]	1,31	0,20	1,15
MN5257540	+	Contig11520 [1:similar to PIR|E86286|E862 F9L1.18 protein - Arabidopsis thaliana] [2:-] [3:-]	1,31	0,20	1,15
MN5177771	+	Contig6917 [1:similar to UP|RUXF_ARATH (Q9SUM2) Probable small nuclear ribonucleoprotein F (snRNP-F) (Sm protein F) (Sm-F) (SmF)] [2:putative small nuclear ribonucleoprotein polypeptide F [2:Oryza sativa (japonica cultivar-group)]] [3:PF01423.10;LSM]	0,00	0,20	1,15
MN5257611	+	Contig11558 [1:] [2:putative nicotianamine aminotransferase A [Oryza sativa (japonica cultivar-group)] dbj|BAD23350.1| putative nicotianamine aminotransferase A [Oryza sativa (japonica cultivar-group)] dbj|BAD23256.1| putative nicotianamine aminotransfer	0,84	0,20	1,15
MN5236817	+	Contig3549 [1:weakly similar to UP|Q9FKC2 (Q9FKC2) Receptor protein kinase-like protein] [2:leucine-rich repeat family protein / protein kinase family protein [Arabidopsis thaliana]] [3:PF00069.12;Pkinase]	1,19	0,20	1,15
MN5175784	+	Contig14148 [1:] [2:unknown [Arabidopsis thaliana] emb|CAB78266.1| putative protein [Arabidopsis thaliana] emb|CAB45970.1| putative protein [Arabidopsis thaliana] ref|NP_192960.1| esterase/lipase/thioesterase family protein [Arabidopsis thaliana] pir||T4	0,84	0,20	1,15
MN5193252	+	Contig9223 [1:] [2:-] [3:-]	0,84	0,20	1,15
MN5256467	+	Contig11842 [1:] [2:putative nicotianamine aminotransferase A [Oryza sativa (japonica cultivar-group)] dbj|BAD23582.1| putative nicotianamine aminotransferase A [Oryza sativa (japonica cultivar-group)]] [3:PF00155.8;Aminotran_1_2]	0,72	0,20	1,15
MN5253983	+	Contig2080* [1:-] [2:-] [3:-]	1,31	0,20	1,15
MN5237051	+	Contig12059* [1:-] [2:-] [3:-]	1,24	0,20	1,15
MN5246210	+	Contig2983 [1:-] [2:-] [3:-]	1,31	0,20	1,14
MN5255369	+	Contig11899* [1:-] [2:-] [3:-]	1,31	0,19	1,14
MN5162994	+	Contig1463 [1:] [2:-] [3:-]	0,00	0,19	1,14
MN5251375	+	Contig11382* [1:-] [2:-] [3:-]	1,31	0,19	1,14
MN5233591	+	Contig9571 [1:similar to UP|Q9LUR1 (Q9LUR1) RING zinc finger protein-like] [2:putative zinc finger protein [Arabidopsis thaliana] pir||T52079 probable zinc finger protein [2:imported] - Arabidopsis thaliana] [3:-]	0,84	0,19	1,14
MN5171434	+	Contig13497* [1:-] [2:-] [3:-]	0,00	0,19	1,14
MN5238052	+	Contig4921* [1:-] [2:-] [3:-]	0,84	0,19	1,14
MN5234334	+	Contig4629 [1:] [2:-] [3:-]	1,31	0,19	1,14

MN5159560	+	Contig3225 [1:weakly similar to UP|Q39684 (Q39684) AX110P] [2:AX110P-like protein [Arabidopsis thaliana] gb|AAM63123.1| AX110P-like protein [Arabidopsis thaliana] emb|CAB78090.1| AX110P-like protein [Arabidopsis thaliana] gb|AAN86188.1| putative AX110P p	1,31	0,19	1,14
MN5257191	+	Contig11864 [1:similar to UP|O49551 (O49551) Adrenodoxin-like protein (MFDX2)] [2:-] [3:-]	1,24	0,19	1,14
MN5237668	+	Contig6178* [1:-] [2:-] [3:-]	1,20	0,19	1,14
MN5164750	+	Contig13712 [1:-] [2:ftsZ1 [2:Marchantia polymorpha] dbj|BAC57986.1| ftsZ1 [2:Marchantia polymorpha]] [3:PF03953.6;Tubulin_C]	0,00	0,19	1,14
MN5238515	+	Contig12111* [1:-] [2:-] [3:-]	0,00	0,19	1,14
MN5235789	+	Contig7931* [1:-] [2:-] [3:-]	1,31	0,19	1,14
MN5192437	+	Contig10697 [1:] [2:CG2913-PB] [3:PF00854.10;PTR2]	1,20	0,19	1,14
MN5193884	+	Contig4215* [1:-] [2:-] [3:-]	0,84	0,18	1,14
MN5196126	+	Contig4119 [1:-] [2:-] [3:-]	1,31	0,18	1,14
MN5158734	+	Contig5511 [1:weakly similar to UP|Q93X17 (Q93X17) Snakin2 precursor] [2:gibberellin-regulated protein GASA2 precursor [2:Arabidopsis thaliana] emb|CAB78084.1| gibberellin-regulated protein GASA2 precursor [Arabidopsis thaliana] sp|P46688|GAS2_ARATH Gibb	1,31	0,18	1,14
MN5164263	+	Contig13682 [1:-] [2:-] [3:-]	1,31	0,18	1,13
MN5252124	+	Contig11054 [1:-] [2:-] [3:-]	0,84	0,18	1,13
MN5162763	+	Contig1590 [1:-] [2:-] [3:-]	1,31	0,18	1,13
MN5252349	+	Contig10963 [1:weakly similar to GB|AAL11556.1|15983376|AF424562 AT3g59080/F17J16_130 {Arabidopsis thaliana;}] [2:putative chloroplast nucleoid DNA binding protein [Oryza sativa (japonica cultivar-group)] dbj|BAD15987.1| putative chloroplast nucleoid D	0,84	0,18	1,13
MN5234378	+	Contig6396 [1:weakly similar to UP|RNC_AGRT5 (Q8UGK2) Ribonuclease III (RNase III)] [2:At1g24450/F21J9_210 [Arabidopsis thaliana] ref|NP_173854.1| ribonuclease III family protein [Arabidopsis thaliana] gb|AAK73956.1| At1g24450/F21J9_210 [Arabidopsis th	0,00	0,18	1,13
MN5236461	+	Contig5066 [1:-] [2:unknown protein [Arabidopsis thaliana] gb|AAM64581.1| NADH] [3:PF06747.2;CHCH]	1,31	0,18	1,13
MN5192925	+	Contig7708 [1:homologue to UP|Q93Y15 (Q93Y15) Beta tubulin] [2:Beta tubulin 1 [2:Lupinus albus] sp|P37392|TBB1_LUPAL Tubulin beta-1 chain (Beta-1 tubulin) pir||S35142 tubulin beta chain - white lupine] [3:PF03953.6;Tubulin_C]	0,84	0,17	1,13
MN5182078	-	Contig8117 [1:similar to GB|AAP75807.1|32189305|BT009657 At4g12590 {Arabidopsis thaliana;}] [2:At4g12590 [Arabidopsis thaliana] gb|AAM64797.1| unknown [Arabidopsis thaliana] emb|CAB53752.1| putative protein [Arabidopsis thaliana] emb|CAB78302.1| putati	0,00	-0,17	1,13
MN5239521	+	Contig7294* [1:-] [2:-] [3:-]	1,31	0,17	1,13
MN5193579	+	Contig9396 [1:similar to UP|TRB1_ARATH (Q39243) Thioredoxin reductase 1 (NADPH-dependent thioredoxin reductase 1) (NTR 1)] [2:putative thioredoxin reductase [Arabidopsis thaliana] sp|Q39242|TRB2_ARATH Thioredoxin reductase 2 (NADPH-dependent thioredoxi	0,84	0,17	1,12
MN5165199	+	Contig13509 [1:] [2:-] [3:-]	0,83	0,17	1,12
MN5235269	+	Contig5875* [1:-] [2:-] [3:-]	1,20	0,17	1,12
MN5258840	+	Contig11495 [1:homologue to UP|Q8DJ14 (Q8DJ14) Tll1417 protein] [2:putative transducin / WD-40 repeat protein [Oryza sativa (japonica cultivar-group)] dbj|BAC16061.1| putative transducin / WD-40 repeat protein [Oryza sativa (japonica cultivar-group)]] [	0,71	0,17	1,12

Annexe 3 – Protocoles

A. Arborea Spruce Microarray Hybridization Protocol using Alexa Fluor® Dyes

Version 2.0
August 2005

This protocol is designed for the preparation and hybridization of slides with targets prepared using the Arborea Indirect Labeling Protocol, and is based on the recommendations of the manufacturer of the slides.

** All solutions (including water) should be filtered on a 0,45 µm membrane.*
**Always wear gloves when manipulating slides.*
**Avoid exposure of labeled targets and hybridized slides to light.*

- Array prep: set heating block to 95°C
- Heat the Pre-Hyb buffer at 65°C and Hyb solution at 42°C in the hybridization ovens.
- Set up centrifuge to take 96-well plates; set rotor and rpm to 1600
- Optional: scan slides to determine background. Average intensity <500 is excellent; <1000 is ok.

Pre-Hybridization

Pre-Hyb solution: [5X SSC, 0.1% SDS, 0.2mg/ml BSA, 0.1mg/ml Herring Sperm DNA]

For 1 L:

- 20X SSC : 250 ml
- 10% SDS : 10 ml
- 10 mg/ml BSA : 20 ml
- Nanopure water : 710 ml
- Herring Sperm DNA : 10 ml

*Filter on 0,45 µm membrane

- Make sure that the slides are free of any dust. If not, use filtered air to clean the slides.
- Incubate slides in pre-hyb solution at 65°C for 2 to 2,5 hours.
- Transfer slides to 0,1X SSC, room temperature, 30 sec.
- Transfer slides to 0,1X SSC, room temperature, 30 sec.
- Transfer slides to water, room temperature 30 sec.
- Transfer slides to water, room temperature 30 sec.
- Transfer slides to boiling water, incubate 3 minutes.
- Quickly place slides in the staining kit box with a kimwipe in bottom.
- Centrifuge at 1600 rpm for 2 minutes to dry the slides.

- Keep dried slides in the staining kit box to avoid dust while preparing target.
  * *It's recommended to use pre-hybridized slides as soon as possible. Do not keep dried more than 1 hour. Keep in a dessicator chamber until use.*

Coverslip wash:

- Place coverslips in water, incubate at room temperature 30 sec with continual shaking.
- Transfer slides to isopropanol, incubate at room temperature 10 sec with continual shaking.
- Quickly place slides in 50ml conical tubes with a new kimwipe stuffed in bottom.
- Centrifuge at 1600 rpm for 2 minutes.
- Keep tubes closed until they are used to protect coverslips from dust.

Preparing Target for Hybridization:

Hyb solution: [50% formamide, 5XSSC, 0.1% SDS, 0.1mg/ml Herring Sperm DNA]

For 50 ml:

- 20X SSC : 12.5 ml
- 10% SDS : 0.5 ml
- 10 mg/ml Herring Sperm DNA : 0.5 ml
- Nanopure water : 11.5 ml
- Formamide : 25 ml

*Filter on 0,45 µm membrane

- Dry the purified cDNAs in a speed vac at low heat until the volume is reduced to 3,5µL. If the volume is less than 3,5µL, complete with 10mM EDTA.
  There is some (invisible) material precipitated on the side of the tube, make sure to get it!
- Heat at 95°C about 3 min
- Incubate on ice 1 min. Quick spin.
- Add 52,5 µl Hyb solution (warmed to 42°C) to targets, vortex, quick spin and keep at 42°C until ready to add to slide.
  **The volume of Hyb solution to add to denatured targets mentioned here is appropriate for the 24X60 Erie Lifter Slips used with the Arborea Spruce array.*

Adding Target to Slide:

- Note bar code on slide and target to be hybridized.
- Inspect slides; remove any dust with filtered air.
- Gently breathe on slide to locate subarrays.
- Put slide in bottom half of a Corning LL hybridization chamber (those with more internal space, appropriate for Erie Lifter Slips), remembering where subarrays are.

- Gently place the coverslip on the slide over the subarrays, the side with white printed edges down at 1 mm from the barcode sticker.
- Add 15 µl of warm Hyb solution (or water) to each well at the extremities of the chamber.
- Add target to one edge of the coverslip, letting it wick underneath.

  ** Let the solution cool for 10 sec in the pipet tip before dispensing. Cooler solution will move more evenly under the coverslip. Also, dispensing the target/hyb solution too fast will produce big air bubbles that are very hard to get rid of afterward.*
- Assemble the Corning chamber.
- Place in blue tip box with a little prehyb solution (or water) in bottom.
- Seal tip box with masking tape.
- Set in hybridization oven, making sure that the box is level.
- Incubate overnight at 42°C.

➢ Place wash solution 1 and 2 and 0,5X SSC, 0,1% SDS solution in hybridization oven.
➢ *Proceed to Arborea Spruce Microarray Wash Protocol.*

B. Arborea Spruce Microarray Wash Protocol

Version 1.1
April 2005

This protocol is designed for the washing of Spruce cDNA microarray slides following hybridization using the Arborea Microarray Hybridization Protocol and the Arborea Indirect Labeling Protocol. This protocol is based on Jurgen Ehlting's Arabidopsis oligo array hybridization protocol (GenomeBC Forestry, Nov. 2003), with minor modifications.

**All solutions (including water) should be filtred on a 0,45 µm membrane.*
**Always wear non-latex gloves when manipulating slides.*
**Avoid exposure of hybridized slides to light and moisture.*
**Take care that the slides never dry between washes.*

- Heat Array Wash solutions 1, 2 and 3 at 42°C.
- Set up centrifuge to take 96-well plates; set rotor and rpm to 1600.
- Turn on scanner lasers 15 minutes prior to scan.

Array Wash 1: [2X SSC, 0.5% SDS]

For 1 L:

- 20X SSC : 100 ml
- 20% SDS : 25 ml
- Nanopure water : to 1L

*Filter on 0,45 µm membrane

Array Wash 2: [0.5X SSC, 0.5% SDS]

For 1 L:

- 20X SSC : 25 ml
- 10% SDS : 50 ml
- Nanopure water : to 1L

*Filter on 0,45 µm membrane

Array Wash 3: [0.5X SSC, 0.1% SDS]

For 1 L:

- 20X SSC : 25 ml
- 10% SDS : 10 ml
- Nanopure water : to 1L

*Filter on 0,45 µm membrane

0.1 X SSC:

For 1 L:

- 20X SSC : 5 ml
- Nanopure water : to 1L

*Filter on 0,45 µm membrane

Washes

- Gently open hybridization chambers, taking care not to move the slide and coverslip.
- Gently float off coverslip in Array Wash 1 solution (heated at 42°C), in a Coplin jar.
- Transfer slides in Array Wash 1 at 42°C, 15 min, shaking every few minutes.
- Transfer slides in 0,5X SSC, 0,1% SDS, 15 min at 42°C, shaking every few minutes.
- Repeat.
- Transfer slides in 0.1 X SSC, room temperature, 1 min, shaking.
- Repeat 0.1X SSC, room temperature, 1 min, shaking.
- Transfer slides on a staining kit slide holder submerged in water. Agitate for 20 seconds at room temperature.
- Quickly place slide holder in a staining kit box with a kimwipe in bottom, and centrifuge at 1600 rpm (~ 500*g*) for 2 min to dry the slides.

  **Avoid letting the slides dry with SSC on it, as salt deposited on the slide will produce fluorescence.*

- Place dried slides in a slide box in the dark and keep in a desiccation chamber until ready to scan.

Proceed to Arborea Microarray Scanning Protocol