LUIZ EDUARDO VIEIRA DEL BEM

“EVOLUÇÃO DE FAMÍLIAS MULTIGÊNICAS E REDES DE REGULAÇÃO EM PLANTAS”

CAMPINAS
2013

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

LUIZ EDUARDO VIEIRA DEL BEM

"EVOLUÇÃO DE FAMÍLIAS MULTIGÊNICAS E REDES DE REGULAÇÃO EM PLANTAS"

Tese apresentada ao Instituto de Biologia da UNICAMP para obtenção do Título de Doutor em Genética e Biologia Molecular, na área de Genética Vegetal e Melhoramento.

Orientador: Prof. Dr. Michel Georges Albert Vincentz
Coorientador: Dr. Renato Vicentini dos Santos

CAMPINAS,
2013

FICHA CATALOGRÁFICA ELABORADA POR
MARA JANAINA DE OLIVEIRA – CRB8/6972
BIBLIOTECA DO INSTITUTO DE BIOLOGIA - UNICAMP

D376e

Del Bem, Luiz Eduardo Vieira, 1984-
Evolução de famílias multigênicas e redes de regulação em plantas / Luiz Eduardo Vieira Del Bem. – Campinas, SP: [s.n.], 2013.

Orientador: Michel Georges Albert Vincentz.
Coorientador: Renato Vicentini dos Santos.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Família multigênica. 2. Regulação da expressão gênica. 3. Plantas - Evolução. 4. Evolução molecular. 5. Filogenia. I. Vincentz, Michel Georges Albert, 1958-. II. Santos, Renato Vicentini. III. Universidade Estadual de Campinas. Instituto de Biologia. IV. Título.

Informações para Biblioteca Digital

Título em Inglês: Evolution of multigenic families and genetic networks in plants
Palavras-chave em Inglês:
Multigene family
Gene expression regulation
Plant - Evolution
Molecular evolution
Phylogeny
Área de concentração: Genética Vegetal e Melhoramento
Titulação: Doutor em Genética e Biologia Molecular
Banca examinadora:
Michel Georges Albert Vincentz [Orientador]
Jörg Kobarg
Sérgio Furtado dos Reis
Douglas Silva Domingues
Camila Caldana
Data da defesa: 29-01-2013
Programa de Pós Graduação: Genética e Biologia Molecular

Campinas, 29 de janeiro de 2013

BANCA EXAMINADORA

Prof. Dr. Michel Georges Albert Vincentz (Orientador) ______________________ Assinatura

Dra. Camila Caldana ______________________ Assinatura

Dr. Douglas Silva Domingues ______________________ Assinatura

Dr. Jörg Kobarg Jörg Kobarg ______________________ Assinatura

Prof. Dr. Sergio Furtado Dos Reis ______________________ Assinatura

Dra. Katlin Brauer Massirer ______________________ Assinatura

Dr. Carlos Augusto Colombo ______________________ Assinatura

Dra. Prianda Rios Laborda ______________________ Assinatura

SUMÁRIO

Resumo

O sequenciamento de um número crescente de genomas completos tem transformado a biologia. Mais especificamente, no campo da biologia evolutiva, tem se tornado possível endereçar perguntas centrais sobre o funcionamento ultimato dos mecanismos de transformação genética, com potencial impacto em todos os campos da biologia, assim como na filosofia. Esta tese está dividida em dois aspectos importantes da evolução de genomas: o processo de duplicação e fixação de genes duplicados, que é a base do surgimento de famílias multigênicas, e a evolução de redes de regulação, que determinam as relações de causalidade nos processos celulares. Os dois aspectos se relacionam à evolução da complexidade, tanto no que tange o conteúdo gênico dos seres vivos quanto nas interações mecanistica entre os genes via seus produtos (RNAs e proteínas basicamente).

No primeiro aspecto abordamos a evolução de dois mecanismos biológicos que depende da ação integrada entre proteínas de famílias distintas: o mecanismo de síntese e degradação do polissacarídeo de parede celular xiloglucano, e o ciclo das chaperonas calreticulina/calnexina envolvidas no controle de qualidade de proteínas sintetizadas no retículo endoplasmático. Nossos trabalhos mostraram que uma forma primordial de xiloglucano, mais simples, surgiu antes da conquista do meio terrestre pela linhagem das plantas, ao contrário do que se imaginava, e que o ciclo calreticulina/calnexina é produto da subfuncionalização em eucariotos basais de uma chaperona ancestral, além do surgimento de funções específicas na família da calreticulina em plantas terrestres. O interesse em evolução de famílias multigênicas nos levou a desenvolver um método (Phylexpress) para análise de ortologia em larga escala, bem como permitir a integração de dados de expressão na tentativa de entender a dinâmica evolutiva da expressão gênica em famílias multigênicas. Utilizamos nosso método para revisitar o conteúdo gênico dos ESTs públicos de cana-de-açúcar, como prova de conceito, numa análise comparativa com o proteoma predito de sorgo. Nossos resultados mostram uma cobertura em termos de ortólogos para apenas ~58% do proteoma predito de sorgo em contrates com estimativas anteriores, com métodos mais simples, que chegaram a 90% do proteoma hipotético de cana.

Para abordar a dinâmica evolutiva de redes de regulação, realizamos medições, em escala genômica, das alterações nos níveis de mRNAs de plântulas de sorgo e arroz em resposta a tratamentos de curta duração (2hrs) com sinais exógenos de ABA (hormônio vegetal) e dos açúcares glicose e sacarose. Utilizamos dados públicos e experimentalmente comparáveis de *Arabidopsis thaliana* em resposta aos mesmos sinais para realizar comparações que revelassem respostas conservadas ou divergentes entre ortólogos. Além disso, buscamos entender a dinâmica evolutiva das respostas transcricionais num contexto de duplicação gênica em famílias multigênicas, onde há diversos genes potencialmente redundantes do ponto de vista bioquímico/estrutural. Nossa abordagem sugere que redes de regulação gênica em eucariotos complexos evoluem majoritariamente de forma neutra, pois parecem apresentar uma taxa de divergência constante, que independe da rede (disparada por cada um dos diferentes sinais) e das espécies envolvidas. Nossos dados são complementares e potencialmente confirmadores de modelos recentes de evolução não-adaptativa em redes de regulação gênica. Concluímos que a evolução da complexidade em sistemas biológicos está parcialmente ligada à diminuição da eficiência da seleção, causada majoritariamente por números populacionais efetivos restritivos presentes nas linhagens de eucariotos complexos (vertebrados e plantas terrestres).

Abstract

The availability of complete sequences of a growing number of genomes is transforming biology. More specifically, in the field of evolutionary biology, it became possible to address central questions on the ultimate mechanisms underlying genetic changes. It has a broad impact on biology and philosophy as well. This thesis deals with two important aspects of genome evolution: the process of gene duplication and fixation of duplicated genes, which is the basis of the origins of multigenic families, and the evolution of genetic regulatory networks that determines the causality of the cellular processes. Both aspects are related to the evolution of complexity regarding the gene content of living forms and the mechanistic interaction between the gene products (mainly RNAs and proteins).

In the first aspect we studied the evolution of two biological mechanisms depending on the integrated function of proteins from distinct families. The mechanism of synthesis and remobilization of xyloglucan, a plant cell wall polysaccharide, and the calreticulin/calnexin cycle of protein folding that takes place in the endoplasmic reticulum. Our work showed that a primordial form of xyloglucan already existed before the land conquest by plants. We propose that the calreticulin/calnexin cycle is the product of subfuncionalization of an ancestral eukaryotic chaperone, and plants evolved specific calreticulin functions due to gene duplication. Our interest in the evolution of multigenic families impelled the development of Phylexpress, a method dedicated to large-scale orthology analyses. It can integrate expression data in the context of multigenic families with the goal of understand the evolutionary dynamics of gene expression. We used Phylexpress to revisit the gene content of the publicly available sugarcane ESTs as a proof of concept. Our results showed that the ESTs sampled orthologs for just ~58% of the predict sorghum proteome, in contrast with previous estimations acconting for 90% of the hypotethical sugarcane proteome.

In order to approach the evolutionary dynamics of regulatory networks, we measured global changes in gene expression of sorghum and rice plantlets in response to short-term treatments (2hrs) with exogenous ABA (plant hormone) and the sugars glucose and sucrose. We took public data from comparable experiments using *Arabidopsis thaliana* in order to unravel conserved and divergent responses across orthologs. Furthermore, we analyzed the evolutionary change in transcriptional responses in a context of gene duplications in multigenic families, leading to a set of potentially redundant genes in terms of biochemical/structural properties. Our approach suggests that gene regulatory networks in complex eukaryotes evolve mainly neutrally, in a constant rate that is independent of the analyzed network (triggered by each one of the signals) and the species. Our data is complementary and potentially confirmatory of recent models of non-adaptive evolution in regulatory networks. We concluded that the evolution of the complexity in biological systems is partially connected to the attenuation of the efficiency, mainly due to low effective population sizes present in the lineages that gave rise to complex eukaryotes (vertebrates and land plants).

Agradecimentos

Aos meus pais, Luiz e Marcia, e ao meu irmão Rodrigo, sem os quais nada teria sido possível.

Ao meu orientador, por mais de nove anos, e amigo Prof. Michel Vincentz, por todo o aprendizado e amizade.

Ao meu coorientador e amigo, Renato Vicentini, pelos extensos momentos de trabalho e grande companheirismo.

A todos os amigos que fiz em 11 anos de Unicamp, os quais seriam impossíveis enumerar.

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento, via bolsa de doutorado, através do processo 2008/09105-1 e aos diversos fomentos fornecidos ao nosso laboratório em todos estes anos.

1. Introdução

1.1 Desenvolvimento histórico da teoria da evolução biológica

A diversidade biológica intriga pensadores desde a antiguidade. Talvez tenha sido o grego Aristóteles quem primeiro se perguntou se havia uma ordem biológica capaz de explicar a diversidade da vida (*Scala naturae* – os organismos podem ser ordenados de forma contínua e linear do mais simples ao mais complexo). Aristóteles também foi o pioneirismo em tentar explicar como se dava a herança biológica, segundo ele através da mistura de características parentais num cenário onde haveria herança de caracteres adquiridos ao longo da vida. Desde muito cedo na história do pensamento identificou-se que deveria haver uma relação entre o mecanismo de herança, que explicaria porque pais e filhos se parecem, com a base da diversidade biológica, que explica porque indivíduos da mesma espécie não são idênticos (variação intraespecífica) e porque há espécies que se assemelham ainda que permaneçam isoladas reprodutivamente (variação interespecífica).

A primeira sistematização deste problema, na tentativa de conceber uma teoria que unificasse herança e variação, coube a Jean-Baptiste Lamarck no começo do século XIX. Há um claro paralelo entre a teoria lamarckiana da 'herança dos caracteres adquiridos' (1809) e a concepção aristotélica clássica. Lamarck formalizou o conceito de variação na biologia e assumiu que as criaturas vivas mudavam constantemente **para** (conceito teleológico – há um propósito ou finalidade necessária que explica porque algo é como é – *A função delimita a forma*) se adaptar ao seu meio e condições de vida. Ficou claro a Lamarck a necessidade de um mecanismo de criação de variação biológica, e em sua teoria este era produto da interação das espécies com seu meio ambiente. As modificações em resposta ao ambiente ocorridas ao longo da vida de cada indivíduo poderiam ser herdadas verticalmente. Formalmente considera-se a teoria lamarckiana como a primeira teoria da evolução, ainda que este conceito apareça deste o período clássico. Em comum a todas as teorias evolutivas seguintes, até os conceitos mais modernos que serão apresentados posteriormente, há a premissa central do conceito de evolução: todas as espécies vivas que habitam o planeta hoje não surgiram em sua forma atual, apenas representam o resultado de modificações contínuas ao longo dos períodos geológicos.

Os naturalistas Charles Darwin e Alfred Wallace (1858) foram os primeiros a conceber, de forma independente e simultânea, uma teoria evolutiva não-teleológica, isto é, onde não há necessidade de evocar a finalidade para explicar a existência de qualquer entidade biológica (*A forma delimita a função*). O conceito central da teoria darwiniana (e de Wallace) reside no conceito de **seleção natural**. As espécies se modificariam no tempo não em resposta ao ambiente, e sim pela perpetuação diferencial de variações interespecíficas. O meio aparece como um agente seletivo que atua sobre a variação biológica existente em populações naturais. A natureza se torna palco da competição (ou luta) pela existência, onde variações menos favorecidas reprodutivamente seriam progressivamente substituídas por variações mais favorecidas reprodutivamente seja qual fosse sua base mecanística. Tal força evolutiva (seleção natural) produziria modificações graduais em uma determinada direção definida. Historicamente a seleção natural foi o primeiro mecanismo de evolução capaz de explicar, pelo menos parcialmente, a diversidade biológica. Havia, porém, dois problemas que não foram resolvidos pela teoria darwiniana: 1 – qual é o mecanismo de herança? 2 - qual é a base do surgimento da variação biológica?

Num dos trabalhos menos prestigiados imediatamente da história da ciência moderna, o austríaco Gregor Mendel (1865), descobre em experimentos com ervilhas a segregação das características hereditárias e funda as bases da moderna ciência da genética. A genética forneceu (ainda que provavelmente desconhecida de Darwin e contemporâneos) a resposta para o mecanismo de herança. Ao contrário do que se acreditou desde Aristóteles não havia mistura entre as características dos parentais num cruzamento. Mendel explicou a variação fenotípica através da interação entre **alelos** (ainda que este termo tenha sido cunhado posteriormente) em diferentes combinações (genótipos; modelo de dominância) e demonstrou que a segregação dos alelos seguia uma proporção bem definida matematicamente (3 fenótipo dominante : 1 fenótipo recessivo num cruzamento de heterozigotos). Era o começo do uso de conceitos estatísticos na biologia. Ainda que o mecanismo proposto por Mendel tenha sido confirmado desde então ele ainda não era capaz de explicar a base do surgimento da variação biológica. Tal problema começou a encontrar uma resposta no trabalho de Hugo de Vries (um dos redescobridores dos trabalhos de Mendel, ao lado de Karl

Correns e Erich Tschermak) na virada do século XX. de Vries observou que plantas da espécie *Oenothera lamarckiana*, que haviam sido recentemente introduzidas na Europa vindas dos EUA, apresentavam variações fenotípicas claras que eram muito raras (por exemplo fenótipo anão, fenótipo latifoliado entre outros). Ele transferiu vários espécimes para um jardim onde permitiu que se reproduzissem por autofecundação. Percebeu assim que os variantes fenotípicos transmitiam suas características às plantas filhas. Com base neste trabalho de Vries cunhou o termo **mutação**. Ao contrário do que foi predito na teoria darwiniana, de Vries sugeriu que a variação biológica poderia surgir num salto, ao invés de ocorrer de forma gradual. Uma planta de tamanho selvagem poderia dar origem a uma planta de porte anão sem que houvesse toda uma gama de tamanhos progressivamente menores ao longo do tempo. O debate entre o saltacionismo de de Vries versus o gradualismo darwiniano só seria resolvido pelos trabalhos de genética populacional, em especial de J.B.S Haldane e Ronald Fisher entre 1918 e 1932, que culminariam na síntese neodarwiniana nas décadas de 1930 e 1940. Já a base físico-química da mutação só seria solucionada com o modelo da dupla hélice do DNA de Watson e Crick em 1953: mutações são alterações herdáveis na sequência do DNA.

Francis Galton, um primo de Darwin, havia atacado o problema da hereditariedade de um ponto de vista inovador. Ele questionou a visão mecanicistica da herança apresentada por Darwin, a pangenesis, teoria que assumia que "partículas" produzidas por todos os órgãos de um ser-vivo eram transmitidas via células sexuais de uma geração para a outra (as células sexuais e seu papel na herança haviam sido identificadas muito antes com a descoberta do ovo (óvulo) de mamíferos por William Harvey e as linhagens germinais, feminina por Renier de Graaf e masculina por Johan Ham e Anthonie van Leeuwenhoek, todos no século XVII). Galton realizou experimentos onde injetou sangue de coelhos brancos em casais de coelhos cinza esperando obter progênies malhadas. Ao contrário do esperado os coelhos produziram quatro gerações de progênies cinza, demonstrando que o sangue dos coelhos branco introduzido nos progenitores não afetou a herança do caractere 'cor' presente nos parentais. Galton concluiu que as células sexuais continham os caracteres necessários para o desenvolvimento da prole e que em tais células os condicionantes dos caracteres biológicos não seriam afetados pelas condições de vida dos parentais. Os trabalhos de August Weissmann com camundongos no final do século XIX apoiavam a ideia de

Galton: a ‘teoria da continuidade do plasma germinativo’ segundo a qual a informação hereditária contida na linhagem germinal não seria afetada pelo ambiente. Desta forma no começo do século XX, com a redescoberta das leis mendelianas da hibridização, ficou claro que a base da evolução deveria incluir modificações (mutações) nas unidades hereditárias (genes – termo cunhado por William Bateson em 1905, de onde se origina a palavra genética) das linhagens germinais, e somente assim estas seriam herdáveis e passíveis de seleção.

A síntese entre a genética e a teoria da evolução darwiniana começou com o equilíbrio (lei) de Hardy-Wienberg ($p^2 + 2pq + q^2 = 1$, sendo p e q frequências de dois alelos numa população diploide) em 1908. De forma independente eles demonstraram que numa população mendeliana, assumindo cinco premissas, a frequência dos alelos na geração inicial se manteria constante independente do número de gerações que se passasse. As cinco premissas são:

1 – Tamanho da população deve ser infinito com número igual de machos e fêmeas

2 – Os cruzamentos devem ocorrer ao acaso

3 – O número de descendentes de qualquer cruzamento deve ser igual

4 – A população deve ser isolada reprodutivamente de outras populações

5 – Novos alelos não podem se originar a partir dos alelos iniciais

Os estudos subsequentes concentraram-se justamente em quebrar tais premissas com objetivo de entender como isso afetaria as frequências alélicas ao longo das gerações. O equilíbrio de Hardy-Weinberg forneceu a primeira hipótese nula da teoria da evolução ao mostrar um cenário onde não havia evolução (definida como ‘**mudanças nas frequências alélicas**’). Diversos evolucionistas trabalharam nestes problemas, com especial destaque para J.B.S Haldane, Ronald Fisher e Sewall Wright. Entre as décadas de 1910 e 1930 estes pioneiros da genética de populações confirmaram a compatibilidade entre a seleção natural darwiniana e as leis de Mendel, conciliando o gradualismo com o surgimento repentino de mutações observado por de Vries uma década antes. Eles identificaram as forças básicas que atuam em populações mendelianas e são capazes de alterar a frequência dos alelos presentes numa população inicial. Ao quebrar a premissa referente ao tamanho populacional infinito (e a equiproporção de machos e fêmeas) encontrou-se a **deriva genética**, uma força capaz de alterar as frequências alélicas ao acaso baseada na aleatoriedade da meiose e da recombinação (por exemplo, num cruzamento *A1A2* x *A1A2* que gera dois descendentes há uma chance de 6,25% de que nenhum deles herde o alelo *A2* simplesmente por acaso da segregação ocorrida na meiose). A quebra das premissas 2 e 3 levam à **seleção sexual** (2 – caso especial de seleção natural) e/ou **endogamia** (cruzamentos consanguíneos preferenciais, porém isso não altera frequências alélicas, somente genotípicas, não sendo portanto uma força capaz de gerar evolução) e à **seleção natural** (3 – definida como: '**reprodução diferencial de um genótipo**'). Se a população não for isolada reprodutivamente, ou seja, houver troca de alelos entre populações, surge a **migração gênica** que é capaz de modificar as frequências alélicas a depender apenas do número e da composição genética dos migrantes e do tamanho da população receptora. E finalmente, se um alelo puder dar origem a outro alelo temos a definição de **mutação**, que é, em último nível, a única força capaz de introduzir diversidade numa população isolada.

Em 1902 Walter Sutton propôs que os cromossomos ocorriam em pares de homólogos, cada um deles vindo de um dos pais e que cada um deles poderia carregar informações hereditárias diferentes. Esta foi a primeira tentativa de explicar fisicamente a base das leis de Mendel. Na busca pela explicação do fenômeno da ligação gênica (*genetic linkage*), que contrariava a 2ª lei de Mendel (segregação independente dos caracteres), Morgan e seus alunos demonstraram que os grupos de ligação correspondiam aos cromossomos, estruturas presentes no núcleo celular de eucariotos que já haviam sido previamente identificadas. Através de análises em larga escala de **recombinação** usando *Drosophila* eles demonstraram que as unidades hereditárias mendelianas, batizadas por Bateson de **genes**, eram fisicamente localizadas nos cromossomos de forma linear. Quanto mais próximos dois genes num dado cromossomo menor era a frequência de recombinação entre eles, e quanto mais distantes, maior era a frequências de recombinantes. Uma das mais importantes descobertas neste campo foram os genes ligados ao sexo em *Drosophila*. O macho de *Drosophila* carrega apenas um cromossomo X, portanto é haploide para os genes ligados ao sexo, já a fêmea carrega dois cromossomos X, sendo diploide para os genes ligados ao sexo. Estudos com mutações no cromossomo X de *Drosophila* e seu padrão de segregação entre os sexos sugeria fortemente que os genes ligados ao sexo estavam contidos no cromossomo sexual. Houve uma grande resistência na aceitação da teoria cromossômica da herança, sendo o próprio Bateson um dos principais opositores. Tal disputa só seria resolvida definitivamente em 1953, com a publicação do modelo Watson-Crick da estrutura do DNA.

1.2 Bases moleculares de evolução

A variação genética, que é a "matéria prima" do processo evolutivo, é gerada por uma combinação de mutação e recombinação (Drake, 2006; Miller, 2005). As mutações podem ser causadas pontualmente por substituições de nucleotídeos e/ou inserções/deleções basicamente devido a erros de replicação do DNA (Pray, 2008). Mutações de larga escala envolvem duplicações gênicas, de segmentos cromossômicos, de cromossomos inteiros ou até duplicações do genoma inteiro (Van de Peer, 2004). A recombinação pode gerar novas combinações de mutações originadas em diferentes linhagens, além de ter um papel na duplicação gênica e de segmentos cromossômicos via erros de recombinação homóloga (Iraqui *et al*, 2012).

Do ponto de vista da seleção natural mutações de qualquer tipo podem ser categorizadas de três maneiras: mutações benéficas ou vantajosas (aumentam o valor adaptativo do indivíduo portador, são extremamente raras), mutações deletérias (diminuem o valor adaptativo, maior parte das mutações em regiões codificantes) ou mutações neutras (não afetam a capacidade reprodutiva, sua proporção depende do tamanho efetivo da população). Mutações benéficas podem ser eventualmente fixadas por seleção (darwiniana) positiva, enquanto mutações deletérias serão geralmente eliminadas por seleção purificadora (ou negativa). Mutações neutras, por causarem efeitos marginais ou nulos no valor adaptativo, podem ser eliminadas ou fixadas apenas por deriva genética (Eyre-Walker e Keightley, 2007). Esta última conclusão, apesar de simples, é a base da teoria neutra (Kimura, 1983). De forma simplificada, o acúmulo de mutações neutras num segmento homólogo de DNA é indicador da divergência, em gerações, entre linhagens (relógio molecular). Isto é baseado no fato da taxa de fixação de alelos neutros (por deriva genética) ser numericamente igual à taxa de mutação (Kimura, 1983).

1.3 Duplicação gênica e a origem de famílias multigênicas

Duplicação gênica é um processo essencial na diversificação e adaptação biológicas. Genes duplicados representam adição de variabilidade sob a qual uma combinação de mutação e seleção natural pode levar ao surgimento de novas funções (Alvares-Buylla *et al.*, 2000; Lawton-Rauh, 2003; Vandepoele *et al.*, 2003; Kellis *et al.*, 2004). É o processo de acúmulo de genes duplicados (especialmente via erro de recombinação homóloga e duplicações genômicas) e a divergência evolutiva entre as cópias que explica a existência de famílias multigênicas (Wendel, 2000; Bennetzen, 2002; Kellis *et al.*, 2004). Este processo evolutivo explica, pelo menos em parte, o formato atual dos genomas eucarióticos onde a maior parte das categorias de genes está presentes em múltiplas cópias (Lespinet *et al.*, 2002; Lynch, 2002). No genoma da eudicotiledônea *Arabidopsis thaliana*, por exemplo, 65% dos genes apresentam duas ou mais cópias (Wendel, 2000).

Um gene duplicado apresenta, inicialmente, redundância funcional em relação à cópia da qual se originou. A partir de uma duplicação gênica, quatro podem ser os destinos das cópias (Figura 1). A primeira possibilidade é que as cópias mantenham a função original (**redundância funcional**). As cópias podem evoluir para um estado onde a função que era desempenhada pelo gene ancestral passe a ser desempenhada pela ação conjunta das cópias (**subfuncionalização**). Um das cópias pode manter a função original do gene ancestral enquanto a outra acumula mutações mais rapidamente e atinge um estado funcional diferente do ancestral (**neofuncionalização**). Por último, o acúmulo de mutações *non-sense* (*frameshifts* ou inserção de *stop codons*) pode tornar uma das cópias um pseudogene, que por definição é uma sequência sem atividade biológica de DNA genômico que apresenta similaridade e descende evolutivamente de genes funcionais (Lawton-Rauh, 2003).

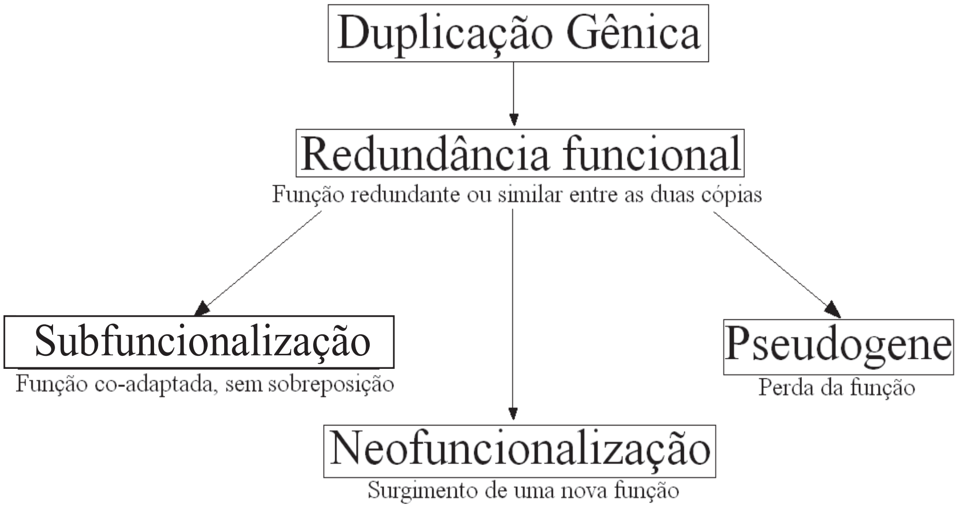

Duplicação Gênica
Redundância funcional
Função redundante ou similar entre as duas cópias
Subfuncionalização
Função co-adaptada, sem sobreposição
Pseudogene
Perda da função
Neofuncionalização
Surgimento de uma nova função

Figura 1. Possíveis destinos de genes duplicados segundo uma trajetória temporal. Após a duplicação, as duas cópias geradas apresentam redundância funcional que pode se manter ao longo da evolução, ou pode haver subfuncionalização, neofuncionalização ou a formação de um pseudogene (Modificado de Lawton-Rauh, 2003).

Estudos comparativos, como entre as leveduras *Kluyveromyces waltii* e *Saccharomyces cerevisiae* (Kellis *et al*., 2004), mostraram que na maioria das duplicações gênicas (95% dos casos neste trabalho) uma das cópias sofre evolução acelerada enquanto a outra permanece sob seleção de purificação (eliminação de mutações não-sinônimas). Isso permitiria uma rápida divergência funcional entre as cópias duplicadas. Este resultado sugere que o mecanismo de neofuncionalização explica a maior parte dos genes duplicados que são retidos pela evolução. A presença de pseudogenes em todos os genomas sequenciados demonstra que duplicações gênicas nem sempre são retidas evolutivamente.

1.4 Homologias moleculares: Ortólogos e Parálogos

O grande número de genomas completamente sequenciados na atualidade permitiu diversas abordagens comparativas no estudo da evolução de famílias multigênicas. Um dos aspectos centrais neste campo é o estabelecimento de relações evolutivas entre genes de diferentes genomas, em um sistema de genes homólogos (Bennetzen, 2002;

Pennacchio, 2003; Vincentz *et al*., 2004; Figura 2), que inclui **ortólogos** e **parálogos**. Ortólogos são genes homólogos que divergiram após de um evento de especiação e são derivados de uma cópia ancestral existente no último ancestral como entre as espécies em questão. Parálogos são genes homólogos resultantes de um evento de duplicação gênica restrito a apenas uma linhagem evolutiva (Tatusov *et al*., 1997; Thornton e DeSalle, 2000; Fitch, 2000; Meyrowitz, 2002).

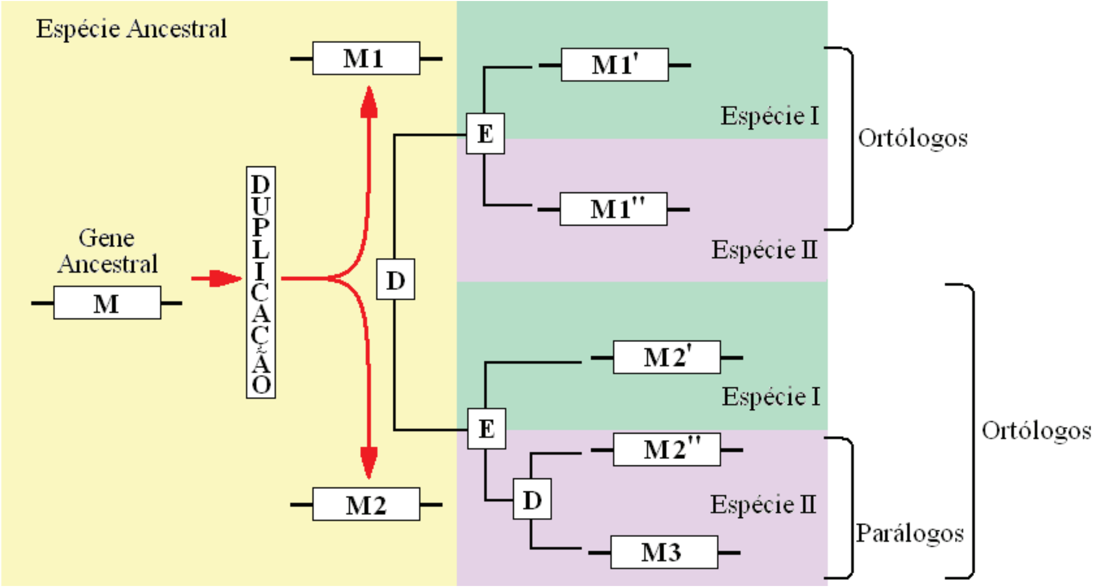


Figura 2. Representação das relações evolutivas de ortologia e paralogia entre genes. Um gene ancestral M sofre uma duplicação (D) originando os parálogos M1 e M2 no genoma da espécie ancestral. Um evento de especiação (E) gera dois grupos de ortólogos, M1 e M2 nas espécies I e II. Na espécie II, M2'' e M3 são parálogos gerados por uma duplicação ocorrida depois da separação entre as espécies I e II.

As relações de ortologia entre genes de famílias multigênicas podem ser avaliadas através da definição de grupos de genes ortólogos. Cada grupo de genes ortólogos é assumido como sendo resultado da evolução de um gene ancestral através de eventos e duplicação e especiação. Considerando que todos os genes dentro de um mesmo grupo de ortólogos se originaram de uma única cópia gênica em um dado momento da história evolutiva, o estudo comparativo entre eles é importante para entender como a evolução atua na diversificação ou manutenção de uma função ancestral (Tatusov *et al*., 1997). Grupos de ortólogos são de grande valor para estudos de taxas evolutivas e de duplicação gênica (Henikoff *et al*., 1997).

1.5 Evolução molecular em famílias multigênicas

Uma família multigênica é um grupo de genes que descende de um mesmo gene ancestral, portanto é esperado que as diversas cópias tenham similaridades em sequência de DNA, sequência e estrutura proteicas e funcionalmente (Nei e Rooney, 2005). Historicamente, os processos evolutivos que atuam em famílias multigênicas foram alvo de controvérsia. O paradigma evolutivo de famílias multigênicas antes da década de 1970 era o das hemoglobinas e mioglobina (Ingram, 1961). Os genes que codificam tais proteínas são filogenéticamente relacionados e divergiram gradualmente conforme os genes duplicados adquiriram novas funções. Este modelo de evolução foi chamado de **evolução divergente** (*divergent evolution*; Figura 3a).

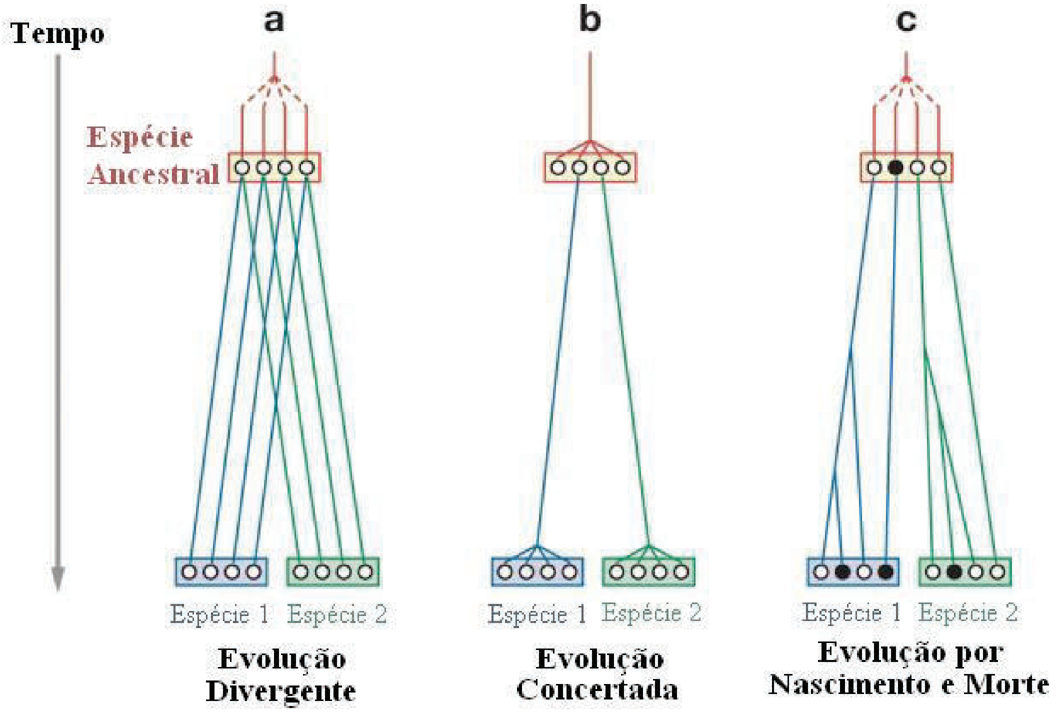


Figura 3. Três modelos diferentes para evolução de famílias multigênicas. Círculos brancos representam genes funcionais fixados e os pretos pseudogenes (extraído de Nei e Rooney, 2005).

Durante a década de 1970 pesquisas em *Xenopus* (anfíbio) mostraram que seus genes de RNAs ribossomais (rRNAs) ocorriam em um grande número de cópias *in tandem* e que as sequências de nucleotídeos intergênicas eram mais similares dentro de uma espécie do que entre duas espécies relacionadas (*X. laevis* x *X. mulleri*; Brown *et al*., 1972). Em plantas rRNAs também estão dispostos *in tandem* e apresentam o mesmo padrão evolutivo (Muir e Schlötterer, 1999). Estas observações não podiam ser explicadas pelo modelo de evolução divergente, portanto um novo modelo foi criado chamado de **evolução orquestrada** (*concerted evolution*). Neste modelo era necessário um mecanismo que homogeneizasse as diferentes cópias dentro de cada genoma enquanto permitisse a divergência entre espécies ou populações (Figura 3b). Uma mutação ocorrida em uma das cópias se espalharia através de todos os genes membros desta família através da ocorrência repetida de recombinações desiguais e/ou por conversão gênica (recombinação não-recíproca na qual o segmento de DNA de um gene receptor é copiado de um gene doador). Este processo leva à homogeneização das sequências gênicas e intergênicas dos rRNAs. No entanto, ao contrário do que se pensou inicialmente, não se pode excluir a ação de seleção purificadora nas sequências transcritas, que também contribui para a homogeneização das cópias. Atualmente a evolução de famílias de rRNAs é encarada como um misto de seleção purificadora e evolução orquestrada (Nei e Rooney, 2005).

O aparente sucesso do modelo de evolução orquestrada levou muitos autores a acreditar que a maioria das famílias multigênicas evoluía desta forma (Hood *et al*., 1975; Ohta, 1981). Conforme o número de sequências gênicas e proteicas aumentava percebeu-se que o modelo de evolução orquestrada não se aplicava à maioria e outro modelo foi proposto chamado **nascimento-e-morte** (Nei e Hughes, 1992). Neste modelo novos genes são criados por duplicação gênica e alguns dos genes duplicados são mantidos no genoma por longos períodos de tempo enquanto outros genes são eliminados ou se tornam pseudogenes através do acúmulo de mutações *non-sense* (Figura 3c). O modelo de nascimento-e-morte consegue explicar o surgimento de novas funções e mostrou-se aplicável à maioria das famílias multigênicas (Hughes e Nei, 1990 [o termo birth-and-death ainda não havia sido cunhado]; Ota e Nei, 1994; Zhang *et al*., 1998). Mesmo famílias gênicas altamente conservadas em termos de sequência como a das ubiquitinas e histonas parecem ter evoluído por nascimento-e-morte. Seu alto grau de conservação se deve primariamente à seleção purificadora ao invés de evolução orquestrada.

Em angiospermas o desenvolvimento dos órgãos florais é controlado por fatores de transcrição do tipo MADS-box e várias classes destes genes se mostraram essenciais para o desenvolvimento de flores (Ma e de Pamphilis, 2000; Theissen, 2001; Weigel e

Meyerowitz, 1994). Análises filogenéticas dos genes MADS-box mostraram que os elementos que controlam o florescimento tiveram origem a aproximadamente 650 milhões de anos atrás, muito antes do surgimento das flores (Nam *et al*., 2003; Tanabe *et al*, 2005). Se pensarmos que os registros fósseis mais antigos para angiospermas e gimnospermas têm, respectivamente, 150 e 300 milhões de anos, aparentemente os genes ancestrais dos MADS-box que controlam o desenvolvimento floral existem por um longo período de tempo antes da evolução das flores. Especula-se que este grupo de genes originalmente controlava o desenvolvimento dos estágios haploide e diploide em algas-verdaes (Tanabe *et al*, 2005). MADS-box também existem em animais e fungos. Em animais eles controlam o desenvolvimento muscular, por exemplo. No processo de evolução das gimnospermas e angiospermas, diferentes MADS-box parecem ter evoluído no controle do desenvolvimento dos órgãos reprodutivos (Nam *et al*., 2003).

1.6 Filogenética aplicada ao estudo de famílias multigênicas

Árvores filogenéticas de genes são comumente empregadas no estabelecimento de relações de ortologia e paralogia em famílias multigênicas. Os métodos mais utilizados para construção de árvores filogenéticas incluem os métodos de distância, de parcimônia e de verossimilhança. Os métodos de reconstrução são baseados na topologia da árvore e no comprimento dos ramos de forma que a árvore resultante possa ser testada estatisticamente (Nei e Kumar, 2000).

O método de **máxima parcimônia** foi originalmente desenvolvido para caracteres morfológicos (MP; Henning, 1966) sendo rapidamente utilizado para construção de árvores com dados de substituição de amino ácidos (Eck e Dayhoff, 1967). Neste método um alinhamento de quatro ou mais sequências é utilizado para reconstruir a sequência ancestral mais provável. As mudanças mutacionais são assumidas como ocorrendo em todas as direções entre os quatro nucleotídeos ou 20 amino ácidos. O menor número de substituições que explica a diversidade encontrada nos dados é computado para cada topologia potencialmente correta. A topologia que requer o menor número de substituições é então escolhida como a melhor árvore. A base teórica deste método é a ideia filosófica de William de Ockham, segundo a qual a melhor hipótese para explicar um processo é aquela que requer o menor número de eventos a serem assumidos.

Na ausência de substituições retrocedentes ou paralelas (ausência de homoplasia) em cada posição considerada e sendo grande o número de nucleotídeos ou aminoácidos alinhados para cada posição o método de MP deverá produzir a topologia correta. Na prática, no entanto, as sequências são alvo de mutações retrocedentes e paralelas e o número de nucleotídeos ou amino ácidos alinhados pode não ser muito grande. Neste caso a MP tende a dar topologias incorretas (Nei e Kumar, 2000). Felsenstein (1978) mostrou que quando a taxa de substituição de nucleotídeos varia muito entre as linhagens evolutivas em análise, o método de MP gera topologias incorretas mesmo quando um número infinito de nucleotídeos é analisado. Sob certas condições, mesmo quando a taxa de substituição for constante para todas as linhagens, os métodos de MP podem gerar resultados incorretos (Hendy e Penny, 1989; Zharkikh e Li, 1992; Takezaki e Nei, 1994; Kim, 1996). Nestes casos os ramos longos ou curtos da árvore tendem a se agrupar na árvore reconstruída por MP. Estes fenômenos são chamados de "*long-branch attraction*" (Hendy e Penny, 1989) e "*short-branch attraction*" (Nei, 1996). Outro aspecto negativo do método de MP é que o tempo de computação necessário para analisar uma grande quantidade de sequências é sempre muito longo (Nei e Kumar, 2000).

O método de Máxima Verossimilhança (MV; Cavalli-Sforza e Edwards, 1967) originalmente foi desenvolvido para inferências filogenéticas baseadas em dados de frequência gênica. Felsenstein (1981) desenvolveu um algoritmo que utilizava sequências de nucleotídeos para reconstruir árvores filogenéticas. Kishino *et al*. (1990) adaptou o algoritmo para utilizá-lo com sequências de amino ácidos usando a matriz experimental de substituição de amino ácidos de Dayhoff *et al*. (1978).

No método de MV o parâmetro a ser considerado não é a topologia, mas sim o comprimento dos ramos para cada topologia. Este método apresenta menor variância do que os outros métodos e tende a ser robusto, tem boa base estatística para lidar com dados de sequência, no entanto é um método proibitivo para análises que envolvem grande quantidade de sequências devido ao seu grande consumo de memória (Nei e Kumar, 2000).

Métodos de distância computam as distâncias evolucionárias para todos os pares de sequências de forma que a árvore filogenética resultante é construída considerando as relações ao longo destes valores de distância (Nei e Kumar, 2000). Vários são os métodos que permitem construir árvores filogenéticas à partir de dados de distância. O mais simples é o UPGMA (*Unweighted pair-group method using arithmetic averages*; Sokal e Michener, 1958; Sneath e Sokal, 1973). Originalmente desenvolvido para taxonomia numérica, este método produz árvores razoavelmente boas quando as frequências de substituição de nucleotídeos ou aminoácidos são mais ou menos constantes (Nei *et al*., 1983; Takezaki e Nei, 1996). UPGMA tem como intenção reconstruir árvores de espécies e não árvores de genes. Erros topológicos ocorrem frequentemente quando a taxa de substituição não é constante ou quando número de sequências ou de nucleotídeos alinhados é pequeno (Nei e Kumar, 2000).

Outro método de distância bastante utilizado é o método de evolução mínima (*minimum evolution*; ME; Edwards e Cavalli-Sforza, 1963; Kidd e Sgaramella-Zonta, 1971). Neste método o somatório total das distâncias (S) representadas pelo comprimento total dos ramos de uma árvore é calculado para todas as topologias plausíveis e então a árvore com o menor valor de S é escolhida como a melhor árvore (Nei e Kumar, 2000).

Para contornar o problema do grande tempo computacional despendido para processar filogenias utilizando ME, Saitou e Nei (1987) desenvolveram um método de construção de árvores eficiente e que se baseia no mesmo princípio de ME. Este método não examina todas as topologias possíveis, mas a cada estágio de agrupamento das sequências o princípio da evolução mínima é utilizado. Este método, que consiste numa simplificação do ME, é chamado *Neighbor-Joining* (NJ).

Um dos mais importantes conceitos no método de NJ é o de vizinhos, que são definidos como duas sequências que estão conectadas por um só nó numa árvore não-enraizada. Por exemplo, 1 e 2, na árvore da Figura 4, são vizinhos pois estão conectados por somente um nó A. Da mesma forma 5 e 6 são vizinhos conectados pelo nó B. Todos os outros pares não são vizinhos. Entretanto, 1 e 2 podem ser combinados e considerados como um único táxon, então (1-2) e 3 passam a serem vizinhos. É possível definir a topologia da árvore pelo agrupamento sucessivo dos vizinhos e pela produção de novos vizinhos. A topologia da árvore da Figura 4 pode ser descrita pelos seguintes pares de vizinhos: (1,2), (5,6), (1-2, 3) e (1-2-3, 4) (Nei e Kumar, 2000).

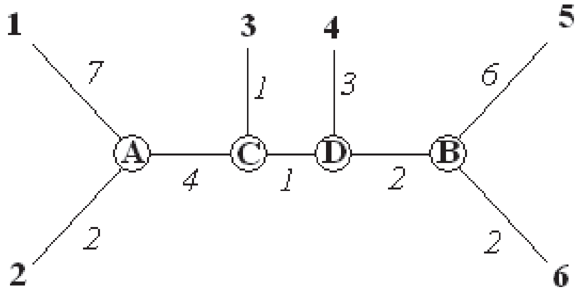


Figura 4. Filogenia de seis sequências com tamanho de ramos conhecidos (extraído de Nei e Kumar, 2000).

O método de NJ é rápido comparado aos outros, portanto é apropriado a grandes conjuntos de dados, permite linhagens com diferentes tamanhos de ramos e substituições múltiplas, no entanto ele mostra apenas uma topologia possível (Nei e Kumar, 2000).

1.7 Controle da expressão gênica em eucariotos

O genoma de uma célula contém, em sua seqüência de DNA, as unidades informacionais (genes) capazes de controlar a síntese de milhares de proteínas e RNAs não-codificantes. O genoma da planta modelo *Arabidopsis thaliana*, por exemplo, contém aproximadamente 28 mil genes codificantes de proteína (www.arabidopsis.org), além de diversos genes produtores de RNAs não-codificantes que exercem funções importantes na regulação do crescimento, desenvolvimento e respostas a estresses (Matik, 2001; Billoud *et al*., 2005; Ben Amor *et al*., 2008; Wu *et al*., 2012). Estes genes, apesar de estarem presentes em todas as células de cada organismo, não são expressos ao mesmo tempo, tampouco em todos os órgãos e tecidos, sendo ao contrário, precisamente regulados no tempo e no espaço. **Controle da expressão gênica** é o termo usado para os mecanismos que definem o momento, o local e a quantidade que cada gene será expresso.

O controle da expressão gênica pode ocorrer em diferentes níveis: transcrição (McKnight & Kingsbury, 1982; Singh, 1998; Kaufmann *et al*., 2010), processamento e controle da estabilidade dos mRNAs, tradução dos mRNAs (Van der Kelen *et al*., 2009), modificações pós-traducionais das proteínas, multimerização em complexos protéicos, controle da localização celular e regulação da estabilidade das proteínas. Estes diferentes níveis de regulação operam simultaneamente e de forma integrada permitindo um fino controle da expressão gênica.

A modulação da transcrição gênica depende de fatores reguladores de transcrição (elementos *trans*), que são proteínas que se ligam a pequenas seqüências (geralmente entre 5 e 10nts) de DNA (elementos *cis*) contidas na região promotora dos genes (Latchaman, 1990), aumentando ou reduzindo a taxa de iniciação de transcrição pela RNA polimerase (Lee & Young, 2000; Beckett, 2001; Warren, 2002) (Figura 5).

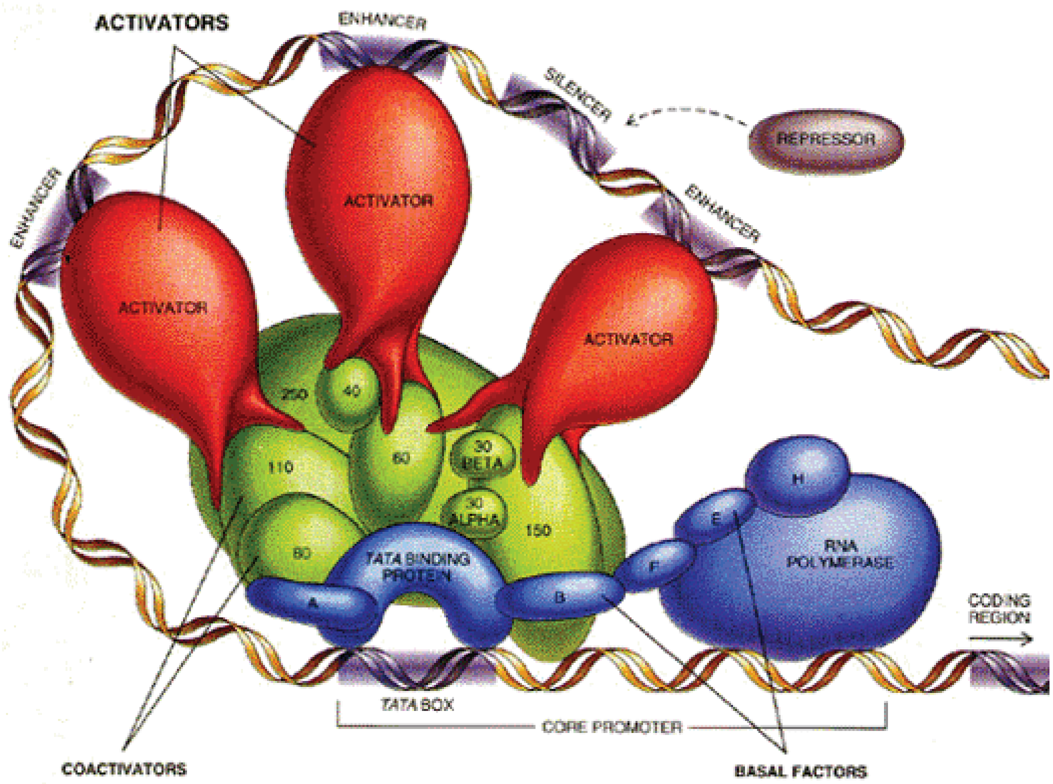


Figura 5. Representação esquemática do complexo de transcrição, contendo os fatores basais, co-ativadores, ativadores e repressores de transcrição. (Extraído de http://berkeley.edu/news/features/1999/12/09_3dimage.html)

O controle da expressão gênica também pode ocorrer através de mecanismos pós-transcricionais envolvendo a modulação da estabilidade dos mRNAs e da taxa de tradução. Quanto à estabilidade dos mRNAs, estes podem ser classificados como estáveis (meia-vida de algumas horas), muito estáveis (meia-vida de dias) e instáveis (meia-vida de aproximadamente 1 hora) (Jonhson et al, 1998). Os componentes moleculares que controlam a estabilidade de mRNAs podem ser alocados em três categorias: 1. Componentes do sistema basal de degradação de mRNA; 2. Fatores (elementos *trans*) e sequências específicas (elementos *cis*) que controlam a estabilidade inerente dos mRNA; e 3. Elementos transdutores de estímulos externos e/ou condições fisiológicas modulando a estabilidade de mRNAs específicos (miRNAs, por exemplo; Gutiérrez et al, 1999). O controle da tradução efetuado por miRNAs foi recentemente descrito em plantas por Brodersen *et al*., 2008 (geralmente miRNAs tem como função a clivagem do mRNA alvo) e desempenha um papel importante na regulação da expressão gênica através da atenuação das taxas de tradução de mRNAs alvos.

1.8 Mecanismos sensores e vias de sinalização dos açúcares glicose e sacarose e do hormônio ABA em plantas

Plantas são organismos sésseis e foto-autotróficos, capazes de capturar e transformar energia luminosa em energia química, utilizada principalmente para síntese de açúcares capazes de sustentar todos os processos metabólicos do organismo. Os açúcares produzidos pelos tecidos fotossintéticos (fonte) são transportados para os órgãos dreno, como raízes, zonas meristemáticas e órgãos reprodutivos através do floema. Na otimização do crescimento e desenvolvimento, as plantas ajustam a produção de açúcares com a demanda dos tecidos dreno, mas também integram eficientemente o metabolismo de carboidratos com a disponibilidade de nutrientes minerais, bem como estresses bióticos e abióticos. Plantas desenvolveram redes regulatórias interconectadas nas quais açúcares tem um papel central (Forde, 2002; Rolland *et al*., 2006; Rook *et al.*, 2006; Gutiérez *et al*., 2007). Açúcares são sinais metabólicos chave no controle da expressão gênica (Koch, 1996; Price *et al*., 2004; Li *et al*., 2006) e na modulação de diferentes estágios do desenvolvimento incluindo embriogênese, germinação, desenvolvimento da plântula, crescimento das raízes, florescimento e outro processos importantes como fotossíntese, senescência e respostas a estresses (Smeekens, 2000; Rolland *et al*., 2002; Moore *et al*., 2003; Gibson, 2005).

Uma série de mecanismos sensores e sinalizadores de açúcares foram descritos. O dissacarídeo sacarose, que é o principal produto da fotossíntese e a forma transportada de açúcar, regula especificamente a tradução de um grupo restrito de fatores de transcrição do tipo bZIP de *Arabidopsis* (Wiese *et al*., 2004). Um destes genes, *AtbZIP11*, possivelmente media a interação entre carbono e nitrogênio (Hanson *et al*., 2007). Um dos produtos da hidrólise de sacarose é glicose, que parece ser o principal sinalizador entre os metabólitos de açúcar. A caracterização do mutante *glucose insensitive 2* (*gin2*) de *Arabidopsis* provê evidências para atividade sensor e de sinalização dependente de HXK1 (hexoquinase 1), porém desacoplada de sua atividade de fosforilação, que medeia à repressão de genes relacionados à fotossíntese, bem como controla o crescimento de *Arabidopsis* (Moore *et al*., 2003). O mecanismo molecular responsável pela repressão transcricional glicose-dependente do gene *CAB2* (chlorophyll a/b protein) está sendo decifrado e envolve um complexo nuclear de HXK1 que se liga diretamente ao promotor do gene *CAB2*. (Cho *et al.*, 2006). Outros mecanismos sensores de glicose foram descritos (para uma revisão veja Rolland *et al*., 2002 e 2006). Um deles é uma via dependente de glicólise que requer a atividade catalítica de HXK e regula a expressão dos genes de resposta a patógenos *PR1* and *PR2* (Xiao *et al.*, 2000). Uma terceira via está envolvida na regulação de um grupo de genes codificantes para chalcone synthases e invertases de parede celular e é independente da atividade de HXK1 (Roitsch, 1999; Xiao *et al.*, 2000). Evidências genéticas também indicam que o mecanismo sensor e transdutor de glicose, independente de atividade de hexoquinase, envolve receptor do tipo proteína G (GPCR) existente em plantas (Ullah *et al*., 2002; Chen JG *et al*., 2003; Chen & Jones, 2004; Chen Y *et al*., 2006; Huang *et al*., 2006). Recentemente foi reportada a interação entre a via de proteína G e o transporte de hexose localizado no complexo de golgi (Wang *et al*., 2006).

A caracterização de mutantes das vias de sinalização de glicose em *Arabidopsis* tem revelado a relação entre glicose e as vias de sinalização dos hormônios ABA (Zhou et al., 1998; Arenas-Huertero et al., 2000; Laby et al., 2000; Huisjer et al., 2000; Rook et al., 2001; Brocard et al., 2002; Cheng et al., 2002; Broccard-Gifford et al., 2004) e etileno (Zhou et al., 1998; Gibson et al., 2001; Cheng et al., 2002). Vem sendo proposto que a inibição do desenvolvimento inicial por alta concentração de glicose é HXK1-dependente e requer a biosíntese e subsequente transdução de sinal de ABA, um hormônio vegetal responsável pela transdução de sinal de estresses como dessecação levando ao fechamento estomático (Zhou et al., 1998; Arenas-Huertero et al., 2000; Cheng et al., 2002). Dentro desta cascata regulatória, o gene *ABA2*/*GIN1* que codifica para uma dehidrogenase/redutase de cadeia curta envolvida na síntese de ABA (Cheng *et al*., 2002) e o fator de transcrição do tipo APETALA 2 *ABI4*/*GIN6* envolvido na sinalização de ABA (Finkelstein et al. 1998; Arenas-Huertero et al., 2000) parecem ter um papel central (para revisão veja León & Shenn, 2003; Rolland *et al*., 2006). ABI4 também atua como repressor no promotor de um gene rbcS em resposta a glicose ou ABA (Acevedo-Hernández *et al*., 2005). ABI4 também foi mostrado como ativador sinérgico do gene ADP-glucose pyrophosphorilase por ABA e glicose, que configura um exemplo de indução por glicose que é modulada por ABA (Rook *et al*., 2001; Li *et al*., 2006). A modulação da expressão gênica de forma sinérgica entre glicose e ABA também foi demonstrado para o gene *AtbZIP63* de *Arabidopsis* (Matiolli *et al.*, 2011). Mutações no gene *ABI5* (bZIP), outro componente da via de sinalização de ABA (Finkelstein & Linch, 2000), também conferem um fenótipo de insensibilidade a glicose (GIN; Arenas-Huertero et al., 2000; Brocard et al., 2002). Outros bZIPs homólogos a ABI5 e capazes de ligar ao elemento responsivo a ABA ("ABA-responsive element" – ABRE) também participam da sinalização de glicose (Kang et al., 2002; Kim et al., 2004). Demonstrou-se pelo menos um conjunto de sete bZIPs sendo rapidamente regulados por glicose, possivelmente constituindo importantes intermediários na cascata transcricional ativada por glicose (Li *et al*., 2006). Parece claro que o controle transcricional é um aspecto essencial da cascata regulatória mediada por glicose, bem como a interação entre os sinais glicose e ABA que parece ser crucial para a integração das respostas mediadas pela disponibilidade de carbono (energia) e estresse abiótico.

Recentemente um conjunto de genes relacionado ao conteúdo de sacarose em cana-de-açúcar foi reportado. Alguns destes genes são elementos de sinalização de açúcares e ABA (Papini-Terzi *et al*., 2009). Mostramos que cinco genes, de 24, de resposta rápida a glicose e sacarose em cana-de-açúcar possuem ortólogos em arabidopsis que respondem de forma conservada em um tratamento curto com estes açúcares (3% de glicose, 0,5% de sacarose; Li *et al*., 2006; Papini-Terzi *et al*., 2009). Este fato mostra que alguns elementos de transdução de sinal de açúcares, entre eles kinases e fatores de transcrição, possuem resposta transcricional conservada entre as duas principais linhagens de angiospermas (*Arabidopsis* – eudicot e cana – monocot). Possivelmente estes genes representam pontos essenciais de controle nestas vias de transdução de sinais que foram fixados antes da divergência entre eudicotiledôneas e monocotiledôneas.

1.9 Redes de regulação gênica

A vida de uma célula pode ser entendida como o produto de programas de expressão envolvendo a regulação transcricional coordenada de milhares de genes. Tais programas de expressão gênica dependem do reconhecimento de sequências específicas em promotores por proteínas regulatórias, essencialmente fatores de transcrição. Desta forma a coleção de proteínas regulatórias associadas aos seus genes alvo pode ser descrita como uma rede de regulação transcricional. Redes de regulação da expressão gênica controlam desde simples rotas bioquímicas até processos complexos do ponto de vista espaço-temporal como o desenvolvimento do corpo de seres multicelulares (animais e plantas terrestres especialmente). Em termos físicos, basicamente estas redes se constituem de milhares de sequências modulares de DNA. Cada um destes módulos recebe e integra múltiplas entradas (***inputs***) na forma de proteínas regulatórias (que podem ser ativadores ou repressores da expressão dos genes alvo) que reconhecem sequências alvo específicas. O resultado final é o controle preciso da expressão dos genes em um genoma (***outputs***). Redes de regulação representam explicitamente a causalidade em processos biológicos, como respostas a estresses ambientais e desenvolvimento (Figura 6). Elas explicam como sequências genômicas controlam a expressão de conjuntos de genes que progressivamente adaptam a fisiologia da célula a uma nova condição ou então geram padrões de desenvolvimento levando a múltiplos estágios de diferenciação celular (Davidson, 2005, Karlebach e Shamir, 2008; Tkačik e Walczak, 2011).

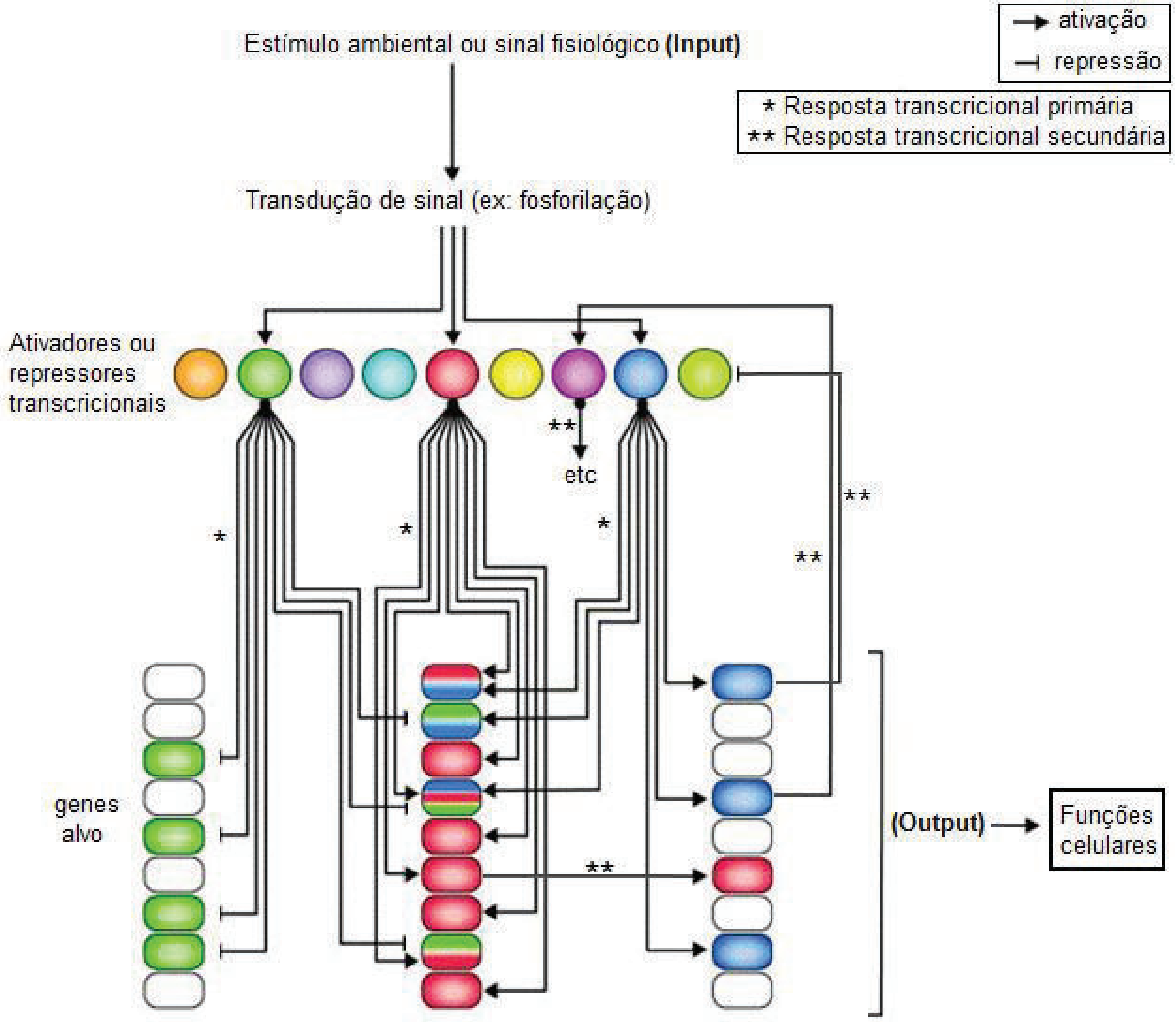

Estímulo ambiental ou sinal fisiológico (Input)
ativação
repressão
* Resposta transcricional primária
** Resposta transcricional secundária
Transdução de sinal (ex: fosforilação)
Ativadores ou repressores transcricionais
**
etc
*
*
*
**
**
**
genes alvo
(Output)
Funções celulares

Figura 6. Esquema ilustrativo de uma rede de regulação da expressão gênica. Um sinal ou estímulo (*input*) é percebido e ativa uma via de transdução de sinal, que geralmente é constituída por quinases que ativam fatores de transcrição via fosforilação. Estes fatores podem ser ativadores ou repressores da expressão gênica. No esquema as setas significam ativação enquanto as barras significam repressão da expressão gênica dos genes alvo. Na rede vemos a resposta primária (*) e uma resposta secundária (**) causada pela ação dos elementos que foram transcritos na resposta inicial. O *output* da rede é o conjunto de funções celulares desempenhadas pelos elementos induzidos e pela eliminação ou atenuação da expressão dos elementos reprimidos.

Dentro deste contexto alguns conceitos centrais de redes surgem:

1 – A regulação da expressão gênica se apresenta como um sistema de processamento lógico. Cada módulo regulatório contido no genoma recebe múltiplos *inputs* e os processa. Este processamento pode ser matematicamente representado como combinações de funções lógicas (por exemplo, funções “*and*”, “*or*”, “*switch*”, “*nor*”,

etc). Em termos de sistemas, uma rede regulatória consiste da assembleia dessas unidades processadoras de informação. Essencialmente redes deste tipo podem ser entendidas grosso modo (e não de forma literal) como sendo análogas a dispositivos computacionais, as funções dos quais são condicionadas pelos *inputs* que recebem.

2 – Causalidade na regulação do genoma: As razões pelas quais genes são expressos no tempo e no espaço são descritas pela **arquitetura** ou **topologia** de redes. A topologia de uma rede é definida pelo agregado total dos padrões lógicos de interação entre os genes que a constituem. As funções biológicas que emergem do funcionamento de tais redes só pode ser explicada por sua topologia (interação entre genes), sendo essas funções difíceis de detectar ao nível de qualquer de seus genes individuais. Exemplos disto são circuitos multigênicos que atuam na produção de ciclos de *feedback* positivo e negativo. Desta forma é crítico determinar a topologia de redes regulatórias. Uma forma de decifrar a topologia de redes é através de perturbações experimentais destes sistemas (por exemplo, inserindo algum *input* exógeno ao sistema) seguidas de medições do efeito alcançado na transcrição de genes individuais (por exemplo, técnicas globais de medição da expressão gênica como *microarrays* ou RNA-seq). Os modelos produzidos desta forma podem ser testados experimentalmente via biologia molecular determinando diretamente a função de uma sequência ou módulo regulatório (uso de mutantes, por exemplo).

3 – Estrutura de redes: Redes de regulação são composições de diversos tipos de subcircuitos não homogêneos. Cada um realizando um tipo específico de função. Este conceito é especialmente importante, pois determina os princípios mecanisticos por traz de uma rede. Alguns subcircuitos são recorrentes em uma diversidade de contextos biológicos e processos (por exemplo, a maior parte dos subcircuitos de transdução de sinal contém subcircuitos de lógica semelhante). Ainda assim, análises evolutivas mostram que diversos subcircuitos são muito flexíveis e maleáveis. Até pequenas modificações nas conexões entres os genes em tal contexto pode produzir diversidade morfológica (desenvolvimento) entre grupos relacionados de animais.

4 – Reengenharia de sistemas biológicos: Para redesenhar sistemas de controle genéticos para fins intelectuais ou práticos, é necessário entender o fluxo de causalidade em redes gênicas. Tal entendimento requer uma mistura interdisciplinar de teoria e experimentos, biologia molecular e computacional, alta tecnologia e formas sofisticadas

de manipular sistemas biológicos. Uma vez decifrada e controlada experimentalmente a riqueza embutida no controle genômico certamente nos dará evidências sobre o funcionamento de processos biológicos que estamos apenas começando a definir.

Em plantas, especificamente, o estudo de redes regulatórias tem explicado diversos aspectos de sua biologia. Como exemplo o desenvolvimento do plano do corpo de plantas (*body plan*; Alvarez-Buylla et al, 2007), determinação do destino celular no desenvolvimento de flores (Espinosa-Soto et al, 2004), resposta a seca (Valliyodan e Nguyen, 2006), desenvolvimento de fruto (Mounet *et al*, 2009), transição do ápice vegetativo para tecidos germinativos (Dong *et al*, 2012), desenvolvimento de raízes (Brady *et al*, 2011; Bruex *et al*, 2012), o papel de miRNAs (Meng *et al*, 2011) e *long noncoding* RNAs (Kim e Sung, 2011) em redes regulatórias, desenvolvimento de tricomas e pelos radiculares (Ishida *et al*, 2008), entre outros.

1.10 Evolução de redes de regulação, universalidades evolutivas e a emergência de sistemas biológicos complexos

As explicações evolutivas para a origem da modularidade em redes e dos programas de desenvolvimento geralmente assumem intrinsecamente uma vantagem seletiva (Force *et al*., 2005). Um intenso debate vem se transcorrendo no estudo teórico da evolução biológica à cerca deste tema. Seria a topologia das redes genéticas, que em último nível governam as funções biológicas dos seres vivos, produto do acúmulo de pequenas variações topológicas que dariam vantagens seletivas (adaptacionais) ou meros efeitos colaterais (**propriedades emergentes**) das forças evolutivas não-adaptativas (deriva genética, mutação e recombinação)?

Opiniões se dividem neste assunto. Segundo Michael Lynch (2005): "A vasta maioria dos biólogos envolvidos em estudos evolutivos interpretam virtualmente todos os aspectos da biodiversidade em termos adaptativos. Esta visão estreita tem se tornado insustentável à luz das recentes observações do sequenciamento de genomas e da teoria de genética populacional." Um exemplo de tal visão pode ser encontrado em Dennett (1995): "Seleção natural deve ser tratada como a explicação padrão para fenótipos complexos, a menos que alguém possa mostrar que ela não tem um papel". A visão mais moderna que vem se desenvolvendo entre alguns evolucionistas, de que forças não-adaptativas teriam um papel extremamente importante na evolução de características complexas tem incomodado geneticistas mais tradicionais. Pugliucci (2007) se vale de uma afirmação anterior de Popper (1977): "A teoria neodarwiniana é estritamente uma teoria de genes, ainda que o fenômeno que tem de ser explicado é aquele da transformação da forma", para posicionar que: "... ainda que necessária, a genética de populações não é nem próxima de suficiente para entender como fenótipos evoluem." Uma forma de entender se há mais razão em um destes lados em debate seria buscar por universalidades evolutivas. Comportamentos genéticos e evolutivos que, independente da história seletiva de cada linhagem, sejam constantes entre todas as formas de vida. Um conjunto crescente de dados vindos da revolução genômica tem apoiado a existência de universalidades evolutivas. Se estes dados realmente apontam para tal, seria difícil ou virtualmente impossível, que tais universalidades fossem produto da seleção natural, que por definição é circunstancial e por esta razão, desde sua formulação inicial por Darwin, vem sendo utilizada para explicar a divergência entre as linhagens evolutivas. Os seres-vivos seriam diferentes, pois enfrentaram um conjunto histórico de seleção independente e único. O que há, então, de universal na evolução de genomas?

Koonin (2011) propõe que as universalidades evolutivas sejam tratadas como leis da evolução, num sentido comparável às leis da física moderna. As universalidades já encontradas incluem:

1 – a distribuição log-normal das taxas evolutivas entre genes ortólogos (Grishin et al., 2000; Drummond *et al*., 2008; Wolf *et al.*, 2009).

2 – Distribuições do tipo "lei da potência" (*power-law*) no número de parálogos em famílias gênicas e no *node-degree* em redes biológicas (Barabási e Oltvai, 2004; Karev *et al.*, 2002; Koonin *et al*., 2002; Huynen e Nimwegen *et al*., 1998).

3 – Correlação negativa entre a taxa evolutiva de um gene e seu nível expressão (e abundância da proteína resultante) (Pal *et al.*, 2001; Krylov *et al*., 2003; Drummond *et al*., 2005; Drummond *et al*., 2006).

4 – Dependência (ou correlação) diferencial entre número de genes num genoma e números de genes em famílias gênicas de classes funcionais diferentes (sem

dependência ou correlação: componentes do sistema de tradução; dependência linear: enzimas e elementos do metabolismo; dependência quadrática: componentes de sistemas de transdução de sinal e genes reguladores) (van Nimwegen, 2003; Molina e van Nimwegen, 2009).

Uma representação gráfica dos dados que suportam as universalidades apresentadas acima pode ser visto na Figura 7. Koonin (2011) conclui disto que "universalidades evolutivas são propriedades emergentes de conjuntos gênicos, não características selecionadas".

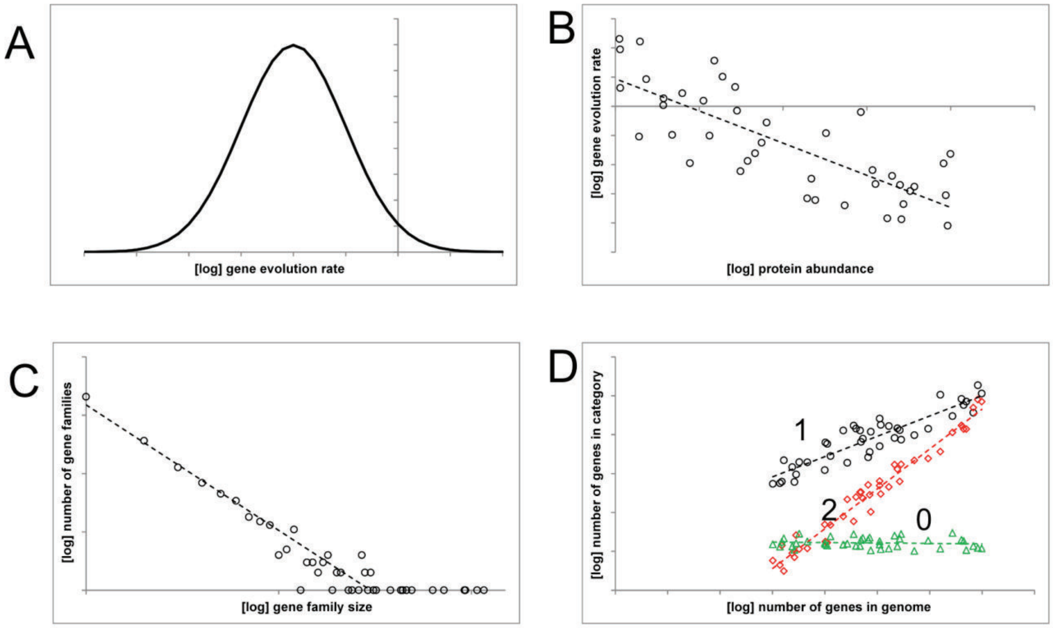


Figura 7. Universalidades da evolução de genomas molecular. A figura mostra versões idealizadas de dependências universais e distribuições. (A) Distribuição log-normal da taxa evolutiva de genes ortólogos. (B) Anticorrelação entre o nível de expressão de genes (abundância das proteínas) e a taxa evolutiva das sequências codificantes. (C) Distribuição *power law-like* do número de parálogos em famílias gênicas. (D) Diferentes incrementos no número de genes de diferentes classes funcionais com relação ao número total de genes em um genoma (0 – sem dependência, típico de componentes do sistema de tradução; 1 – dependência linear, caracterísitca de enzimas metabólicas; 2 – dependência quadrática, característica de componentes regulatórios e de transdução de sinal) (Extraído de Koonin *et al*, 2011).

O processo e o curso da evolução são, no entanto, criticamente afetados por contingências históricas e envolvem constante ajuste seletivo (Jacob, 1977; Koonin, 2011 b). Portanto uma teoria física completa da evolução (ou qualquer outro processo com contribuição substancial de processos históricos) seria inconcebível. Entretanto, a universalidade de vários padrões simples de evolução genômica, e a capacidade de modelos matemáticos simples explicarem estas universalidades (como o modelo de nascimento-e-morte de Nei apresentado em seções anteriores, Koonin, 2011), sugere que "leis da biologia evolutiva", comparáveis em status com as leis da física, podem ser possíveis.

Uma teoria não-adaptativa (não envolve seleção positiva) de genética de populações foi proposta por Lynch (2007a e 2007b) para explicar a emergência de redes de regulação e por consequência o surgimento da complexidade biológica via aumento da complexidade de redes. Esta teoria mostra que muitas das características qualitativas das redes genéticas já conhecidas podem ser explicadas sem a necessidade de evocar nenhum tipo de processo adaptativos, deixando em aberto a questão de quanto a seleção positiva é necessária ou mesmo suficiente para explicar o surgimento da complexidade biológica (representada especialmente por vertebrados e plantas terrestres).

Observações qualitativas sugerem que a complexidade de redes regulatórias e de interação proteína-proteína aumenta de procariotos (Archaea e Bacteria) para eucariotos unicelulares e destes para eucariotos multicelulares, com *loops* autoregulatórios simples sendo mais comuns em seres unicelulares e *loops* com múltiplos componentes sendo mais comuns em eucariotos multicelulares (Thieffry *et al*., 1998; Lee *et al*., 2002; Wuchty e Almaas, 2005). Ainda mais, é uma questão em aberto o quanto redes regulatórias complexas são um pré-requisito para a evolução de fenótipos complexos, ou o quanto a arquitetura genômica das espécies multicelulares é simplesmente mais permissiva à emergência passiva de novas conexões em redes (Lynch, 2007 b).

Dado o grande número de fatores de transcrição na maioria dos seres vivos e sua ligação a sítios-alvo simples e pequenos (geralmente <10nt) que são sujeitos a mutações estocásticas capazes de mudar o status de ligação de um dado fator de transcrição (que podem ser definidos matematicamente como taxas de mutação: taxa de surgimento = μ_g ou inativação de um sítio de ligação = μ_l) há muitos mecanismos plausíveis para o surgimento de variação em redes regulatórias por processos neutros (Lynch, 2007 c; Johnson e Porter, 2000; Force *et al*., 2005; Haag e Molla, 2005). A taxa de inativação mutacional de um domínio de fator de transcrição pode ser definiada por $\mu_l = n\mu$, onde n é o número de nucleotídeos em um domínio de ligação ao DNA (geralmente de 5-

10nt) e μ é a taxa de mutação por nucleotídeo. A taxa de ganho de um domínio de ligação de um fator de transcrição, de forma neutra, depende do espaço ou quantidade de nucleotídeos com potencial de regular um gene (L). L pode varia em ordem de magnitude entre espécies, em bactérias raramente é maior que 10^2nt (universo de bases reguladoras em potencial de um operon) e em eucariotos pode ser tão alto quanto 10^4 - 10^6nt (espaço de DNA que pode ser ocupado por módulos reguladores controlando um a expressão de um dado gene). Desta forma o número potencial de sítios para o ganho de um domínio de ligação de fator de transcrição é dado por L-n+1, assumindo a não ocorrência *a priori*. Assumindo uma distribuição igualitária entre os quatro nucleotídeos, a probabilidade da ocorrência de domínios de ligação em potencial que diferem do domínio funcional por apenas um nucleotídeo é dada por $n.0{,}75.0{,}25^{n-1}$. Desta forma, com a expectativa de conversão mutacional de tal sítio potencial em um sítio ativo é $\mu/3$ (um em três possibilidades de mutação, por exemplo, se a base é A e a mutação para T ativa o sítio, as mutações de A para C ou G devem ser excluídas do cálculo). Logo μ_g=$(L$-$n+1)n\mu/4^n$ é chance mutacional da formação de um novo domínio de ligação de um fator de transcrição de forma neutra numa sequência potencialmente reguladora de um gene. Um parâmetro para a avaliação relativa das grandezas da perda de domínios versus o ganho de domínios reguladores em genes é a razão μ_l/μ_g, definida por $\alpha \approx 4^n/(L$-$n+1)$. Por exemplo, se $\alpha = \mu_l/\mu_g = 1.0$, a chance de ganhar um domínio de ligação de um fator de transcrição é a mesma de perder um domínio existente, se α = 10.0 a chance de perder um domínio existente é 10 vezes maior que ganhar um novo domínio, se α = 0.1 a chance de ganhar um domínio é 10 vezes maior que perder um domínio existente. A Figura 8 mostra um modelo para recrutamento de um regulador *upstream* num alelo genérico ilustrando conceitualmente a ação de μ_l e μ_g. Se um gene que não pode ser regulado (extremos na figura 8) for eliminado por seleção negativa, fica claro que num modelo genérico mínimo só é possível haver três estados para um alelo: auto-regulação (s), regulação dependente de um elemento *upstream* (u) ou ambas, sendo o alelo redundantemente regulado (r). Colocando esta construção teórica numa perspectiva de genética de populações, Lynch (2007 b) calculou as frequências de equilíbrio entre estas três configurações mínimas para um sistema mendeliano (Figura 9).

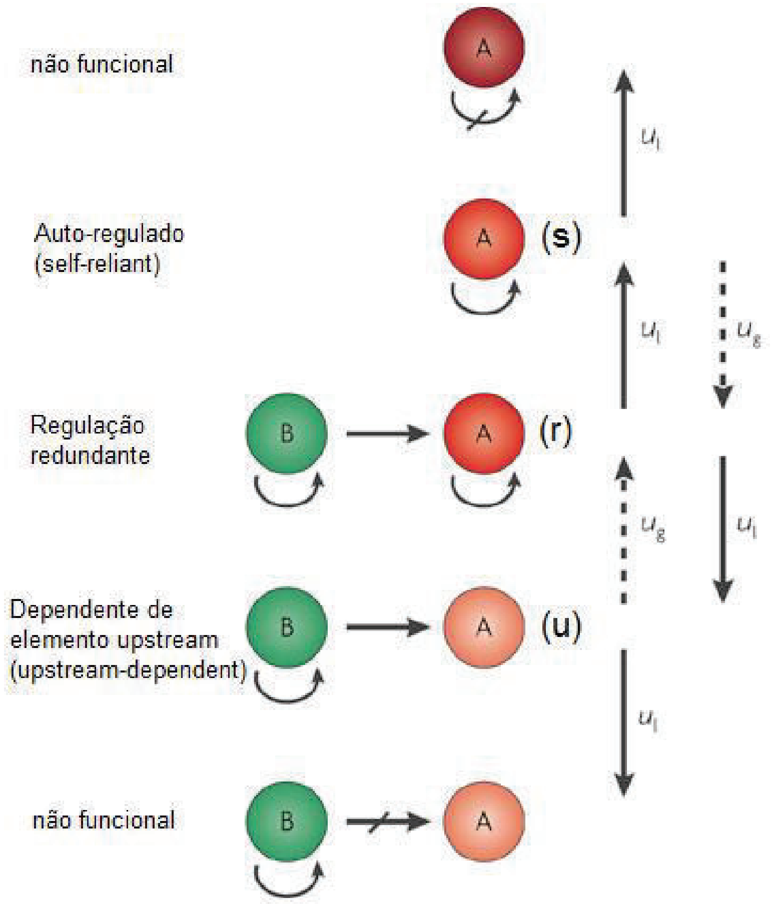


Figura 8. Modelo para recrutamento de um regulador *upstream*. O gene B é assumido como essencial, porém isso é independente de sua função sobre A, portanto está fixado na população em questão. Alelos não-funcionais de A são assumidos como sendo letais. As setas contínuas representam a perda de um domínio regulador e as setas pontilhadas representam o ganho de um domínio regulador (Adaptado de Lynch, 2007b).

Posicionando este modelo num contexto de genética de populações, para uma dada razão $\alpha=\mu_l/\mu_g$ o equilíbrio de frequências é dado por $p_{(s)}=p_{(u)}= \alpha/(1+2\alpha)$ para os alelos auto-regulados e os *upstream* dependentes e $p_{(r)}=1(1+2\alpha)$ para o alelo redundantemente regulado. Desta forma se a chance de ganhar um domínio regulador for progressivamente maior do que perder (contexto eucariótico multicelular) ocorrerá a fixação de alelos redundantemente regulados. Assim como se a chance de perder um domínio for progressivamente maior do que de ganhar um novo domínio (contexto de Bacteria e Archaea) ocorrerá a eliminação dos alelos redundantemente regulados em favor de um equilíbrio entre os sistemas mais simples de regulação ($p_{(s)}=p_{(u)}=0,5$). Com α=1 os três tipos de alelos encontram um equilíbrio na igualdade de suas frequências (Figura 9).

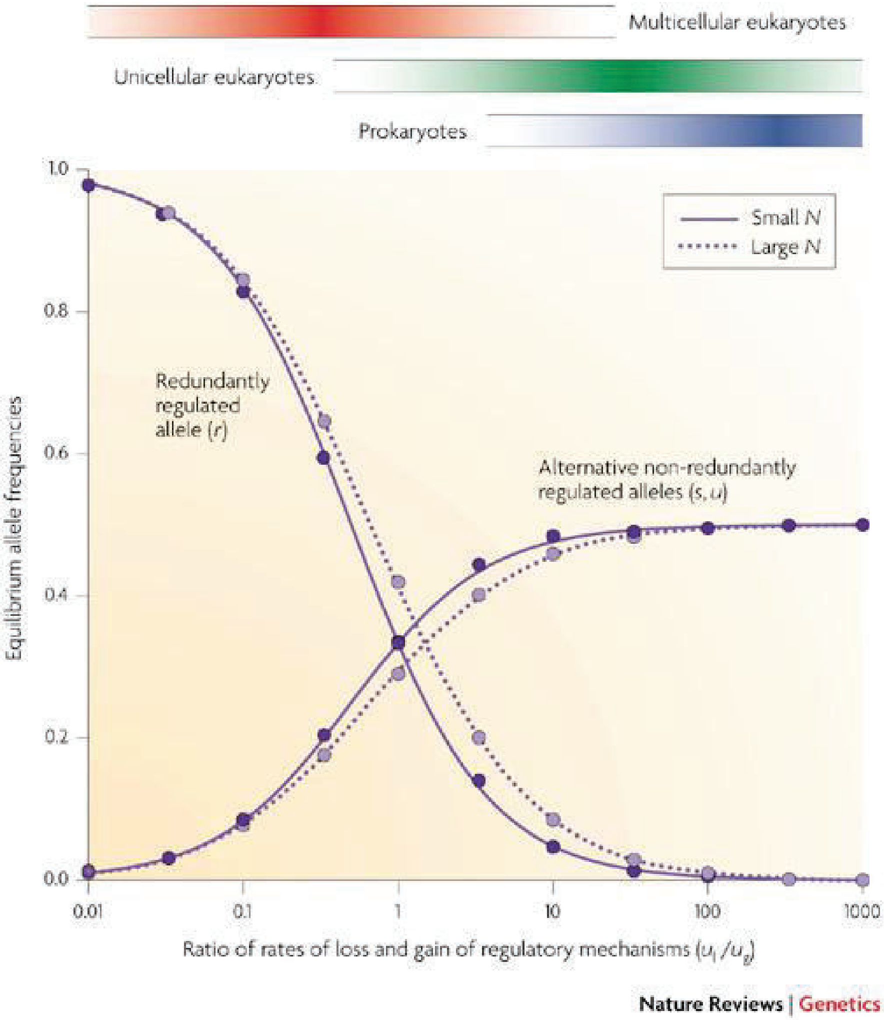


Figura 9. A ocorrência de modos alternativos de regulação gênica depende fortemente da razão μ_l/μ_g. A aproximação dos valores de μ_l/μ_g para grupos filogenéticos reais aparece no topo da imagem: vermelho para eucariotos multicelulares, verde para eucariotos unicelulares e azul para procariotos (extraído de Lynch, 2007 b).

É importante notar que variando o *Ne*, (número populacional efetivo, aparece como *N* na figura 8; Wright, 1931 e 1938) que pode ser definido como o número de indivíduos intercruzantes numa população idealizada que apresenta o mesmo efeito de perda ou fixação de alelos por deriva genética que a população original (de *N* indivíduos), o comportamento das frequências de equilíbrio pouco muda. Logo, ao que tudo indica, o aumento da complexidade regulatória (regulação redundante, ou acúmulo de domínios reguladores em promotores) dependeria apenas da razão entre ganhar ou perder um domínio regulador. E esta razão é dependente apenas da quantidade de DNA potencialmente regulador (não-codificante e/ou não-gênico) num genoma, considerando uma taxa de mutação constante (o que pode não ser o caso). Em bactérias, por exemplo, há pouco DNA não-codificante, logo este modelo teórico impõe, como propriedade emergente do sistema, a eliminação de alelos regulados por múltiplos fatores de transcrição. Num contexto com abundância de DNA não-codificante, como em eucariotos multicelulares, esta teoria impõe um acúmulo, não adaptativo, de alelos regulados por múltiplos fatores de transcrição. Os eucariotos unicelulares ocupariam uma posição intermediária neste cenário (Topo da Figura 9).

Num cenário onde a redundância regulatória sobre um gene é a regra, duplicações gênicas poderiam facilmente levar à neofuncionalização pela combinação do ganho de domínios regulatórios exclusivos entre as cópias e à subfuncionalização como através da perda de domínios ancestrais diferentes entre as cópias. Isto acontece devido a alelos redundantemente regulados terem uma pequena vantagem seletiva devido à mutação (que é numericamente igual a μ_l). Se uma mutação elimina um sítio de regulação num alelo redundantemente regulado, ainda seria possível haver expressão, comparado a alelos regulados por um único sítio de regulação que se tornariam não-funcionais com a mesma mutação e seriam contra-selecionados. Se $1/Ne > \mu_l$ ($1/Ne$ é a chance da eliminação de um alelo por deriva genética) a vantagem conferida pela regulação redundante seria tão pequena a ponto de não ser influenciada por seleção positiva, e a população vai evoluir para um estado onde o tipo de regulação depende só de α (redes redundantes para $\alpha << 1$ e redes simples para $\alpha >> 1$). Em contraste, se $1/Ne < \mu_l$ (*Ne* é grande, por exemplo em bactérias), a acumulação de elementos reguladores *upstream* vai ser inibida de forma seletiva pelo prejuízo mutacional conferido pelo acréscimo de nucleotídeos reguladores (aumento de μ_l num cenário de seleção muito eficiente). Desta forma, enquanto um *Ne* pequeno deve promover a elongação passiva de redes genéticas (pelo acréscimo de genes alvo a redes já existentes), um *Ne* grande tem o efeito oposto, gerando redes menores, com menos conexões e menos redundantes. Isso não quer dizer que o incremento de redes não pode ocorrer em populações muito grandes (bactérias, grande *Ne*), porém para tanto tais modificações em redes devem conferir uma grande vantagem seletiva imediata (Lynch 2007 a).

Em resumo, não há evidências em nenhum nível de organização biológica de que a seleção natural é uma força que direciona sistemas biológicos à complexidade, visto que a vida na Terra é dominada por formas unicelulares simples tanto em número de espécies, quanto em número de indivíduos em populações e ainda à diversidade genética interna a cada população (causa primária de um grande *Ne*). Em contraste, há substanciais evidências que a redução da eficiência da seleção natural é um fator determinante na evolução da complexidade genômica (Lynch, 2007c; Lynch, 2007b). A eficiência da seleção natural pode ser reduzida com a diminuição do número populacional efetivo (*Ne*). Se $1/Ne >> s$ (*Ne* muito baixo), onde *s* é a vantagem seletiva de um alelo em comparação a outro (A1 vs A2, por exemplo), então o equilíbrio de frequências entre dois alelos será atingido basicamente pela pressão de mutação de um no outro (a razão $\mu_{\mathbf{A1\text{->}A2}}/\mu_{\mathbf{A2\text{->}A1}}$), a seleção teria pouca ou nenhuma força sobre alelos que conferem pequenas vantagens ou desvantagens seletivas. Este é o caso de seres multicelulares (humanos e *Arabidopsis* tem *Ne* ~10^4, em contraste com *E. coli* que varia de 10^6-10^7) onde a seleção seria mais fraca em aumentar a frequência de mutações com vantagens seletivas pequenas comparadas à chance de serem eliminadas por deriva (1/*Ne*). Desta forma, o incremento da complexidade de redes de regulação, que parecem ligadas à complexidade de eucariotos multicelulares não pode ser explicada sem recorrer a forças não-adaptativas atuando ao nível do DNA (mutação e recombinação) e ao nível populacional (deriva genética). Um conjunto crescente de dados suporta a ideia de que muitos aspectos da complexidade genômica em espécies multicelulares (abundância de elementos de transposição, abundância de introns, abundância de UTRs, regulação gênica modular, redes com alta conectividade, alta taxa de manutenção de genes duplicados) tem origem nestes processos não-adaptativos que representam pouco mais do que resultados passivos de um ambiente genético-populacional único (cromossomos lineares, evolução do sexo, alta taxa de recombinação, *Ne* baixos, alta taxa de duplicação gênica, expansão de DNA não-codificante) apresentado por tais linhagens, representadas especialmente por plantas terrestres e vertebrados.

2. Objetivos gerais

Os objetivos desta tese se dividem em dois aspectos centrais que se relacionam no campo da evolução de genomas. O primeiro aspecto é o estudo de evolução de famílias multigênicas e o segundo o estudo da evolução de redes de regulação transcricional. Todos os genomas analisados bem como os experimentos realizados (na parte de redes) foram centrados no reino Viridiplantae, as plantas verdes.

2.1 Objetivos específicos

Capítulo I: Evolução de famílias multigênicas em plantas

Dois trabalhos abordaram a evolução de sistemas complexos, onde diferentes tipos de genes participam de um processo biológico: a maquinaria de síntese e degradação de xiloglucano (polissacarídeo de parede celular de plantas) de angiospermas, que integra nove funções enzimáticas diferentes, e o sistema de controle de qualidade de proteínas sintetizadas no retículo endoplasmático pelo ciclo das chaperonas calnexina e calreticulina. Estes trabalhos, e outros apresentados como anexo com abordagens similares, motivaram o desenvolvimento da ferramenta Phylexpress destinada a possibilitar análises filogenéticas em larga escala e permitir o estudo de padrões de resposta transcricional numa perspectiva comparativa. Como prova de conceito, utilizamos nosso método para uma análise de ortologia em larga escala entre os ESTs públicos de cana-de-açúcar e o proteoma de sorgo. Desta forma fomos capazes de revisitar as estimativas de conteúdo gênico nos ESTs de cana-de-açúcar inicialmente feitas com métodos mais simples.

Capítulo II: Evolução de redes de regulação em plantas

O objetivo deste segundo capítulo resume-se em entender numa perspectiva filogenética a evolução nas redes transcricionais atuantes na resposta de curto prazo (2hrs) a sinais exógenos comuns a todas as plantas terrestres. Utilizamos para tal os açúcares glicose e sacarose e o hormônio vegetal ABA. Medimos as respostas transcricionais globais a tais sinais em plântulas de arroz e sorgo e utilizamos dados públicos de plântulas de *Arabidopsis thaliana* para realizar uma análise comparativa das populações de genes responsivos transcricionalmente aos mesmos sinais entre estas três espécies.

Capítulo I

Evolução de famílias multigênicas em plantas

Parte 1: Evolução do mecanismo de síntese e degradação de xiloglucano

Del Bem and Vincentz *BMC Evolutionary Biology* 2010, **10**:341
http://www.biomedcentral.com/1471-2148/10/341

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Evolution of xyloglucan-related genes in green plants

Luiz Eduardo V Del Bem[1*], Michel GA Vincentz[1,2*]

Abstract

Background: The cell shape and morphology of plant tissues are intimately related to structural modifications in the primary cell wall that are associated with key processes in the regulation of cell growth and differentiation. The primary cell wall is composed mainly of cellulose immersed in a matrix of hemicellulose, pectin, lignin and some structural proteins. Xyloglucan is a hemicellulose polysaccharide present in the cell walls of all land plants (Embryophyta) and is the main hemicellulose in non-graminaceous angiosperms.

Results: In this work, we used a comparative genomic approach to obtain new insights into the evolution of the xyloglucan-related enzymatic machinery in green plants. Detailed phylogenetic analyses were done for enzymes involved in xyloglucan synthesis (xyloglucan transglycosylase/hydrolase, α-xylosidase, β-galactosidase, β-glucosidase and α-fucosidase) and mobilization/degradation (β-(1→4)-glucan synthase, α-fucosyltransferases, β-galactosyltransferases and α-xylosyl transferase) based on 12 fully sequenced genomes and expressed sequence tags from 29 species of green plants. Evidence from Chlorophyta and Streptophyta green algae indicated that part of the Embryophyta xyloglucan-related machinery evolved in an aquatic environment, before land colonization. Streptophyte algae have at least three enzymes of the xyloglucan machinery: xyloglucan transglycosylase/hydrolase, β-(1→4)-glucan synthase from the celullose synthase-like C family and α-xylosidase that is also present in chlorophytes. Interestingly, gymnosperm sequences orthologs to xyloglucan transglycosylase/hydrolases with exclusively hydrolytic activity were also detected, suggesting that such activity must have emerged within the last common ancestor of spermatophytes. There was a positive correlation between the numbers of founder genes within each gene family and the complexity of the plant cell wall.

Conclusions: Our data support the idea that a primordial xyloglucan-like polymer emerged in streptophyte algae as a pre-adaptation that allowed plants to subsequently colonize terrestrial habitats. Our results also provide additional evidence that charophycean algae and land plants are sister groups.

Background

The cell shape and morphology of all plant tissues are a consequence of cell division and expansion throughout the plant's life cycle. Structural modifications in the primary cell wall (PCW) are key processes in the regulation of cell growth and differentiation. The PCW is a complex dynamic structure that shows spatial and temporal variability in composition and organization. Cell shape, size, and cell-cell adhesion are processes that rely on the coordinated action of enzymes involved in the synthesis, deposition, reorganization and selective disassembly of cell wall components. The ability to selectively modify the wall architecture is a major part of many processes such as cell growth, organ abscission, vascular differentiation, fruit softening and the response to pathogens [1-3].

The PCW consists primarily of cellulose immersed in a matrix of hemicellulose, pectin, lignin and some structural proteins [4,5]. Xyloglucan (XyG) is a well-characterized hemicellulose polysaccharide present in the cell walls of all spermatophytes [6]. Xyloglucan has also recently been found in the cell walls of non-vascular and seedless vascular plants [7]. XyG can also be stored as a reserve in cotyledons of many eudicots, such as nasturtium [8,9], *Tamarindus indica* [10], *Copaifera langsdorffii* [11] and *Hymenaea courbaril* [12].

* Correspondence: lev.del.bem@gmail.com; mgavince@unicamp.br
[1]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), CP 6010, CEP 13083-875, Campinas, SP, Brazil
Full list of author information is available at the end of the article

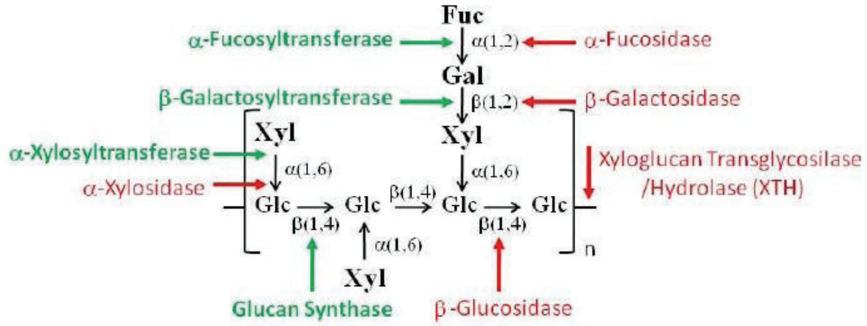


Figure 1 Schematic representation of Xyloglucan (XyG) structure and related enzymatic activities. An oligosaccharide of XyG (XXFG) is represented. β-(1-4)-glucan synthase produce the glucan backbone, α-Xylosyl Tranferase (XXT) acts on transfer xylose residues to the main backbone, β-Galactosyl Tranferase transfers galactose residue to xylose and α-Fucosyl Transferase transfer fucose residue to galactose. Xyloglucan Transglycosylase/Hydrolase (XTH) acts on hydrolysis of XyG oligosaccharides and/or XyG transglycosylation. α-Xylosidase removes the xylose residues, β-Galactosidase removes the galactose, α-fucosidase removes the fucose and β-Glucosidase mobilizes glucose monosaccharide from the main glucan backbone.

Xyloglucans have a main β-D-(1→4)-glucan backbone (denoted as G) generally branched with α(1→6)-linked D-xylopyranosyl (denoted as X) or β-D-galactopyranosyl (1→2)-D-xylopyranosyl residues (denoted as L). The presence of terminal fucosyl α-L-(1→2) units linked to branching β-D-galactosyl residues (denoted as F; for an example see Figure 1 and for nomenclature of XyG oligosaccharides see [13]) is the main difference between seed reserve XyG and structural XyG from the PCW of eudicot tissues [14].

Two main substitution patterns (XXXG and XXGG) occur in storage and structural eudicot XyG, although oligosaccharides containing five or six repeats (XXXXG and XXXXXG) have also been found in the XyG of seeds from the tropical tree *Hymenaea courbaril* ([15,16]). The archetypical seed XyG of *Tamarindus indica* consists of XXXG, XXLG, XLXG and XLLG in a molar ratio of 1.4:3:1:5.4, respectively. However, these polysaccharides are not identical among plant groups. For example, the moss *Physcomitrella patens* and the liverwort *Marchantia polymorpha* synthesize XXGGG- and XXGG-type XyGs, respectively, with side chains that contain a β-D-galactosyluronic acid and a branched xylosyl residue. In contrast, hornworts synthesize XXXG-type XyGs that are structurally homologous to the XyGs synthesized by many seed-bearing and seedless vascular plants [7].

XyG is degraded *in vivo* by five hydrolases: β-galactosidase, α-xylosidase, β-glucosidase, xyloglucan transglycosylase/hydrolase (XTH) and α-fucosidase (Figure 1; [17,18]). Although the machinery involved in XyG degradation is relatively well-characterized [17], important details of the biosynthesis of this hemicellulose remain poorly understood. A number of enzymes participate in XyG biosynthesis, including β-(1→4)-glucan synthase, α-fucosyltransferases, β-galactosyltransferases and α-xylosyltransferases (Figure 1; [19]). Recently, only two genes for α-xylosyltransferases (*XXT1* and *XXT2*) were found to be essential for the biosynthesis of XyG in *Arabidopsis* [20]. The double-mutant *xxt1*/*xxt2* lacks detectable XyG and has aberrant root hairs, but is viable and has almost normal development. This finding challenges conventional models for the functional organization of PCW components [20].

An evolutionary analysis of the XyG-related machinery could provide new insights into the origin of this polymer during plant evolution, as well as information on the context in which it occurred. This knowledge could help to explain the role of XyG in plant adaptive features. In this work, we describe a comprehensive evolutionary analysis of the multigenic families of glycosyl hydrolases (β-galactosidase, α-xylosidase, β-glucosidase, XTH and α-fucosidase) and transferases (β-(1→4)-glucan synthase, α-fucosyltransferases, β-galactosyltransferases and XXT) involved in the biosynthesis, modification and degradation of XyG (Figure 1). Our results indicate that the XyG machinery is present in all embryophytic genomes and possibly emerged from the last common ancestor of the streptophytes (Charophyta algae + embryophytes). This inference suggests that the essential enzymes involved in XyG biosynthesis and turnover originated before land colonization by plants. This conclusion indicates that XyG is more than just a structural and mechanical molecule. Our data also provide additional evidence that streptophyte algae and land plants (Embryophyta) are sister groups.

Results and Discussion

Identification and phylogenetic analysis of XyG-related genes in green plant genomes

In order to identify genes related to XyG synthesis (β-(1→4)-glucan synthase, α-fucosyltransferases, β-galactosyltransferases and α-xylosyltransferases) and mobilization/modification (β-galactosidase, α-xylosidase, β-glucosidase, XTH, α-fucosidase; Figure 1) in green plants, we used previously characterized protein sequences as queries to perform blast searches using a self-employed algorithm (Additional File 1; see Methods for a complete list of the protein sequences used as queries). We also generated a sequence database containing the complete predicted proteomes and transcriptomes for 12 species (Viridiplantae 1.0 containing 365,187 protein sequences and Viridiplantae_nt 1.0 containing 403,380 EST sequences, respectively), including angiosperms (eudicots and monocots), seedless tracheophytes (Lycophyta), non-vascular plants (Bryophyta), and green algae (Chlorophyta). In addition, searches were also run against an EST database (ViridiESTs 1.0 containing 402,770 assembled EST sequences) that included sequences belonging to 29 species from taxonomic groups lacking complete genome information, such as basal and non-eudicot/monocot angiosperms, gymnosperms (Pinophyta, Cycadophyta, Ginkgophyta and Gnetophyta), seedless Tracheophyta (Pterydophyta), non-vascular plants (Marchantiophyta) and Streptophyta algae.

Using this strategy, we identified 862 XyG-related genes that included 293 XTH sequences, 133 β-galactosidases, 53 β-glucosidases, 24 α-xylosidases, 91 β-(1→4)-glucan synthases, 79 α-fucosyltransferases, 108 β-galactosyltransferases and 45 XXTs. We found two evolutionarily unrelated clusters of XyG-related α-fucosidase, one containing 22 sequences homologous to *Arabidopsis ATFXG1* (At1g67830 - TAIR; [18]) and the other containing 14 sequences homologous to *Lilium longiflorum EBM II* (BAF85832 - GenBank; [21]). ESTs with less than 40% of the protein-based query coverage were excluded.

The relationships between genes can be represented as a system of homologous families that include orthologs and paralogs [22]. Orthologs are genes in different species that evolved from a common ancestral gene through speciation whereas paralogs are genes sharing a common ancestral gene that duplicated within the genome [23]. Orthologs normally retain their original function during evolution whereas paralogs can evolve new functions that may or may not be related to the original one. Consequently, the identification of orthologs is critical for the reliable prediction of gene functions in newly sequenced genomes. This identification is equally important for phylogenetic analysis because interpretable phylogenetic trees can generally be constructed only within sets of orthologs [23,24]. A complete list of orthologs is also a prerequisite for meaningful comparisons of genome organization [22].

The **po**ssible **g**roups of **o**rthologs (PoGOs) were established by using two phylogenetic analyses that involved the amino acid sequences. The first analysis was based on the p-distance (number of differences/number of aligned residues) and PAM 001 matrix [25] and used the neighbor-joining tree building method (NJ; [26]) while the second analysis was based on the maximum likelihood (ML; [27]). We also sought for shared derived ancestral intron positions in *Arabidopsis* (eudicot), sorghum (monocot), *Selaginella* (Lycophyte) and *Physcomitrella* (Bryophyte) since this information was useful in inferring the evolutionary relationships between homologous groups. The results for intron positions generally agreed with the phylogenetic analysis (data not shown). This combination of analyses yielded a comprehensive evolutionary profile of the enzymes involved in XyG synthesis and turnover. In the following sections, we present evidence that the complete set of enzymes involved in XyG synthesis and mobilization is present among all embryophytic lineages and that some of them emerged within streptophytes (XTH and β-(1→4)-glucan synthase) and chlorophyte algae (α-xylosidase).

XTH originated in the last common ancestor of streptophytes, before land colonization, and was amplified through several lineage-specific events in embryophytes

Two hundred and ninety-three XTH homologous sequences were identified among green plants. Phylogenetic analysis of these sequences resulted in 21 PoGOs and a group of paralogs among streptophytes (A to V in Figure 2; Additional File 2; Additional File 3). While no XTH sequence was found among the four complete genomes of green algae (chlorophytes), at least two land plant XTH founder genes were identified in the last common ancestor of streptophytes (PoGOs U and V in Figure 2). One of these genes was represented by *CvXTH1* (previously identified as *Chara2* [28]) from *Chara vulgaris* (class Charophyceae). This ancestral gene persisted in all streptophyte lineages and was represented by the single PoGO V (Figure 2). The other founder gene is represented by PoGO U (Figure 2), that integrates the *CpXTH1* gene in the charophyte alga *Closterium peracerosum* (class Zygnemophyceae) and consisted exclusively of XTH sequences from marchantiophyte and streptophyte algae. PoGO U most likely gave rise to 19 PoGOs and a group of paralogs (group K in Figure 2). These PoGOs and paralogs included all previously reported *Arabidopsis*, rice, and poplar XTH genes (Figure 2; [6,29,30]). Since molecular and morphological data suggest that streptophyte algae are sister groups of land plants [31,32], and since liverworts such as *Marchantia* are the most basal embryophytes [33], it is plausible that PoGO U-related orthologs have been lost in bryophytes and tracheophytes (Figure 2). PoGO U and the homologous group that emerged from it were

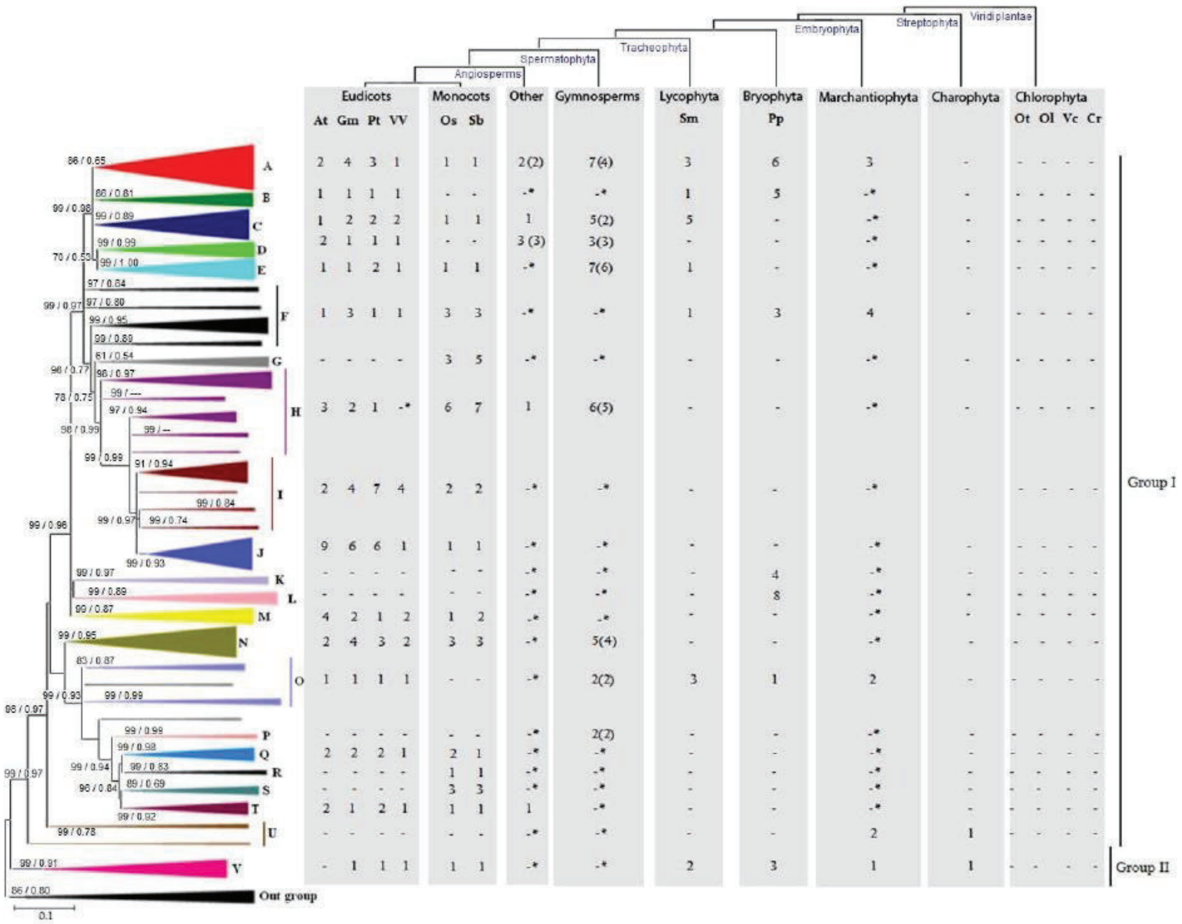


Figure 2 Evolutionary profile of XTH in green plants. Phylogenetic tree showing XTH PoGOs in green plants. The topology was obtained by Neighbor-Joining method with genetic distances calculated by p-distance. The bootstrap values higher than 50% are shown for 1000 replicates along with Maximum-likelihood (ML) aLRT values higher than 0.50 (in the figure [**NJ/ML**]). Each triangle represents a PoGO and the amount of sequences belonging to each PoGO is proportional to triangles' size. The number of genes by species is given (At - *Arabidopsis thaliana*, Gm - *Glycine max*, Pt - *Populus trichocarpa*, Vv - *Vitis vinifera*, Os - *Oryza sativa*, Sb - *Sorghum bicolor*, Sm - *Selaginella moellendorffii*, Pp - *Physcomitrella patens patens*, Ot - *Ostreococcus tauri*, Ol - *Ostreococcus lucimarinus*, Vc - *Volvox carteri*, Cr - *Chalmydomonas reinhardtii*). Outgroup is formed by fungus sequences. Accession numbers for all genes in the tree are described in Additional File 3 and the detailed tree is shown in Additional File 2. * - No genes found within a PoGO that lacks genomic information, only ESTs. ( ) - Given number of different species within a PoGO.

defined as Group I while the homologous genes from PoGO V were identified as Group II (Figure 2).

Together, these results indicate that PoGOs U and V share a common origin, which implies that the first XTH gene duplication and maintenance occurred before land colonization by plants. This conclusion is supported by the detection of XyG in all groups of land plants [7] and the presence of XyG transglycosylation activity in the charophyte alga *Chara vulgaris* [28]. Thus, XTH apparently originated after the divergence of chlorophyte and streptophyte algae. XyG was therefore probably absent in the more ancestral Viridiplantae lineages represented by chlorophyte algae and emerged as a new cell wall component in streptophytes. Since chlorophyte algae occur mainly in salt-water wheres streptophyte algae are mainly fresh-water, we suggest that XyG provided a selective advantage in the colonization of fresh-water habitats. In addition, the ability of XyG to confer mechanical strength [20] may have been particularly advantageous in allowing streptophytes to colonize terrestrial habitats. Successful land colonization by plants has apparently been limited to a sister lineage of streptophyte algae that gave rise to all embryophyte groups [32]. These conclusions provide additional support for the suggestion that the acquisition of XyG by streptophyte algae was an important factor in land colonization [34].

The PoGOs in Group I consisted of genes from all major embryophyte lineages. PoGOs A, F, and O that contained XTH genes from the marchantiophyte *Marchantia polymorpha* (liverwort) and PoGO B that contained bryophyte (moss) XTH genes appeared to have emerged from four genes in the last common ancestor of embryophytes (Figure 2). Thus, up to four ancestral XTH genes were related to early non-vascular land plants that were present at least 475 million years ago (based on the current fossil record) [35]. This conclusion supports the notion that gene duplication in the XTH family and its resulting selective advantages is an ancient phenomenon among land plant lineages.

A striking feature of Group I was the extensive amplification of XTH genes among angiosperms, i.e., 33 genes in *Arabidopsis*, 34 in soybean, 35 in poplar, 19 in grape, 30 in rice, and 32 in sorghum, all of which were distributed among 17 PoGOs (Figure 2). However, differential patterns of amplification and/or gene losses were observed among land plants. For instance, PoGOs B, D, and O were present among eudicots but were not detected in monocots (Figure 2). The presence of bryophyte and lycophyte XTH genes in PoGO B, gymnosperm XTHs in PoGO D, and embryophyte XTH genes in PoGO O indicates that gene losses from these PoGOs occurred specifically in the monocot lineage (Figure 2). On the other hand, PoGOs G, R, and S were restricted to monocots (Figure 2), which suggests that these groups emerged after the divergence of eudicots and monocots. Although the abundance of XTH genes in the rice genome [30] was initially considered unusual because of the small content of XyG in the PCW of most grasses [4,14], XyG can account for up to 10% of the wall mass in grass tissues during growth [36] and XTH activity may be more important for grasses than previously thought [30,37].

PoGO L and the set of paralogs genes that formed Group K were restricted to bryophytes whereas PoGO P was restricted to gymnosperms. The simplest explanation for this is that these lineage-specific acquisitions may be related to functional specialization and/or novelties. These lineage-specific differences suggest that distinct patterns of selective pressure acted on the XTH genes in different lineages, and may partly explain the differential abundance of XyG, i.e., 10-20% of the PCW dry weight in eudicots compared to <5% in graminaceous monocots [4,14,38,39]) and the different patterns of XyG substitution and structure, e.g., presence of galacturonic acid in bryophyte and marchantiophyte XyG [7].

PoGO N contained *Arabidopsis* proteins encoded by *At-XTH31* and *At-XTH32*, which are involved exclusively in XyG hydrolysis and lack transglycosylation activity [6]. XTH with exclusively hydrolytic activity may have derived from transglycosylating proteins as a new feature of angiosperms [6]. However, this conclusion may need to be reevaluated in the light of the data presented here. Indeed, sequences from several taxonomic groups of flowerless seed plants (pinophytes, gnetophytes and cycadophytes; Figure 2) were included in PoGO L, suggesting that XTHs with exclusively hydrolytic activity emerged at least in the last common ancestor of the spermatophytes. This inference is further supported by the occurrence of hydrolytic activity in fast growing tissues such as meristems (*Arabidopsis*; [40]) or during specific developmental stages such as germinating seeds (tomato; [41]) or in physiological processes such as the mobilization of endosperm reserves (*Hymenaea courbaril*; [17,42]), all of which are key phenomena in seed plants and originated at least 300 million years ago, as suggested by the cycadophytes fossil record [43].

The moss *Physcomitrella* genome contained 30 XTH genes, a number comparable to that found in angiosperms (33 in *Arabidopsis* and 30 in rice). This elevated number of genes may reflect lineage-specific genome duplications in mosses [44]. Of these 30 genes, 27 were classified in Group I and three in Group II (Figure 2), and could be divided into six PoGOs and a paralog group (Group K, Figure 2). Two of these groups, PoGO L and the paralog group K, were bryophyte-specific while the other five PoGOs (A, B, F, O and V) were shared by tracheophytes (Figure 2; Additional File 3). The emergence of these lineage-specific XTH genes in bryophytes could be related to the presence of a specific type of XyG containing β-D-galactosyluronic acid and a branched xylosyl residue [7] that is not shared with tracheophytes [7,37].

The vascular seedless *Selaginella* had only 16 XTH genes (14 in Group I and two in Group II) that were divided into seven PoGOs conserved among other tracheophyte lineages (Figure 2). All *Selaginella* XTH genes occurred in PoGOs shared by angiosperms. The retention of these genes by angiosperms suggests that the early set of tracheophyte XTH genes was conserved in higher taxa, whereas the basic XyG pattern of XXXG emerged in hornworts and is shared by all tracheophytes [7,37]. The appearance of PoGOs C and E in *Selaginella* suggested that the last common ancestor of tracheophytes carried at least two additional XTH genes when compared to the last common ancestor of embryophytes (Figure 2). If each PoGO shared between different lineages is considered to be representative of founder genes then during their evolution the number of green plant XTHs gradually expanded from two ancestral genes in streptophyte algae to five in early embryophytes (*Physcomitrella* - Bryophyta), seven in early tracheophytes (*Selaginella* - Lycophyta), and 18 in angiosperms (eudicots and monocots). This increasing number of PoGOs suggests an important role for XyG in the evolution from non-vascular land plants to angiosperms.

The genes in PoGO V have not previously been reported to be XTH, perhaps because the model plant *Arabidopsis* lacks genes in this PoGO. Indeed, PoGO V included genes from all other complete embryophyte genomes and also *M. polymorpha* and the charophyte alga *C. vulgaris* (Figure 2; Additional File 3). A *C. vulgaris* cDNA sequence encoding a protein encompassing the main XTH catalytic site (DEIDFEFLG) has been isolated and may correspond to the XyG transglycosylation activity identified in growing tissues of this alga [28]. As in angiosperms, the *C. vulgaris* transglycosylase activity may be involved in adjustment of the PCW during growth [28].

We also searched for genes similar to XTH in animal and fungus genomes. Although no proteins similar to XTH were identified in the animal genomes, we found three glycosyl hydrolases in the complete *Saccharomyces cerevisiae* genome (*Utr2* - NP_010874, *Crr1p* - NP_013314 and *Crh1p* - NP_011705; GenBank) and five in *Aspergillus nidulans* (XP_662119, XP_664552, XP_660657, XP_658537 and XP_661518; GenBank) that were similar to plant XTHs (Additional File 2). When these fungus sequences were analyzed together with those for XTHs from all groups of plants they formed an outgroup (Figure 2). The high bootstrap support (86%) suggested a possible single origin for these fungi hydrolases. However, it is unclear whether these fungus genes share a common ancestor with streptophytes XTHs or whether the similarity merely reflects functional convergence from an ancestral eukaryotic glycosyl hydrolase.

β-Galactosidase genes are present in eukaryotes and were notably amplified during the evolution of land plants

One hundred and thirty-three non-redundant β-galactosidase genes were identified in the embryophyte lineages analyzed. These β-galactosidase genes were organized into 10 PoGOs that were divided into two homologous groups (Figure 3A; Additional File 4). Group I contained only PoGO J composed of animal, plant and fungus genes (Figure 3A; Additional File 5). On the other hand, Group II contained nine PoGOs present exclusively in plants (Figure 3A; Additional File 5). The presence in PoGO J of genes from all major eukaryotic lineages suggested that the plant-specific β-galactosidases in Group II must have derived from PoGO J after the divergence of plants from the fungus/animal lineage. PoGO J can therefore be considered to be representative of the ancestral β-galactosidase gene. Remarkably, all embryophytes had a single gene, except for *Physcomitrella*, which had two, possibly because of moss-specific genome duplication (Figure 3A; Additional File 5). In contrast, the β-galactosidases genes of Group II showed significant duplication events during plant evolution. β-Galactosidases from the non-vascular plants *Physcomitrella* (a bryophyte) and *Marchantia* (a marchantiophyte) and the vascular seedless *Selaginella* were restricted to PoGO F, indicating that this PoGO most probably emerged from PoGO J (Group I) genes in the last common ancestor of embryophytes (Figure 3A).

PoGOs C and E formed part of the β-galactosidase genes from gymnosperms and angiosperms, indicating that these PoGOs probably emerged from PoGO F in the last common ancestor of spermatophytes. PoGOs A, B, G, H, and I apparently emerged exclusively in angiosperms (Figure 3A; Additional File 5). However, the lack of a complete genome for gymnosperms means that the presence of genes belonging to this series of PoGOs in flowerless seed plants cannot be discarded.

PoGO D consisted exclusively of gymnosperm EST sequences from pinophytes, gnetophytes, and ginkgophytes. Based on the tree topology, the PoGO D genes probably emerged from PoGO C after the divergence of angiosperms and gymnosperms (Figure 3A). An alternative hypothesis is that these genes were selectively lost in the angiosperm lineage.

No β-galactosidase-coding genes were detected in Chlorophyta algae. This situation may reflect the loss of these genes from the *Chlamydomonas*, *Volvox*, *Ostreococcus tauri* and *O. lucimarinus* genomes, possibly because of the lack of selective pressure to maintain these enzymes.

Land plant β-glucosidases active on XyG form two PoGOs: an ancestral one shared by all embryophytes and a derived one restricted to angiosperms

By using the protein sequence of the well characterized *Tropaeolum majus* (eudicot) β-glucosidase (CAA07070 - GenBank; [45]) as a query in our pipeline (Additional File 1) we identified 53 non-redundant plant sequences that were classified into two PoGOs (A and B; Figure 3B; Additional File 6; Additional File 7). PoGO A included the *Tropaeolum* β-glucosidase and genes from all major embryophyte lineages (Figure 3B). PoGO A also contained four *Arabidopsis* paralog genes (Additional File 7) coding for proteins characterized as XyG hydrolytic β-glucosidases present in apoplastic fluid [46]. The genes in PoGO B probably emerged from PoGO A and were detected exclusively among angiosperms (Figure 3B). *Arabidopsis* had four paralogs genes in PoGO B (Additional File 7) that are arranged *in tandem* in chromosome 3 whereas rice had a single gene (Additional File 7). Functional characterization of the PoGO B genes is still lacking.

No gene similar to XyG-active β-glucosidase was detected in the genomes of green algae, fungi or animals. Surprisingly, we found very similar sequences in

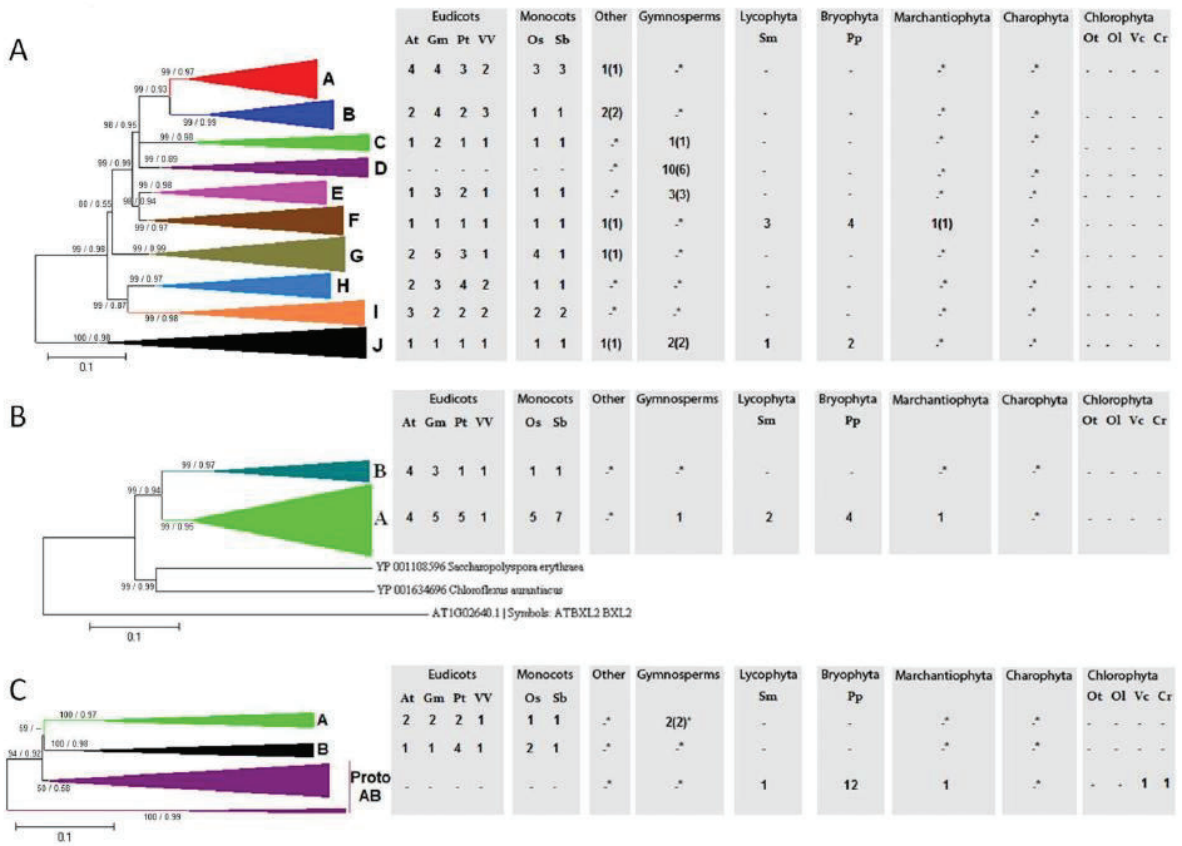

A
Eudicots At Gm Pt VV | Monocots Os Sb | Other | Gymnosperms | Lycophyta Sm | Bryophyta Pp | Marchantiophyta | Charophyta | Chlorophyta Ot Ol Vc Cr
A 4 4 3 2 | 3 3 | 1(1) | .* | - | - | .* | .* | - - - -
B 2 4 2 3 | 1 1 | 2(2) | .* | - | - | .* | .* | - - - -
C 1 2 1 1 | 1 1 | .* | 1(1) | - | - | .* | .* | - - - -
D - - - - | - - | .* | 10(6) | - | - | .* | .* | - - - -
E 1 3 2 1 | 1 1 | .* | 3(3) | - | - | .* | .* | - - - -
F 1 1 1 1 | 1 1 | 1(1) | .* | 3 | 4 | 1(1) | .* | - - - -
G 2 5 3 1 | 4 1 | 1(1) | .* | - | - | .* | .* | - - - -
H 2 3 4 2 | 1 1 | .* | .* | - | - | .* | .* | - - - -
I 3 2 2 2 | 2 2 | .* | .* | - | - | .* | .* | - - - -
J 1 1 1 1 | 1 1 | 1(1) | 2(2) | 1 | 2 | .* | .* | - - - -
99 / 0.97
99 / 0.93
99 / 0.99
98 / 0.95
99 / 0.98
99 / 0.99
99 / 0.89
99 / 0.98
80 / 0.55
98 / 0.94
99 / 0.97
99 / 0.98
99 / 0.99
99 / 0.97
99 / 0.87
99 / 0.98
100 / 0.98
0.1
B
Eudicots At Gm Pt VV | Monocots Os Sb | Other | Gymnosperms | Lycophyta Sm | Bryophyta Pp | Marchantiophyta | Charophyta | Chlorophyta Ot Ol Vc Cr
B 4 3 1 1 | 1 1 | .* | .* | - | - | .* | .* | - - - -
A 4 5 5 1 | 5 7 | .* | 1 | 2 | 4 | 1 | .* | - - - -
99 / 0.97
99 / 0.94
99 / 0.95
99 / 0.99
YP 001108596 Saccharopolyspora erythraea
YP 001634696 Chloroflexus aurantiacus
AT1G02640.1 | Symbols: ATBXL2 BXL2
0.1
C
Eudicots At Gm Pt VV | Monocots Os Sb | Other | Gymnosperms | Lycophyta Sm | Bryophyta Pp | Marchantiophyta | Charophyta | Chlorophyta Ot Ol Vc Cr
A 2 2 2 1 | 1 1 | .* | 2(2)* | - | - | .* | .* | - - - -
B 1 1 4 1 | 2 1 | .* | .* | - | - | .* | .* | - - - -
Proto AB - - - - | - - | .* | .* | 1 | 12 | 1 | .* | - - 1 1
100 / 0.97
59 / -
100 / 0.98
94 / 0.92
50 / 0.68
100 / 0.99
0.1

Figure 3 Evolutionary profiles of β-galactosidase (A), XyG-active β-glucosidase (B) and α-xylosidase (C) genes in green plants. The topologies were obtained by Neighbor-Joining method with genetic distances calculated by p-distance. The bootstrap values higher than 50% are shown for 1000 replicates along with Maximum-likelihood (ML) aLRT values higher than 0.50 (in the figure [**NJ**/**ML**]). Accession numbers for all genes in the trees are described in Additional File 5 for β-galactosidase, Additional File 7 for β-glucosidase and Additional File 9 for α-xylosidase. The detailed trees are shown in Additional File 4 for β-galactosidase, Additional File 6 for β-glucosidase and Additional File 8 for α-xylosidase. The tree containing the 100 first blastp hits of *AtXYL1* is shown in Additional File 10.

some bacterial species. However, phylogenetic analysis of the evolutionary relationship between our set of plant β-glucosidases and the two most similar bacterial sequences (YP_001634696 from *Chloroflexus aurantiacus* and YP_001108596 from *Saccharopolyspora erythraea*; GenBank; Additional File 6) was inconclusive, although it is possible that these bacterial genes may share a common origin with plant XyG-active β-glucosidases. Several explanations could account for this scenario. First, the ancestral genes may have survived only in bacteria and streptophytes, having been lost in the fungal/metazoan group and in Viridiplante from chlorophytes. Second, these β-glucosidases could have a bacterial origin and were transmitted horizontally from the ancestral cyanobacterial endosymbiont, which gave rise to chloroplasts, to earlier Viridiplantae, but were specifically lost in chlorophytes. Third, the similarity between the plant XyG-active β-glucosidase and bacterial genes may simply be a case of convergent evolution from distinct ancestral hydrolases.

Plant α-xylosidase emerged before the divergence between chlorophyte and streptophyte algae and is evolutionarily related to eukaryote α-glucosidases

At least 24 sequences significantly similar to the well-characterized α-xylosidase *AtXYL1* from *Arabidopsis* (*At1g68560*; [47]) and *Tropaeolum majus* (CAA10382 - GenBank; [48]) were identified in the Viridiplantae species (Figure 3C). These genes were grouped into a single set of homologous sequences that were further organized into three PoGOs, of which PoGO A was spermatophyte-specific and PoGO B was restricted to angiosperms (Figure 3C; Additional File 8; Additional File 9). These two PoGOs probably emerged from PoGO Proto AB (Figure 3C; Additional File 8; Additional File 9) that includes genes from more ancestral land plants such as *Physcomitrella* (seven genes), and *Selaginella* (one gene) (Additional File 9). Finally, PoGO Proto AB included genes from the green algae *Chlamydomonas* and *Volvox* (Figure 3C; Additional File 8; Additional File 9). Unexpectedly, the Prasinophyceae algae *Ostreococcus tauri* and *O. lucimarinus* had no genes in PoGO Proto AB, suggesting that α-xylosidase genes were specifically lost in these organisms after their divergence from other Viridiplantae lineages.

An interesting feature of α-xylosidases was the extensive gene duplication in the *Physcomitrella* genome, which contained at least 12 genes compared to three in *Arabidopsis* and a single gene in *Selaginella* (Figure 3C; Additional File 9). This greater number of genes suggests that α-xylosidase gene duplication and fixation in these basal embryophytes may have conferred some selective advantage possibly related to the ecological role of mosses. A plausible explanation for the evolutionary development of α-xylosidase could be that a single ancestral gene in green algae (represented by PoGO Proto AB) eventually gave rise to spermatophyte-specific PoGO A and angiosperm-specific PoGO B (Figure 3C).

To improve our understanding of the origin of plant α-xylosidases, we extended our analysis to all genes that shared any similarity with the query sequences from *Arabidopsis* and *Tropaeolum* (e-value < e^{-4}), including sequences obtained from searches against bacteria and the fungal/metazoan group (Additional File 10). In *Arabidopsis*, the genes most closely related to α-xylosidases were α-glucosidases *RSW3* (radial swelling 3; At5g63840), which shared 27% identity with *AtXYL1* (245 out of 903 amino acids; e-value = $1e^{-81}$), and *HGL1* (heteroglycan glucosidase 1; AT3G23640), which shared 33% identity with *AtXTYL1* (185 out of 559 amino acids; e-value = $6e^{-79}$). A phylogenetic analysis that integrated α-xylosidase homologues and *RSW3* and *HGL1* homologues from green plants with the most closely related corresponding sequences from fungi, animals and bacteria (Additional File 10) showed that the *RSW3* and *HGL1* genes formed a single PoGO within plants. The *HGL1* PoGO included genes from embryophytes and the most similar non-plant sequences were from bacteria. No genes from fungi or animals were included in the *HGL1* PoGO. The *RSW3* PoGO included genes from all Viridiplantae lineages, including the Prasinophyceae algae *O. tauri* and *O. lucimarinus*. *rsw3* is a temperature-sensitive mutant of *Arabidopsis* that has radially swollen roots and a deficiency in cellulose deposition. *RSW3* is thought to process N-linked glycans in the endoplasmatic reticulum, as part of the quality control pathway to ensure correct protein folding [49]. In our analysis, the *RSW3* gene shared high similarity with the catalytic α-subunit of fungal and animal glucosidase II. Together, these findings suggest that the plant-specific α-xylosidase involved in XyG mobilization evolved from an ancestral eukaryotic α-glucosidase gene, represented here by *RSW3* PoGO (Additional File 10). This finding also supports the idea that neofunctionalization could be the main process responsible for the switch in substrate specificity from α-glucosidase to α-xylosidase during the evolution of glycosyl hydrolases.

XyG β-(1→4)-glucan synthase belongs to the celulose synthase-like C gene family present in streptophyte algae

Ninety-one genes (threshold of e^{-04}) related to the *Arabidopsis* celulose synthase-like (CSL) gene *AtCSLC4* (*At3g28180*) encoding a β-(1→4)-glucan synthase [19] were identified by searching the Viridiplantae database. The threshold of e^{-04} used to define significant similarity throughout this study allowed the recovery of genes that formed the CSL C [50,51] and CSL A [51] groups, as well as a group of chlorophyte genes that behaved as an outgroup to CSL C and CSL A in our analysis (Figure 4A; Additional File 11; Additional File 12). This finding supports the suggestion that CSL C and CSL A resulted from a duplication event of an ancestral green plant gene present in chlorophytes, and agrees with a recently published report [51]. This ancestral gene is represented in our analysis by PoGO H, which contained a single copy in each of the green algae genomes analyzed (*Volvox*, *Chlamydomonas*, *O. tauri* and *O. lucimarinus*; Figure 4A; Additional File 12).

Based on an analysis of complete genomes from land plants and chlorophyte algae, Yin *et al.* [51] concluded that the CSL C and CSL A groups were the products of an ancestral gene duplication in earlier embryophytes. This conclusion may have to be re-evaluated since, as shown here, CSL C included a gene from the streptophyte alga *Chara globularis* (Charophyta), indicating that this group emerged before land colonization by plants (Figure 4A). These data raise the interesting possibility that the gene duplication event that resulted in CSL A and CSL C had occurred in early streptophytes. A recent work [52] has detected a XyG-like polymer containing glucose and xylose in the streptophyte algae *Spirogyra* (Class Zygnematophyceae). In contrast, Popper and Fry [53] reported the absence of XyG in the cell walls of charophycean algae such as *Chara*, *Nitella*, *Coleochaete* and *Klebsormidium*. Although these partially contradictory results indicate that more research is needed to understand the composition of the PCW in streptophyte algae, it seems plausible that the CLS C gene from *C. globularis* could be involved in the synthesis of a XyG-like polymer, as occurs in *Spirogyra*.

Members of CLS A (*AtCSLA9* [At5g03760 - PoGO D], *AtCSLA2* [At5g22740 - PoGO E] and *AtCSLA7* [At2g35650 - PoGO F]) have β-mannan synthase activity when expressed in S2 *Drosophila* cells supplied

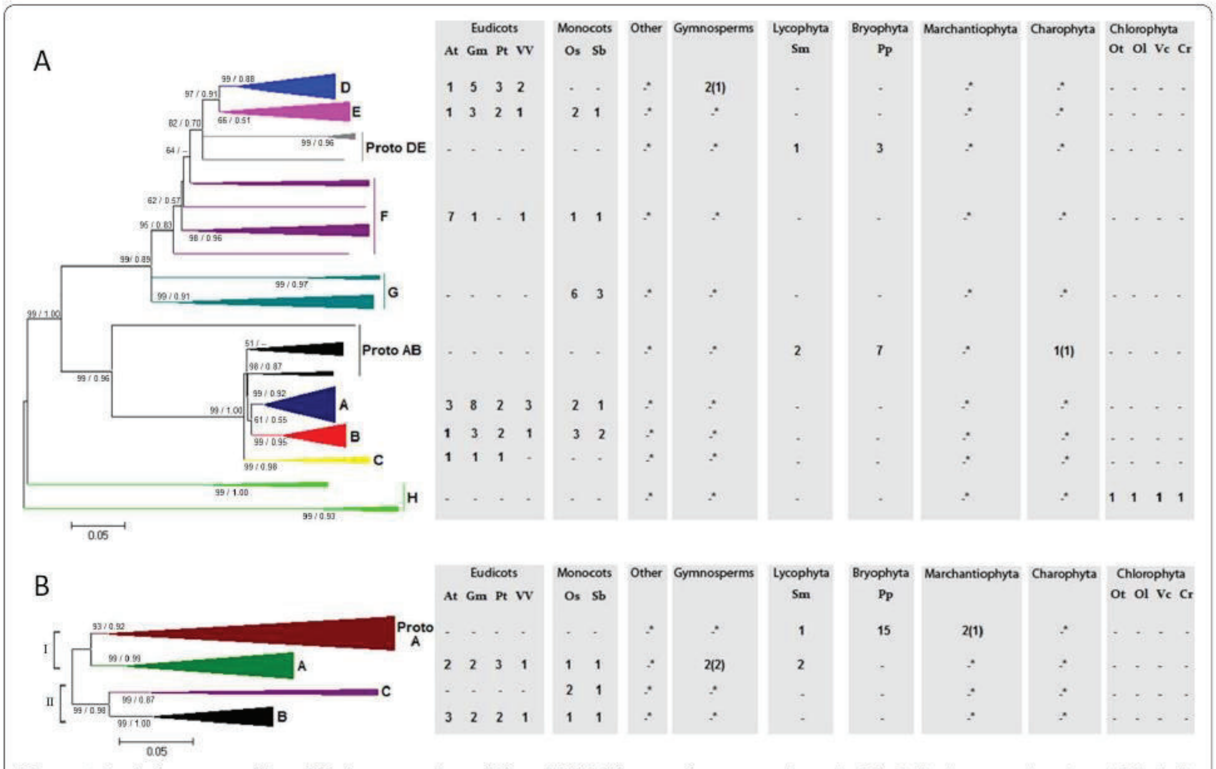


Figure 4 Evolutionary profiles of β-glucan synthase (A) and XXT (B) genes in green plants. A: CSL C (β-glucan synthase) and CSL A (β-mannan and β-glucomannan synthases) PoGOs. The topologies were obtained by Neighbor-Joining method with genetic distances calculated by p-distance. The bootstrap values higher than 50% are shown for 1000 replicates along with Maximum-likelihood (ML) aLRT values higher than 0.50 (in the figure [**NJ**/**ML**]). Accession numbers for all genes in the trees are described in Additional File 12 for β-Glucan Synthase and Additional File 14 for XXT. The detailed trees are shown in Additional File 11 for β-Glucan Synthase and Additional File 13 for XXT.

with GDP-mannose [54]. Interestingly, the proteins encoded by these three genes were also able to produce β-glucomannan when supplied with GDP-mannose and GDP-glucose, and unexpectedly AtCSLA9 (PoGO D) produced β-glucan when supplied with GDP-glucose; this latter activity is believed to be the main function of CSL C. Thus, it appears that CSL A can synthesize at least three different polymers, β-mannan, β-glucomannan and β-glucan. Conversely, to date, only a single activity (β-glucan synthesis) has been described for CSL C members. These observations raise the question of which activity is performed by the ancestral PoGO H members in chlorophytes. Only mannans, glucoronic acids, mannuronic acids and 3-O-methyl rhamnose have been detected in the cell walls of chlorophytes [55], thus supporting the view that β-mannan synthase is the basic or primordial activity of the plant CSL A and C ancestral group represented by chlorophyte PoGO H [37,51]. This conclusion further supports the hypothesis that the specific β-glucan synthase activity used in XyG synthesis emerged from mixed activity (β-mannan/β-glucomannan/β-glucan synthase) proteins in the course of streptophyte evolution.

The CSL C group included five genes from *Arabidopsis*, poplar and rice, 12 from soybean, four from grape, three from sorghum, two from *Selaginella* and seven from *Physcomitrella.* The group was further divided into four PoGOs (A, B, C, and Proto AB; Figure 4A). PoGOs A and B included angiosperm genes whereas PoGO C was restricted to eudicots. PoGO A included the *Arabidopsis AtCSLC4* gene that was shown to encode a β-(1→4)-glucan synthase involved in XyG biosynthesis [19]. PoGO Proto AB included *Selaginella*, *Physcomitrella* and *C. globularis* genes, with the angiosperm PoGOs A and B probably resulting from the duplication of an original PoGO Proto AB gene within the angiosperm lineage.

The CSL A group included genes from all embryophyte lineages. The genes in this group were divided into five additional PoGOs (D, E, Proto DE, F and G; Figure 4A). PoGO D was spermatophyte-specific but lacked genes from monocots, which suggests these genes were specifically lost in the monocot lineage.

PoGO D included *Arabidopsis AtCSLA9*, the protein product of which has important β-mannan synthase activity, as well as β-glucomannan and β-glucan synthase activities [54]. PoGO E, which is more closely related to PoGO D, was restricted to angiosperms and included *Arabidopsis AtCSLA2*, which has prevalent β-mannan synthase activity and β-glucomannan synthase activity, but almost no β-glucan synthase activity. PoGO Proto DE, which contained sequences from *Selaginella* and *Physcomitrella*, was considered as an outgroup to PoGOs D and E (Figure 4A). This finding suggested that PoGOs D, E and Proto DE had a common origin in the last common embryophyte ancestor. PoGO F was restricted to angiosperms and contained *Arabidopsis AtCSLA7*, which has mainly β-mannan synthase activity and lacks β-glucan synthase activity [54]. PoGO F had an apparent *Arabidopsis*-specific gene duplication pattern that resulted in seven paralogs whereas the grape, soybean, sorghum and rice genomes possess a single gene. The relevance of these lineage-specific gene duplication events remains to be investigated. Another striking feature of the CSL A group was the monocot-specific PoGO G.

Based on the evidence presented here, we conclude that XyG-specific β-glucan synthases in CSL C evolved from an ancestral β-mannan synthase represented by PoGO H, the ancestral group of the CSL C and A families. The presence of CSL C genes in charophytes is strong evidence that XyG emerged prior to the colonization of land by early embryophytes. This conclusion agrees with the recent detection of XyG in the cell walls of Charophycean algae [37].

α-Xylosyltransferases (XXT) are present in all land plant lineages but absent from chlorophyte algae

An analysis of 45 XXT genes resulted in the recognition of two homologous groups among embryophyte XXTs (Groups I and II in Figure 4B). Group I was the most ancient and included PoGO Proto A that contained genes from *Marchantia* (Marchantiophyta), *Physcomitrella* (Bryophyta) and *Selaginella* (Lycophyta), and PoGO A that included *Selaginella*, gymnosperm and angiosperm genes (Figure 4B; Additional File 13; Additional File 14). We suggest that PoGO A emerged from Proto A by gene duplication in the last common ancestor of tracheophytes. PoGO A included the *Arabidopsis XXT1* and *XXT2* genes [20]. The *xxt1/xxt2* double mutant lacks detectable XyG, but the only apparent phenotypes associated with these mutations were aberrant root hair development, slow growth, and a slightly smaller stature at maturity [20]. This result challenges the conventional model for the structure of the PCW in eudicot and non-graminaceous monocots, which states that XyG is the principal load-bearing structure [4,14,56-60]. In the light of this traditional PCW model a plant lacking XyG would not be viable or at least would have a very deleterious phenotype, which apparently is not the case, at least in *Arabidopsis*.

Group II contains *Arabidopsis XXT5* (At1g74380), which participates in XyG synthesis [61]. *xxt5* exhibits a phenotype similar to that described for the *xxt1/xxt2* double mutant and consists of short root hairs with bubble-like extrusions at the tip. In addition, the main root cell morphology was altered in the *xxt5* mutant and the level of XyG was reduced. Unexpectedly, although *XXT5* was expressed in the *xxt1/xxt2* double mutant no XyG was detected in these mutants, possibly indicating an epistatic effect whereby the activity of either *XXT1* or *XXT2* is required before *XXT5* can act [20]. It will be interesting to evaluate whether the Group I XXTs of other species have this type of epistatic effect on genes in Group II.

Two types of evolutionarily unrelated plant α-fucosidases are active against XyG oligosaccharides

Two genes were found to encode XyG-active α-fucosidases: the *Arabidopsis* gene *AtFXG1* (At1g67830; [18]), which belongs to the largely unknown GDSL-motif lipase/hydrolase family protein (Figure 5A; Additional File 15; Additional File 16) and the *Lilium longiflorum* gene *EBM II* (BAF85832 - GenBank; [21]; Figure 5B; Additional File 16). Interestingly, these proteins shared no similarity with each other (e-value = 8.3) and were therefore evolutionarily unrelated but converged functionally to fulfill a similar enzymatic activity.

The *Arabidopsis* genes most closely related to *AtFXG1* (Additional File 15; Additional File 16) lacked any functional information, and we will therefore focus here on the PoGO that included *AtFXG1* (PoGO A; Figure 5A). PoGO A contained another *Arabidopsis* gene, *At3g26430*, and included genes from other angiosperms, as well as gymnosperm ESTs, which lead to the conclusion that this group emerged in the last common ancestor of the Spermatophyta. To extend our understanding of the evolutionary origin of *AtFXG1* homologues in plants, we performed a broader phylogenetic analysis (Additional File 15) that encompassed the first 100 blast hits obtained with an *AtFXG1*-encoded protein sequence query run against the Viridiplantae 1.0 dataset using an e-value threshold of e^{-4} (see Methods). This analysis identified PoGOs 1 and 2 that emerged before angiosperm divergence (Additional File 15). These two PoGOs included *Selaginella* (PoGOs 1 and 2) and *Physcomitrella* (PoGO 2) with angiosperm genes, which suggests that they emerged at least in the last common ancestor of tracheophytes and land plants, respectively. No gene similar to *AtFXG1* was detected in chlorophytes, suggesting that hydrolases of this type were restricted to embryophyte lineages. Based on the tree

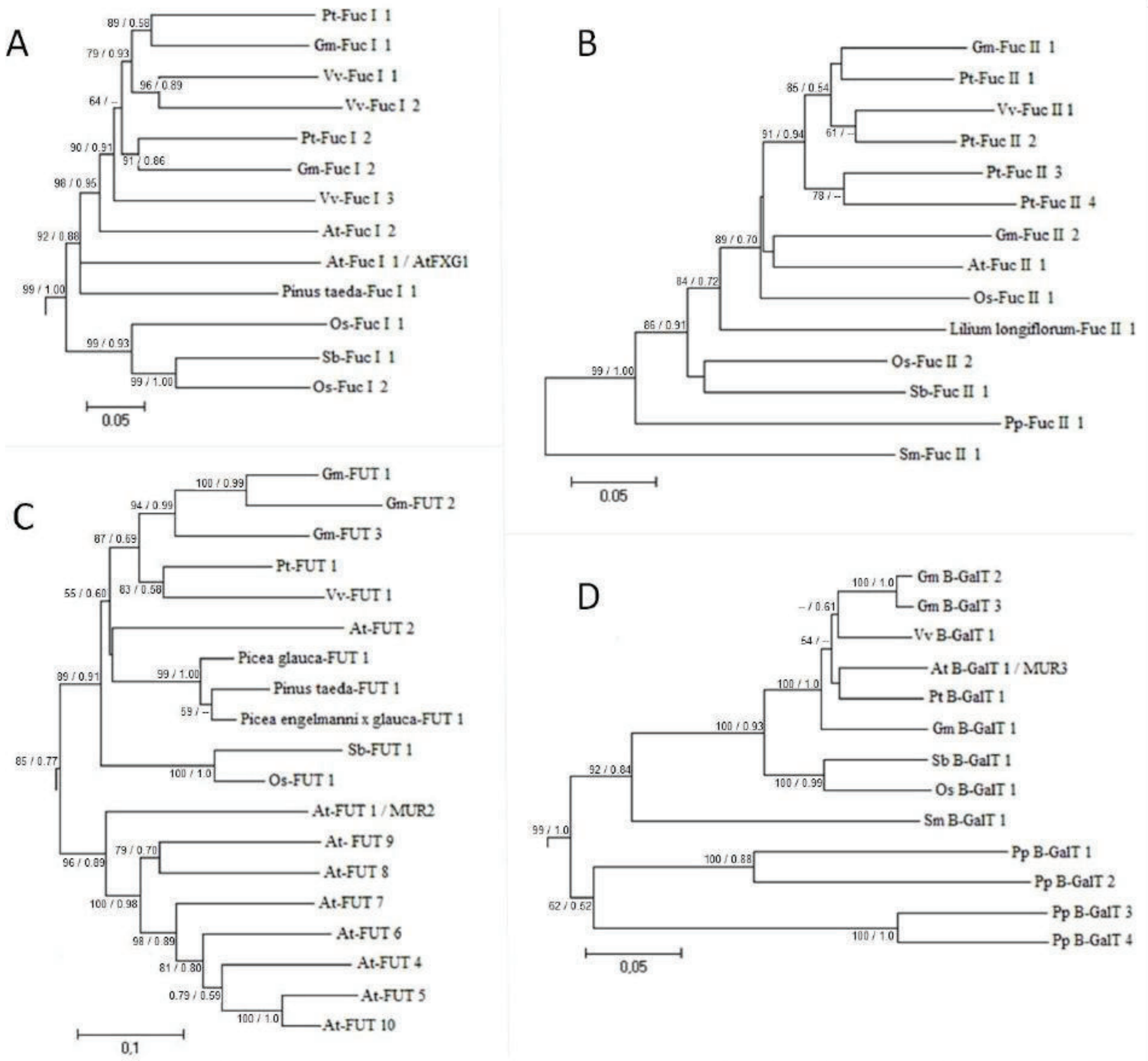


Figure 5 Evolutionary profiles of α-fucosidase type I (A), α-fucosidase type II (B), α-fucosyltransferase (C) and β-galactosyltransferase (D) genes in green plants. A: *Arabidopsis AtXYL1* orthologs. B: *Lilium longiflorum EBM II* orthologs. C: *Arabidopsis Mur2* orthologs. D: *Arabidopsis Mur3* orthologs. The topology was obtained by Neighbor-Joining method with genetic distances calculated by p-distance. The bootstrap values higher than 50% are shown for 1000 replicates along with Maximum-likelihood (ML) aLRT values higher than 0.50 (in the figure [**NJ**/**ML**]). Accession numbers for all genes in the trees are described in the following Supplemental Tables: A and B - Additional File 16, C - Additional File 18 and 18D - Additional File 20. The detailed trees are shown in the following Additional Files: A - Additional File 15, C - Additional File 17 and 17D - Additional File 19.

topology shown in Additional File 15, it is likely that PoGO A XyG-active α-fucosidases emerged from PoGO α. The functional characterizations of genes from PoGOs α and β should improve our understanding of the diversification of α-fucosidase from GDSL-motif lipase/hydrolase family protein.

Lilium longiflorum EBM II homologues among green plants formed a single PoGO (PoGO B; Figure 5B; Additional File 16) that was unrelated to the GDSL-motif lipase/hydrolase gene family. PoGO B arose at least in the last common ancestor of embryophytes. *Arabidopsis* had a single gene in PoGO B (*At4g34260*) that was recently confirmed to encode a protein with XyG α-fucosidase activity (*AtFuc95A*; [62]). Green algae genomes contained no genes similar to PoGO B members, suggesting that α-fucosidases homologous to *EBM II* are limited to land plants, in a manner similar to *AtFXG1* homologues.

α-Fucosyltransferases orthologs to *Mur2* are present among spermatophytes and share similarity with uncharacterized embryophyte genes

To determine the evolutionary profile of XyG α-fucosyltransferase in plants we used the protein encoded by the functionally characterized *Arabidopsis Mur2* gene (At2g03220; [63]) as a query. This strategy identified 82 possible *Mur2* homologous genes among embryophytes (Additional File 17). Because there is little functional information for this family, we will limit our discussion to PoGO A, which contains *Mur2* and includes genes from angiosperms and gymnosperms (Figure 5C; Additional File 18).

The presence of gymnosperm sequences suggested that PoGO A must have emerged in the last common ancestor of spermatophytes. This finding agreed with the presence of fucosylated XyG exclusively among spermatophytes [55]. Since the *Arabidopsis mur2* mutant contains <2% of wild-type fucosylated XyG [63] it is likely that the protein encoded by *Mur2* is the principle activity responsible for the transfer of fucosyl residues to XyG. In contrast, the *Arabidopsis* genome contains another set of nine genes that share high similarity with *Mur2* (Additional File 17), of which eight paralogs are present in PoGO A (Figure 5C; Additional File 18). The role played by these genes remains unclear and it will be interesting to understand the genetic interaction between these genes and *Mur2*.

β-Galactosyltransferases emerged in early land plants and share similarity with an extensive group of poorly characterized genes in green plants

The protein sequence of the well characterized β-galactosyltransferase gene *Mur3* (At2g20370) from *Arabidopsis* [64] was used as a query to search for homologous genes among green plants. Madson *et al.* [64] showed that *Mur3* has sequence similarity to animal exostosins, which are proteins involved in biosynthesis of the extracellular matrix. Our search revealed 191 genes that were possibly homologous to *Mur3*, none of which has been functionally characterized (Additional File 19). Within this extensive group of genes, several from chlorophytes (27 from *Chlamydomonas*, 16 from *Volvox*, three from *O. tauri* and two from *O. lucimarinus*) could represent the ancestral exostosin-like genes in plants (Additional File 19) from which the XyG galactosyltranferase activity probably evolved.

Functional information about this large family is restricted to *Mur3*, which is included in the embryophyte-specific PoGO A (Figure 5D; Additional File 20). *Mur3*-encoded protein acts specifically on the third xylose residue in the XXXG core structure of XyG, implying that other related enzymes transfer the galactosyl residues to the second xylose residue [64]. The candidate genes associated with the latter activity in *Arabidopsis* must be *At2g29040*, *At4g13990* and *At2g32750*, which are the genes most closely related to *Mur3* (Figure 5D).

Conclusions

The comparative genomic analysis of enzymes involved in XyG synthesis and turnover described here has provided a few key conclusions about the evolution of this polymer in green plants. Evidence from non-XyG-bearing chlorophyte and streptophyte green algae indicates that part of the embryophyte XyG-related machinery (XTH, β-[1→4]-glucan synthase from the CSL C family and α-xylosidase) evolved in an aquatic environment, before land colonization by plants. This conclusion agrees with a recent report by Sørensen *et al.* [37] who used a combination of monosaccharide linkage analysis, CoMPP and immunolabeling to detect XyG in the PCW of some Charophycean algae, including Charales, Coleochaetales, and Zygnematales. Although Popper and Fry [53] detected no XyG in the cell wall of streptophyte algae such as *Chara*, *Nitella*, *Coleochaete* and *Klebsormidium*, the presence of an XyG-like polymer containing glucose and xylose was also reported in the alga *Spirogyra* (a streptophyte from the Class Zygnematophyceaes; [52]). In addition, XyG endotransglycosylase (XTH) activity has been detected in growing tissues of *Chara* [28]. Together, these observations suggest that in Charophycean algae a XyG-like polymer may be part of the PCW structure and that the mechanism by which hemicellulose is transglycosylated to adapt the PCW to cellular growth is conserved among streptophytes.

Streptophyta algae have at least three enzymes involved in XyG synthesis and turnover that are homologous to those of embryophytes, namely, XTH, β-(1→4)-glucan synthase and α-xylosidase (Figure 6). Homologous of α-xylosidase are present in Chlorophyta algae that completely lack XyG. Overall, our findings support the idea that a primordial XyG-like polymer emerged before land colonization by plants. The selective advantage conferred by this polymer may have been related to cell-cell attachment features within streptophytes multicellular algae rather than to mechanical structure [52]. Once the land was colonized, XyG was definitively incorporated into the PCW, as exemplified by the presence of XyG in basal land plants [7].

Our evolutionary data highlight the great functional plasticity of XyG glycosyl hydrolases (GHs) and XyG glycosyl transferase (GTs) in the course of green plant evolution. For example, α-xylosidase activity possibly emerged from α-glucosidase, β-(1→4)-glucan synthase-specific enzymes possibly emerged from enzymes with β-mannan activity, and α-fucosidase type I possibly emerged from the GDSL-motif lipase/hydrolase family.

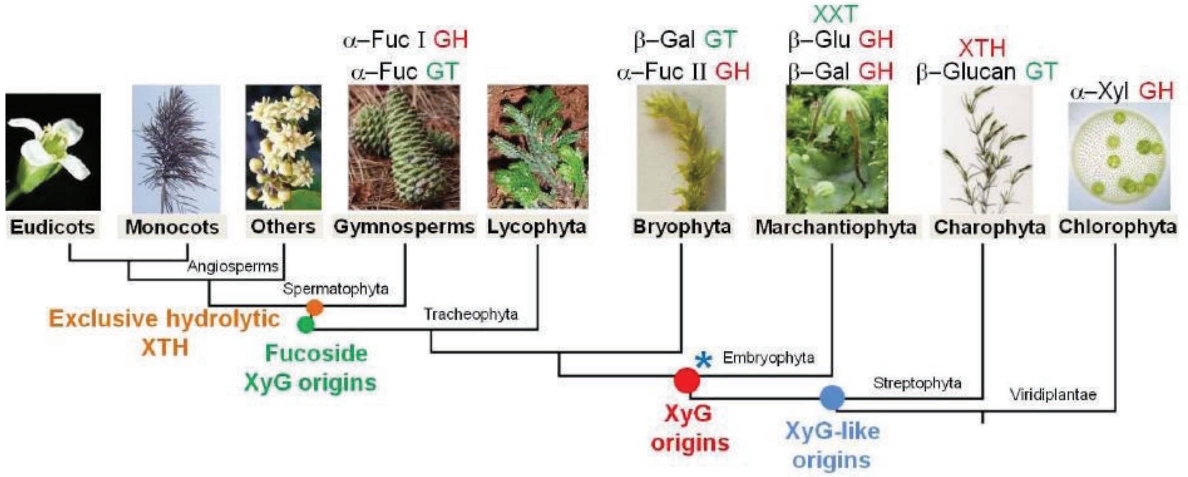


Figure 6 Evolutionary model of XyG-related genes emergence in Viridiplantae kingdom. The model shows the ancient origins we could trace back of each XyG-related gene families and the major events of XyG evolution in plants. XyG-like (containing only glucose and xylose) emerged in streptophytes algae, XyG (containing glucose, xylose, and galactose) emerged in early embryophytes and the fucosylated XyG emerged in the last common ancestor of spermatophytes. (*) indicates the possible origins of the ancestral genes that gave rise to Spermatophytes α-Fucosyl Transferase (*Mur2* orthologs) and α-Fucosylase type I (*AtXYL1* orthologs). GH - Glycosyl Hydrolases and GT - Glycosyl Transferases.

There was a positive correlation between the number of founder genes in XyG-related gene families, defined by the number of PoGOs, and the growing complexity of the PCW. For instance, the number of PoGOs involving XTH sequences from streptophyte algae was limited to two but increased to 12 PoGOs shared amongst eudicots and monocots (Figure 2). Overall, the higher number of PoGOs found to include angiosperm XyG-related genes compared to other plant groups was probably related to the high degree of specialization (expression pattern and/or functional novelties) among gene copies in angiosperm species.

In contrast, there was no clear correlation between the gene copy numbers of XyG GHs and GTs and the amount of XyG in PCW. For example, the moss *Physcomitrella*, as well as rice and sorghum (i.e, graminaceous monocots), in which XyG accounts for <5% of the PCW dry-weight, had 30, 30 and 32 XTH genes, respectively, a number similar to that observed in eudicots (33 in *Arabidopsis*, 35 in poplar and soybean) in which XyG accounts for 10-20% of the PCW dry-weight. We speculate that remodeling of the PCW by the selective turnover and transglycosylation of XyG may be important, even for species with low amounts of XyG.

The role of XyG in embryophyte PCW remains unclear, although recent work has shown that *Arabidopsis* mutant plants with undetectable XyG have an almost normal development [20]. More research is needed to improve our knowledge of the mechanical structure of the PCW and the relationships among its components.

Our data suggest that the colonization of land by plants was marked by a notable increase in the sophistication of the machinery required for XyG biosynthesis and turnover when compared to the pathways present in Streptophyta algae. This finding suggests that complex systems involving several enzymes may evolve in a stepwise manner, with each new step providing some selective advantage, as seen in the galactosylation of embryophyte XyG and fucosylation of spermatophyte XyG (Figure 6). XTHs with exclusively hydrolytic activity emerged by the neofunctionalization of an enzyme with mixed activity (transglycosylase/hydrolase) in the last common ancestor of spermatophytes. Finally, our data provide additional evidence to support the idea that Streptophyta algae and land plants are sister groups because they share XyG-related enzymes (XTH and CLS C) that are not present in chlorophytes (Figure 6).

Methods

Green plant sequence datasets

We generated a dataset of green plant proteins (Viridiplantae 1.0) that included 365,187 protein sequences from several completed genomes (*Arabidopsis thaliana*, version 7.0 - http://www.arabidopsis.org; *Populus trichocarpa*, version 1.1 - http://genome.jgi-psf.org/poplar/poplar.home.html; *Glycine max*, version 0.1 - http://www.phytozome.net/soybean.php; *Oryza sativa*, version 5.0 - http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml; *Sorghum bicolor*, version 1.4 - http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html; *Selaginella* *moellendorffii*, version 1.0 - http://genome.jgi-psf.org/Selmo1/Selmo1.home.html; *Physcomitrella patens patens*, version 1.1 - http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html; *Volvox carteri*, version 1.0 - http://genome.jgi-psf.org/Volca1/Volca1.home.html; *Chlamydomonas reinhardtii*, version 3.0 - http://genome.jgi-psf.org/chlre3/chlre3.home.html; *Ostreococcus lucimarinus*, version 2.0 - http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.home.html and *Ostreococcus tauri*, version 2.0 - http://genome.jgi-psf.org/Ostta4/Ostta4.home.html.

In addition, 402,770 assembled ESTs from *Adiantum capillus-veneris* (7,715), *Amborella trichopoda* (6,649), *Ceratodon purpureus* (1,044), *Ceratopteris richardii* (4,492), *Chamaecyparis obtusa* (4,061), *Closterium peracerosum* (1,716), *Cryptomeria japonica* (9,098), *Cycas rumphii* (4,335), *Ginkgo biloba* (4,178), *Gossypium hirsutum* (70,667), *Liriodendron tulipifera* (7,087), *Marchantia polymorpha* (10,721), *Nuphar advena* (8,144), *Persea americana* (6,700), *Physcomitrella patens patens* (45,149), *Picea abies* (5,204), *Picea engelmannii × glauca* (14,201), *Picea glauca* (49,412), *Picea sitchensis* (25,425), *Pinus pinaster* (13,067), *Pinus taeda* (78,873), *Populus trichocarpa* (31,082), *Pseudotsuga menzienssi* (12,074), *Saruma henryi* (6,956), *Selaginella lepidophylla* (2,861), *Taiwania cryptomerioides* (778), *Tortula ruralis* (7,689) and *Welwitschia mirabilis* (6,680) were downloaded from the TIGR Plant Transcript Assemblies Database [65] and pooled to form the ViridiEST 1.0 database.

Identification of XyG-related genes

To identify XTH, α-xylosidase, β-galactosidase, β-glucosidase, α-fucosidase, β-(1→4)-glucan synthase, α-fucosyltransferases, β-galactosyltransferases and XXT we ran Blast [66] searches against Viridiplantae 1.0 and ViridiEST 1.0. We also performed online Blast searches against the genomic sequences of complete plant genomes using the NCBI http://www.ncbi.nlm.nih.gov/, Phytozome http://www.phytozome.net/ and JGI Eukaryotic Genomes http://genome.jgi-psf.org/ databanks to ensure an exhaustive search. However, no protein sequences were incorporated from the genomic sequence searches if they were not already present in predicted proteomes of the different genome initiatives.

The amino acid sequences of previously reported genes were used as queries. For XTH queries we used 33 *A. thaliana* [29] and 29 *O. sativa* [30] proteins. For β-galactosidase we used 17 *A. thaliana* and 15 *O. sativa* proteins [67]. For XyG-active β-glucosidase we used the well-characterized *Tropaeolum majus* β-glucosidase (CAA07070 - GeneBank; [45]) as the query. For α-xylosidases we used queries from *A. thaliana* (*AtXYL1*, At1g68560 - TAIR; [47]) and *T. majus* (CAA10382 - GeneBank; [48]). For α-fucosidase we used the protein sequences encoded by *AtFXG1* (At1g67830 - TAIR; [18]) and *EBM II* (BAF85832 - GeneBank; [21]) from *A. thaliana* and *Lilium longiflorum*, respectively. For β-(1→4)-glucan synthase we used the protein sequence encoded by *CSLC4* from *A. thaliana* [19]. For α-fucosyltransferases the query was the *AtFUT1* (*Mur2* [63]). For β-galactosyltransferases we used the protein encoded by *Mur3* (At2g20370) from *Arabidopsis* as the query [68]. For XXT we used the *Arabidopsis* genes *XXT1* (At3g62720; [20]), *XXT2* (At4g02500; [20]) and *XXT5* [61].

The complete bioinformatics pipeline that was designed to perform similarity searches and used to produce non-redundant nucleotide and amino acid sequence data-sets is detailed in Additional File 1. We developed two programs used in the pipeline: BTF ("Blast to Fasta") and ETTool ("ESTs Translator Tool"). BTF reads the Blast results, and places the resulting subjects in a Fasta file. ETTool reads the tblastn results of protein queries against EST databases and selects only the blocks of amino acids that aligned between the queries and EST subjects; these blocks were transferred to a Fasta file.

Phylogenetic analysis

The amino acids sequences were aligned with ClustalW [69] using the default parameters and then adjusted manually. All phylogenetic analyses were done using MEGA4.0 [70]. Phylogenetic distance tree topologies were obtained by the neighbor-joining method [26] with distances calculated by the PAM 001 distance matrix [25] and p-distances using 1000 bootstrap replicates. Maximum likelihood analyses were done in PhyML 3.0 [27] using the LG substitution model and an LTR statistical test [27]. All sequences used in this study are available upon request.

Identification of possible groups of orthologs (PoGOs)

The detailed evolutionary analysis of XyG-related gene families allowed the identification of PoGOs. A PoGO was defined by the following criteria: (1) members of a PoGO were assumed to have a monophyletic origin, indicated by a bootstrap support greater than 50%; (2) a PoGO possessed at least one representative gene from *A. thaliana* and/or *O. sativa*, assuming that the putative complete set of genes for these organisms had been identified. In the case of a PoGO being restricted to some lineage, e.g., mosses or gymnosperms, the presence of sequences from at least two species of the same lineage in this PoGO was required.; and (3) the inferred phylogeny should be consistent with the known phylogeny of plant species [71].

Additional material

Additional file 1: Bioinformatics search protocol. The pipeline was used in construction of non-redundant protein data-sets. The arrows with asterisks represent manually conducted processes. The e-value cutoffs were $1e^{-4}$ for Blastp and tBlastn. Our own programs (BTF and ETTool, both written in JAVA®) were developed for this protocol (available upon request). False positives from B1 and B2 protein sets were eliminated from the alignment by visual confrontation with reference sequences. A NJ tree was generated using B3 and B4 sets together with reference sequences. Redundant sequences and alternative splicing isoforms were eliminated by manual inspection of resulting tree. The final non-redundant protein data-sets obtained were used in our analyses.

Additional file 2: Detailed phylogenetic analysis of XTH gene family in green plants. PoGOs names and color scheme are the same of Figure 2. The topology was inferred by Neighbor-Joining (NJ) method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown.

Additional file 3: XTH Possible Groups of Orthologs (PoGOs) in green plants. Classification of XTH PoGOs by taxonomic ranking and the complete list of gene IDs.

Additional file 4: Detailed phylogenetic analysis of β-galactosidase gene family in green plants. PoGOs names and color scheme are the same of Figure 3A. The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown.

Additional file 5: β-galactosidase Possible Groups of Orthologs (PoGOs) in green plants. Classification of β-galactosidase PoGOs by taxonomic ranking and the complete list of gene IDs.

Additional file 6: Detailed phylogenetic analysis of β-glucosidase gene family in green plants. Description: PoGOs names and color scheme are the same of Figure 3B. The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown.

Additional file 7: β-glucosidase Possible Groups of Orthologs (PoGOs) in green plants. Classification of β-glucosidase PoGOs by taxonomic ranking and the complete list of gene IDs.

Additional file 8: Detailed phylogenetic analysis of α-xylosidase gene family in green plants. PoGOs names and color scheme are the same of Figure 3C. The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown.

Additional file 9: α-xylosidase Possible Groups of Orthologs (PoGOs) in green plants. Classification of α-xylosidase PoGOs by taxonomic ranking and the complete list of gene IDs.

Additional file 10: Phylogenetic analysis of α-xylosidase related homologous groups in Eukaryotes and Bacteria. All sequences analyzed were selected using *AtXYL1* from *Arabidopsis* (*At1g68560*) as query in blast searches with e-value cutoff of e^{-4}. The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown.

Additional file 11: Detailed phylogenetic analyses of CSL-A and CSL-C (β-Glucan Synthase) gene families in green plants. PoGOs names and color scheme are the same of Figure 4A. The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown.

Additional file 12: CSL C (β-glucan synthase) and CSL A Possible Groups of Orthologs (PoGOs) in green plants. Classification of CSL C and A PoGOs by taxonomic ranking and the complete list of gene IDs.

Additional file 13: Detailed phylogenetic analysis of α-xylosyl transferase (XXT) gene family in green plants. PoGOs names and color scheme are the same of Figure 4B. The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown.

Additional file 14: XXT Possible Groups of Orthologs (PoGOs) in green plants. Classification of XXT PoGOs by taxonomic ranking and the complete list of gene IDs.

Additional file 15: Detailed phylogenetic analysis of α-fucosidase type I gene family in green plants. *AtFXG1* (At1g67830) PoGO is marked (Figure 5A). The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown. This analysis allowed identification of PoGOs 1 and 2 which integrate *Selaginella* and *Physcomitrella* genes, suggesting that they emerged at least in the last common ancestor of land plants and represent the ancestral groups. The function of these enzymes is largely unknown.

Additional file 16: α-fucosidase I and II Possible Groups of Orthologs (PoGOs) in green plants. Classification of α-fucosidase I and II PoGOs by taxonomic ranking and the complete list of gene IDs.

Additional file 17: Detailed phylogenetic analysis of α-fucosyltransferases in green plants. *Arabidopsis Mur2* PoGO is marked (Figure 5C). The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown. *Mur2* are present among spermatophytes and share similarity with uncharacterized gene from *Physcomitrella* and *Selaginella*, suggesting that the genes that gave rise to *Mur2* orthologs emerged in early land plants.

Additional file 18: α-fucosyltransferase Possible Groups of Orthologs (PoGOs) in green plants. Classification of α-fucosyltransferase PoGOs by taxonomic ranking and the complete list of gene IDs.

Additional file 19: Detailed phylogenetic analysis of β-galactosyltransferases in green plants. *Arabidopsis Mur3* PoGO is marked (Figure 5D). The topology was inferred by NJ method with 1000 bootstraps replicates and the genetic distances were calculated using p-distance. Bootstrap values higher than 50% are shown. Several genes from chlorophytes (27 from *Chlamydomonas*, 16 from *Volvox*, three from *Ostreococcus tauri*, and two from *O. lucimarinus*) could represent the ancestral plant exostosin-like genes from which the XyG galactosyl tranferase activity probably evolved. This analysis includes animal exostosin.

Additional file 20: β-galactosyltransferase Possible Groups of Orthologs (PoGOs) in green plants. Classification of β-galactosyltransferase PoGOs by taxonomic ranking and the complete list of gene IDs.

Abbreviations

CSL: cellulose synthase-like; ESTs: expressed sequence tags; GH: glycosyl hydrolase; GT: glycosyl transferase; ML: maximum likelihood; NJ: neighbor joining; PCW: primary cell wall; PoGO: possible group of orthologs; XTH: xyloglucan transglycosylase/hydrolase; XyG: xyloglucan; XXT: α-xylosyl transferase.

Acknowledgements

The authors are grateful to Dr. Harry Brumer (Stockholm, Sweden) and Dr. Pedro Coutinho (Marseille, France) for kindly provide poplar XTH's sequences and to Guilherme F. de Mello (Campinas, Brazil) for bioinformatics collaboration in development of in-house tools used in this study. All green plant images in Figure 6 were taken from the internet http://www.wikipedia.org, except for the *Selaginella* (Lycophyta) pictured by LEV Del Bem. This work was supported by "Fundação de Amparo à Pesquisa do Estado de São Paulo" (FAPESP) and "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES).

Author details

[1]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), CP 6010, CEP 13083-875, Campinas, SP, Brazil.
[2]Departamento de Biologia Vegetal, Instituto de Biologia, Universidade

Estadual de Campinas (UNICAMP), CP 6109, CEP 13081-970, Campinas, SP, Brazil.

Authors' contributions

LEVDB idealized the research, performed all the analysis and wrote the manuscript. MGAV is a group leader, has intellectual input in all presented results and conclusions and manuscript elaboration. Both authors read and approved the final manuscript.

Received: 1 May 2010 Accepted: 5 November 2010
Published: 5 November 2010

References

1. Fry SC: **Cross-linking of matrix polymers in the growing cell walls of angiosperms.** *Annual Review of Plant Physiology, Palo Alto* 1986, **37**:165-186.
2. Carpita N, McCann M: **The cell wall.** In *Biochemistry and Molecular Biology of Plants.* Edited by: Buchanan BB, Gruissem W, Jones RL. American Society of Plant Physiologists, Rockville, Maryland; 2000:52-108.
3. Albert M, Werner M, Proksch P, Fry SC, Kaldenhoff R: **The cell wall-modifying xyloglucan endotransglycosylase/hydrolase LeXTH1 is expressed during the defence reaction of tomato against the plant parasite Cuscuta reflexa.** *Plant Biol (Stuttg)* 2004, **6(4)**:402-407.
4. Carpita NC, Gibeaut DM: **Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth.** *Plant Journal* 1993, **3(1)**:1-30.
5. Cosgrove DJ: **Relaxation in a high-stress environment: the molecular bases of extensible cell walls and cell enlargement.** *Plant Cell* 1997, **9(7)**:1031-1041.
6. Baumann MJ, Eklöf JM, Michel G, Kallas AM, Teeri TT, Czjzek M, Brumer H: **Structural evidence for the evolution of xyloglucanase activity from xyloglucan endo-transglycosylases: biological implications for cell wall metabolism.** *Plant Cell* 2007, **19(6)**:1947-1963.
7. Peña MJ, Darvill AG, Eberhard S, York WS, O'Neill MA: **Moss and liverwort xyloglucans contain galacturonic acid and are structurally distinct from the xyloglucans synthesized by hornworts and vascular plants.** *Glycobiology* 2008, **18(11)**:891-904.
8. Edwards M, Dea IC, Bulpin PV, Reid JS: **Purification and properties of a novel xyloglucan-specific endo-(1-4)-beta-D-glucanase from germinated nasturtium seeds (Tropaeolum majus L.).** *J Biol Chem* 1986, **261(20)**:9489-9494.
9. Edwards M, Bowman YJ, Dea IC, Reid JS: **A beta-D-galactosidase from nasturtium (Tropaeolum majus L.) cotyledons. Purification, properties, and demonstration that xyloglucan is the natural substrate.** *J Biol Chem* 1988, **263(9)**:4333-4337.
10. Reis D, Vian B, Darzens D, Roland JC: **Sequential patterns of intramural digestion of galactoxyloglucan in tamarind seedlings.** *Planta* 1987, **170(1)**:60-73.
11. Buckeridge MS, Rocha DC, Reid JSG, Dietrich SMC: **Xyloglucan structure and post-germinative metabolism in seeds of *Copaifera langsdorfii* from savanna and forest populations.** *Physiologia Plantarum (Copenhagen)* 1992, **86**:145-151.
12. Tiné MAS, Cortelazzo AL, Buckeridge MS: **Xyloglucan mobilisation in cotyledons of developing plantlets of *Hymenaea courbaril* L. (Leguminosae-Caesalpinoideae).** *Plant Science* 2000, **154(2)**:117-126.
13. Fry SC, York WS, Albersheim P, Darvill A, Hayashi T, Joseleau JP, Kato Y, Lorences EP, Maclachlan GA, McNeil M, Mort AJ, Reid JSG, Seitz HU, Selvendran RR, Voragen AGJ, White AR: **An unambiguous nomenclature for xyloglucan-derived oligosaccharides.** *Physiologia Plantarum* 1993, **89(1)**:1-3.
14. Hayashi T: **Xyloglucans in the Primary Cell Wall.** *Annual Review of Plant Physiology and Plant Molecular Biology* 1989, **40**:139-168.
15. Buckeridge MS, Crombie HJ, Mendes CJ, Reid JS, Gidley MJ, Vieira CC: **A new family of oligosaccharides from the xyloglucan of Hymenaea courbaril L. (Leguminosae) cotyledons.** *Carbohydr Res* 1997, **303(2)**:233-237.
16. Tiné MAS, Silva CO, de Lima DU, Carpita NC, Buckeridge MS: **Fine structure of a mixed-oligomer storage xyloglucan from seeds of *Hymenaea courbaril*.** *Carbohydrate Polymers* 2006, **66(4)**:444-454.
17. Santos HP, Purgato E, Mercier H, Buckeridge MS: **The Control of Storage Xyloglucan Mobilization in Cotyledons of *Hymenaea courbaril* L.** *Plant Physiology* 2004, **135**:287-299.
18. De La Torre F, Sampedro J, Zarra I, Revilla G: ***AtFXG1*, an Arabidopsis Gene Encoding alpha-L-Fucosidase Active against Fucosylated Xyloglucan Oligosaccharides.** *Plant Physiology* 2002, **128**:247-255.
19. Cocuron JC, Lerouxel O, Drakakaki G, Alonso AP, Liepman AH, Keegstra K, Raikhel N, Wilkerson CG: **A gene from the cellulose synthase-like C family encodes a {beta}-1,4 glucan synthase.** *PNAS* 2007, **104**:8550-8555.
20. Cavalier DM, Lerouxel O, Neumetzler L, Yamauchi K, Reinecke A, Freshour G, Zabotina OA, Hahn MG, Burgert I, Pauly M, Raikhel NV, Keegstra k: **Disruption of two *Arabidopsis thaliana xylosyltransferase* genes results in plants deficient in xyloglucan, a major primary cell wall component.** *Plant Cell* 2008, **20**:1519-1537.
21. Ishimizu T, Hashimoto C, Takeda R, Fujii K, Hase S: **A Novel 1,2-L-Fucosidase Acting on Xyloglucan Oligosaccharides is Associated with Endoß-Mannosidase.** *Journal of Biochemistry* 2007, **142(6)**:721-729.
22. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**:631-637.
23. Fitch WM: **Homology a personal view on some of the problems.** *Trends Genet* 2000, **16**:227-31.
24. Thornton JW, DeSalle R: **Gene family evolution and homology: genomics meets phylogenetics.** *Annu Rev Genomics Hum Genet* 2000, **1**:41-73.
25. Dayhoff MO, Schwartz RM, Orcutt BC: **A model of evolutionary change in proteins.** *Atlas of protein sequence and structure* 1978, **5(suppl 3)**:345-351.
26. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4(4)**:406-25.
27. Guindon S, Gascuel O: **PhyML - A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Systematic Biology* 2003, **52(5)**:696-704.
28. Van Sandt VST, Stieperaere H, Guisez Y, Verbelen JP, Vissenberg K: **XET Activity is Found Near Sites of Growth and Cell Elongation in Bryophytes and Some Green Algae: New Insights into the Evolution of Primary Cell Wall Elongation.** *Annals of Botany* 2007, **99(1)**:39-51.
29. Rose JKC, Braam J, Fry SC, Nishitani K: **The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: current perspectives and a new unifying nomenclature.** *Plant and Cell Physiol* 2002, **43(12)**:1421-1435.
30. Yokoyama R, Rose JKC, Nishitani K: **A surprising diversity and abundance of xyloglucan endotransglucosylase/hydrolases in rice: Classification and expression analysis.** *Plant Physiology* 2004, **134(3)**:1088-1099.
31. Karol KG, McCourt RM, Cimino MT, Delwiche CF: **The closest Living Relatives of Land Plants.** *Science* 2001, **294(5550)**:2351-2353.
32. Becker B, Birger M: **Streptophyte algae and the origin of embryophytes.** *Annals of Botany* 2009, **103(7)**:999-1004.
33. Boisselier-Dubayle MC, Lambourdière J, Bischler H: **Molecular phylogenies support multiple morphological reductions in the liverwort subclass Marchantiidae (Bryophyta).** *Mol Phylogenet Evol* 2002, **24**:66-77.
34. Sørensen I, Domozych D, Willats WGT: **How have plant cell walls evolved?** *Plant Physiology* 2010, **153**:366-372.
35. Wellman CH, Osterloff PL, Mohiuddin U: **Fragments of the Earliest Land Plants.** *Nature* 2003, **425**:282-285.
36. Gibeaut DM, Pauly M, Bacic A, Fincher GB: **Changes in cell wall polysaccharides in developing barley (Hordeum vulgare) coleoptiles.** *Planta* 2005, **221**:729-738.
37. Penning BW, Hunter CT III, Tayengwa R, Eveland AL, Dugard CK, Olek AT, Vermerris W, Koch KE, McCarty DR, Davis MF, Thomas SR, McCann MC, Carpita NC: **Genetic Resources for Maize Cell Wall Biology.** *Plant Physiology* 2009, **151**:1703-1728.
38. Fry SC: **The structure and functions of xyloglucan.** *J Exp Bot* 1989, **40**:1-11.
39. O'Neill MA, York WS: **The composition and structure of plant primary cell walls.** In *The Plant Cell Wall.* Edited by: Rose JKC. Boca Raton, FL: CRC Press; 2003:1-54.
40. Grandjean O, Vernoux T, Laufs P, Belcram K, Mizukami Y, Traas J: **In vivo analysis of cell division, cell growth, and differentiation at the shoot apical meristem in *Arabidopsis*.** *The Plant Cell* 2004, **16**:74-87.
41. Pinto LVA, Da Silva EAA, Davide AC, Mendes De Jesus VA, Toorop PE, Hilhorst HWM: **Mechanism and Control of *Solanum lycocarpum* Seed Germination.** *Annals of Botany* 2007, **100**:1175-1187.
42. Brandão AD, Del Bem LEV, Vincentz M, Buckeridge MS: **Expression pattern of four storage xyloglucan mobilization-related genes during seedling development of the rain forest tree *Hymenaea courbaril* L.** *Journal of Experimental Botany* 2009, **60**:1191-1206.

43. Norstog KJ, Nicholls TJ: **The biology of the Cycads.** Cornell University Press, Ithaca, New York; 1997, 363.

44. Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R: **An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens.** *BMC Evol Biol* 2007, **7**:130.

45. Crombie HJ, Chengappa S, Hellyer A, Reid JSG: **A xyloglucan oligosaccharide-active, transglycosylating beta-D-glucosidase from the cotyledons of nasturtium (*Tropaeolum majus* L) seedlings–purification, properties and characterization of a cDNA clone.** *Plant Journal* 1998, **15(1)**:27-38.

46. Iglesias N, Abelenda JA, Rodiño M, Sampedro J, Revilla G, Zarra I: **Apoplastic glycosidases active against xyloglucan oligosaccharides of *Arabidopsis thaliana*.** *Plant Cell Physiol* 2006, **47(1)**:55-63.

47. Sampedro J, Sieiro C, Revilla G, González-Villa T, Zarra I: **Cloning and Expression Pattern of a Gene Encoding an alpha-Xylosidase Active against Xyloglucan Oligosaccharides from Arabidopsis.** *Plant Physiology* 2001, **126**:910-920.

48. Crombie HJ, Chengappa S, Jarman C, Sidebottom C, Reid JSG: **Molecular characterisation of a xyloglucan oligosaccharide-acting alpha-D-xylosidase from nasturtium (*Tropaeolum majus* L.) cotyledons that resembles plant 'apoplastic' alpha-D-glucosidases.** *Planta* 2002, **214(3)**:406-413.

49. Burn JE, Hurley UA, Birch RJ, Arioli T, Cork A, Williamson RE: **The cellulose-deficient Arabidopsis mutant *rsw3* is defective in a gene encoding a putative glucosidase II, an enzyme processing N-glycans during ER quality control.** *Plant J* 2002, **32**:949-960.

50. Richmond TA, Somerville CR: **The cellulose synthase superfamily.** *Plant physiology* 2000, **124(2)**:495-498.

51. Yin Y, Huang J, Xu Y: **The cellulose synthase superfamily in fully sequenced plants and algae.** *BMC Plant Biology* 2009, **9**:99.

52. Ikegaya H, Hayashi T, Kaku T, Iwata K, Sonobe S, Shimmen T: **Presence of xyloglucan-like polysaccharide in Spirogyra and possible involvement in cell-cell attachment.** *Phycological Research* 2008, **56(3)**:216-222.

53. Popper ZA, Fry SC: **Primary cell wall composition of bryophytes and charophytes.** *Ann Bot* 2003, **91**:1-12.

54. Liepman AH, Nairn CJ, Willats WGT, Sørensen I, Roberts AW, Keegstra K: **Functional genomic analysis supports conservation of function among cellulose synthase-like A gene family members and suggest diverse roles of mannans in plants.** *Plant Physiol* 2007, **143**:1881-1893.

55. Sarkar P, Bosneaga E, Auer M: **Plant cell walls throughout evolution: towards a molecular understanding of their design principles.** *Journal of Experimental Botany* 2009, **60(13)**:3615-3635.

56. Fry SC, Miller JK: **Toward a working model of the growing plant cell wall: Phenolic cross-linking reactions in the primary cell walls of dicotyledons.** In *Plant Cell Wall Polymers: Biogenesis and Degradations.* Edited by: Lewis NG, Paice MG. ACS Symposium series 299, American Chemical Society, Washington, DC; 1989:33-46.

57. McCann MC, Roberts K: **Architecture of the primary cell wall.** In *The Cytoskeletal Basis of Plant Growth and Form.* Edited by: Lloyd CW. Academic Press, New York; 1991:109-129.

58. Passioura JB, Fry SC: **Turgor and cell expansion: beyond the Lockhard equation.** *Aust Plant Physiol* 1992, **19**:565-576.

59. Veytsman BA, Cosgrove DJ: **A model of cell wall expansion based on thermodynamics of polymer networks.** *Biophysical journal* 1998, **75(5)**:2240-2250.

60. Somerville C, Bauer S, Brininstool G, Facette M, Hamann T, Milne J, Osborne E, Paredez A, Persson S, Raab T, Vorwerk S, Youngs H: **Toward a Systems Approach to Understanding Plant Cell Walls.** *Science* 2004, **306(5705)**:2206-2211.

61. Zabotina OA, van de Ven WT, Freshour G, Drakakaki G, Cavalier D, Mouille G, Hahn MG, Keegstra K, Raikhel NV: **Arabidopsis *XXT5* gene encodes a putative alpha-1,6-xylosyltransferase that is involved in xyloglucan biosynthesis.** *Plant Journal* 2008, **56(1)**:101-15.

62. Léonard R, Pabst M, Bondili JS, Chambat G, Veit C, Strasser R, Altmann F: **Identification of an Arabidopsis gene encoding a GH95 alpha1,2-fucosidase active on xyloglucan oligo- and polysaccharides.** *Phytochemistry* 2008, **69(10)**:1983-1988.

63. Vanzin GF, Madson M, Carpita NC, Raikhel NV, Keegstra K, Reiter WD: **The *mur2* mutant of *Arabidopsis thaliana* lacks fucosylated xyloglucan because of a lesion in fucosyltransferase *AtFUT1*.** *PNAS* 2002, **99(5)**:3340-3345.

64. Madson M, Dunand C, Li X, Vermab R, Vanzin GF, Caplanb J, Shouea DA, Carpita NC, Reiter WD: **The *MUR3* Gene of Arabidopsis Encodes a Xyloglucan Galactosyltransferase That Is Evolutionarily Related to Animal Exostosins.** *The Plant Cell* 2003, **15**:1662-1670.

65. Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, Wu H, Rabinowicz PD, Town CD, Buell CR, Chan AP: **The TIGR Plant Transcript Assemblies database.** *Nucleic Acids Research* 2007, , **35 Database:** D846-51.

66. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.

67. Ahn YO, Zheng M, Bevan DR, Esen A, Shiu SH, Benson J, Peng HP, Miller JT, Cheng CL, Poulton JE, Shih MC: **Functional genomic analysis of Arabidopsis thaliana glycoside hydrolase family 35.** *Phytochemistry* 2007, **68(11)**:1510-1520.

68. Tedman-Jones JD, Lei R, Jay F, Fabro G, Li X, Reiter WD, Brearley C, Jones JDG: **Characterization of Arabidopsis *mur3* mutations that result in constitutive activation of defence in petioles, but not leaves.** *The Plant Journal* 2008, **56(5)**:691-703, (13).

69. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Research* 1994, **22(22)**:4673-80.

70. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Molecular Biology and Evolution* 2007, **24**:1596-1599.

71. Vincentz M, Bandeira-Kobarg C, Gauer L, Schlogl P, Leite A: **Evolutionary pattern of angiosperm bZIP factors homologous to the maize Opaque2 regulatory protein.** *J Mol Evol* 2003, **56**:105-116.

doi:10.1186/1471-2148-10-341
Cite this article as: Del Bem and Vincentz: **Evolution of xyloglucan-related genes in green plants.** *BMC Evolutionary Biology* 2010 **10**:341.

Capítulo I

Evolução de famílias multigênicas em plantas

Parte 2: Evolução do ciclo de calnexina e calreticulina no reino Viridiplantae

Genetica
DOI 10.1007/s10709-010-9544-y

The evolutionary history of calreticulin and calnexin genes in green plants

Luiz Eduardo V. Del Bem

Received: 26 May 2010 / Accepted: 27 December 2010

Abstract Calreticulin and calnexin are Ca^{2+}-binding chaperones localized in the endoplasmic reticulum of eukaryotes acting in glycoprotein folding quality control and Ca^{2+} homeostasis. The evolutionary histories of calreticulin and calnexin gene families were inferred by comprehensive phylogenetic analyses using 18 completed genomes and ESTs covering the major green plants groups, from green algae to angiosperms. Calreticulin and calnexin possibly share a common origin, and both proteins are present along all green plants lineages. The calreticulin founder gene within green plants duplicated in early tracheophytes leading to two possible groups of orthologs with specialized functions, followed by lineage-specific gene duplications in spermatophytes. Calnexin founder gene in land plants was inherited from basal green algae during evolution in a very conservative copy number. A comprehensive classification in possible groups of orthologs and a catalog of calreticulin and calnexin genes from green plants are provided.

Keywords Calreticulin · Calnexin · Chaperones · Evolution · Green plants

Introduction

The highly conserved eukaryotic Ca^{2+}-binding proteins calreticulin (CRT) and calnexin (CNX) are the central players in the so-called CRT/CNX cycle of glycoprotein folding quality control. CRT possesses a C-terminal domain with a (K/H)DEL endoplasmic reticulum (ER) retrieval signal (Michalak et al. 2009), while CNX is a ER membrane-bound protein (Jin et al. 2009). CRT/CNX cycle is part of the N-glycan-dependent quality control mechanism that takes place in the ER lumen (Hammond et al. 1994). Glycan processing starts with its transfer to Asn residues in nascent proteins within ER. Several subsequent glycosyl hydrolysis exposes the $Glc_1Man_9GlcNAc_2$ epitope that is then recognized by CRT and CNX that specifically bind monoglucosylated polymannose glycans (Ware et al. 1995; Caramelo and Parodi 2008).

CNX and CRT were first described in plants in 1993 and 1998, respectively (Huang et al. 1993; Crofts and Denecke 1998). Plant's CRTs were further classified into two groups of homologs, CRT1/2 and CRT3, which were initially thought to be resulting from a gene duplication event occurring before the divergence between monocots and eudicots (Persson et al. 2003). A recent work (Jin et al. 2009) suggested that CRT3 group is present in basal land plants and *Arabidopsis* CRT3-specific function on the retention of defective brassinosteroid receptor EFR in the ER, which is a specific function of plant's CRT3 without functional overlapping with CRT1 and 2 (Christensen et al. 2010). EFR accumulation and signaling are impaired in *Arabidopsis crt3* mutant, affecting the immune response to the bacterial epitope elf18 (Saijo et al. 2009) suggesting a role for CRT3 in bacterial pathogen-associated molecular pattern (PAMP), while CRT1 and 2 are possibly involved

Electronic supplementary material The online version of this article (10.1007/s10709-010-9544-y) contains supplementary material, which is available to authorized users.

L. E. V. Del Bem (✉)
Centro de Biologia Molecular e Engenharia Genética,
Universidade Estadual de Campinas (UNICAMP),
Av. Cândido Rondon 400–Cidade Universitária,
CP 6010, Campinas CEP 13083-875, Brazil
e-mail: lev.del.bem@gmail.com

Published online: 11 January 2011

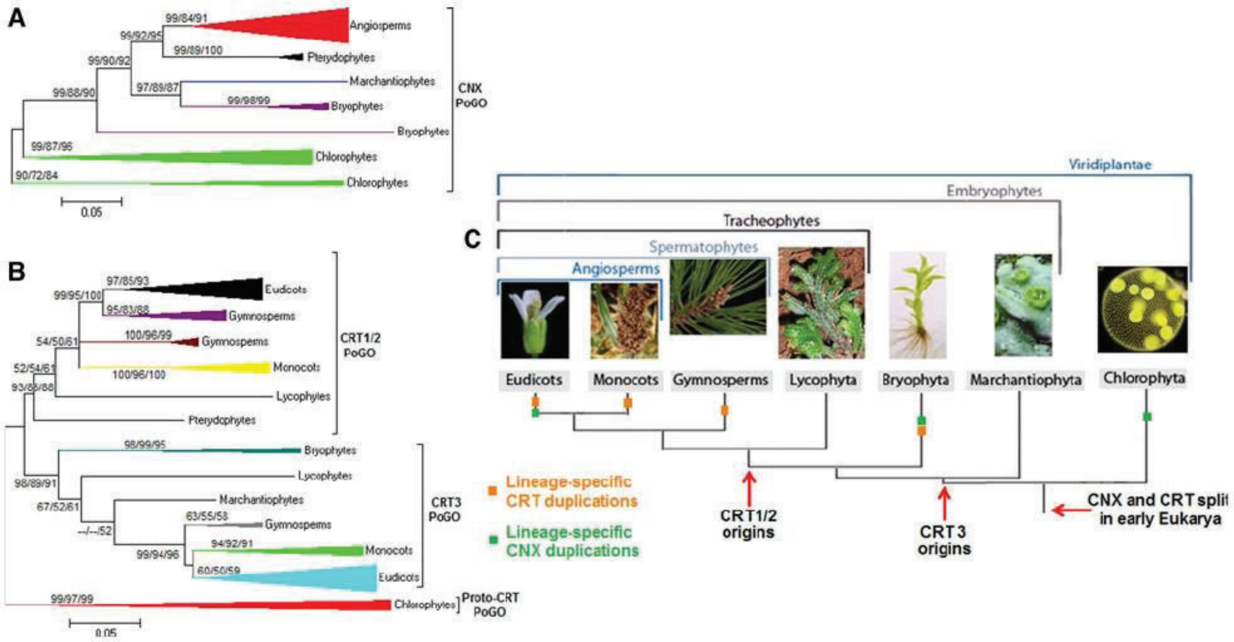


Fig. 1 Phylogenetic trees and evolutionary profile of *CNX* and *CRT* genes in green plants. **a** Phylogenetic tree showing the evolutionary relationship between plants *CNXs*. Tree topology is a consensus from NJ, MP, and Bayesian analyses. Bootstrap values and posterior probabilities from the original trees higher than 50% are shown (NJ/MP/Bayesian). *Triangles* represent compacted groups of orthologs that appear in detail in Supplemental Fig. 1. **b** Phylogenetic tree showing the evolutionary relationship between plants *CRTs*. Tree topology is a consensus from NJ, MP, and Bayesian analyses. Bootstrap values and posterior probabilities from the original trees higher than 50% are shown (NJ/MP/Bayesian). *Triangles* represent compacted groups of orthologs that appear in detail in Supplemental Fig. 2. **c** Evolutionary profile of *CNX* and *CRT* genes in green plants. The *arrows mark* duplication events shared along the descendent lineages, and the *squares mark* lineage-specific duplication events

ebi.ac.uk/interpro/IEntry?ac=IPR018124), which are indicative of a possible common origin. The fact that CRTs and CNXs genes are present along animals and plants along with my results showing the presence of those genes in genomes of green algae such as *Micromonas*, *Volvox*, *Chlorella* and *Ostreococcus* (Fig. 1a, b; Supplemental Figs. 1 and 2; Supplemental Table 1) strongly indicate that they originated by an ancestral gene duplication prior to the divergence between Chlorophyta and Embryophyta. This duplication event could even take place in early eukaryotes.

CNX genes in green plants were further classified in a single possible group of orthologs (PoGO) that integrate genes from Chlorophyta algae to angiosperms (Fig. 1a; Supplemental Table 1; Supplemental Fig. 1). This PoGO generally remained as a single-copy gene in a very diverse taxonomic ranking of green plants such as in the Chlorophyta *Volvox carteri* or even in the monocots sorghum and rice. This observation suggests that single-copy green plant's *CNX* genes probably retained the ancestral eukaryotic function that is thought to be related to glycoprotein folding quality control (Schrag et al. 2001). In contrast, soybean and the moss *Physcomitrella* genomes, probably due to recent large-scale genome duplications, contain four *CNX* paralogs (Supplemental Table 1). I also analysed *Arabidopsis thaliana* (eudicot), sorghum (monocot), *Physcomitrella patens patens* (moss), and *Volvox carteri* (green algae) *CNX* genes for shared intron positions within their coding sequences (Supplemental Fig. 3). This analysis helped support the suggested phylogenetic relationship between green plants *CNX* genes.

Differently from *CNX*, *CRT* genes in Viridiplantae kingdom have diversified specifically in land plants by an ancient event of gene duplication in the last common ancestor of Tracheophyta (Fig. 1b; Supplemental Table 1; Supplemental Fig. 2). While chlorophytes' *CRTs* formed a single PoGO, land plant's *CRTs* were further divided into *CRT1/2* and *CRT3* PoGOs in agreement with the previous literature (Persson et al. 2003). *CRT3* PoGO is embryophyte-specific, which means that genes from this group emerged as a single gene in early land plants' genomes evolving directly from chlorophyte's ancestral single-copy *CRT* gene (PoGO Proto-CRT in Fig. 1b). I interpreted *CRT1/2* PoGO as been derived from a *CRT3* PoGO gene by an ancestral duplication taking place in the last common ancestor of tracheophytes, as evidenced by the presence of a *CRT* gene from *Selaginella moellendorffii* in both PoGOs and the absence of Marchantiophyta and Bryophyta genes in *CRT1/2* PoGO. The proposed phylogenetic classification in more general chaperone functions (Li et al. 2009; Christensen et al. 2010). Plant's CRTs were also implicated in several physiological processes such as virus defense (Chen et al. 2005), ER calcium buffering (Persson et al. 2001; Christensen et al. 2010), plasmodesma cell–cell transport (Baluska et al. 1999; Laporte et al. 2003), and stress response and tolerance (Jia et al. 2008). In the following sections, I will show detailed phylogenetic analyses of *CRT* and *CNX* genes in Viridiplantae kingdom using 18 completed genomes and ESTs from diverse lineages such as green algae, basal non-vascular and vascular land plants, gymnosperms and angiosperms.

Methods

Comparative sequence analyses

Predicted proteomes for *Arabidopsis thaliana*, version 8.0–http://www.arabidopsis.org; *Arabidopsis lyrata*, version 1.0–http://genomeportal.jgi-psf.org/Araly1/Araly1.home.html; *Populus trichocarpa*, version 1.1–http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html; *Glycine max*, version 1.0–http://www.phytozome.net/soybean.php; *Ricinus communis*, version 0.1–http://castorbean.jcvi.org/downloads.php; *Oryza sativa*, version 5.0–http://rice.plantbiology.msu.edu; *Sorghum bicolor*, version 1.4–http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html; *Selaginella moellendorffii*, version 1.0–http://genome.jgi-psf.org/Selmo1/Selmo1.home.html; *Physcomitrella patens patens*, version 1.1–http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html; *Volvox carteri*, version 1.0–http://genome.jgi-psf.org/Volca1/Volca1.home.html; *Chlamydomonas reinhardtii*, version 4.0–http://genome.jgi-psf.org/Chlre4/Chlre4.home.html; *Ostreococcus lucimarinus*, version 2.0–http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.home.html; *Ostreococcus tauri*, version 2.0–http://genome.jgi-psf.org/Ostta4/Ostta4.home.html; *Ostreococcus sp.* RCC809, version 2.0–http://genome.jgi-psf.org/OstRCC809_2/OstRCC809_2.home.html; *Micromonas pusilla* CCMP1545, version 2.0–http://genome.jgi-psf.org/MicpuC2/MicpuC2.home.html; *Micromonas sp.* RCC299, version 2.0–http://genome.jgi-psf.org/MicpuN2/MicpuN2.home.html; *Chlorella vulgaris*, version 1.0–http://genome.jgi-psf.org/Chlvu1/Chlvu1.home.html and *Chlorella sp.* NC64A–http://genome.jgi-psf.org/ChlNC64A_1/ChlNC64A_1.home.html were downloaded and pooled together (Viridiplantae 3.0–530.234 sequences). I performed Hidden Markov Model (HMM) searches using HMMER3 software (http://hmmer.janelia.org/) against Viridiplantae 3.0 in order to identify possible CRT and CNX homologs with an e-value threshold of e^{-5}. I used two different alignments as queries in two independent HMM searches, one of them containing *Arabidopsis* CRT1, 2 and 3 protein sequences and the other containing *Arabidopsis* CNX1 and CNX2 protein sequences.

Maize cDNA sequences were obtained from MAGI (http://magi.plantgenomics.iastate.edu/), and ESTs used in this study were downloaded from TIGR Plant Transcript Assemblies (http://plantta.jcvi.org/) and included 202.387 assembled ESTs (unisequences) from *Ceratopteris richardii* (4.492), *Cycas rumphii* (4.335), *Ginkgo biloba* (4.178), *Marchantia polymorpha* (10.721), *Picea abies* (5.204), *Picea glauca* (49.412), *Picea sitchensis* (25.425), *Pinus pinaster* (13.067), *Pinus taeda* (78.873), and *Welwitschia mirabilis* (6.680). ESTs and cDNAs presenting less than 30% of protein query coverage were discarded.

Phylogenetic analyses

All significantly similar sequences found by HMM were automatically recovered using an in-house algorithm (Del Bem and Vincentz 2010) and manually checked. Sequences were aligned using MAFFT 6.717b (Katoh and Toh 2008) under L–INS–i parameters, and all gaps were removed. Three phylogenetic methods were used to infer the presented trees. Neighbor joining (NJ; Saitou and Nei 1987) using PAM 001 matrix to calculate the genetic distances (Dayhoff et al. 1978) and maximum parsimony (MP; Eck and Dayhoff 1966) under default parameters, both conducted in MEGA 4.0 software (Tamura et al. 2007). Bayesian analyses using the Markov chain Monte Carlo technique were performed using MrBayes3 software (Ronquist and Huelsenbeck 2003) under WAG model of protein substitution (Whelan and Goldman 2001). The Bayesian log-likelihood scores were found to stabilize after 10.000 generations. Therefore, I discarded the initial 10.000 generation trees and sampled one out of every 100 generations from the remaining 0.99 million generations (9,900 trees) to calculate posterior probabilities of each branch. The alignments used contained 329 and 338 amino acids of CNXs and CRTs homologs, respectively. The consensus trees presented in Fig. 1a, b and Supplemental Figs. 1 and 2 were constructed with the three different analyses by *consense* software from PHYLIP suite (Felsenstein 1989) using the majority-rule consensus tree method. The resulting consensus trees were drawn with MEGA4 tree display tool (Tamura et al. 2007). PoGOs were defined as described in Del Bem and Vincentz (2010).

Results and discussion

CRT and CNX possess two conserved domains: the calcium-binding P-domain 'InterPro IPR009033' (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR009033) and the lectin-like N-Domain 'InterPro IPR018124' (http://www.

Schrag JD, Bergeron JJM, Li Y, Borisova S, Hahn M, Thomas DY, Cygler M (2001) The structure of calnexin, an ER chaperone involved in quality control of protein folding. Mol Cell 8(3):633–644

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Ware FE, Vassilakos A, Peterson PA, Jackson MR, Lehrman MA, Williams DB (1995) The molecular chaperone calnexin binds $Glc_1Man_9GlcNAc_2$ oligosaccharide as an initial step in recognizing unfolded glycoproteins. J Biol Chem 270:4697–4704

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18(5):691–699

Capítulo I

Evolução de famílias multigênicas em plantas

Parte 3: Phylexpress (apresentado em formato de artigo ainda em elaboração).

Phylexpress: a phylogenetic method for high-throughput ortholog establishment and evaluation of genes expression evolutionary dynamics

Luiz Eduardo V. Del Bem[1*§], Renato Vicentini[1*] and Michel G. A. Vincentz[1,2§]

[1] Centro de Biologia Molecular e Engenharia Genética – Universidade Estadual de Campinas, Campinas – SP, Brazil.

[2] Departamento de Biologia Vegetal – Universidade Estadual de Campinas, Campinas – SP, Brazil.

[*] These authors contributed equally to this work. [§] Corresponding authors

Introduction

Comparative genomic approaches dramatically improved our understanding of molecular evolutionary processes that gave rise to the complexity of eukaryotic genome organization and function (Koonin, 2009). Furthermore, high-throughput parallel quantification gene expression technologies such as microarrays and second-generation sequencing generated an unprecedented collection of information related to organism spatial-temporal gene expression patterns for several experimental conditions and evolutionary lineages providing a way to explore how gene expression controls growth and development (Schulze and Downward, 2001; Reinartz *et al.*, 2002; Barrett *et al.*, 2010). This combination of biological information led to the identification of several unknown expressed genes (Vandepoele and Van de Peer, 2005) and provided the possibility to address questions on the evolution of comparable RNA profiles responses across species (Mustroph *et al*, 2010).

The task of comparing genomes content and RNA profiles relies partly on the establishment of sequence homology between different organisms. Homologous genes could be further classified into two fundamentally different types of homologues:

orthologs and paralogs (Koonin, 2005). Orthologs are genes from different species that evolved vertically from a single gene in an ancestral species, while paralogs evolved within a single genome by duplication events that are not shared by other species (Tatusov *et al*, 1997; Koonin, 2005). Orthologs could retain the same ancestral function across different species, whereas paralogs may evolve to new functions, even if related to the original one. Thus, identification of orthologs is critically important for gene function prediction in newly sequenced genomes and for gene function information transfer between species. Complete repertoires of orthologs are also a prerequisite for any meaningful comparison of genome organization and gene content (Tatusov *et al*, 1997; Bennetzen, 2002; Pennacchio, 2003; Vincentz *et al*., 2004; Paterson et al., 2010; Del Bem and Vincentz, 2010).

Complementary comparative approaches relying on co-linearity analysis and syntheny inference were essential to unravel complex patterns of polyploidization followed by chromosomal rearrangements, gene loss and gene transposition along angiosperm genome evolution (Tang et al., 2008; Woodhouse et al., 2010). Together, these comparative methods allow delineating the evolutionary fate of ancestral genetic tool kits which ultimately may help to further understand the output of natural variation.

The capacity of evolutionary-related species to grow and develop in different environments is essentially due to the diversification of an ancestral gene pool driven by the interplay of mutational processes and natural selection (negative and positive selection) or genetic drift (Fucile *et al*., 2008; Koonin, 2009; Jackson *et al*., 2009). Large scale chromosomal or even whole genome duplications (WGD) together with tandem duplications are among the mutational mechanisms that represent important sources of increasing genome complexity and genetic novelties (Dehal and Boore, 2005; Vision *et al*, 2000; Kellis *et al*, 2004; Lynch and Conery, 2000; Jaillon *et al*.,

2007) mainly through sub-functionalization or neo-functionalization processes (Force et al., 1999; Lynch and Force 2000; Conant and Wolfe, 2008).

Both pathway may imply cis- or trans-regulatory changes that promote rewiring of regulatory or metabolic networks potentially leading to biological diversity (Prince and Pickett, 2003; Blanc and Wolfe, 2004; Moore and Purugganan, 2005; Doebley and al., 2006; Adams, 2007; Chen and Rajewsky, 2007; Ha et al., 2007; Woolfe and Elgar, 2007; Lynch and Wagner, 2008; Wang et al., 2009; Rosin and Kramer, 2009; Blackman et al., 2010). Similarly, increasing evidences indicate that cis-mediated regulatory alteration of key genes controlling growth and/or development contribute to the evolution of specific traits which may improve adaptation or were important for domestication (Hittinger and Carroll, 2007; Alonso-Blanco et al., 2009; Rebeiz et al., 2009; Schwartz et al., 2009; Tirosh et al., 2009; Wittkopp and al., 2009; Chan et al., 2010). Regarding this issue, it is also noteworthy that an important source of regulatory polymorphism can be powered by transposable elements activity as has been illustrated in rice (Naito et al., 2009).

To get further insight into the evolution of regulatory features related to specific endogenous or environmental signals, a survey of the ever increasing set of genomic-scale transcriptomics from green plant species should be quite informative. For instance, evaluating the degree of gene expression (transcriptional network) divergence across lineages in response to specific signals may provide new perspectives on how expression profile changes can impact the diversification of biological processes (Rifkin et al., 2003; Wohlbach et al., 2009; Li and Johnson, 2010; MacManus et al., 2010, Mustroph *et al*., 2010). With this aim in mind we developed Phylexpress, a bioinformatics tool for large scale orthology establishment that could integrates expression information across orthologs. This later aspect is intended to find conserved hubs within transcriptional networks, as well as to help understanding genetic networks evolutionary plasticity. Our method is a hybrid between similarity-based and tree-based methods. In fact, Phylexpress first employs similarity searchers to find groups of putative homologous genes and then to assign orthologs by considering their phylogenetic relationships represented in trees.

Using Phylexpress to analyze public available microarray data allowed to identify 23 **po**ssible **g**roups of **o**rthologs (PoGOs) containing genes form both rice and *Arabidopsis* (39 from *Arabidopsis* and 30 from rice) that exhibited conserved transcriptional responses to exogenous plant hormone auxin. These genes may represent conserved hubs within IAA signaling transduction that were fixed before the split of eudicots and monocots.

Results & Discussion

Phylexpress overview

Phylexpress is a tree building method aiming to define orthologs. The method uses blastp to assign to an inputted protein sequence up to the 50 most similar (threshold cutoff of < e^{-5}) sequences from a selected database (i.e. a single species or a combination of species predicted proteome). The selected sequences are aligned together with the inputted sequence by MAFFT 6.0 (Katoh and Toh, 2008) and the resulting alignment is fueled to PhyML 3.0 to perform a maximum likelihood (ML) phylogenetic analysis which is displayed in a tree. Currently seven green plants genomes – the eudicots *Arabidopsis thaliana* and *Populus trichocarpa* (poplar), the monocots rice and sorghum, the basal land plants *Selaginella moellendorffii* (Lycophyta) and *Physcomitrella patens patens* (Bryophyta) and the green algae

Chlamydomonas reinhardtii (Chlorophyta) can be searched for orthologs relationship. This set of species covers the main events in plants evolution, including the emergence of green plants (Viridiplantae), the evolution of land plants (Embryophyta), the emergence of plants with conducting vessels (Tracheophyta) and the main split event in the angiosperm (Magnolyophyta) lineage (eudicots and monocots). The ML tree output is then used by Phylexpress to assign the orthologs/paralogs to the inputted sequence according to the following steps. First the method finds the inputted sequence within the tree, and then looks to each node connecting the inputted sequence to one or more sequences of other species following the taxonomic ranking mentioned above. For example, if the user is interested in determining the angiosperms orthologs of a given sequence, Phylexpress will choose the minimum number of nodes connecting the inputted sequence to sequences from *Arabidopsis*, poplar, rice and sorghum. The node used to define a PoGO must have more than 50% of statistical support in and aLRT test (similar to bootstrap test but faster) performed by PhyML. Once a PoGO was assigned to a given node, the program crops the respective branches from the original tree and displays it in a new sub-tree containing only the orthologs assigned to the inputted sequence. An example of orthology assignment in three different taxonomic rankings, using the *Arabidopsis* bZIP gene *AtbZIP3* is shown in figure 1.

Additionally, Phylexpress allows performing searches with coding nucleotide (nt) sequences. In such a case, the program will perform a blastx search step against the selected database. The program will then look for the alignments of best hit, select the aligned block(s) as the predicted protein sequence (or partial sequence in case of ESTs). By performing this step it is possible to avoid discard phylogenetic information due to frame shift mutations that are commonly present in ESTs collections. All subsequent phylogenetic trees are produced from amino acids (aa) alignments.

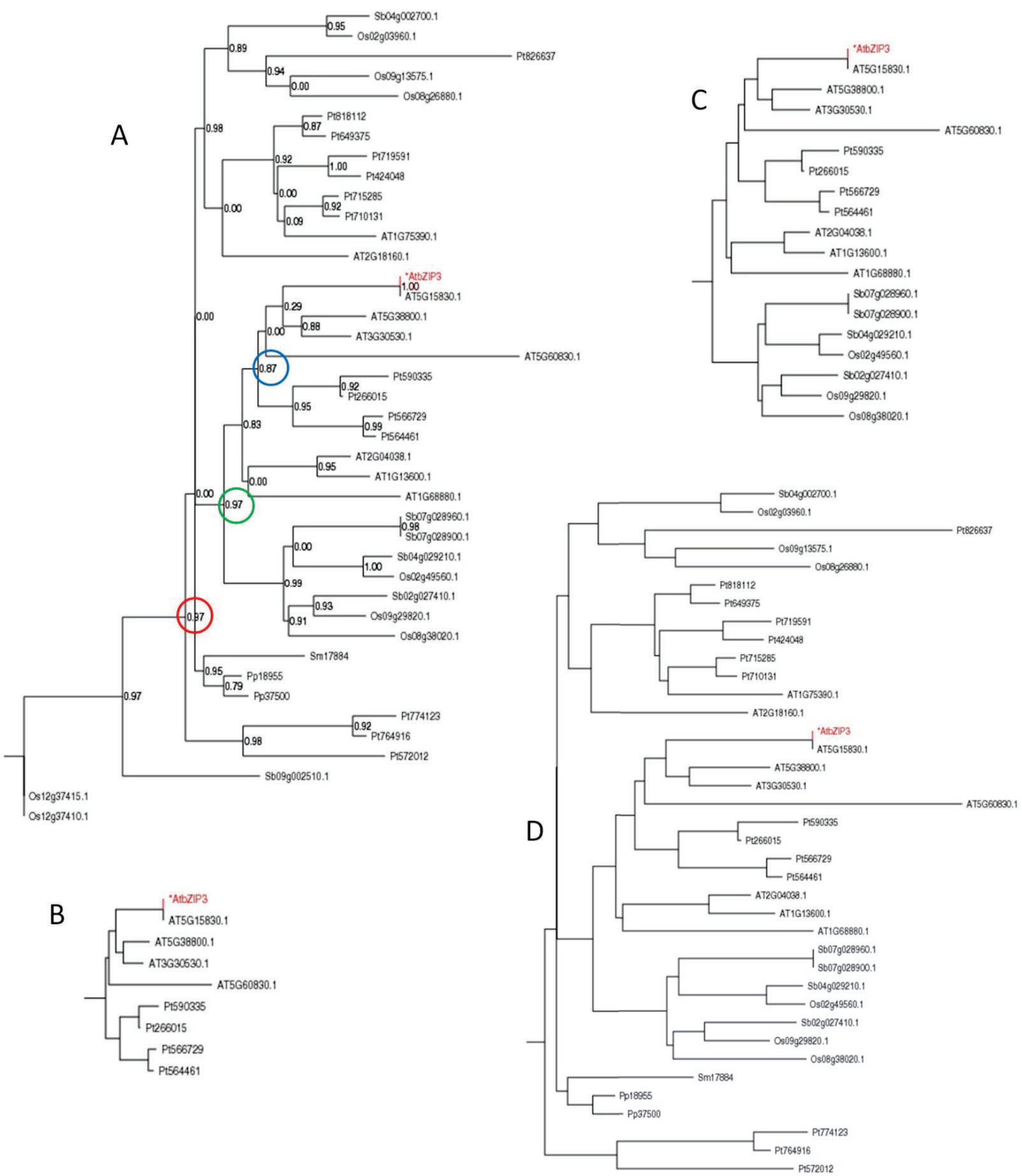


Figure 1. ***Arabidopsis thaliana AtbZIP3*** **orthology assignment in diverse taxonomic ranking.** The blastp was performed against all plants predicted proteomes included in Phylexpress (7 species; *Arabidopsis* – At, poplar – Pt, rice – Os, sorghum – Sb, *Selaginella* – Sm, *Physcomitrella* – Pp and *Chlamydomonas* - Cr). **A –** Original ML tree produced by Phylexpress with the inputted AtbZIP3 protein (highlighted in red). AtbZIP3 most similar sequences (e-value cutoff < e^{-5}) and aLRT statistical support values are shown. The highlighted aLRT represents the nodes selected by Phylexpress (aLRT value must be greater than 0.50) to assign the orthology related to eudicots (in blue), angiosperms (in green) and Embryophyta (in red). **B –** AtbZIP3 eudicots assigned orthologs (At + Pt). **C** – *AtbZIP3* angiosperms orthologs (At + Pt + Os + Sb). **D** – AtbZIP3 Embryophyta orthologs (At + Pt + Os + Sb + Sm + Pp). The method was unable to find algae orthologs in *Chlamydomonas*.

Another feature of Phylexpress is the capacity to perform an evolutionary comparison of two sets of genes from different organisms and organize them into PoGOs. This scheme allows, for instance to evaluate the evolutionary conservation/divergence of transcriptionally regulatory networks. The complete workflow of Phylexpress is shown in figure 2. The first step is to input the protein sequences, each one being separated in up- or down- regulated categories. Phylexpress will blastp reciprocally the protein sequences from both organisms. Each sequence that does not have a significant hit (e-value>e^{-5}) is discarded and these sequences are assumed to represent divergent responses specific to each species or lineage. All sequences possessing a significant hit (e-value<e^{-5}) are selected for further analyses. The retained sequences will individually serve as query to search the chosen databases and form the corresponding PoGOs as described above. Subsequently, Phylexpress recovers the expression information assigned to each sequence (up- or down- regulation) and searches for PoGOs integrating at least one gene from each species under study sharing the same regulation profile. These genes are considered orthologs that share mRNA levels fluctuation responses. An example of such analysis using microarray data of differentially expressed genes from *Arabidopsis* and rice in response to exogenous auxin is presented in the following sections.

Phylexpress could also be used to establish large scale sequence orthology relationship among several species. In contrast to pairwise comparison methods like blast or InParanoid (http://inparanoid.sbc.su.se), Phylexpress could assign orthologs in several species in a multidirectional analysis. Doing the searches in a multidirectional way allows recovering orthologs even when some species lost the corresponding ortholog.

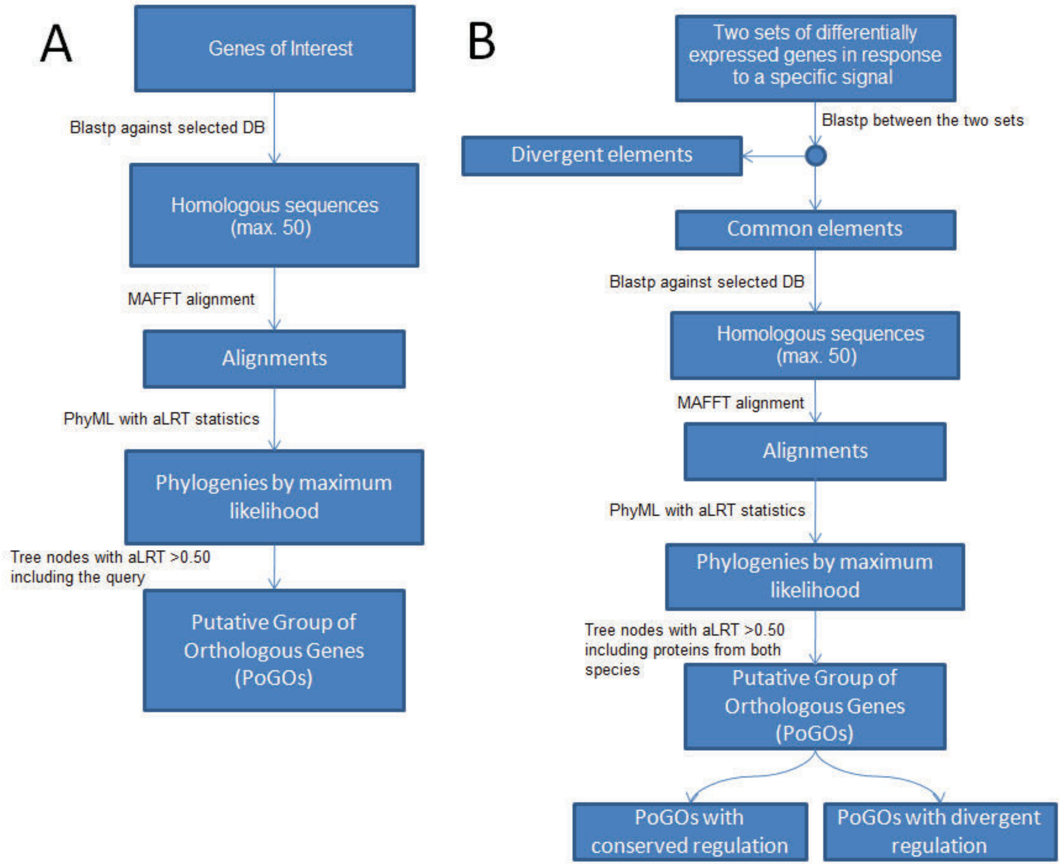

A
Genes of Interest
Blastp against selected DB
Homologous sequences (max. 50)
MAFFT alignment
Alignments
PhyML with aLRT statistics
Phylogenies by maximum likelihood
Tree nodes with aLRT >0.50 including the query
Putative Group of Orthologous Genes (PoGOs)
B
Two sets of differentially expressed genes in response to a specific signal
Blastp between the two sets
Divergent elements
Common elements
Blastp against selected DB
Homologous sequences (max. 50)
MAFFT alignment
Alignments
PhyML with aLRT statistics
Phylogenies by maximum likelihood
Tree nodes with aLRT >0.50 including proteins from both species
Putative Group of Orthologous Genes (PoGOs)
PoGOs with conserved regulation
PoGOs with divergent regulation

Figure 2. Main workflow of Phylexpress. A - Flow for a one sequence or a batch of sequences. **B -** Strategy used when two sets of differential expressed genes are inputted to performed PoGOs analysis and investigate conserved and divergent regulation between orthologs genes.

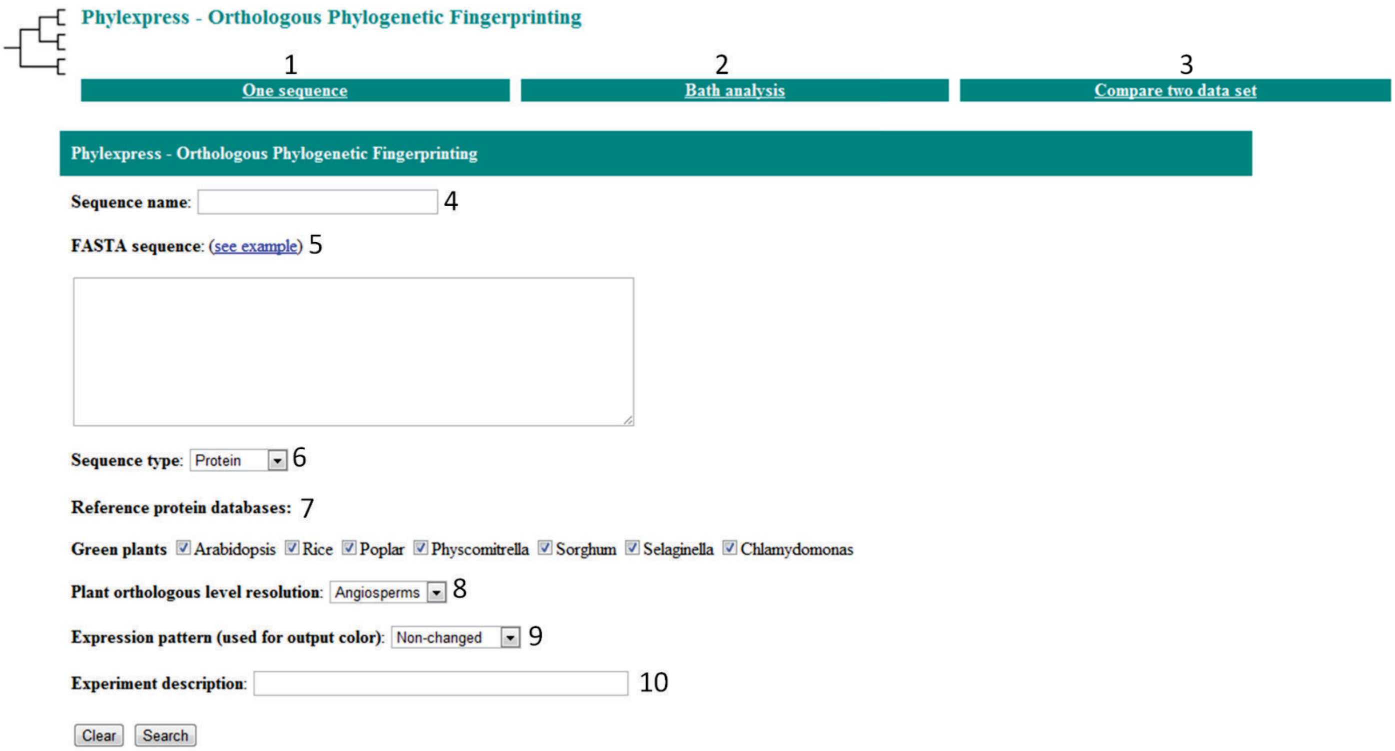
In the case of multidirectional analysis (each against each sequence as input) Phylexpress will produce a tree for each protein, which means that each protein from a specific orthologs group is expected to recover the same PoGO when used as queries. Let us consider now the case where an *Arabidopsis* protein At1 used as query was assigned to a PoGO containing two paralogs (At2 and At3) and three rice orthologs (Os1, Os2 and Os3), the algorithm will then look for the trees produced with the other five genes (At2, At3, Os1, Os2 and Os3) used as queries. A proposed PoGO will be considered as robust only if at least in 70% of the trees generated by each of its member will produce the same PoGO (integrating the same proteins). In our hypothetical example, to be retained, the same PoGO proposed for At1 must be recovered with at least four other members of the PoGO (At2, At3, Os1, Os2 and Os3) giving a reproducibility value of 5 in 6 (~83%) or 6 in 6 (100%) output trees. If the reproducibility is limited to 4 out of 6 trees (66.6%) the initial PoGO would be discarded for being under the 70% threshold level. Once a PoGO is validated by this analysis an average value of aLRT statistics is calculated based on the values of each individual tree.

Accessing Phylexpress robustness

In order to further test the assignment of PoGO by Phylexpress we compared results from whole gene family's trees from the literature with the results obtained with Phylexpress. We used a re-interpreted tree (interpreted in 13 angiosperms PoGOs) from Nam *et al*., (2004) presenting all *Arabidopsis* and rice MIKC-type MADS-Box transcription factors (TF) and a β-Galactosidase tree from Del Bem and Vincentz (2010) presenting nine angiosperms PoGOs. Phylexpress recovered the same *Arabidopsis* and rice genes for 11 out of 13 PoGOs of MIKC-type MADS-Box (~84.6%) and the same *Arabidopsis*, rice, sorghum and poplar genes for 8 out of 9 PoGOs of β-Galactosidase (~88.9%). The PoGOs that could not be identified by Phylexpress is essentially due to insufficient aLRT statistics values supporting the ancestral node of the PoGO.

Phylexpress web server

Phylexpress web server could be accessed in http://sysbiol.cbmeg.unicamp.br/phylo/. In the web interface, the user can upload sequences and chooses parameters such as the sequence type (aa or nt) and the proteomes datasets to conduct the analysis (Figure 3).



Phylexpress - Orthologous Phylogenetic Fingerprinting

Phylexpress - Orthologous Phylogenetic Fingerprinting

Sequence name:

FASTA sequence: (see example)

Sequence type: Protein

Reference protein databases:

Green plants ☑ Arabidopsis ☑ Rice ☑ Poplar ☑ Physcomitrella ☑ Sorghum ☑ Selaginella ☑ Chlamydomonas

Plant orthologous level resolution: Angiosperms

Expression pattern (used for output color): Non-changed

Experiment description:

Figure 3. Phylexpress web server. The initial page of Phylexpress web server is shown. The numbers are highlighting the different options of usage and parameter of the algorithm: **1** – One sequence mode, used to identify orthologs of a single sequence; **2** – Batch analyzes, used to indentify orthologs of a list of sequences (max. 100 seqs); **3** – Compare two data sets, used for identify orthologs between to lists of genes; **4** and **5** – Giving name and input sequence in the "one sequence mode"; **6** – Kind of sequence (aa or nt); **7** – Currently available datasets, the user could select or deselect each one by clicking the boxes; **8** – Taxonomy ranking of interest, used to tell the program which orthologs to retrieve; **9** – If there are some information of gene regulation it is possible to inform and a color scheme will be used to drawn the final tree (red for induction and green for repression; optional parameter in "one sequence mode"). The same schemes of colors are used in the "compare two data sets" mode. **10** – Description of the experiment (optional).

It is possible to analyze only one sequence against available databases, or a set of sequences (limited to a maximum of 100 sequences). Another option is available to perform identification of PoGOs between two different sets of sequences uploaded by user. Phylexpress provides to the user all blast results, FASTA files including all sequences used in the analyzes, alignment files in Phylip format, phylogenetic trees in newick and graphic format, and visualizations of the PoGOs identified. To perform high-throughput analyzes (more than 100 sequences) it is possible to download the Phylexpress PERL® script and run it locally.

Evaluation of conserved auxin-responsive patterns between rice and **Arabidopsis**

Microarrays measure the expression of large numbers of genes simultaneously and can be used to delve into interaction networks involving many genes at a time. An important question is to which extent transfer of knowledge about the expression of genes gleaned in one model organism can be transferred to other species. This question is ultimately related to the degree of divergence of transcriptional regulatory network among different lineage. In order to examine such issue, microarray data from equivalent experimental conditions can be compared in the context of orthologs relationship.

We used the orthology-establishment option of Phylexpress to find orthologs between rice and *Arabidopsis* that share similar responsiveness to exogenous auxin (up- or down-regulated). The expression data from microarray experiments of auxin-responsive genes in both species were taken from the literature (Goda *et al*., 2004 and Nemhauser *et al*., 2006 for *Arabidopsis* and Jain & Khurana, 2009 for rice). Microarray data from rice were produced from a pool of plant tissues, including mature and young leaves, plantlets' roots (7-days old) and several panicle and seed developmental stages treated with exogenous auxin during one and three hours. Goda *et al* (2004) used 7-days old WT (Col-0) *Arabidopsis* plantlets treated with exogenous auxin during a kinetics of 24 hours (15 and 30 minutes, 3, 12 and 24 hours) and Nemhauser *et al* (2006) performed a kinetics experiment where WT (Col-0) plantlets were treated with auxin and then collected after 30, 60 and 180 minutes.

Since the experimental conditions between rice and *Arabidopsis* were partly comparable, we decided to investigate the degree of similarity between the responses of these two species to auxin. Rice gene expression data revealed 225 up- and 73 down-regulated genes (Jain & Khurana, 2009), while in *Arabidopsis* 176 up- and 100 down-regulated genes were found by Goda *et al*. (2004) and 430 up- and 356 down-regulated genes were reported by Nemhauser *et al.* (2006). Quite interestingly, among these auxin-responsive genes, only 23 PoGOs including at least one rice and one *Arabidopsis* genes regulated similarly by auxin were identified (Table 1), suggesting a significant divergence between the auxin-related network in *Arabidopsis* and rice. These 23 PoGOs contained 39 *Arabidopsis* and 30 rice auxin-responsive genes (which will be referred to as auxin conserved) some of which are archetypical auxin-responsive genes such as *GH3* IAA synthase, *ARF* TFs, *Aux/IAA* transcription regulators, *SAUR*, the auxin efflux-carrier *PIN1* and the putative TF *LBD* (Paponov *et al*., 2009). Amongst the 39 *Arabidopsis* and 30 rice genes analyzed we found 9 Arabidospsis (25.6%) and 7 rice (23.3%) known TFs which are co-regulated by auxin and constitute six PoGOs (Table 1). This high proportion suggests a statistically significant enrichment of TFs (χ^2=18.28; p<0.005) when compared to the *Arabidopsis* whole genome that has 27.235 protein-coding loci (TAIR, www.arabidopsis.org) and 1.968 TFs (7.23%; Iida *et al*., 2005). Another relevant finding is that for 13 out of the 39 conserved auxin-regulated *Arabidopsis* genes (~33.3%), mutants expressing a clear phenotype are known (Table 2; information on mutants taken from TAIR, www.arabidopsis.org) suggesting that these elements play critical roles in auxin-signaling.

Table 1. Orthologs between *Arabidopsis* and rice sharing auxin-mediated regulation.

PoGO replicability	Average aLRT	Rice genes	*Arabidopsis* genes	IAA responsiveness	*Arabidopsis* gene name (TAIR 9.0)
0.75	1.0	Os01g55940 Os07g40290	AT1G59500 [1] AT2G14960 [1,2] AT2G23170 [1,2] AT4G37390 [2]	Up-regulated	*GH3.4* *GH3.1* *GH3.3* *GH3.2*
1.0	0.92	Os01g71310	AT3G63440 [1]	Up-regulated	*CKX6*
1.0	0.95	Os02g21040	AT3G59080 [1]	Up-regulated	- -
1.0	0.90	Os10g41060	AT4G17350 [1,2] AT4G16670 [2]	Up-regulated	- -
1.0	0.95	Os04g46350 Os10g41230	AT4G17460 [1] AT4G37790 [1] AT5G47370 [1,2]	Up-regulated	*HAT1* *HAT22* *HAT2*
1.0	0.91	Os08g43760	AT4G35200 [1]	Up-regulated	-
1.0	0.97	Os03g12660	AT3G50660 [1]	Up-regulated	*DWF4 - CYP90B1*
1.0	0.88	Os11g04720	AT2G41310 [1]	Down-regulated	*ARR3*
1.0	0.97	Os11g32030	AT3G26760 [1] AT4G03140 [1,2]	Up-regulated	- -
1.0	0.83	Os04g56500	AT2G43060[1,2]	Up-regulated	*ILI1 BINDING BHLH 1*

1.0	0.93	Os02g50960 Os06g12610	AT1G73590 [1]	Up-regulated	*PIN1*
1.0	0.88	Os05g41760	AT1G28370 [1] AT1G53170 [1]	Up-regulated	*ERF11* *ERF8*
1.0	0.95	Os12g06080	AT3G25730 [1]	Up-regulated	*EDF3*
1.0	0.92	Os01g57610 Os05g42150	AT4G27260 [1,2]	Up-regulated	*GH3.5 / WES1*
1.0	0.98	Os04g43910	AT2G28350 [2] AT4G30080 [1,2]	Up-regulated	*ARF10* *ARF16*
1.0	0.86	Os04g44150	AT4G21200 [2]	Up-regulated	*GA2OX8*
1.0	1.0	Os01g29150 Os02g11020 Os06g39880	AT2G26710 [1,2]	Up-regulated	*BAS1*
1.0	0.97	Os07g23570 Os07g44140	AT2G46950 [2]	Up-regulated	*CYP709B2*
1.0	0.95	Os03g43400	AT3G23050 [1,2] AT4G14550 [1]	Up-regulated	*IAA7* *IAA14*
1.0	0.92	Os09g31990	AT4G37240 [1]	Up-regulated*	-
1.0	0.97	Os01g18360	AT3G62100 [1]	Up-regulated*	*IAA30*

1.0	0.83	Os09g37430	AT1G29430 [1]	Up-regulated*	*SAUR*
			AT1G29440 [1]		*SAUR*
			AT1G29450 [1]		*SAUR*
			AT1G29460 [1]		*SAUR*
			AT1G29490 [1]		*SAUR*
			AT1G29500 [1]		*SAUR*
			AT1G29510 [1]		*SAUR*
1.0	0.89	Os02g57490	AT2G42430 [1] [2]	Up-regulated*	*ASL18/LBD16*

[1]Nemhauser *et al.* (2006); [2]Goda *et al.* (2004); *PoGos found in the pairwise comparison of rice x *Arabidopsis* (2x2).

Table 2. Phenotypes of *Arabidopsis* mutants for auxin-responsive genes whose rice orthologs are co-regulated by auxin

Arabidopsis gene IDs	Gene name (TAIR 9.0)	Mutant phenotype
AT3G50660	*DWF4 - CYP90B*	Dwarf phenotype. Nakamoto, et al.(2006)
AT2G41310	*ARR3*	Roots are shorter and have significantly fewer lateral roots than the WT in the absence of exogenous cytokinin; hypersensitivity to cytokinin. Alonso, et al.(2004)
AT1G73590	*PIN1*	Naked inflorescence stem and higher vascular density in leafs. Alonso-Peral, et al.(2006)
AT4G27260	*GH3.5 / WES1*	Longer hypocotyls than WT (Park *et al.* 2007) and hypersensitivity to auxin (ABRC - www.arabidopsis.org)
AT2G28350	*ARF10*	Developmental defects in leaves, flowers and siliques. Yellowish cauline leaves and sepals. Flowers have contorted and elongated petals and produce twisted siliques. Seeds and plants are hypersensitive to ABA in a dose-dependent manner. Liu et al., (2007)
AT4G21200	*GA2OX8*	In short day conditions mutant formed fewer rosette leaves before bolting and a greater number of cauline leaves. Schomburg, et al.(2003)
AT2G26710	*BAS1*	Less responsive to light and early flowering. Increased levels of 6-deoxotyphasterol and cataseterone, which are products of the brassiolide pathway. Turk EM, et al.(2005)
AT3G23050	*IAA7*	Less sensitive to brassinolide than WT. Stronger inhibition of root elongation by brassinolide compared to WT. Nakamura, et al.(2006)
AT4G14550	*IAA14*	Absence of lateral roots. Defective response to auxin stimulus in the root. Fukaki, et al.(2002)
AT3G62100	*IAA30*	Significantly less shoot apical meristem somatic embryogenesis than WT. Zheng, et al.(2009)
AT2G42430	*ASL18/LBD16*	Fewer lateral roots in auxin-treated mutant than in the wild type. Okushima, et al.(2007)
AT2G14960	*GH3.1*	Hypersensitivity to auxin (ABRC - www.arabidopsis.org).
AT4G37390	*GH3.2*	Hypersensitivity to auxin (ABRC - www.arabidopsis.org).

Our analysis also provide evidence that a substantial proportion of the *Arabidopsis* vs rice conserved auxin-responsive orthologs are involved in hormone cross-talk pathways as illustrated by the following examples. Cytokinin oxidase/dehydrogenase enzyme is involved in cytokinin degradation and we found that the Arabidopsis *AtCKX6* gene encoding a cytokinin oxidase/dehydrogenase isophorm as well as its corresponding rice ortholog (*Os01g71310*) are up-regulated by auxin. In addition, the A-type response regulator *ARR3* that modulates sensitivity to cytokinin (To *et al*., 2004; Ren *et al*., 2009) in *Arabidopsis* roots and its rice ortholog (*Os11g04720*) were found to be both down-regulated by auxin. The *Arabidopsis* genes *DWF4/CYP90B* and *BAS1* and their rice orthologs (*Os03g12660* and *Os01g29150, Os02g11020, Os06g39880*, respectively) were shown to be induced by auxin. The *Arabidopsis DWF4/CYP90B1* and *BAS1* genes are involved in brassinosteroid metabolism and signaling. *DWF4/CYP90B1* encodes a 22α hydroxylase whose reaction is a rate-limiting step in brassinosteroid biosynthetic pathway (Nakamoto *et al*., 2006; Fujita *et al*., 2006) and *BAS1* (Turk *et al*., 2005) is a member of the cytochrome p450 family that serves as a control point between multiple photoreceptor systems and brassinosteroid signal transduction. The *Arabidopsis* AUX/IAA transcription regulator *IAA7* and its rice ortholog *Os03g43400* are part of the auxin-induced genes. *IAA7* is probably involved in brassinosteroid signaling pathway since the corresponding null mutant was described as hyposensitive to brassinosteroid (Nakamura *et al*., 2006).

We also identified the *Arabidopsis GA2OX8* gene encoding a gibberellin 2-oxidase which acts specifically on C-20 gibberellins (Schomburg *et al*., 2003) as being induced by auxin as is the case of its rice ortholog *Os04g44150*. Other genes possibly involved in hormone cross-talk are the *Arabidopsis* ARF TF *ARF10* and its rice orthologs *Os04g43910*. *ARF10* is potentially involved in ABA signaling as suggest by the hypersensitivity to ABA during germination and post-germination of its corresponding mutant (Liu *et al*., 2007). Finally, the Arabidopsis ERF/AP2-Type TFs *ERF8*, *ERF11* and *EDF3* which are likely to be important ethylene transduction pathway elements (Riechmann and Meyerowitz, 1998), and their corresponding rice orthologs were found to be auxin-induced in a conserved regulatory scheme of ethylene signaling by auxin. These observations raise the interesting possibility that some aspects of the regulatory pattern of cross-talk nodes between hormones may have been conserved after species divergence emphasizing their functional importance and suggesting the existence of selection processes resulting in the conservation of their regulatory features.

Conclusions

We presented Phylexpress, a bioinformatic tool that allows orthology establishment of single sequences or even whole proteomes and organizes lists of differentially expressed genes from different species into co-responsive groups of orthologs. Phylexpress is available as a web version (sysbiol.cbmeg.unicamp.br/phylo/) and the full program script could be downloaded for high-throughput local analyzes.

Using Phylexpress to compare the RNA-levels responses to exogenous auxin between rice and *Arabidopsis* we observed a significant enrichment of TFs and genes with corresponding mutants leading to obvious phenotype, indicating that comparing signal-mediated RNA profiles alteration in different species in an evolutionary perspective is a way to pinpoint important elements of a transcriptional regulatory network. We propose that regulatory conservation of orthologs reflects functionally important nodes in regulatory networks. This conclusion may be further supported by the assumption that at least part of the functionally important core set of genes within a signaling network would be evolutionary-conserved maintaining ancestral response to a signal along species divergence. On the other hand, species-specific transcriptional responses are in some part responsible for the molecular basis of biological diversity.

Methods

Data collection

The Protein sequences available in Phylexpress web server were download from each genome initiative project: *Arabidopsis thaliana*, version 9.0 – www.arabidopsis.org; *Populus trichocarpa*, version 1.1 – http://genome.jgi-psf.org/poplar/poplar.home.html; *Oryza sativa*, version 5.0 – www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml; *Sorghum bicolor*, version 1.4 – http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html; *Selaginella moellendorffii*, version 1.0 – http://genome.jgi-psf.org/Selmo1/Selmo1.home.html; *Physcomitrella patens patens*, version 1.1 – http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html; *Chlamydomonas reinhardtii*, version 3.0 – http://genome.jgi-psf.org/chlre3/chlre3 .home.html.

Alignment, phylogenetic analysis and orthologs assignment

All alignments are performed by MAFFT 6.0 (Katoh and Toh, 2008) under the 'auto' parameter. The resulting alignments are inputted in the PhyML 3.0 (Guindon and Gascuel, 2003) and a ML tree is produced with and aLRT test of topology. The PhyML paramenters are: ivar 0, model WAG, opt n, and category 4. For the orthologs assignment, we extract a subtree from the original tree that contain at least one sequence from each species of the evolutionary ranking defined previous (Viridiplantae, Embryophyta, Tracheophyta, angiosperms, eudicots and monocots) and had an aLRT value greater than 0.50 in its the ancestral node.

References

Adams KL (2007). Evolution of duplicate gene expression in polyploid and hybrid plants. J Hered 98: 136–141.

Alonso-Blanco C, Aarts MG, Bentsink L, Keurentjes JJ, Reymond M, Vreugdenhil D, Koornneef M. (2009) What has natural variation taught us about plant development, physiology, and adaptation? Plant Cell.7:1877-1896.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A (2010). NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. Nov 21. [Epub ahead of print]

Bennetzen, J (2002) The rice genome. Opening the door to comparative plant biology. Science 296: 60–63.

Blackman BK, Strasburg JL, Raduski AR, Michaels SD, Rieseberg LH (2010) The Role of Recently Derived FT Paralogs in Sunflower Domestication. Current Biology 20, 629–635.

Blanc G and Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16: 1679-691,

Chan YF, Marks ME, Jones FC, Villarreal G Jr, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jónsson B, Schluter D, Bell MA, Kingsley DM. (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327: 302-305.

Chen K and Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8:93-103

Connant GC and Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9: 938-950.

Dehal, P, Boore, JL (2005) Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLoS Biol 3(10): e314.

Doebley JF, GAut BS and Smith BD (2006) The molecular genetics of crop domestication. Cell 127:1309-1321

Fucile G, Falconer S, Christendat D (2008) Evolutionary Diversification of Plant Shikimate Kinase Gene Duplicates. PLoS Genet 4(12): e1000292.

Fujita S, Ohnishi T, Watanabe B, Yokota T, Takatsuto S, Fujioka S, Yoshida S, Sakata K, Mizutani M (2006). Arabidopsis CYP90B1 catalyses the early C-22 hydroxylation of C27, C28 and C29 sterols. Plant J; 45(5):765-74.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology, 52(5):696-704.

Ha M, Li W-H and Chen J (2007) External Factors accelerate expresion divergence between duplicate genes. Trends Genet. 23: 162-166

Han Woo Lee, Nan Young Kim, Dong Ju Lee and Jungmook Kim (2009). LBD18/ASL20 Regulates Lateral Root Formation in Combination with LBD16/ASL18 Downstream of ARF7 and ARF19 in Arabidopsis. Plant Physiology 151:1377-138.

Hittinger CT, Carroll SB (2007) Gene duplication and the adaptive evolution of a classic genetic switch.Nature 449:677-681.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konogaya, A. and Shinozaki, K. (2005). RARTF: Database and Tools for Complete Sets of Arabidopsis Transcription Factors. DNA Research 12(4):247-256

Jackson AP, Thomas GH, Parkhill J and Thomson NR (2009). Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. BMC Genomics 2009, 10:584.

Katoh and Toh (2008). Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics 9:286-298

Katoh K, Kuma K, Toh H, Miyata T (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res, 33:511-518*.

Kellis, M, Birren, BW, Lander, ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428: 617–624.

Koonin, EV (2005). Orthologs, Paralogs, and Evolutionary Genomics. Annual Rev Genet Vol. 39: 309-338.

Koonin, EV (2009). Darwinian evolution in the light of genomics. Nucl. Acids Res. 37 (4): 1011-1034.

Li H and Johnson AD (2010) Evolution of transcription networks – lessons from yeast. Current Biology 20:R746-R753.

Liu PP, Montgomery TA, Fahlgren N, Kasschau KD, Nonogaki H, Carrington JC (2007). Repression of AUXIN RESPONSE FACTOR10 by microRNA160 is critical for seed germination and post-germination stages. Plant J;52(1):133-46.

Lynch M and Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

LynchM and Conery JS (2000) The evolutionary fate and consequences of duplicated genes. Science 90: 1151-1155.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. (2010) Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res 6: 816-825.

Moore RC, Purugganana MD, (2005) The evolutionary dynamics of plant duplicate genes. Current Opinion in Plant Biology 8:122-128

Mustroph, A, Lee, SC, Oosumi, T, Zanetti, ME, Yang, H, Ma, K, Yaghoubi-Masihi, A, Fukao, T and Bailey-Serres, J (2010). Cross-Kingdom Comparison of Transcriptomic Adjustments to Low-Oxygen Stress Highlights Conserved and Plant-Specific Responses. Plant Physiol. 152:1484-1500.

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 461:1130-4.

Nakamoto D, Ikeura A, Asami T, Yamamoto KT (2006). Inhibition of brassinosteroid biosynthesis by either a dwarf4 mutation or a brassinosteroid biosynthesis inhibitor rescues defects in tropic responses of hypocotyls in the arabidopsis mutant nonphototropic hypocotyl 4. Plant Physiol; 141(2):456-64.

Paponov IA, Teale W, Lang D, Paponov M, Reski R, Rensing SA, Palme K (2009). The evolution of nuclear auxin signalling. BMC Evol Biol, Jun 3;9:126.

Paterson AH, Freeling M, Tang H, Wang X (2010) Insights from the Comparison of Plant Genome Sequences. Annu. Rev. Plant Biol. 2010. 61:349–72

Pennacchio, LA (2003) Insights from human/mouse genome comparisons. Mamm Genome 14: 429–436.

Prince VE and Pickett FB (2003) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet 3: 827-837.

Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB (2009) Stepwise Modification of a Modular Enhancer Underlies Adaptation in a Drosophila Population. Science 326: 1663 – 1667.

Reinartz, J, Bruyns, E, Lin, JZ, Burcham, T, Brenner, S, Bowen B, Kramer, M, and Woychik R. (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Briefings in Functional Genomics and Proteomics 1 (1): 95-104.

Ren B, Liang Y, Deng Y, Chen Q, Zhang J, Yang X and Zuo J (2009). Genome-wide comparative analysis of type-A Arabidopsis response regulator genes by overexpression studies reveals their diverse roles and regulatory mechanisms in cytokinin signaling. Cell Research; 19:1178–1190.

Riechmann JL, Meyerowitz EM (1998). The AP2/EREBP family of plant transcription factors. Biol Chem, 379(6):633-46.

Rifkin SA, Kim J, White KP. (2003) Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet. 2003 33:138-44.

Rosin FM, Kramer EM (2009). Old dogs, new tricks: regulatory evolution in conserved genetic modules leads to novel morphologies in plants. Dev Biol. 332:25-35.

Schomburg FM, Bizzell CM, Lee DJ, Zeevaart JA, Amasino RM (2003). Overexpression of a novel class of gibberellin 2-oxidases decreases gibberellin levels and creates dwarf plants. Plant Cell; 15(1):151-63.

Schulze, A and Downward, J (2001). Navigating gene expression using microarrays — a technology review. Nature Cell Biol Vol 3: E190-E195.

Schwartz C, Balasubramanian S, Warthmann N, Michael TP, Lempe J, Sureshkumar S, Kobayashi Y, Maloof JN, Borevitz JO, Chory J, Weigel D. (2009) Cis-regulatory changes at FLOWERING LOCUS T mediate natural variation in flowering responses of Arabidopsis thaliana. Genetics. 183:723-32.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Syntehny and collinearity in Plant Genomes. Science 320: 486-488 DOI: 10.1126/science.1153917

Tirosh I, Reikhav S, Levy AA, Barkai N (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. Science 324659-662.

To JPC, Haberer G, Ferreira FJ, Deruère J, Mason MG, Schaller GE, Alonso JM, Ecker JR, and Kiebera JJ (2004). Type-A Arabidopsis Response Regulators Are Partially Redundant Negative Regulators of Cytokinin Signaling. Plant Cell; 16(3): 658-671.

Turk EM, Fujioka S, Seto H, Shimada Y, Takatsuto S, Yoshida S, Wang H, Torres QI, Ward JM, Murthy G, Zhang J, Walker JC, Neff MM (2005). BAS1 and SOB7 act redundantly to modulate Arabidopsis photomorphogenesis via unique brassinosteroid inactivation mechanisms. The Plant J; 42(1):23-34.

Vandepoele, K, and Van de Peer, Y (2005). Exploring the plant transcriptome through phylogenetic profiling. Plant Physiol. 137: 31–42.

Vincentz, M, Cara, FA, Okura, VK, da Silva, FR, Pedrosa, GL, Hemerly, AS, Capella, AN, Marins, M, Ferreira, PC, França, SC, Grivet, L, Vettore, AL, Kemper, EL, Burnquist, WL, Targon, ML, Siqueira, WJ, Kuramae, EE, Marino, CL, Camargo, LE, Carrer, H, Coutinho, LL, Furlan, LR, Lemos, MV, Nunes, LR, Gomes, SL, Santelli, RV, Goldman, MH, Bacci, M Jr, Giglioti, EA, Thiemann, OH, Silva, FH, Van Sluys, MA, Nobrega, FG, Arruda, P, Menck, CF (2004). Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. Plant Physiol. 134(3):951-959.

Vision, TJ, Brown, DG, Tanksley, SD (2000) The origins of genomic duplications in Arabidopsis. Science 290: 2114–2117.

Wang X, Gowik U, Tang H, Bowers JE, Westhoff P and Paterson AH (2009) Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. Genome Biology 2009, 10:R68 doi:10.1186/gb-2009-10-6-r68

Witkopp PJ (2010) Variable transcription factor binding: a mechanism of evolutionary change. PLoS Biology 8 doi: 10.1371/journal.pbio.1000342.

Wittkopp PJ, Stewart EE, Arnold LL, Neidert AH, Haerum BK, Thompson EM, Akhras S, Smith-Winberry G, Shefner L.(2009). Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila. Science 326:540-544.

Wohlbach DJ, Thompson DA, Gasch AP, Regev A. (2009) From elements to modules: regulatory evolution in Ascomycota fungi.Curr Opin Genet Dev. 19:571-578.

Woodhouse MR, Pedersen B, Freeling M (2010) Transposed Genes in Arabidopsis Are Often Associated with Flanking Repeats. PLoS Genet 6(5): e1000949. doi:10.1371/journal.pgen.1000949

Woolfe, A, Elgar, G (2007). Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. Genome Biol 8: R53.

Capítulo I

Evolução de famílias multigênicas em plantas

Parte 4: Estudo de conteúdo gênico em ESTs públicos de cana-de-açúcar como prova de conceito da utilização do Phylexpress em larga escala

Tropical Plant Biol. (2012) 5:199–205
DOI 10.1007/s12042-012-9103-z

Gene Content Analysis of Sugarcane Public ESTs Reveals Thousands of Missing Coding-Genes and an Unexpected Pool of Grasses Conserved ncRNAs

R. Vicentini · L. E. V. Del Bem · M. A. Van Sluys · F. T. S. Nogueira · M. Vincentz

Received: 12 December 2011 / Accepted: 12 April 2012 / Published online: 15 May 2012

Abstract Sugarcane is the most important crop for sugar industry and raw material for bioethanol. Here we present a quantitative analysis of the gene content from publicly available sugarcane ESTs. The current sugarcane EST collection sampled orthologs for ~58 % of the closely-related sorghum proteome, suggesting that more than 10,000 sugarcane coding-genes remain undiscovered. Moreover the existence of more than 2,000 ncRNAs conserved between sugarcane and sorghum was revealed, among which over 500 are also detected in rice, supporting the existence of hundreds of conserved ncRNAs in grasses. New efforts towards sugarcane transcriptome sequencing were needed to sample the missing coding-genes as well as to expand the catalog of ncRNAs.

Keywords · ncRNAs · Orthology · Sorghum · Sugarcane · Transcriptome

Introduction

Sugarcane (*Saccharum* spp. L., Poaceae) is a C4 sucrose-accumulating grass, the most important crop for sugar industry (Lam et al. 2009) and probably the most successful bioenergy raw material nowadays been widely used in the production of bioethanol in Brazil (Moore 1995; Goldemberg 2006). Modern *Saccharum* hybrids are highly polyploid and aneuploid with chromosome numbers in somatic cells ranging from 100 to 130. This complex genome is derived from a few crosses between the sucrose-accumulating *Saccharum officinarum* L. (2n=8x=80) and the disease-resistant but low sucrose content *S. spontaneum* L. (2n=5x to 12x=40–128). Cultivars are vegetatively propagated and result from selection in populations derived from crosses between outcrossing heterozygous parents (Daniels and Roach 1987; Grivet et al. 2004; Garcia et al. 2006). Current sugarcane cultivars are estimated to possess 80–90 % of the genome from *S. officinarum* and 10–20 % from *S. spontaneum* (Grivet et al. 1996; Hoarau et al. 2002; D'Hont 2005; Piperidis et al. 2010). Sugarcane's basic monoploid genome ranges between 760 Mb and 930 Mb depending

R. Vicentini and L. E. V. Del Bem are first authors.

Communicated by: Robert Henry

Electronic supplementary material The online version of this article (doi:10.1007/s12042-012-9103-z) contains supplementary material, which is available to authorized users.

R. Vicentini
Systems Biology Laboratory, Center for Molecular Biology
and Genetic Engineering, State University of Campinas,
Campinas, SP, Brazil

L. E. V. Del Bem · M. Vincentz
Plant Genetics Laboratory, Center for Molecular Biology
and Genetic Engineering, State University of Campinas,
Campinas, SP, Brazil

F. T. S. Nogueira
Department of Genetics, Institute of Biosciences,
São Paulo State University,
Botucatu, SP, Brazil

M. A. Van Sluys
Genomes and Transposable Elements Laboratory, Department
of Botany, Institute of Biosciences, University of São Paulo,
Rua do Matão, 277,
05508-090, São Paulo, Brazil

R. Vicentini (✉) · L. E. V. Del Bem (✉)
Center for Molecular Biology and Genetic Engineering,
State University of Campinas,
Av. Cândido Rondon, 400,
Campinas, SP, Brazil CEP: 13083-875
e-mail: shinapes@unicamp.br

L. E. V. Del Bem
e-mail: lev.del.bem@gmail.com

on the cultivar breeding history, which represents more than twice the size of rice genome (389 Mb) and is close to sorghum (730 Mb) (D'Hont & Glaszmann 2001). Analyses of haplotype organization suggest that despite the elevated ploidy sugarcane's monoploid genome is highly conserved with sorghum in terms of gene retention and colinearity (Jannoo et al. 2007; Garsmeur et al. 2011). This result makes sorghum the most obvious model choice for sugarcane genomics.

In the last ten years, several sugarcane ESTs collections have been developed (Casu et al. 2001; Carson and Botha 2002; Carson et al. 2002; Casu et al. 2003; Vettore et al. 2003; Ma et al. 2004; Bower et al. 2005; Gupta et al. 2010). The publicly available sugarcane ESTs were assembled into tentative consensus sequences referred to as the Sugarcane Gene Index, mainly composed by sequences from the Brazilian sugarcane EST project (SUCEST; Vettore et al. 2003). The SUCEST project generated 237,954 ESTs, which were organized into 43,141 putative unique sugarcane transcripts referred to as Sugarcane-Assembled Sequences (SASs). These ESTs were used to develop molecular markers such as microsatellite (SSR) and single nucleotide polymorphisms (SNPs), which were successfully used to produce linkage maps and identify QTL for important agronomical traits (Oliveira et al. 2007; Pastina et al. 2010; Somerville et al. 2010). Whether this ESTs collection represents the complete set of sugarcane genes is unclear since around 60 % of the SASs present an average two-fold redundancy with *Arabidopsis* proteome (Menossi et al. 2008). Whether this degree of redundancy could be attributed to the high degree of ploidy/aneuploidy found in sugarcane genome still needs to be further investigated. In order to improve the assessment of sugarcane genes in public ESTs we performed a comparative analysis of the sugarcane ESTs against sorghum and rice genomes.

Our approach uses orthology assignment based on high-throughput amino acids maximum-likelihood (ML) phylogenetic analysis, to identify sugarcane's sorghum and rice possible orthologs along with a complementary nucleotide mapping of sugarcane sequences against sorghum chromosomes. This strategy estimated the sugarcane sampled genes as corresponding to only ~58 % of the predicted sorghum proteome, and uncovers the possibility that more than two thousand putative non-coding RNAs (ncRNAs) are conserved between sugarcane and sorghum, been a quarter possibly shared by rice.

Results and Discussion

Sugarcane EST Collections

All sugarcane ESTs were compiled as the Sugarcane Gene Index database SoGI (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=s_officinarum). The current version (3.0) of SoGI contains 121,342 unique sequences of which only 7,587 singletons and 1,192 tentative consensuses are composed exclusively of ESTs not generated by SUCEST project (~7 %). Although SoGI integrates all published sequences, its clustering strategy produces redundant clusters. This aspect makes the use of a less redundant assemblage strategy, like the one implemented by SUCEST, more appropriate for an orthology-based analysis. A detailed study on sugarcane's *Adh* genes using SNPs (Grivet et al. 2003) suggested that the SUCEST assembly does not merge the paralogous genes into chimeric clusters, which reinforces the reliability of the SASs in a gene-content study where the discrimination between paralogous genes is of critical importance.

In order to obtain a less redundant dataset that includes sequences that were not sampled by SUCEST, we performed a blastn alignment of the 43,141 SASs against the set of 8,779 sequences from SoGI composed exclusively by ESTs not generated by the SUCEST project. This step resulted in a set of 8,106 sequences lacking detectable similarity to SASs (e-value cutoff$<e^{-5}$). We generated a new dataset that integrates the SAS with the latter set leading to 51,247 consolidated clusters (Fig. 1) that will referred to as CSCs (Consolidated Sugarcane Clusters).

Estimation of the Non-Redundant Protein-Coding Gene set in Sugarcane EST Collections

To further contribute with the prediction of the non-redundant set of genes that were sampled by sugarcane ESTs we focused on assigning orthology of the CSC to proteins of sorghum which is the sugarcane closest related grass species whose genome has been sequenced. Since sugarcane and sorghum lineages have diverged recently (5–9 million years ago; Ming et al. 1998; Jannoo et al. 2007) it is reasonable to assume that the monoploid set of genes from sugarcane and sorghum is highly conserved. This assumption is further supported by a recent study that compared sugarcane and sorghum homologous genomic regions (Garsmeur et al. 2011).

To infer orthologous relationships we designed an algorithm that starts with a blastx search of CSC (51,247 sequences) against the sorghum and rice predicted proteomes (Fig. 1a). This initial step produced blastx alignments (e-value cutoff$<e^{-5}$) against sorghum proteome for 38,405 CSC (~75 %). An additional set of 900 CSC (~1.7 %) produced positive hits exclusively against rice proteins. This first step revealed that ~77 % (39,305) of the CSC collection is most likely derived from protein-coding genes (Fig. 1a). The remaining 11,942 CSC were compared by blastx against NCBI's NR database (e-value cutoff$<e^{-5}$) resulting in 2127 positive hits and 9815 no-hits. Among the positive hits 1,826 (~85 %) CSC had a non-Embryophyta (land plants) sequence as best hit. This later set was considered as corresponding to

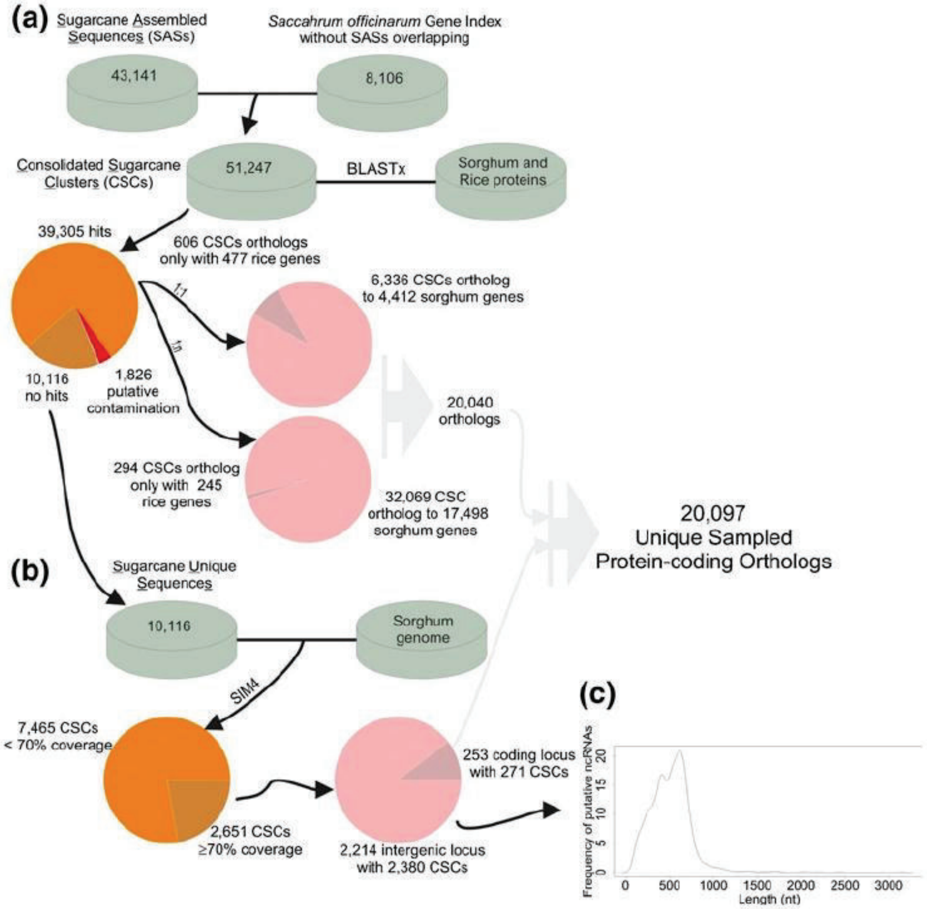

(a)
Sugarcane Assembled Sequences (SASs)
Saccahrum officinarum Gene Index without SASs overlapping
43,141
8,106
Consolidated Sugarcane Clusters (CSCs)
51,247
BLASTx
Sorghum and Rice proteins
39,305 hits
606 CSCs orthologs only with 477 rice genes
1:1
6,336 CSCs ortholog to 4,412 sorghum genes
1:n
10,116 no hits
1,826 putative contamination
20,040 orthologs
294 CSCs ortholog only with 245 rice genes
32,069 CSC ortholog to 17,498 sorghum genes
20,097 Unique Sampled Protein-coding Orthologs
(b)
Sugarcane Unique Sequences
10,116
Sorghum genome
SIM4
(c)
7,465 CSCs < 70% coverage
253 coding locus with 271 CSCs
2,651 CSCs ≥70% coverage
2,214 intergenic locus with 2,380 CSCs
Frequency of putative ncRNAs
0
5
10
15
20
0
500
1000
1500
2000
2500
3000
Length (nt)

Fig. 1 Estimation of the non-redundant gene content of sugarcane ESTs. (**a**) Schematic diagram describing the different filters applied to obtain the list of putative non-redundant sugarcane coding-genes and (**b**) the mapping of sugarcane sequences without detectable protein similarity to the sorghum genome. We identified 20,097 putative non-redundant coding-genes and 2,214 putative non-coding RNAs. (**c**) The sequence size distribution of the ncRNAs shown by frequency plot

non-plant contaminants and removed from subsequent analyses. All the remaining CSC lacking blastx positive hits (10,116; Fig. 1b) were analysed by mapping to sorghum genome and will be further discussed in the next section.

The next step of the algorithm was to separate the CSC that produced a single blastx hit against sorghum or rice proteomes (6,942 or ~17.6 %) from those producing multiple hits (32,363 or 82.4 %). CSC from the first category were directly defined as orthologs to its unique sorghum or rice blastx hit which resulted in the assignment of 6,336 sugarcane CSC to 4,412 sorghum protein-coding orthologs (~43 % redundancy) and 606 sugarcane unigenes lacking sorghum blastx hits which were assigned to 477 rice orthologs (~27 % redundancy) (Fig. 1a).

Each one of the CSC producing multiple blastx hits went through a one-by-one ML phylogenetic analysis along with its 40 first sorghum and rice blastx hits. All phylogenetic analyses were done with amino acid sequences. The CSC were assigned to its closest sorghum or rice ortholog in the resulting phylogenetic trees. Whenever a CSC (4,744 sequences or ~14.6 %) was included in a clade containing multiple sorghum or rice putative paralogs, an additional step was performed in order to define a single ortholog such as to increase the resolution of the analysis (i.e. to limit overestimation). This additional analytic step essentially consisted in producing a distance matrix using WAG plus gamma substitution model among the sequences belonging to such clades and the closest sorghum or rice sequence within the clade was assigned as ortholog to the CSC under analysis. This process allowed us to assign 32,069 CSC to 17,498 sorghum orthologs (~83 % redundancy) and 294 CSC to 245 rice orthologs (~19 % redundancy) (Fig. 1a).

More than half (~53 %) of the CSC that produced a single blastx hit in the first step was assigned to the same sorghum (or rice) orthologs as CSC producing multiple blastx hits. This occurred mainly with CSC containing just a small part of the entire protein that due to our blastx e-value cutoff (e^{-5}) produced just one alignment below the threshold. The final estimation of the coding-gene content of sugarcane public ESTs was obtained by removing the redundancy between the set of orthologs assigned by blastx (single-hit CSC) and the set assigned by ML analysis. Based on the orthology assignment to sorghum and rice, we estimated that the CSC derived from coding-genes (39,305 sequences) sampled 20,040 unique protein-coding orthologs implying ~96 %

of internal redundancy. The total number of sorghum orthologs sampled represents ~58 % of its predicted proteome (34,496 unique coding-genes; http://genome.jgi.doe.gov/Sorbi1/Sorbi1.info.html). This estimation highlights a significant degree of redundancy among the public available sugarcane ESTs collection and points to the necessity of new sequencing efforts.

A Set of Conserved Potential ncRNAs was Revealed by Mapping Sugarcane Unigenes to Sorghum Genome

The 10,116 CSC lacking positive blastx hits against sorghum and rice predicted proteomes were further analyzed by mapping them to sorghum chromosomes using SIM4 (Florea et al. 1998) (Fig. 1b). To limit the number of false positives we applied a filter that recovered the CSC that had a minimum of 70 % of its sequence aligned to the same locus of the sorghum genome (Fig. 1b). Under this criterion 2,651 CSC were retained (7,465 were discarded) and further analysed to define their location relative to the sorghum annotated genes. Among this later set of sequences 271 CSC overlapped with 253 annotated sorghum coding-genes, of which just 56 represented previously unidentified sorghum orthologs (Fig. 1a) and it represents a marginal increment to the first assessment, leading to 20,097 unique sampled protein-coding orthologs (Fig. 1, Table S1). The remaining 2,380 CSC were mapped to 'intergenic' loci within the sorghum genome (Table S2). Any independent CSC overlapping by at least one nucleotide at the same sorghum locus were merged resulting in 2,214 possibly unique ncRNAs loci conserved with sorghum (redundancy of ~7,5 %; Fig. 1b). The size distribution of these non-coding CSCs shows that 54 % of them are longer than 500 pb (Fig. 1c). Furthermore, a blastn search against rice chromosomes (e-value cutoff<e^{-5}) was performed and revealed that 533 out of the 2,214 conserved sugarcane/sorghum ncRNAs (~24 %) were also detected suggesting that at least some of these putative ncRNAs are conserved among grasses and are therefore relevant in grass biology. None of the sugarcane miRNA precursors previously reported (Zanca et al. 2010) were recovered in our analysis due to low sequence alignment with sorghum counterparts.

We found out that ~18 % of the sugarcane/sorghum conserved ncRNA (440 sequences) presented a perfect match with at least one 23-25nt small RNA (sRNA) read from a sugarcane leaf sRNA library (42,218 mapped Illumina® reads out of 2,567,356). When using an arbitrary criterion of at least 15 perfect matched sRNA, only 117 putative sugarcane/sorghum ncRNAs were retained and 63 of them are also detected in rice (Table S2). Whether these putative ncRNAs are the precursors of the perfect matched sRNAs (*cis* action) or they are produced by other loci and act in *trans* remains an open question.

A more detailed analysis of the 13 ncRNA most enriched in perfectly matched sRNAs (i.e., >1,000 sRNAs) revealed a phased distribution of sRNAs (Figure S1). In rice, this kind of pattern was found to derive from miRNAs miR2118- and miR2275-mediated cleavage of a target RNA to define the starting point of the grass-specific Dicer-Like OsDCL3b–mediated production of phased 24-nt sRNAs (Johnson et al. 2009; Song et al. 2011) in a way resembling the biogenesis of the 21-nt *trans*-acting siRNAs (Yoshikawa et al. 2005). The mechanism of biogenesis of the phased 24-nt sRNAs also appears to be conserved in maize (Johnson et al. 2009) and our data suggests it is conserved in sugarcane. The function of these grass-specific 24-nt phased sRNAs is still to be explored.

We compared the whole set of sugarcane putative ncRNAs against the TIGR Plant Repeat Databases (Ouyang and Buell 2004) and only 93 positive hits (~4 %, blastn e-value cutoff < e^{-5}, Table S3) were found. This proportion is near 10 times higher in the sRNA-enriched set of ncRNAs (>15 perfect matched sRNA; 46 out of 117 or ~39 %). Performing the same search for repeats and low complexity sequences using the RepeatMasker software (http://www.repeatmasker.org) we obtained 369 positive hits for the whole set (~15.5 %, Table S4) and 85 among the sRNA-enriched sugarcane ncRNAs (~72 % or ~4.6 times higher). This latter result suggests that the pool of sRNA-enriched ncRNAs is enriched in repetitive and/or transposable element (TE)-derived sequences. Whether the remaining ncRNAs represent new TEs or even Pol IV-transcribed sequences remain to be defined.

Coverage of the Sorghum Exome by Sugarcane ESTs

We have shown in the previous sections that the publicly available sugarcane transcriptome could be linked to 20,097 out of 34,496 (~58 %) sorghum coding-genes. To access the completeness of the sugarcane sequences we mapped all the sugarcane CSC to the sorghum genome using SIM4 and recovered only the best alignment for each CSC. We limited the analysis to the CSC aligned to the sorghum exome that summed 17.654.812 aligned bases. This latter number corresponds to ~40 % of the sorghum exome (48.348.706 nucleotides within 34,496 coding-genes). Normalizing the coverage by the number of sampled sugarcane orthologs (20,097) we found an average coverage of ~63 % relative to the sorghum orthologs.

Conclusions

Our comparative approach leads to the conclusion that the publicly available EST collection for sugarcane accounts with orthologs sampled for at least ~58 % of the predicted sorghum proteome. Significantly, we also found more than two thousand conserved sugarcane/sorghum putative ncRNAs, of which 553 also have some degree of conservation in the rice genome. We were able to show that a subset of these putative ncRNAs has a considerable number of perfect matched 23-25nt sRNAs from a library of sugarcane leaf-expressed sRNAs. Some of these ncRNA may correspond to TEs while the function of most of them remains to be investigated with special attention to their involvement in epigenetic-related processes (Mattick 2001; Mattick 2005; Mattick and Makunin 2006; Mercer et al. 2009; Ben Amor et al. 2009; Matzke et al. 2009; De Lucia and Dean 2011; Zhu and Wang 2012). We also show that the total coverage of the sorghum exome by the sugarcane coding-sequences available up to now is ~40 % and the average coverage of the sampled orthologs is ~63 %. The fact that possibly more than ten thousand sugarcane coding-genes are undiscovered shows the need of new sequencing efforts of sugarcane transcriptome to increase the panel of possible molecular markers and sequence information for sugarcane breeding programs and biotechnological improvement.

Materials and Methods

Public Sequence Datasets

The SASs (Vettore et al. 2003) were obtained from Sugarcane Functional Genomics Database (http://www.sucest-fun.org/) and are available for download, including the option of batch downloading (https://sucest-fun.org/cgi-bin/cane_regnet/sucamet/search_transcript.cgi). The Sugarcane Gene Index sequences were obtained from The Gene Index Project (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=s_officinarum) and the *Sorghum bicolor* (Paterson et al. 2009) and *Oryza sativa* (Yu et al. 2002) complete genomic sequences were downloaded from Phytozome (http://www.phytozome.net/). *Oryza sativa* and *Sorghum bicolor* protein data sets were obtained from the Rice Genome Annotation Project (version 5.0, http://rice.plantbiology.msu.edu) and DOE JGI's *Sorghum bicolor* (version 1.4, http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html), respectively.

Consolidated Sugarcane Clusters Phylogenetic Analyses

Blastx searches using the Consolidated Sugarcane Clusters (CSC; see description in Results and Discussion) as queries were performed against the predicted rice and sorghum proteomes, and the CSC coding-sequences were deduced from the amino acids alignment with the blastx best-match (sorghum or rice). Whenever a CSC was aligned in more than one frame with its blastx best hit (i.e. frameshifts; ~20 % of the protein-coding CSC) we recovered the aligned blocks with e-value below e^{-5} to produce a concatenated sequence keeping the same order of the aligned blocks in relation to its best hit. The average coverage over the sorghum counterparts of these concatenated proteins inferred by this method was 76 %.

The translated CSC were then aligned with the 40 first blastx hits from sorghum and/or rice by MAFFT (Katoh et al. 2005) using default parameters. The phylogenetic relationship of the aligned protein sequences was then inferred by ML using PhyML (Guindon et al. 2010) with WAG plus gamma substitution model and aLTR test.

Estimation of the Non-Redundant Coding-Gene set in Consolidated Sugarcane Clusters

The set of phylogenetic trees generated was analysed by a script that searches for the closest sorghum protein sequence to each of the inputted CSC in a given phylogenetic tree to assign ortologous relationships. Similarly, the CSC that only had blastx hits with rice proteins were assigned to the phylogenetically closest rice sequences. We estimated the redundancy among the CSC by merging different CSC that were assigned as orthologs to the same sorghum or rice protein. A marginal proportion of the whole set of sugarcane coding-genes comes from blastx no-hit sequences that were mapped into a sorghum coding-gene as described below.

Consolidated Sugarcane Clusters Sequences Mapping to Sorghum Genome

We used the SIM4 software (Florea et al. 1998) to align the CSC against sorghum genome. Only CSCs without protein similarity and with >70 % of its total length aligned to a single locus were retained. We removed potential redundancies merging different CSC that overlapped (at least one nucleotide) over their best alignment on the sorghum genome.

Small RNA Library Construction and Bioinformatic Analysis

To evaluate the small RNA landscape of putative sugarcane ncRNAs, we analyzed Illumina® sequences from a small RNA library generated from leaves of 1-month old SP80-3280 sugarcane cultivar plants, grown under greenhouse conditions. Ten micrograms of total RNA, prepared using TRizol reagent (Invitrogen®) according to the manufacturer's instructions, were used to generate a sRNA library following Illumina's modified protocol. The sRNA fraction of 19–28 nt was purified by size fractionation on a 15 % TBE–Urea polyacrylamide gel. A 5`-adenylated single-stranded adapter was first ligated to the 3'-end of the sRNAs using T4 RNA ligase without ATP, followed by a second single-stranded adapter ligation at the 5'-end of the RNA using T4 RNA ligase in the presence of ATP. The resulting products were fractioned on a 10 % TBE–Urea polyacrylamide gel and then used for cDNA synthesis and PCR amplification. The resulting library was sequenced on an Illumina® Genome Analyzer (GA-IIx) following the manufacturer's protocol available at http://www.fasteris.com. Raw sequences were retrieved in a FASTQ formatted file and the adapter sequences were removed using Perl® scripts. After trimming of the adapter sequences, we used the software MAQ (http://maq.sourceforge.net) to map 23–25 nt sRNA reads against the CSC representing the set of putative sugarcane ncRNAs. A total of 42,218 high quality raw sequences of 23 to 25 nucleotides shows perfect match against the sugarcane putative ncRNAs, representing ~1.6 % of the whole sRNA library (Supplemental File 1).

Acknowledgments This work was funded by grants 08/58031-0 (RV) and 08/52071-0 (MV) from FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and LEVDB received a PhD scholarship from FAPESP (2008/09105-1).

References

Ben Amor B, Wirth S, Merchan F, Laporte P, d'Aubenton-Carafa Y, Hirsch J, Maizel A, Mallory A, Lucas A, Deragon JM, Vaucheret H, Thermes C, Crespi M (2009) Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. Genome Res 19:57–69

Bower NI, Casu RE, Maclean DJ, Reverter A, Chapman SC, Manners JM (2005) Transcriptional response of sugarcane roots tomethyl jasmonate. Plant Sci 168:761–772

Carson D, Botha F (2002) Genes expressed in sugarcane maturing internodal tissue. Plant Cell Rep 20:1075–1081

Carson DL, Huckett BI, Botha FC (2002) Sugarcane ESTs differentially expressed in immature and maturing intermodal tissue. Plant Sci 162:289–300

Casu RE, Dimmock CM, Thomas M, Bower N, Knight D (2001) Genetic and expression profiling in sugarcane. Proc Int Soc Sugar Cane Technol 24:542–546

Casu RE, Grof CPL, Rae AL, McIntyre CL, Dimmock CM, Manners JM (2003) Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. Plant Mol Biol 52:371–386

De Lucia F, Dean C (2011) Long non-coding RNAs and chromatin regulation. Curr Opin Plant Biol 14(2):168–173

D'Hont A (2005) Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. Cytogenet Genome Res 109:27–33

D'Hont A, Glaszmann JC (2001) Sugarcane genome analysis with molecular markers, a first decade of research. Proc Int Soc Sugar Cane Technol 24:556–559

Daniels J, Roach BT (1987) Taxonomy and evolution in sugarcane. In: Heinz D (ed) Sugarcane improvement through breeding. Elsevier Press, Amsterdam, pp 7–84

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8:967–974

Garcia AA, Kido EA, Meza AN, Souza HM, Pinto LR, Pastina MM, Leite CS, Silva JA, Ulian EC, Figueira A et al (2006) Development of an integrated genetic map of a sugarcane (*Saccharum* spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. Theor Appl Genet 112:298–314

Garsmeur O, Charron C, Bocs S, Jouffe V, Samain S, Couloux A, Droc G, Zini C, Glaszmann JC, Van Sluys MA et al (2011) High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. New Phytol 189:629–642

Goldemberg J (2006) The ethanol program in Brazil. Environ Res Lett 1:014008

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (*Saccharum* spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142:987–1000

Grivet L, Glaszmann JC, Vincentz M, da Silva F, Arruda P (2003) ESTs as a source for sequence polymorphism discovery in sugarcane: example of the *Adh* genes. Theor Appl Genet 106(2):190–197

Grivet L, Daniels C, Glaszmann JC, D'Hont A (2004) A review of recent molecular genetics evidence for sugarcane evolution and domestication. Ethnobot Res Appl 2:9–17

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321

Gupta V, Raghuvanshi S, Gupta A, Saini N, Gaur A, Khan MS, Gupta RS, Singh J, Duttamajumder SK, Srtivastava S et al (2010) The water-deficit stress- and red-rot-related genes in sugarcane. Funct Integr Genomics 10:207–214

Hoarau JY, Grivet L, Offmann B, Raboin LM, Diorflar JP, Payet J, Hellmann M, D'Hont A, Glaszmann JC (2002) Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). II. Detection of QTLs for yield components. Theor Appl Genet 105:1027–1037

Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann J-C, Arruda P, D'Hont A (2007) Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. Plant J 50:574–585

Johnson C, Kasprzewska A, Tennessen K, Fernandes J, Nan GL, Walbot V, Sundaresan V, Vance V, Bowman LH (2009) Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. Genome Res 19:1429–1440

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518

Lam E, Shine J, da Silva J, Lawton M, Bonos S, Calvino M, Carrer H, Silva-Filho MC, Glynn N, Helsel Z et al (2009) Improving sugarcane for biofuel: engineering for an even better feedstock. Glob Chang Biol Bioenergy 1:251–255

Ma HH, Schulze S, Lee S, Yang M, Mirkov E, Irvine J, Moore P, Paterson A (2004) An EST survey of the sugarcane transcriptome. Theor Appl Genet 108:851–863

Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep 2:986–991

Mattick JS (2005) The functional genomics of noncoding RNA. Science 309:1527–1528

Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15:R17–R29

Matzke M, Kanno T, Daxinger L, Huettel B, Matzke AJM (2009) RNA-mediated chromatin-based silencing in plants. Curr Opin Cell Biol 21(3):367–376

Ming R, Liu SC, Lin YR, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. Genetics 150:1663–1682

Menossi M, Silva-Filho MC, Vincentz M, Van-Sluys M, Souza GM (2008) Sugarcane functional genomics: gene discovery for agronomic trait development. Int J Plant Genomics 2008:458732

Mercer TR, Dinger ME, Mattick JS (2009) Long noncoding RNAs: insights into function. Nat Rev Genet 10:155–159

Moore PH (1995) Temporal and spatial regulation of sucrose accumulation in the sugarcane stem. Aust J Plant Physiol 22:661–679

Oliveira KM, Pinto LR, Marconi TG, Margarido GRA, Pastina MM, Teixeira LHM, Figueira AV, Ulian EC, Garcia AAF, Souza AP (2007) Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum* spp.) commercial cross. Mol Breed 20:189–208

Ouyang S, Buell CR (2004) The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:D360–D363

Pastina MM, Pinto LR, Oliveira KM, Souza KM, Garcia AAF (2010) Molecular mapping of complex traits. In: Henry (ed) Genetics, genomics and breeding of sugarcane. CRC Press, Science Publishers

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Piperidis G, Piperidis N, D'Hont A (2010) Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. Mol Genet Genomics 284:65–73

Somerville C, Youngs H, Taylor C, Davis SC, Long SP (2010) Feedstocks for lignocellulosic biofuels. Science 329:790–792

Song X, Li P, Zhai J, Zhou M, Ma L, Liu B, Jeong DH, Nakano M, Cao S, Liu C, Chu C, Wang XJ, Green PJ, Meyers BC, Cao X (2011) Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J 69:462–474

Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MIs, Henrique-Silva F, Giglioti EA, Lemos MVF, Coutinho LL et al (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res 13:2725–2735

Yoshikawa M, Peragine A, Park MY, Poethig RS (2005) A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes Development. 15; 19(18):2164–2175

Yu J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92

Zanca AS, Vicentini R, Ortiz-Morea FA, Del Bem LEV, da Silva MJ, Vincentz M, Nogueira FTS (2010) Identification and expression analysis of microRNAs and targets in the biofuel crop sugarcane. BMC Plant Biol 10:260

Zhu QH, Wang MB (2012) Molecular functions of long non-coding RNAs in plants. Genes 3(1):176–190

Capítulo II

Evolução de redes de regulação em plantas

Evolução das redes transcricionais de ABA, glicose e sacarose em angiospermas

3. Material e Métodos

3.1 Material vegetal e condições de crescimento

Para os experimentos com arroz utilizamos o cultivar “Formoso” da Embrapa (*Oryza sativa* var. japonica). Para os experimentos com sorgo utilizamos o cultivar ‘rio’ (*Sorghum bicolor* cv. rio). Para o arroz apenas, as sementes foram lixadas com lixa de madeira nº 120 para a retirada da panícula. As sementes foram transferidas para tubos falcon de 15 ml e adicionados 5 ml de etanol 70% e deixadas descansar por 5 min e então o etanol foi descartado e as sementes lavadas com água Milli-Q. Adicionamos 5 ml de Hipoclorito de Sódio 40% e as sementes foram mantidas sob agitação por 20 minutos e então o hipoclorito foi descartado e as sementes lavadas 5 vezes com água Milli-Q. As sementes foram transferidas para placas de petri contendo papel filtro embebido em água para que germinassem. Retiramos o endosperma das plântulas resultantes (~7 dias) e estas foram transferidas para erlenmeyers contendo 20 ml de MS/2 líquido sem açúcar e deixadas sob luz e agitação (50 rpm) constantes por 24 horas antes dos tratamentos. Após este período foram realizados tratamentos com glicose (concentração final de 167mM ou 3%), sacarose (167mM ou 5,7%), manitol (controle osmótico, 167mM ou 3%) e ABA (10μM). Os controles negativos foram tratados com volumes iguais de meio MS/2 sem açúcar. Para cada tratamento realizamos três replicas contendo de 7-10 plântulas.

3.2 Extração de RNA

As plântulas expostas aos tratamentos referidos acima foram imediatamente congeladas em N_2 líquido após 2 horas expostas aos sinais de interesse. O material foi macerado em N2 líquido até que se obtivesse um pó fino. Todos os RNAs totais foram extraídos utilizando o kit ‘RNeasy Mini Kit’ (Qiagen) segundo as recomendações do fabricante. A integridade dos RNAs foi verificada utilizando o equipamento ‘Agilent 2100 BioAnalyzer’.

3.3 Material computacional

O algoritmo Phylexpress foi escrito em PERL® versão 5.10.0 (http://www.perl.org), que consiste numa linguagem orientada a texto e multiplataforma, capaz de rodar em sistemas Windows® e Linux. Nosso algoritmo utilizara para as buscas de seqüências ortólogas por similaridade utilizamos o pacote BLAST 2.18 (http://www.ncbi.nlm.nih.org/BLAST/) e um banco de sequências para gerar as filogenias, chamado Viridiplantae 2.0 de desenvolvimento próprio (aproximadamente 392.000 seqüências únicas), contendo os proteomas preditos para diversos genomas completamente sequenciados do reino Viridiplantae (*Arabidopsis thaliana*, *Populus trichocarpa*, arroz, sorgo, *Selaginella moellendorffii*.*Phyvhomitrella patens patens* e *Chlamydomonnas reinhardtii*). Para os alinhamentos o programa MAFFT v6.717b (http://align.bmr.kyushu-u.ac.jp/mafft/software/). Para a obtenção da topologia filogenética por "Maximum-likelihood" utiliza o programa PhyML 3.0 (http://atgc.lirmm.fr/phyml/).

3.4 Microarranjos de arroz e análise estatística

Utilizamos o GeneChip® Rice Genome Array da Affymetrix® (51K) para os experimentos de expressão gênica em arroz. Para o tratamento estatístico dos dados brutos utilizamos o software R 2.12.1 (http://www.r-project.org/) com a interface gráfica de análise de microarrays affylmGUI (http://www.bioconductor.org/packages/2.0/bioc/html/affylmGUI.html). Para a normalização entre as lâminas utilizamos o método RMA ("*Robust Multichip Average*"). Para comparação entre as populações de diferencialmente expressos utilizamos o software VennMaster 0.37.4 (http://www.informatik.uni-ulm.de/ni/staff/HKestler/vennm/doc.html). Os genes diferencialmente expressos por glicose e sacarose foram encontrados como elementos comuns aos conjuntos de diferencialmente expressos nos contrastes Glicose ou Sacarose x Controle não-tratado e Glicose ou Sacarose x Manitol. Desta forma os genes selecionados precisam aparecer como diferencialmente expressos com relação ao controle não-tratado e no tratamento com manitol. Para o tratamento com ABA obtivemos os diferencialmente expressos através do contraste ABA x controle não-tratado. O corte estatístico utilizado entre as três réplicas biológicas foi de p-valor < 0.005.

3.5 Construção das bibliotecas de mRNA de sorgo para sequenciamento

Os RNAs totais extraídos foram utilizados para construção de bibliotecas para sequenciamento de segunda geração na plataforma Illumina GA®. Utilizamos o kit ‘TruSeq RNA Sample Prep Kits v2’ (Illumina) para as bibliotecas. Resumidamente o mRNA contido nas amostrar originais foi purificado utilizando-se *beads* magnéticas, depois fragmentados e adaptadores para amplificação e sequenciamento foram ligados às extremidades dos fragmentos de mRNA. Os mRNAs oriundos das réplicas de cada experimento foram indexados utilizando os reagentes do referido kit (indexes 2, 4 e 6).

3.6 Sequenciamento das bibliotecas de sorgo

As bibliotecas foram sequenciadas em um equipamento Illumina GA. No total sequenciamos 47 bases para cada *read*, sendo 40 dos fragmentos de mRNA e 7 bases dos indexes para posterior separação dos reads oriundos das triplicatas marcadas por código de base específico (multiplex) via bioinformática.

3.7 Análises de expressão por RNA-Seq de Sorgo

Os reads (40 bp) obtidos para o transcriptoma de cada amostra foram mapeados no genoma de sorgo utilizado como referencia, para isto foi usado o software CLC Genomics Workbench (CLC bio). As sequencias anotadas dos 10 cromossomos de sorgo foram obtidas no GenBank, sendo que os critérios para mapeamento nas regiões gênicas foram: ausência de mismatch e número máximo de hits para o read como sendo 1. A quantificação da expressão gênica foi realizada pelo cálculo do RPKM. Desta forma a expressão de cada gene em cada amostra foi calculada.

A análise estatística foi realizada comparando-se a proporção de reads para cada gene utilizando-se o Kal's Z-test. Após a realização dos testes foram selecionados os genes que apresentavam expressão diferencial, indica pelo p-valor < 0.01 entre os tratamentos e o controle desde que não exista expressão diferencial entre o tratamento e o controle osmótico (manitol). Utilizamos também um threshold de 1-fold change na obtenção dos genes diferencialmente expressos.

4. Resultados e Discussão

4.1 Genes diferencialmente expressos por ABA, glicose e sacarose em sorgo e arroz

Os experimentos foram realizados em triplicata biológica com plântulas de sete dias de arroz e sorgo, em meio líquido e com tratamentos curtos (2hrs) de forma que fossem comparáveis com dados previamente publicados para os mesmos sinais em *Arabidopsis thaliana* (Li *et al*, 2006 para ABA e glicose e Osuna *et al*., 2007 para sacarose). Para arroz utilizamos um *microarray* comercial (GeneChip® Rice Genome Array da Affymetrix® 51K) e para sorgo construímos bibliotecas de mRNA e sequenciamos reads de 40 bases utilizando uma plataforma Illumina GA®. Os resultados de expressão em resposta aos três sinais de interesse foram comparados com um controle negativo não-tratado e os resultados de glicose e sacarose ainda foram comparados com um controle osmótico (manitol). Todos os experimentos foram realizados em triplicatas biológicas e os genes diferencialmente expressos nos *microarrays* de arroz foram selecionados com p-valor < 0,01 e em sorgo arbitramos um *threshold* de 1-*fold change* (expressão de duas vezes a mais ou a menos) e p-valor < 0,01. As intensidades de sinal obtidas nas lâminas de arroz antes e depois da normalização aparecem nas figuras 10 e 11, respectivamente. Uma tabela com o número de *reads* antes e depois do tratamento bioinformático (209 milhões no total, sendo 180 milhões de reads de alta qualidade que foram analisados), bem como o comprimento médio dos reads obtidos pode ser visto na tabela 1. No total obtivemos 1378 (777 induzidos, 601 reprimidos) genes diferencialmente expressos por glicose em arroz e 2190 (1394 induzidos, 796 reprimidos) em sorgo. Em arroz 1271 (870 induzidos, 401 reprimidos) foram diferencialmente expressos por sacarose, sendo 1449 em sorgo (907 induzidos, 542 reprimidos). E os tratamentos com ABA revelaram 1650 (1049 induzidos, 601 reprimidos) genes diferencialmente expressos em arroz e 1003 em sorgo (448 induzidos, 555 reprimidos). Os dados obtidos da literatura para *Arabidopsis thaliana* continham 1752 genes diferencialmente expressos por glicose (983 induzidos, 769 reprimidos), 864 genes diferencialmente expressos por ABA (691 induzidos, 173 reprimidos) e 797 genes diferencialmente expressos por sacarose (368 induzidos, 429 reprimidos).

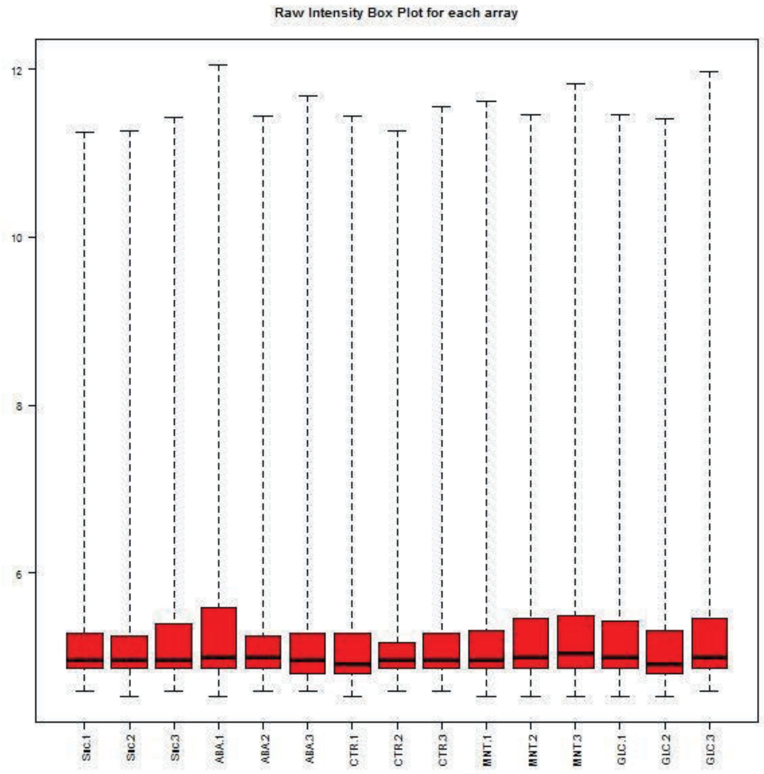


Figura 10. Distribuição da intensidade de sinais das lâminas de array de arroz antes da normalização. São mostradas as três réplicas biológicas para os tratamentos com sacarose (Sac), ABA, controle não-tratado (CTR), manitol (MNT) e glicose (GLC).

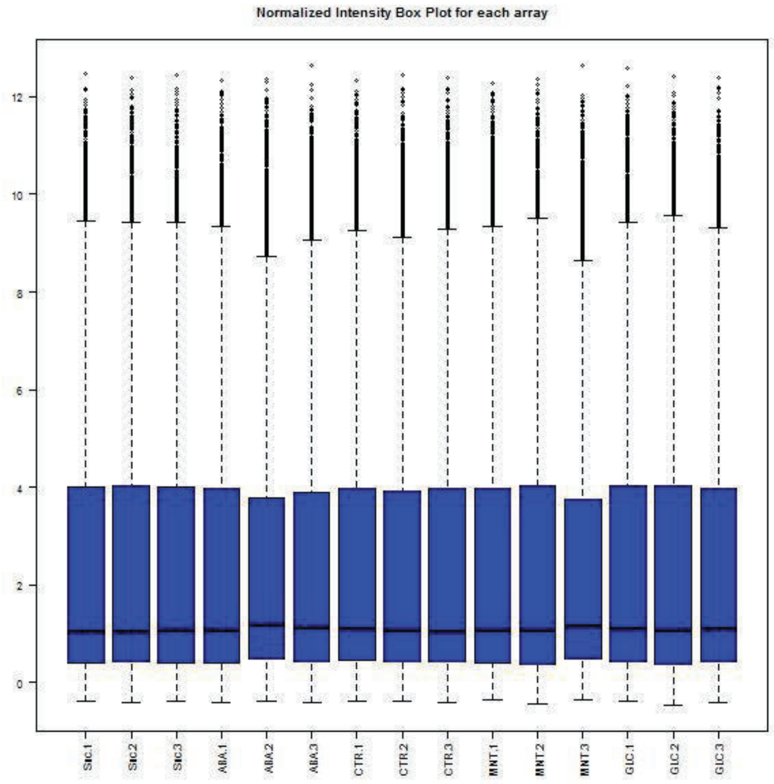
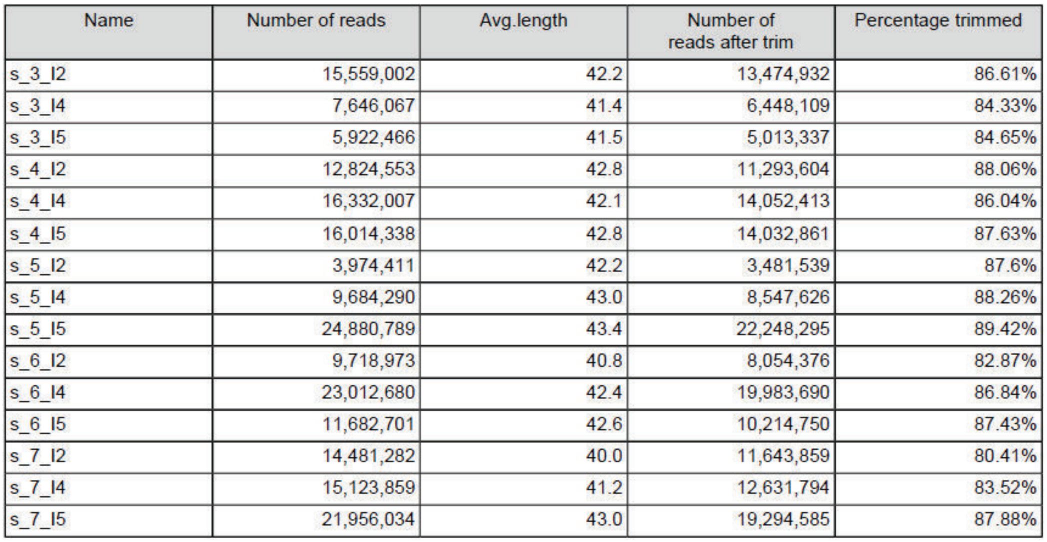


Figura 11. Distribuição da intensidade de sinais das lâminas de array de arroz após a normalização por RMA.



Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed
s_3_I2	15,559,002	42.2	13,474,932	86.61%
s_3_I4	7,646,067	41.4	6,448,109	84.33%
s_3_I5	5,922,466	41.5	5,013,337	84.65%
s_4_I2	12,824,553	42.8	11,293,604	88.06%
s_4_I4	16,332,007	42.1	14,052,413	86.04%
s_4_I5	16,014,338	42.8	14,032,861	87.63%
s_5_I2	3,974,411	42.2	3,481,539	87.6%
s_5_I4	9,684,290	43.0	8,547,626	88.26%
s_5_I5	24,880,789	43.4	22,248,295	89.42%
s_6_I2	9,718,973	40.8	8,054,376	82.87%
s_6_I4	23,012,680	42.4	19,983,690	86.84%
s_6_I5	11,682,701	42.6	10,214,750	87.43%
s_7_I2	14,481,282	40.0	11,643,859	80.41%
s_7_I4	15,123,859	41.2	12,631,794	83.52%
s_7_I5	21,956,034	43.0	19,294,585	87.88%

Tabela 1. Número de reads antes e depois da trimmagem para cada biblioteca. Identificação dos tratamentos: s_3 – ABA, s_4 – Controle não-tratado, s_5 – Glicose, s_6 – Manitol, s_7 – Sacarose. As triplicatas biológicas aparecem identificadas por I2, I4 e I5 para os respectivos tratamentos.

4.2 Análises evolutivas das redes de ABA, glicose e sacarose em angiospermas utilizando grupos de genes ortólogos

O intuito de todos os experimentos apresentados foi obter dados globais de transcrição em resposta aos mesmos sinais entre espécies diferentes de angiospermas. Nossa análise assume como premissas *a priori* que os receptores de cada sinal sejam conservados entre as espécies analisadas, desta forma estaríamos analisando a divergência das populações de genes responsivos dentro da mesma rede, disparada por receptores homólogos.

Utilizamos os grupos de ortólogos do banco de dados de plantas Plaza (http://bioinformatics.psb.ugent.be/plaza/; Van Bel *et al*., 2012; Proost *et al*., 2009) com o intuito de associar os dados obtidos de expressão gênica num contexto de grupos de ortólogos, que são por definição genes de espécies diferentes, que descendem de um gene ancestral comum, separados por um evento de especiação. No próximo tópico mostraremos uma análise em termos de famílias gênicas feito com nosso método Phylexpress. A análise baseada em grupos de ortólogos nos permite ter uma ideia da divergência regulatória entre as espécies utilizadas, sendo a análise focada nas gramíneas sorgo e arroz, tendo *Arabidopsis* como *outgroup* (Figura 12).

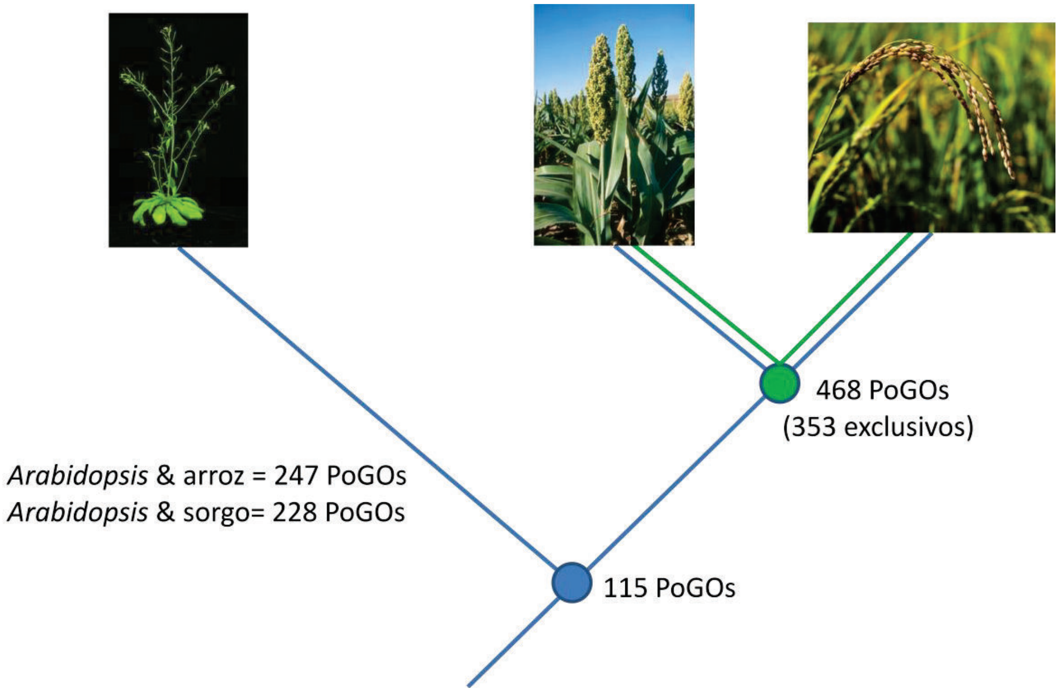

468 PoGOs
(353 exclusivos)
Arabidopsis & arroz = 247 PoGOs
Arabidopsis & sorgo= 228 PoGOs
115 PoGOs

Figura 12. Conservação global da resposta transcricional em angiospermas (redes de ABA, glicose e sacarose). Os números se referem a possíveis grupos de ortólogos onde pelo menos um gene em cada uma das comparações responde no mesmo sentido ao mesmo sinal, o que interpretamos como reflexo da conservação de uma regulação ancestral. Os grupos de ortólogos foram retirados do Plaza (http://bioinformatics.psb.ugent.be/plaza/). O nó azul representa o número de PoGOs com resposta conservada em angiospermas, e o nó verde representa o número para monocotiledôneas. Numericamente a divergência entre *Arabidopsis* e arroz e sorgo não é estatisticamente diferente (χ^2 = 1,46).

Esta análise, que integrou os dados de expressão para todos os sinais, consistiu em encontrar os grupos de ortólogos preditos que possuíam pelo menos um gene em cada par de espécies (ou nas três) que era regulado pelo mesmo sinal no mesmo sentido (induzido ou reprimido). O objetivo era medir a diferença numérica de conservação de redes entre arroz e sorgo (monocot-específico) e arroz, sorgo e *Arabidopsis* (angiospermas). Outro ponto de interesse era avaliar a perda diferencial de genes ancestralmente regulados, que pode ser revelado por grupos de ortólogos com regulação conservada entre *Arabidopsis* (*outgroup*) e arroz, mas que não mais respondem em sorgo, e aqueles grupos de ortólogos que respondem da mesma forma entre *Arabidopsis* e sorgo, mas nenhum gene integrante do grupo responde da mesma forma em arroz. Os dados mostram claramente que o grau de conservação, considerando todos os sinais utilizados, foi ~4 vezes maior entre sorgo e arroz (468 PoGOs, sendo 353 exclusivos) com relação a sorgo, arroz e *Arabidopsis* (115 PoGOs), o que era esperado visto que a divergência entre sorgo e arroz é de ~40 milhões de anos, enquanto a divergência entre ambos e *Arabidopsis* é da ordem de ~130 milhões de anos. O que foi surpreendente em tal análise é que a divergência regulatória entre *Arabidopsis* e arroz (247 PoGOs) foi estatisticamente equivalente à encontrada entre *Arabidopsis* e sorgo (228 PoGOs; χ^2=1,46), o que mostra uma taxa de divergência constante independente da linhagem. Tal divergência constante aponta para uma taxa de perda de regulação (μ_l, ver introdução) constante, o que é similar ao relógio molecular (Kimura, 1983) o que sugere que tal processo possa ser explicado apenas por mutação e deriva genética (como o relógio molecular). Esta conclusão nos impeliu a buscar uma correlação entre a divergência regulatória e a duplicação gênica em famílias multigênicas, pois o processo de subfuncionalização poderia explicar a perda de regulação presente em genes ancestrais.

4.3 Evolução de redes de regulação e duplicação gênica

A aparente taxa constante de perda de regulação gênica poderia ser explicada pelo processo de subfuncionalização (Force *et al*., 2005). A duplicação de um gene ancestral deixaria a nova cópia parcialmente livre da seleção negativa por perda de domínios reguladores no promotor (considerando genes duplicados inteiros, incluindo o promotor). Para abordar tal problema utilizamos nossa ferramenta Phylexpress para uma análise considerando a conservação da regulação versus a fixação de cópias gênicas duplicadas em famílias multigênicas. A análise consistiu em construir árvores filogenéticas utilizando cada um dos genes diferencialmente regulados por nossos sinais de interesse como *query* no método. Em resumo, uma busca por blastp (40 primeiros *hits*) contra sete espécies de plantas (*Arabidopsis*, *Populus*, sorgo, arroz, *Selaginella*, *Physcomitrella* e *Chlamydomonas*) foi feita para cada um dos genes regulados nas espécies de interesse e uma árvore filogenética por *maximum-likelihood* foi construída para cada um deles. Cada árvore foi analisada a partir da folha que continha a *query*, buscando pelos genes mais próximos das outras espécies em análise. Por exemplo, para um gene induzido por ABA em *Arabidopsis* procuramos os genes mais próximos de sorgo e arroz que eram também induzidos por ABA (caso houvesse algum). Os genes de sorgo e arroz, neste exemplo, eram então divididos em categorias conforme o número de duplicações gênicas ocorridas. Se neste exemplo cinco genes de arroz e seis de sorgo foram identificados como sendo induzidos por ABA, eles eram então separados em categorias conforme o grau de conexão na árvore com o gene de *Arabidopsis* para o qual foi realizada a busca (como *query* no blastp) e construída a árvore. O gene mais próximo de cada espécie ao gene de *Arabidopsis* em análise era colocado na categoria 1 (o mais próximo) de sorgo e arroz. O segundo gene mais próximo para cada espécie (sorgo e arroz) eram então colocados na categoria 2 (estas categorias refletem o número de duplicações gênicas fixadas em cada linhagem para a família multigênica em análise), e assim por diante até a categoria 5 em arroz e 6 em sorgo neste exemplo. Este processo foi repetido para todos os genes diferencialmente regulados em nossos experimentos, nas três espécies. Os dados foram separados por sinal e por sentido da resposta (por exemplo, induzidos por glicose, ou reprimidos por sacarose). Uma vez contabilizados todos os genes com regulação conservada, em cada uma das categorias (de 1 a n), para cada um dos pares de espécies (*Arabidopsis* x arroz, *Arabidopsis* x sorgo, sorgo x arroz, sorgo x *Arabidopsis*, arroz x sorgo e arroz x *Arabidopsis*) e sentido de regulação (induzidos ou reprimidos) os números de pares de genes com resposta conservada (em cada categoria) foram dispostos em gráficos (eixo y nos gráficos a seguir) comparados ao número de categorias (número de duplicações gênicas, no eixo x). No total temos seis comparações entre espécies (descritas acima, as comparações são recíprocas) e seis grupos de dados de expressão (glicose, sacarose e ABA, separados por induzidos ou reprimidos). Para uma visão geral somamos todos os números de pares de homólogos com regulação conservada em cada par de espécies (por exemplo, *Arabidopsis* x sorgo) para cada um dos sinais e sentidos de regulação (seis grupos de dados sendo somados). Assim temos uma visão global da frequência em que um dado gene que é regulado por um sinal específico em uma espécie tem homólogos regulados pelo mesmo sinal nas outras espécies, divididos pelo grau de conexão na árvore (número de duplicações gênicas fixadas). Os gráficos individuais, feitos para cada comparação, sinal e sentido de regulação estão apresentados como anexos (por exemplo, *Arabidopsis* vs sorgo, para genes induzidos por ABA). Todos seguem o mesmo padrão, portanto o somatório dos dados para cada par de espécies apresenta também o mesmo padrão (que pode ser visto nas Figuras 13, 14, 15, 16, 17 e 18). Encontramos um padrão de decaimento exponencial (com R^2 em relação a uma exponencial variando entre 0,91 e 0,99) que depende apenas do número de duplicações gênicas fixadas em cada família em cada espécie (Figuras 13, 14, 15, 16, 17 e 18). Em média, cada uma das árvores analisadas apresentava 4,8 genes de *Arabidopsis* (~ grau 5 na escala de duplicação), 5,5 genes de sorgo (entre o grau 5 e 6) e 6,7 em arroz (~ grau 7), ainda que dependendo da família estes números possam ser maiores ou menores. Entendemos que este padrão significa que se um gene ancestral é induzido por um sinal, por exemplo, e este gene começa a se duplicar dando origem a uma família multigênica, a chance de cada uma das cópias filhas manter-se regulada da mesma forma decai exponencialmente conforme a família cresce. A tendência é que genes mais parecidos (em termos de sequência de proteína, que é reflexo da ancestralidade entre os genes) dentro de uma família se comportem de forma exponencialmente mais parecida do que genes mais divergentes. Tal conclusão aponta para uma taxa de decaimento constante da perda da regulação ancestral em famílias multigênicas.

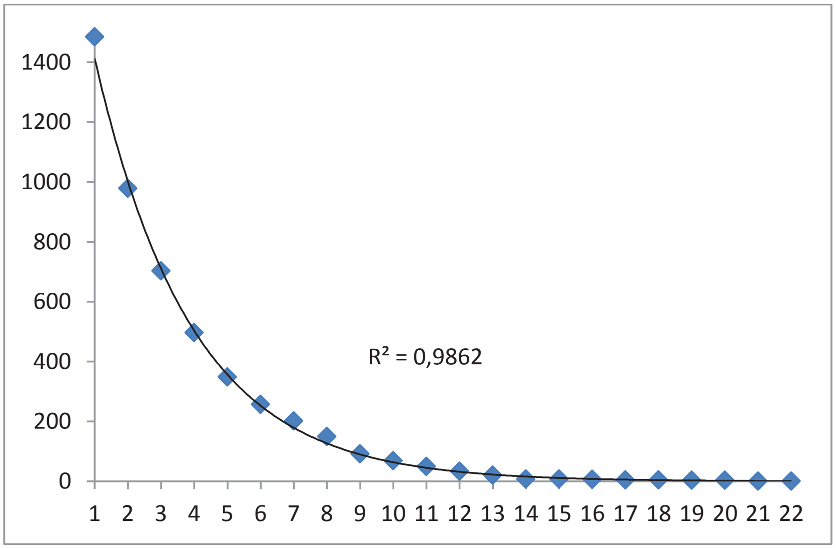


Figura 13. Comparação global entre o número de genes de *Arabidopsis thaliana* que possuem homólogos de arroz que respondem aos mesmos sinais no mesmo sentido e o ranking de proximidade filogenética destes homólogos de arroz em relação aos genes de *Arabidopsis thaliana*. A correlação entre os dados e uma curva de decaimento exponencial é apresentada (R^2).

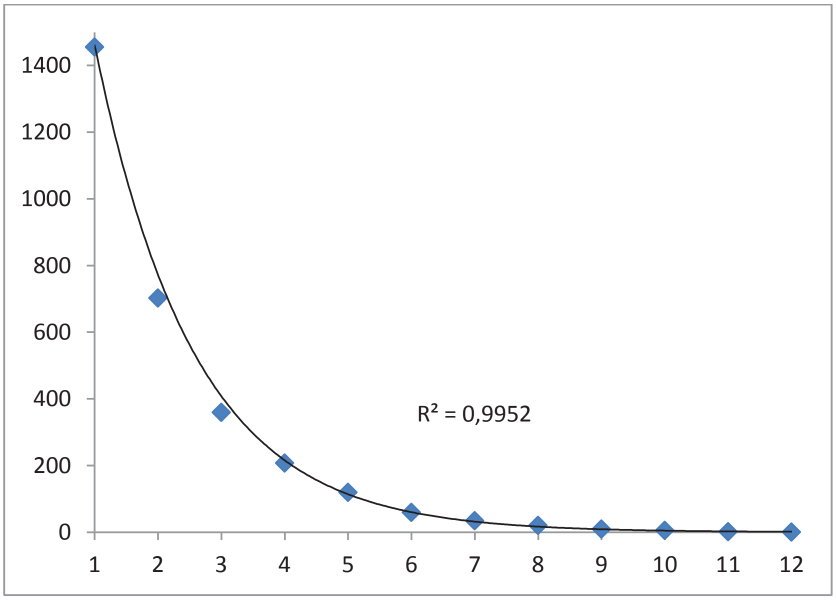


Figura 14. Comparação global entre o número de genes de *Arabidopsis thaliana* que possuem homólogos de sorgo que respondem aos mesmos sinais no mesmo sentido e o ranking de proximidade filogenética destes homólogos de sorgo em relação aos genes de *Arabidopsis thaliana*. A correlação entre os dados e uma curva de decaimento exponencial é apresentada (R^2).

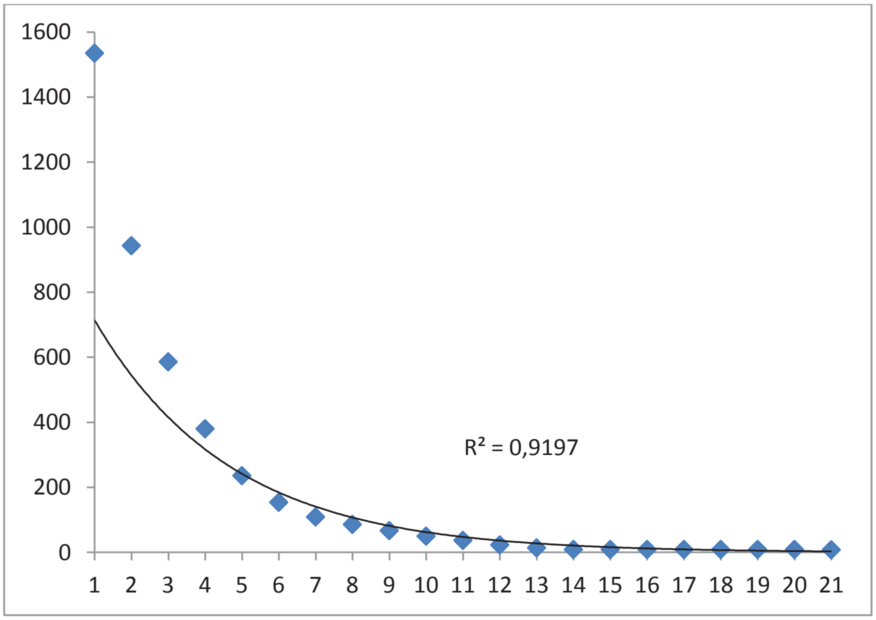


Figura 15. Comparação global entre o número de genes de arroz que possuem homólogos de *Arabidopsis thaliana* que respondem aos mesmos sinais no mesmo sentido e o ranking de proximidade filogenética (número de duplicações) destes homólogos de *Arabidopsis thaliana* em relação aos genes de arroz. A correlação entre os dados e uma curva de decaimento exponencial é apresentada (R^2).

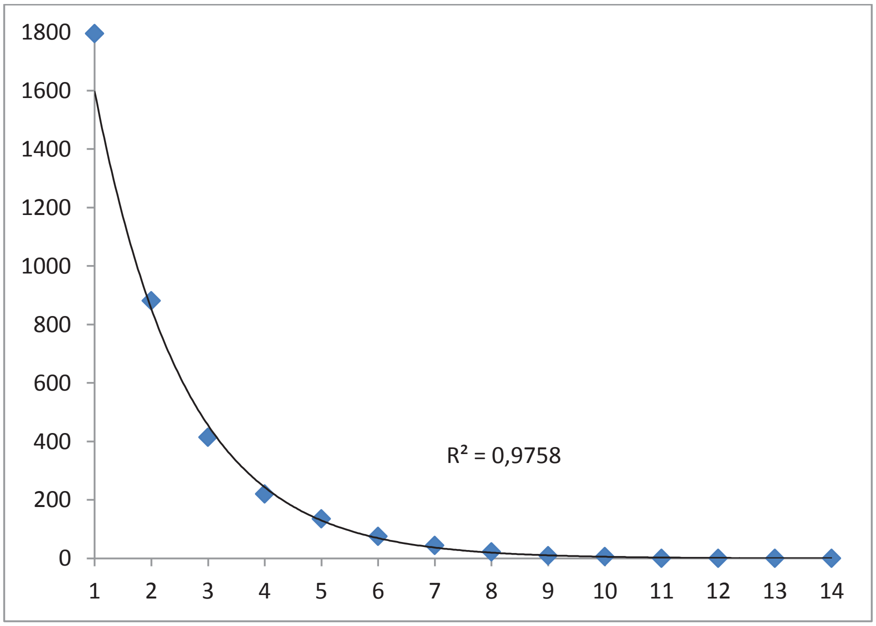


Figura 16. Comparação global entre o número de genes de arroz que possuem homólogos de sorgo que respondem aos mesmos sinais no mesmo sentido e o ranking de proximidade filogenética (número de duplicações) destes homólogos de sorgo em relação aos genes de arroz. A correlação entre os dados e uma curva de decaimento exponencial é apresentada (R²).

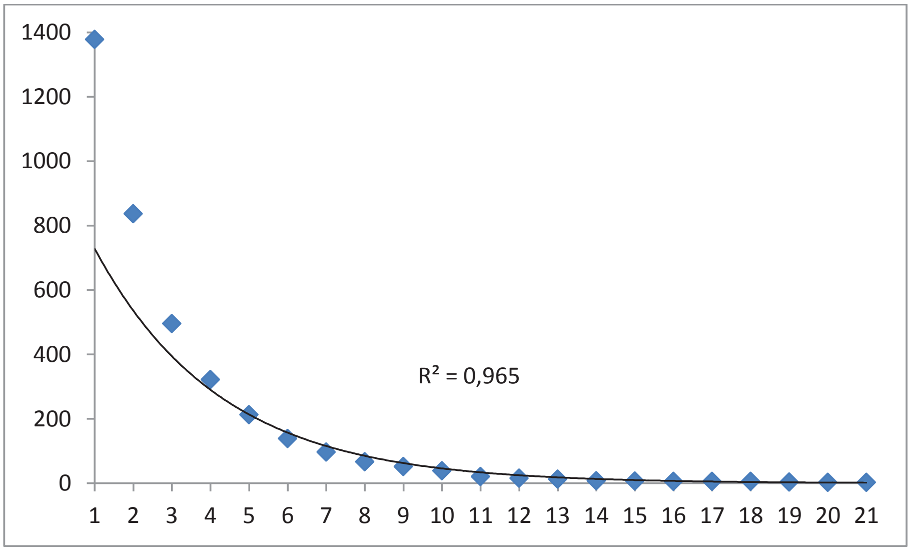


Figura 17. Comparação global entre o número de genes de sorgo que possuem homólogos de *Arabidopsis thaliana* que respondem aos mesmos sinais no mesmo sentido e o ranking de proximidade filogenética (número de duplicações) destes homólogos de *Arabidopsis thaliana* em relação aos genes de sorgo. A correlação entre os dados e uma curva de decaimento exponencial é apresentada (R²).

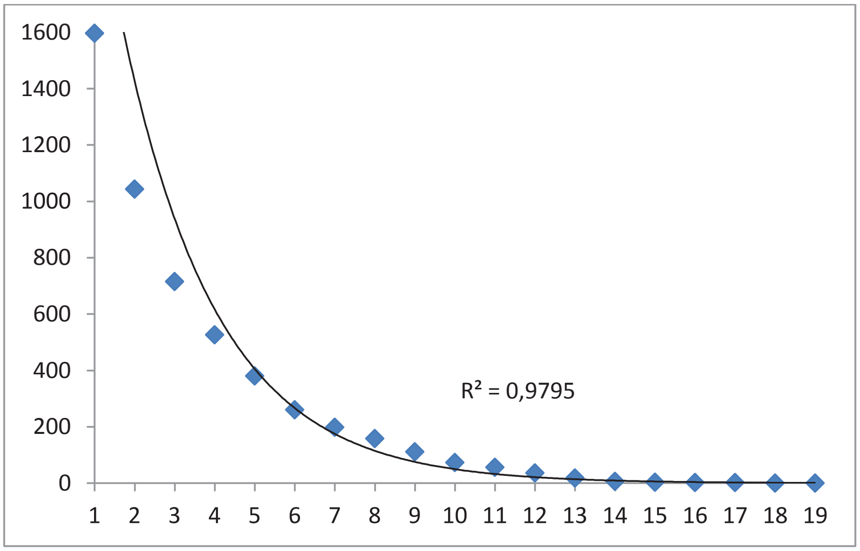


Figura 18. Comparação global entre o número de genes de sorgo que possuem homólogos de arroz que respondem aos mesmos sinais no mesmo sentido e o ranking de proximidade filogenética (número de duplicações) destes homólogos de arroz em relação aos genes de sorgo. A correlação entre os dados e uma curva de decaimento exponencial é apresentada (R^2).

Se a interpretação está correta indica que a expansão de cópias em famílias multigênicas é um fator determinante na diversificação de redes de regulação. Tal conclusão é apoiada por uma das universalidades evolutivas propostas por Koonin, 2011 (baseada nos trabalhos de Van Nimwegen, 2003 e Molina e van Nimwegen, 2009). Genes de categorias funcionais distintas expandem o número de cópias de forma diferente, porém constante entre quaisquer espécies analisadas. Genes de maquinarias celulares basais como elementos da tradução permanecem em número constante independente do número total de genes em cada genoma. O número de enzimas cresce de forma linear conforme o número total de genes aumenta, porém os números de fatores de transcrição e quinases (elementos centrais em redes) crescem de forma quadrática (exponencial) conforme os genomas aumentam em número de genes (via duplicações locais ou de genoma inteiro). Isso significa que os elementos mais críticos em redes regulatórias aumentam exponencialmente conforme genomas complexos (muitos genes) emergem. Esta taxa exponencial de aumento do número de cópias de genes reguladores chave de redes permitiria uma divergência regulatória exponencial, visto que nossos dados sugerem que a diversificação das redes também se comporta segundo uma taxa exponencial. Sendo assim uma rede ancestral regulada por um único fator de transcrição que respondia a glicose, por exemplo, pode, conforme a família expande perder progressivamente a regulação por glicose nas cópias filhas. Segundo a teoria populacional desenvolvida por Lynch (2007a e 2007b) em contextos genômicos eucarióticos como plantas terrestres e vertebrados (maior complexidade da vida na Terra) a chance de um gene ganhar um elemento regulador em seu promotor é maior do que perder um domínio. Desta forma os genes filhos desta família gênica hipotética de fatores de transcrição passariam progressivamente a responder a outros sinais, sendo integrada em outras redes, aumentando a complexidade regulatória das espécies descendentes de forma passiva, aparentemente dependendo muito pouco de processos guiados por seleção positiva. Nossa interpretação dos dados (que é compatível com os modelos apresentados por Lynch e Koonin) revela uma taxa constante de divergência regulatória em redes, o que independe da rede (testamos três sinais diferentes) o que parece novamente (como no caso da análise usando grupos de ortólogos) dependente majoritariamente de deriva genética, mutação (incluindo duplicações gênicas ou duplicações do genoma inteiro) e recombinação.

5.1 Conclusões

Nossos dados apoiam a ideia de que processos não-adaptativos são majoritários na evolução de redes de regulação em eucariotos complexos. A taxa de divergência em grupos de ortólogos sugere que um número equivalente de domínios regulatórios em promotores foi perdido entre *Arabidopsis* e arroz e *Arabidopsis* e sorgo, o que sugere uma taxa de perda constante que independe da história seletiva de cada linhagem. Há uma correlação negativa entre o número de grupos de ortólogos com regulação conservada e a divergência filogenética das espécies, suportado pela conservação quatro vezes maior, em termos numéricos entre arroz e sorgo (divergência de ~40 milhões de anos) em comparação com *Arabidopsis*, arroz e sorgo (divergência de ~130 milhões de anos). Encontramos um gene de arroz e oito genes únicos de sorgo sem similaridade com qualquer outro gene de outra espécie de planta sendo diferencialmente expressos em nossos experimentos, o que demonstra que genes recém-surgidos podem ser integrados a redes ancestrais.

Mais a fundo, nossos resultados sugerem que há uma taxa de divergência regulatória constante em famílias multigênicas em expansão. Dado um gene ancestral integrante de uma rede, a chance dos genes descendentes deste, através do processo de duplicação gênica, ocupar a mesma posição na rede ancestral decai exponencialmente conforme a família gênica aumenta em número de cópias. Este processo parece constante e independente da rede (sinais diferentes) e da divergência entre as espécies analisadas. Novamente os resultados sugerem que a diversificação de redes apresenta taxas constantes que dependem de parâmetros simples, como o número de duplicações gênicas fixadas em cada linhagem ao longo do processo evolutivo. Nossos resultados ajudam a fortalecer a hipótese de que a vida complexa é uma consequência da atenuação da força da seleção positiva em relação à deriva genética.

6. Conclusões gerais

A tese apresentada teve como propósito contribuir no estudo da evolução de famílias multigênicas e redes de regulação gênica em plantas verdes, que ao lado dos animais, representam as formas de vida mais complexas conhecidas. Demonstramos a evolução gradual da maquinaria de síntese e degradação de xiloglucano sugerindo que este polímero se originou numa configuração mais simples antes da conquista do meio terrestre pelas plantas e a subfuncionalização de uma chaperona ancestral que deu origem ao ciclo da calreticulina e calnexina, com duplicações gênicas linhagem-específicas permitindo neofundionalizações únicas deste ciclo em plantas. Desenvolvemos um método de análise filogenética em larga escala, o Phylexpress, que foi utilizado para refinar a estimativa do conteúdo gênico nos ESTs disponíveis de cana-de-açúcar (mostramos que temos apenas 58% de correspondência com o proteoma predito de sorgo) e no estudo de evolução de redes. Realizamos experimentos de expressão gênica (em resposta a ABA, glicose e sacarose) em larga escala em arroz e sorgo com objetivo de estudar a dinâmica da divergência regulatória em organismos complexos. Nossos dados apontam para uma taxa de divergência constante em redes de regulação gênica, o que apoia uma visão emergente sobre a evolução da complexidade. A evolução da complexidade seria permitida pela atenuação da seleção positiva frente ao incremento da deriva genética, devido a condições de declínio no número populacional efetivo (*Ne*), o que geraria a fixação passiva de características complexas (introns, transposons, UTRs, genes com controle transcricional modular e em oposição a operons bacterianos, cromossomos sexuais e redes de regulação gênica com motivos complexos e alta redundância) e pelo incremento recombinatorial que emerge passivamente no surgimento de cromossomos lineares. Em populações com alta eficácia de seleção (*Ne* grande) a complexidade seria contra-selecionada pelo incremento do risco de inativação mutacional de alelos com arquitetura mais complexa. Isto explica o fato de sistemas genéticos simples, como bactérias e Archaea, serem a forma dominante de vida na Terra. Suas populações tem *Ne* milhares de vezes maiores que eucariotos complexos, uma diversidade de espécies incomparável aos seres complexos, uma adaptabilidade única conferida pelo grande *Ne* que maximiza enormemente o surgimento de mutações benéficas, além de ocuparem virtualmente todos os ambientes e nichos ecológicos já explorados na Terra. Frente a este cenário, sistemas biológicos complexos seriam majoritariamente o efeito colateral não-adaptativo de condições populacionais genéticas únicas e raras que atenuam a eficiência da seleção positiva em oposição à visão neodarwinista de que sistemas complexos seriam atingidos pelo acúmulo de caracteres positivamente selecionados à partir de sistemas simples.

7. Perspectivas

Como perspectiva desta tese está a publicação do artigo descrevendo o método Phylexpress, bem como a liberação do uso *online* e aberto da ferramenta. Para tanto pretendemos incluir uma análise de ortologia entre os sete genomas vegetais que compõe nosso banco de dados, que ficou de fora da tese por motivos do alto tempo computacional envolvido.

Outra perspectiva é aprofundar o trabalho sobre evolução de redes de regulação para publicação. Gostaríamos de incluir mais um *outgroup* a análise, o musgo *Physcomitrella patens patens*, o que permitiria ampliar as conclusões. Também pretendemos finalizar uma análise de ortologia, similar à que apresentamos utilizando o Plaza, com nosso método. Esta análise também não foi incluída devido ao tempo computacional envolvido. Outros aspectos relevantes são incluir uma análise de correlação entre a conservação da regulação e a idade filogenética de aparecimento dos genes, bem como analisar a conservação da regulação em genes cópia única versus genes pertencentes a famílias e também a conservação da regulação entre diferentes categorias funcionais (elementos de transdução de sinal, enzimas e maquinarias basais celulares).

8. Bibliografia

1. Lamarck, JB (1809). Philosophie zoologique ou exposition des considérations relatives à l'histoire naturelle des animaux. Premiere edition, L'Imprimerie de Duminil-Lesueur.

2. Darwin, CR & Wallace AR (1858). On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. Journal of the Proceedings of the Linnean Society of London. Zoology 3: 45-50.

3. Mendel, G (1865). Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereines in Brünn. Vol. IV: 3-47.

4. Watson, JD & Crick, FHC (1953). A structure for deoxyribose nucleic acid. Nature (3), 171:737-738.

5. Drake JW (2006). Chaos and order in spontaneous mutation. Genetics; 173:1-8.

6. Miller, RA (2005). Evaluating evidence for aging. Science; 310(5747):441-443.

7. Pray, L (2008). Major molecular events of DNA replication. Nature Education 1(1).

8. Eyre-Walker A & Keightley PD (2007). The distribution of fitness effects of new mutations. Nature Reviews Genetics 8:610-618.

9. Kimura M (1983). The neutral theory of molecular evolution. Cambridge University Press.

10. Alvarez-Buylla ER; Pelaz S, Liljegren SJ, Gold SE, Burgeff, C, Ditta GS, Ribas de Pouplana L, Martinez-Castilla L, Yanofsky MF (2000). An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. PNAS 97(10):5328-5333.

11. Lawton-Rauh A (2003). Evolutionary dynamics of duplicated genes in plants. Mol Phylogenet Evol. 29(3):396-409.

12. Vandepoele K, Simillion C, Van de Peer, Y (2003). Evidence that rice and other cereals are ancient aneuploids. Plant Cell. 15(9):2192-2202.

13. Kellis M, Birren BW, Lander ES (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature (428):617-624.

14. Lespinet O, Wolf YI, Koonin EV, Aravind L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 12(17):1048-1059.

15. Lynch M (2002). Gene duplication and evolution. Science 297(5583):945-947.

16. Wendel JF (2000). Genome evolution in polyploids. Plant Mol Biol 42(1):225-249.

17. Bennetzen J (2002). Opening the door to comparative plant biology. Science (296):60-63.

18. Pennacchio LA (2003). Insights from human/mouse genome comparisons. Mammalian Genome. (14): 429-436.

19. Vincentz M, Cara FA, Okura VK, da Silva FR, Pedrosa GL, Hemerly AS, Capella AN, Marins M, Ferreira PC, Franca SC, Grivet L, Vettore AL, Kemper EL, Burnquist WL, Targon ML, Siqueira WJ, Kuramae EE, Marino CL, Camargo LE, Carrer H, Coutinho LL, Furlan LR, Lemos MV, Nunes LR, Gomes SL, Santelli RV, Goldman MH, Bacci M Jr, Giglioti EA, Thiemann OH, Silva FH, Van Sluys MA, Nobrega FG, Arruda P, Menck CF (2004). Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. Plant Physiol. 134(3): 951-959.

20. Tatusov RL, Koonin EV & Lipman DJ (1997). A genomic perspective on proteins families. Science, 278: 631-637.

21. Thornton J & DeSalle R (2000). Gene family evolution and homology: genomics meets phylogenetics. Annu. Rev. Genomics Hum. Genet., 1:41-73.

22. Fitch WM (2000). Homology: a personal view on some of the problems. TIG 16(5): 227-231.

23. Meyerowitz EM (2002). Plants compared to animals: the broadest comparative study of development. Science (295): 1482 – 1485.

24. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L (1997). Gene families: the taxonomy of protein paralogs and chimeras. Science (278): 609-614.

25. Nei M & Rooney AP (2005). Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121-52.

26. Ingram VM (1961). Gene evolution and the haemoglobins. Nature. 189:704-8.

27. Brown DD, Wensink PC, Jordan E (1972). A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J Mol Biol. 63(1):57-73.

28. Hood L, Campbell JH, Elgin SC (1975). The organization, expression, and evolution of antibody genes and other multigene families. Annu Rev Genet. 9:305-53.

29. Ohta T (1981). Further study on the genetic correlation between members of a multigene family. Genetics. 99(3-4):555-71.

30. Nei, M & Hughes, AL (1992) in 11th Histocompatibility Workshop and Conference, eds. Tsuji, K., Aizawa, M. & Sasazuki, T. (Oxford Univ. Press, Oxford), Vol. 2, pp. 27–38.

31. Hughes AL & Nei M (1990). Evolutionary relationships of class II major-histocompatibility-complex genes in mammals. Mol Biol Evol. 7:491-514.

32. Ota T & Nei M (1994). Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol Biol Evol. 11:469-82.

33. Zhang J, Rosenberg HF, Nei M (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A. 95:3708–3713.

34. Ma H, de Pamphilis C (2000). The ABCs of flower evolution. Cell. 101:5–8

35. Theissen G (2001). Genetics of identity. Nature. 29;414(6863):491.

36. Weigel D & Meyerowitz EM (1994). The ABCs of floral homeotic genes. Cell. 78:203-209

37. Nam J, Kim J, Lee S, An G, Ma H, Nei M. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. PNAS 101(7):1910-1915.

38. Tanabe Y, Hasebe M, Sekimoto H, Nishiyama T, Kitani M, Henschel K, Münster T, Theissen G, Nozaki H, Ito M (2005). Characterization of MADS-box genes in charophycean green algae and its implication for the evolution of MADS-box genes. Proc Natl Acad Sci U S A; 102(7):2436-41.

39. Nei M & Kumar S (2000). Molecular Evolution and Phylogenetics. Oxford University Press.

40. Henning W (1966). Phylogenetic systematics. University of Illinois Press, Urbana.

41. Eck RV & Dayhoff, MO (1967). Atlas of protein sequence and structures. National Biomedical Research Foudation, Silver Springs, MD.

42. Felsenstein J (1978). Cases in which parsimony or compatibility methods will be positively misleading- Syst. Zool. 27:401-410.

43. Hendy M & Penny D (1989). A framework for the quantitative study of evolutionary trees. Syst. Zool. 38(4), 297–309.

44. Zharkikh A & Li WH (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol Biol Evol. 9(6):1119-1147.

45. Takezaki N & M Nei (1996). Genetic Distances and Reconstruction of Phylogenetic Trees From Microsatellite DNA - Genetics. 144 (1) 389-399.

46. Kim CB, Moon SY, Gelder SR, Kim W (1996). Phylogenetic relationships of annelids, molluscs, and arthropods evidenced from molecules and morphology. J Mol Evol. 43(3):207-15.

47. Nei M (1996). Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet. 30:371-403.

48. Cavalli-Sforza LL & Edwards AWF (1967). Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. (19): 233-257.

49. Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. (17): 368-376.

50. Kishino H, Miyata T, Hasegawa M (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. Journal of Molecular Evolution, (31) 151-160.

51. Dayhoff MO, Schwartz RC, Orcutt BC (1978). A model of evolutionary change in proteins. In: Atlas of protein sequence and structure (M.O. Dayhoff, ed.), pp. 301-310. National Biomedical Research Fundation, Silver Spring, MD.

52. Sokal RR & Michener CD (1958). A statistical method for evaluating systematic relationships. Univ. Kansas Sci. v.38:pt.2.

53. Sneath PHA & Sokal RR (1973). Numerical taxonomy. The principles and practice of numerical classification. - A Series of Books in Biology. pp. xv + 573 pp.

54. Nei M, Tajima F, Tateno Y (1983). Accuracy of estimated phylogenetic trees from molecular data. Journal of Molecular Evolution. Vol 19, Issue 2, pp 153-170.

55. Edwards AWF & Cavalli-Sforza LL (1963). The reconstruction of evolution. Heredity. 18:553.

56. Kidd KK & Sgaramella-Zonta LA (1971). Phylogenetic analysis: concepts and methods. Am J Hum Genet; 23(3):235-52.

57. Saitou N & Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. (24): 184-204.

58. Billoud B, De Paepe R, Baulcombe D, and Boccara M (2005). Identification of new small non-coding RNAs from tobacco and Arabidopsis. Biochimie *87*, 905–910.

59. Ben Amor B, Wirth S, Merchan F, Laporte P, d' Aubenton-Carafa Y, Hirsch J, Maizel A, Mallory A, Lucas A, Deragon JM, et al. (2009). Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. Genome Res. *19*, 57–69.

60. McKnight SL & Kingsbury R (1982). Transcriptional control signals of a eukaryotic protein-coding gene. Science *217*, 316–324.

61. Singh KB, Foley RC, Oñate-Sánchez L (2002). Transcription factors in plant defense and stress responses. Current Opinion in Plant Biology *5*, 430–436.

62. Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, Kumar V, Serrano-Mislata A, Madueno F, Krajewski P, Meyerowitz EM, et al. (2010). Orchestration of Floral Initiation by APETALA1. Science *328*, 85–89.

63. Van Der Kelen K, Beyaert R, Inzé D, De Veylder L (2009). Translational control of eukaryotic gene expression. Crit. Rev. Biochem. Mol. Biol. *44*, 143–168.

64. Lee TI & Young RA (2000). Transcription of Eukaryotic Protein-Coding Genes. Annual Review of Genetics *34*, 77–137.

65. Beckett D (2001). Regulated assembly of transcription factors and control of transcription initiation. Journal of Molecular Biology *314*, 335–352.

66. Johnson MA et al (1998) Determinants of mRNA stability in plants. In: J. Bailey-Serres and D.R. Gallie Editors, A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants American Society of Plant Physiologists pag. 40–53.

67. Gutiérrez RA, MacIntosh GC, and Green PJ (1999). Current perspectives on mRNA stability in plants: multiple levels and mechanisms of control. Trends in Plant Science *4*, 429–438.

68. Brodersen P, Sakvarelidze-Achard L, Bruun-Rasmussen M, Dunoyer P, Yamamoto YY, Sieburth L, Voinnet O (2008). Widespread Translational Inhibition by Plant miRNAs and siRNAs. Science *320*, 1185–1190.

69. Forde BG (2002). Local and long-range signaling pathways regulating plant responses to nitrate. Annual Review of Plant Biology. 53:203-224.

70. Rolland F, Baena-Gonzalez E and Sheen J (2006). Sugar sensing and signaling in plants: conserved and novel mechanisms. Annual Review of Plant Biology. 57:675-709.

71. Rook F, Hadingham SA, Li Y, Bevan M (2006). Sugar and ABA response pathways and the control of gene expression.Plant Cell Environ. 29:426-434.

72. Gutiérez RA, Lejay LV, Dean D, Chiaromonte F, Shasha DE and Coruzzi GM (2007). Qualitative network models and genome-wide expression datadefine carbon/nitrogen-responsive molecular machines in *Arabidopsis*. Genome Biology. 8:R7.

73. Koch KE (1996) Carbohydrates modulate gene expression in plants. Annual Review of Plant Physiology and Plant Molecular Biology. 47:509-540.

74. Price J, Laxmi A, Saint Martin S and Jang J-C (2004) Global transcription profiling reveals multiple sugar signal transduction mecanisms in Arabidopsis. The Plant Cell. 16:2128-2150.

75. Li Y, Lee KK, Walsh S, Smith C, Hadingham S, Sorefan K, Cawley G and Bevan MW (2006) Establishing glucose- and ABA-regulated transcription networks in *Arabidopsis* by microarray analysis and promoter classification using a Relevance Vector Machine. Genome Research. 16:414-427.

76. Smeekens S (2000) Sugar-induced signal transduction in plants. Annual Review of Plant Physiology and Plant Molecular Biology. 51:49-81.

77. Rolland F, Moore B and Sheen J (2002) Sugar sensing and signaling in plants. The Plant Cell. Suplement:185-205.

78. Moore B, Zhou L, Rolland F, Hall Q, Cheng W-H, Liu Y-X, Hwang I, Jones T and Sheen J (2003) Role of the Arabidopsis glucose sensor HXK1 in nutrient, light and hormonal signaling. Science. 3000:332-336.

79. Gibson SI (2005) Control of plant development and gene expression by sugar signaling. Current Opinion in Plant Biology. 8:93-102.

80. Wiese A, Elzinga N, Wobbes B and Smeekens S (2004) A conseved upstream open reading frame mediates sucrose-induced repression of translation. The Plant Cell. 16:1717-1729.

81. Hanson J, Hanssen M, Wiese A, Hendriks MMWB, Smeekens S (2007) The sucrose regulated transcription factor bZIP11 affects amino acid metabolism by regulating the expression of *ASPARAGINE SYNTHETASE1* and *PROLINE DEHYDROGENASE2*. Plant J 53: 935-949.

82. Cho Y-H, Yoo S-D and Sheen J (2006) Regulatory functions of nuclear hexokinase1 complex in glucose signaling. Cell. 127:579-589.

83. Xiao W, Sheen J and Jang JC (2000) The role of hexokinase in plant sugar signal transduction and growth and development. Plant Molecular Biology. 44:451-461.

84. Roitsch, T. 1999. Source-sink regulation by sugar and stress. Curr. Opin. Plant Biol. 2: 198–206.

85. Ullah H, Chen JG, Wang S and Jones AM (2002) Role of a heterotrimeric G protein in regulation of *Arabidopsis* seed germination. Plant Physiology. 129:897-907.

86. Chen JG, Willard FS, Huang J, Liang J, Chasse SA, Jones AM and Siderovski DP (2003) A seven-transmembrane RGSprotein that modulates plant cell proliferation. Science. 301:1728–1731.

87. Chen JG and Jones AM (2004) *AtRGS1* function in *Arabidopsis thaliana*. Methods in Enzymology. 389:338-350.

88. Chen Y, Ji F, Xie H, Liang J and Zhang, J (2006) The regulator of G-protein signaling proteins involved in sugar and abscisic acid signaling in Arabidopsis seed germination. Plant Physiology. 140:302–310.

89. Huang J, Taylor JP, Chen JG, Uhrig JF, Schnell DJ, Nakagawa T, Korth L and Jones AM (2006) The plastid protein THYLAKOID FORMATION1 and the plasma membrane G-protein GPA1 interact in a novel sugar-signaling mechanism in Arabidopsis. The Plant Cell. 18:1226-1238.

90. Wang HX, Weerasinghe RR, Perdue TD, Cakmakci NG, Taylor JP, Marzluff WF and Jones AM (2006) A Golgi-localized Hexose Transporter Is Involved in Heterotrimeric G Protein-mediated Early Development in *Arabidopsis.* Molecular Biology of the Cell. 17:4257–4269.

91. Zhou L, Jang J-C, Jones TL and Sheen J (1998) Glucose and ethylene signal transduction crosstalk revealed by an *Arabidopsis* glucose-insensitive mutant. Proceedings of National Academy of Sciences of USA. 95:10294-10299.

92. Arenas-Huertero F, Arroyo A, Zhou L, Sheen J and León P (2000) Analysis of Arabidopsis glucose insensitive mutants, gin5 and gin6, reveals a central role of the plant hormone ABA in the regulation of plant vegetative development by sugar. Genes & Development. 14:2085-2096.

93. Laby RJ, Kincaid MS, Kim D and Gibson SI (2000) The Arabidopsis sugar-insensitive mutants sis4 and sis5 are defective in abscisic acid synthesis and response. The Plant Journal. 23:587-596.

94. Huijser C, Kortstee A, Pego J, Weisbeek P, Wisman E and Smeekens S (2000) The Arabidopsis *SUCROSE UNCOUPLED-6* gene is identical to *ABSCISIC ACID INSENSITIVE-4*: involvement of abscisic acid in sugar responses. The Plant Journal. 23:577-585.

95. Cheng W-H, Endo A, Zhou L, Penney J, Chen H-C, Arroyo A, Leon P, Nambara E, Asami T, Seo M, Koshiba T and Sheen J (2002) A unique short-chain dehydrogenase/reductase in *Arabidopsis* glucose signaling and abscisic acid biosyntesis and functions. The Plant Cell. 14:2723-2743.

96. Finkelstein RR, Wang ML, Lynch TJ, Rao S and Goodman HM (1998) The Arabidopsis abscisic acid response locus ABI4 encodes an APETALA 2 domain protein. The Plant Cell. 10:1043-1054.

97. León P and Sheen J (2003) Sugar and hormone connections. Trends in Plant Science. 8:110-116.

98. Acevedo-Hernández GJ, León P, Herrera-Estrella LR (2005) Sugar and ABA responsiveness of a minimal *RBCS* light-responsive unit is mediated by direct binding of ABI4. Plant J 43 43: 506-519.

99. Rook F, Corke F, Card R, Munz G, Smith C and Bevan MW (2001) Impaired sucrose-induction mutants reveal the modulation of sugar-induced starch biosynthetic gene expression by abscisic acid signalling. The Plant Journal. 26:421-433.

100. Matiolli CC, Tomaz JP, Duarte GT, Prado FM, Del Bem LEV, Silveira AB, Gauer L, Correa LGG, Drumond RD, Viana AJC, Di Mascio P, Meyer C, Vincentz, M (2011). The Arabidopsis bZIP Gene

AtbZIP63 Is a Sensitive Integrator of Transient Abscisic Acid and Glucose Signal Plant Physiology (Bethesda), v. 157, p. 692-705.

101. Finkelstein RR and Lynch TJ (2000) The Arabidopsis abscisic acid response gene *ABI5* encodes a basic leucine zipper transcription factor. The Plant Cell. 12:599-609.

102. Brocard IM, Lynch TJ and Finkelstein RR (2002) Regulation and role of the Arabidopsis *Abscisic Acid-Insensitive 5* gene in abscisic acid, sugar, and stress response. Plant Physiology. 129:1533-1543.

103. Kang JY, Choi HI, Im MY and Kim SY (2002) Arabidopsis basic leucine zipper proteins that mediate stress-responsive abscisic acid signaling. The Plant Cell. 14:343-357.

104. Kim S, Kang JY, Cho DI, Park JH and Kim SY (2004) *ABF2*, an ABRE-binding bZIP factor, is an essential component of glucose signaling and its overexpression affects multiple stress tolerance. The Plant Journal. 40:75-87.

105. Papini-Terzi FS, Rocha FR, Vencio RZN, Felix JM, Branco DS, Waclawovsky AJ, Del Bem LEV, Lembke CG, Costa MDL, Nishiyama MY, Vicentini R, Vincentz MGA, Ulian EC, Menossi M, Souza GM (2009). Sugarcane genes associated with sucrose content. BMC Genomics, v. 10, p. 120

106. Davidson EH (2005). Gene regulatory networks for development. Proc Natl Acad Sci U S A;102(14):4936-42.

107. Karlebach G & Shamir R (2008). Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 9(10):770-80.

108. Tkacik G & Walczak AM (2011). Information transmission in genetic regulatory networks: a review. J. Phys. : Condens. Matter 23 153102

109. Alvarez-Buylla ER, Benítez M, Dávila EB, Chaos A, Espinosa-Soto C, Padilla-Longoria P (2007). Gene regulatory network models for plant development. Current Opinion in Plant Biology 2007, 10:83–91.

110. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004). A Gene Regulatory Network Model for Cell-Fate Determination during Arabidopsis thaliana Flower Development That Is Robust and Recovers Experimental Gene Expression Profiles. Plant Cell. 16(11): 2923–2939.

111. Valliyodan B and Nguyen HT (2006). Understanding regulatory networks and engineering for enhanced drought tolerance in plants. Current Opinions in Plant Biology 9:189-195.

112. Mounet F, Moing A, Garcia V, Petit J, Maucourt M, Deborde C, et al. (2009) Gene and Metabolite Regulatory Network Analysis of Early Developing Fruit Tissues Highlights New Candidate Genes for the Control of Tomato Fruit Composition and Development. Plant Physiology, vol. 149 no. 3 1505-1528.

113. Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M (2012). A Gene Regulatory Network Model for Floral Transition of the Shoot Apex in Maize and Its Dynamic Modeling. PLoS ONE 7(8): e43450. doi:10.1371/journal.pone.0043450

114. Brady SM, Zhang L, Megraw M, Martinez NJ, Jiang E, et al. (2011). A stele-enriched gene regulatory network in the Arabidopsis root. Mol Syst Biol. 7:459.

115. Bruex A, Kainkaryam RM, Wieckowski Y, Kang YH, Bernhardt C, et al. (2012) A Gene Regulatory Network for Root Epidermis Cell Differentiation in Arabidopsis. PLoS Genet 8(1): e1002446. doi:10.1371/journal.pgen.1002446

116. Meng Y, Shao C, Chen M (2011). Toward microRNA-mediated gene regulatory networks in plants. Brief Bioinform. 12(6):645-59.

117. Kim ED, Sung S (2012) Long noncoding RNA: unveiling hidden layer of gene regulatory networks. Trends Plant Sci, 17:16-21.

118. Ishida T, Kurata T, Okada K, Wada T (2008). A genetic regulatory network in the development of trichomes and root hairs. Annu Rev Plant Biol. 2008;59:365-86.

119. Force, A, Cresko WA, Pickett FB, Proulx SR, Amemiya C, Lynch M (2005). The Origin of Subfunctions and Modular Gene Regulation. Genetics. 170(1): 433–446.

120. Pigliucci M (2007). Postgenomic Musings. Science : Vol. 317 no. 5842 pp. 1172-1173

121. Popper K & Eccles JC (1977). The Self and Its Brain: An Argument for Interactionism (Springer, Berlin, 1977)

122. Koonin EV (2011) Are There Laws of Genome Evolution? PLoS Comput Biol 7(8): e1002173. doi:10.1371/journal.pcbi.1002173

123. Grishin NV, Wolf YI, Koonin EV (2000) From complete genomes to measures of substitution rate variability within and between proteins. Genome Res 10: 991–1000.

124. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.

125. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ (2009) The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci U S A 106: 7273–7280.

126. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113.

127. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV (2002) Birth and death of protein domains: A simple model of evolution explains power law behavior. BMC Evol Biol 2: 18.

128. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420: 218–223.

129. Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15: 583–589.

130. Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158: 927–931.

131. Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13: 2229–2235.

132. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102: 14338–14343.

133. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23: 327–337.

134. van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends Genet 19: 479–484.

135. Molina N, van Nimwegen E (2009) Scaling laws in functional genome content across prokaryotic clades and lifestyles. Trends Genet 25: 243–247.

136. Jacob F (1977) Evolution and tinkering. Science 196: 1161–1166.

137. Koonin EV (2011 b) The logic of chance: the nature and origin of biological evolution. Upper Saddle River (NJ): FT Press.

138. Lynch M (2007 a) The frailty of adaptive hypotheses for the origins of organismal complexity. PNAS, vol. 104 no. Suppl 1 8597-8604

139. Lynch M (2007 b) The evolution of genetic networks by non-adaptive processes. Nat Rev Genet. 8(10):803-13.

140. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays 20, 433–440.

141. Lee, T. I. *et al*. Transcriptional regulatory networks in *Saccharomyces cerevisiae*.*Science* 298, 799–804 (2002).

142. Wuchty, S. & Almaas, E. Evolutionary cores of domain co-occurrence networks. *BMC Evol. Biol.* 5, 24 (2005).

143. Lynch M (2007 c). The Origins of Genome Architecture. Sinauer Associates Inc; 1 edition

144. Johnson NA & Porter AH (2000). Rapid Speciation via Parallel, Directional Selection on Regulatory Genetic Pathways. Journal of Theoretical Biology Volume 205, Issue 4, Pages 527–542

145. Haag E & Molla M (2005). Compensatory evolution of interacting gene products through multifunctional intermediates. Evolution, 59, pp. 1620-32.

146. Wright S (1931). Evolution in Mendelian Populations. Genetics; 16(2): 97–159.

147. Wright S (1938). Size of population and breeding structure in relation to evolution. Science 87 (2263): 430–431.

148. Osuna D, Usadel B, Morcuende R, Gibon Y, Bläsing OE, Höhne M, Günter M, Kamlage B, Trethewey R, Scheible WR, Stitt M (2007). Temporal responses of transcripts, enzyme activities and metabolites after adding sucrose to carbon-deprived Arabidopsis seedlings. Plant J. 49(3):463-91.

149. Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K (2012). Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiology 158:590-600

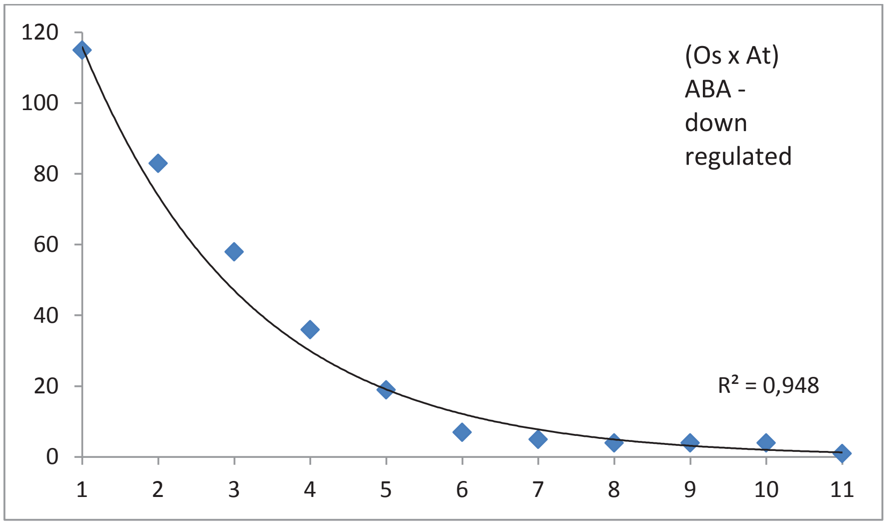
150. Proost S, Van Bel M, Sterk L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K (2009). PLAZA: a comparative genomics resource to study gene and genome evolution in plants. The Plant Cell 21: 3718-3731

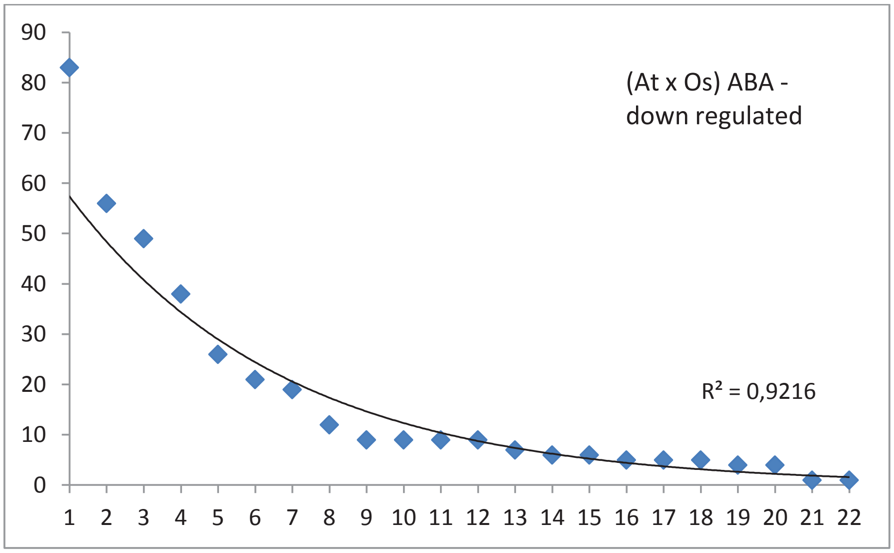
Anexos

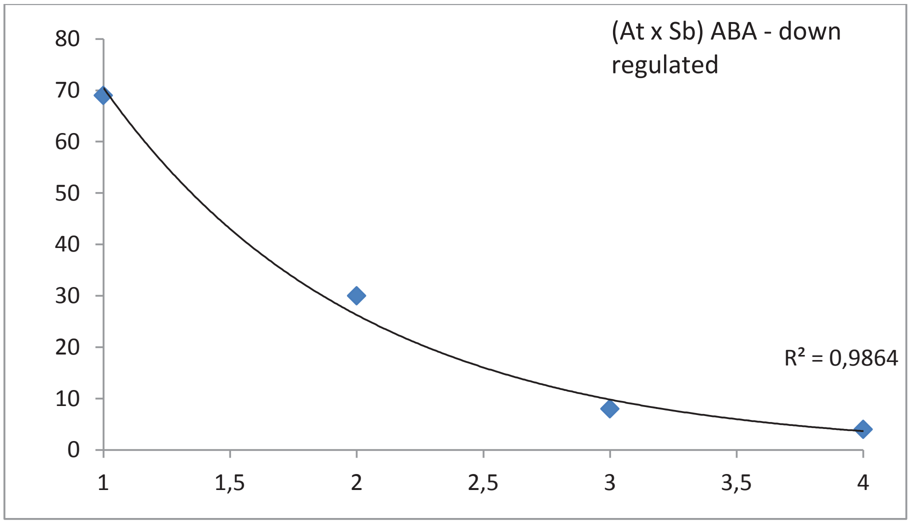
Na seção de anexos apresentamos as distribuições que foram utilizadas para fazer as Figuras 13, 14, 15, 16, 17 e 18 individualizadas. Os gráficos estão identificados pela comparação (por exemplo At x Sb – *Arabidopsis* x sorgo; abreviações: At – *Arabidopsis*, Os – arroz, Sb – sorgo) e pelo sinal (por exemplo ABA – *down regulated*).

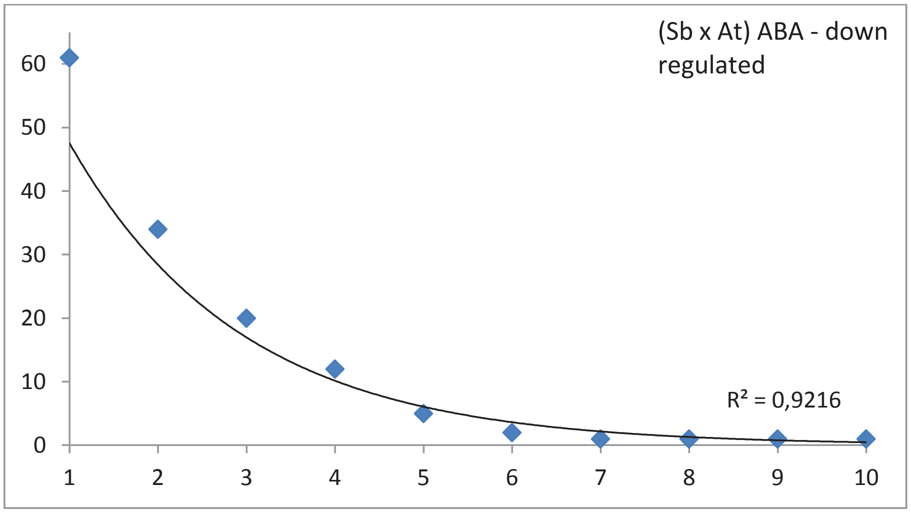
Depois aparecem dois artigos que inspiraram o desenvolvimento do trabalho sobre evolução de xiloglucano (Brandão *et al*., 2008 – Journal of Experimental Botany) e o trabalho que inspirou a análise comparativa da resposta transcricional interespecífica (Papini-Terzi *et al*., 2009 – BMC Genomics).

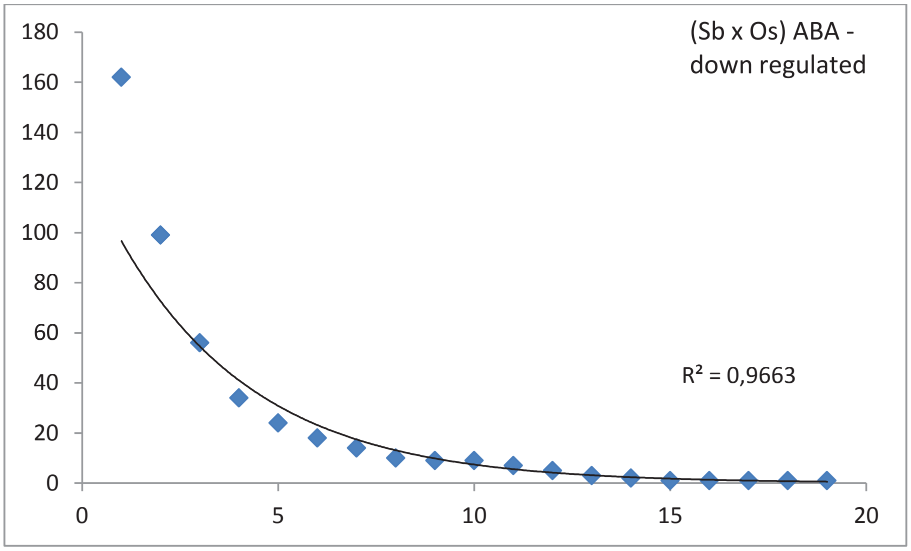
Além destes artigos em anexo e os que foram apresentados no corpo do texto, ao longo deste doutorado também publicamos outros trabalhos, em temas diversos, incluindo **transcriptômica** (Costa, Cardoso, Del Bem et al - Transcriptome analysis of the oil-rich seed of the bioenergy crop *Jatropha curcas* L - BMC Genomics 2010; Cardoso et al - A trancriptomic analysis os gene expression in the venom gland of the snake *Bothrops alternatus* (urutu) - BMC Genomics 2010), **proteômica** (Campos et al - Proteome analysis of castor bean seeds - Pure and Applied Chemistry – 2010), **biologia molecular bacteriana** (Ribeiro et al - The small heat shock proteins from *Acidithiobacillus ferrooxidans*: gene expression, phylogenetic analysis, and structural modeling - BMC Microbiology – 2011), **genética fisiológica vegetal** (Matiolli, Tomaz et al - The *Arabidopsis* bZIP Gene AtbZIP63 Is a Sensitive Integrator of Transient Abscisic Acid and Glucose Signals - Plant Physiology – 2011), **biologia molecular de microRNAs** (Zanca et al - Identification and expression analysis of microRNAs and targets in the biofuel crop sugarcane - BMC Plant Biology 2010) e **paleobotânica** (Faria et al - *Lycopodiopsis derbyi* Renault from the Corumbataí Formation in the state of São Paulo (Guadalupian of Paraná Basin, Southern Brazil): New data from compressed silicified stems - Review of Palaeobotany and Palynology – 2009).

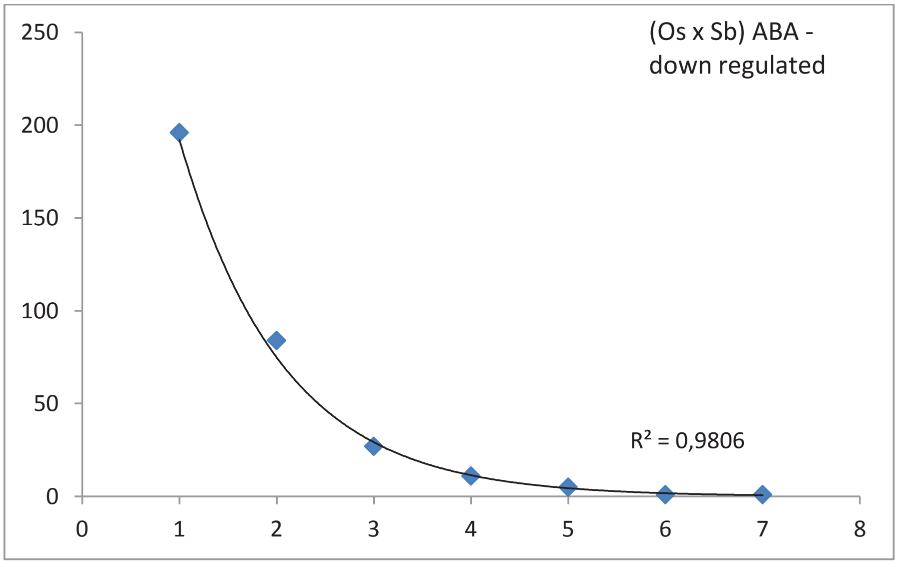

(Os x At) ABA - down regulated
R² = 0,948
0
20
40
60
80
100
120
1
2
3
4
5
6
7
8
9
10
11

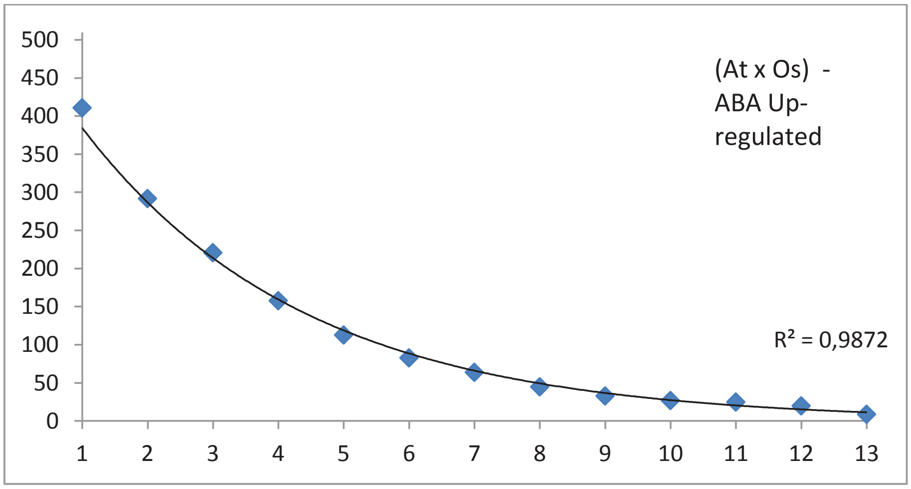

(At x Os) ABA - down regulated
R² = 0,9216
0
10
20
30
40
50
60
70
80
90
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

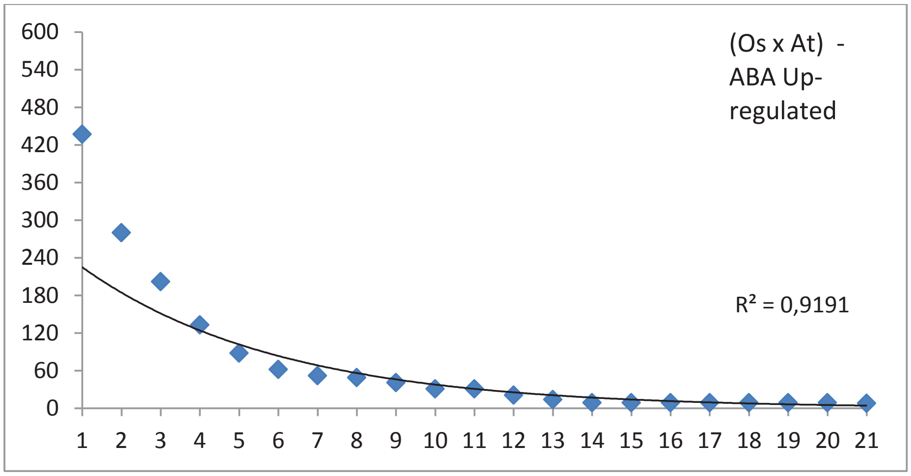

(At x Sb) ABA - down regulated
R² = 0,9864
0
10
20
30
40
50
60
70
80
1
1,5
2
2,5
3
3,5
4

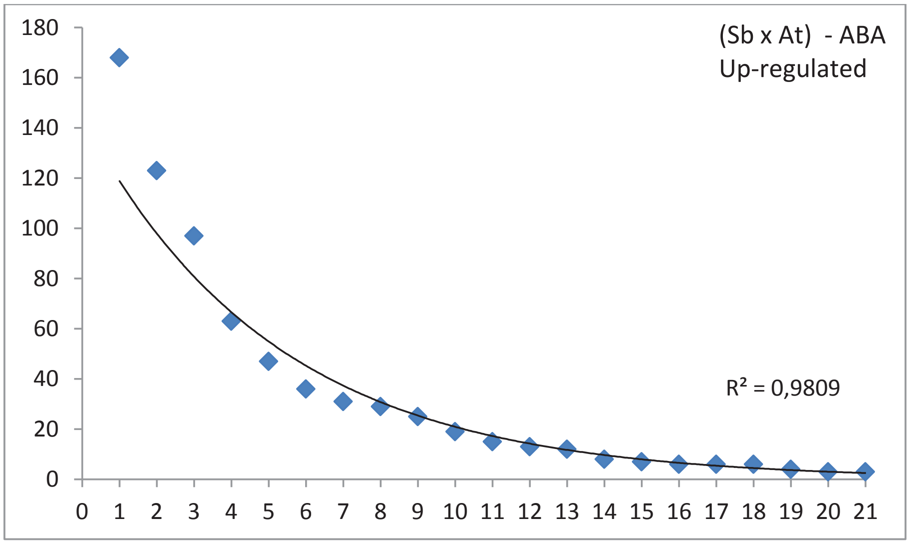

(Sb x At) ABA - down regulated
R² = 0,9216
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10

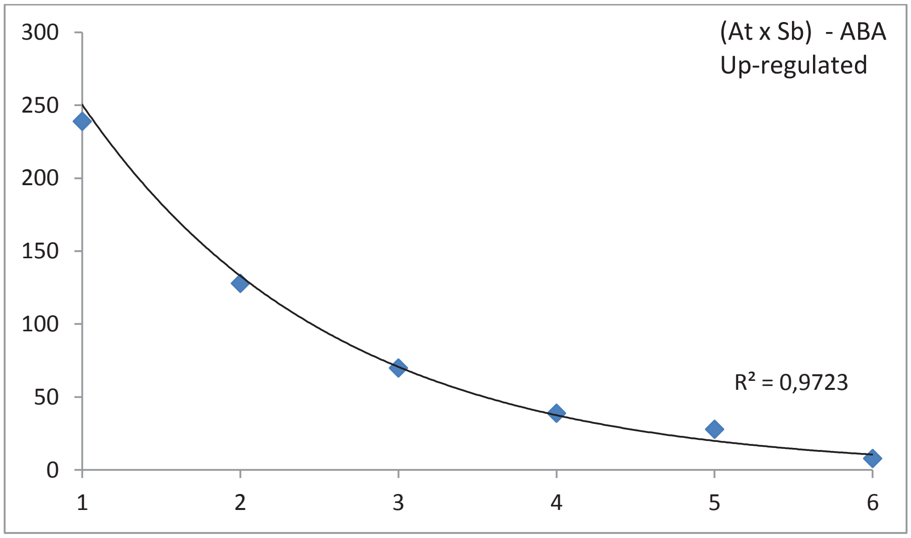

(Sb x Os) ABA - down regulated
R² = 0,9663
180
160
140
120
100
80
60
40
20
0
0
5
10
15
20

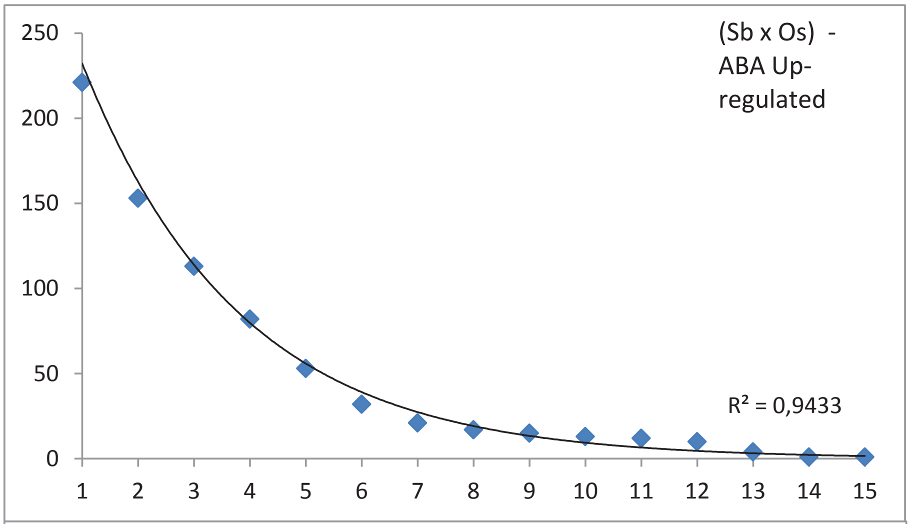

(Os x Sb) ABA - down regulated
R² = 0,9806
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8

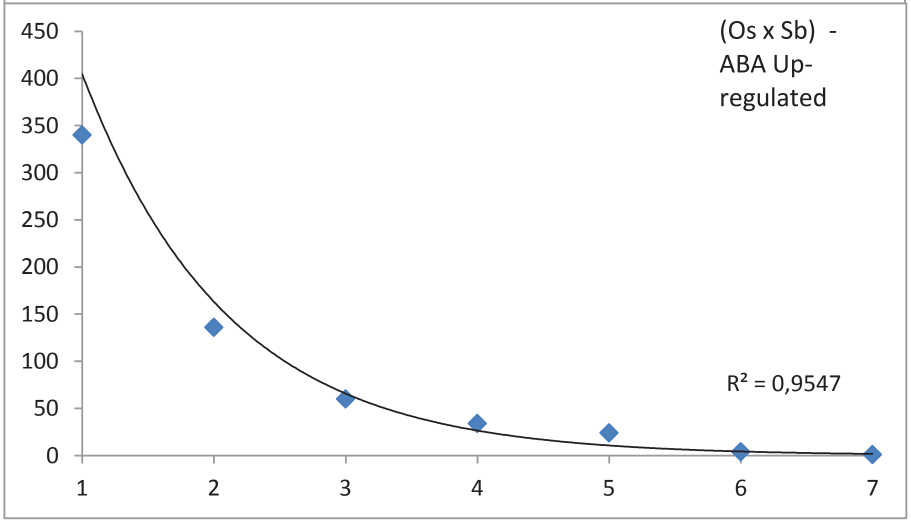

500
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
(At x Os) - ABA Up-regulated
R² = 0,9872

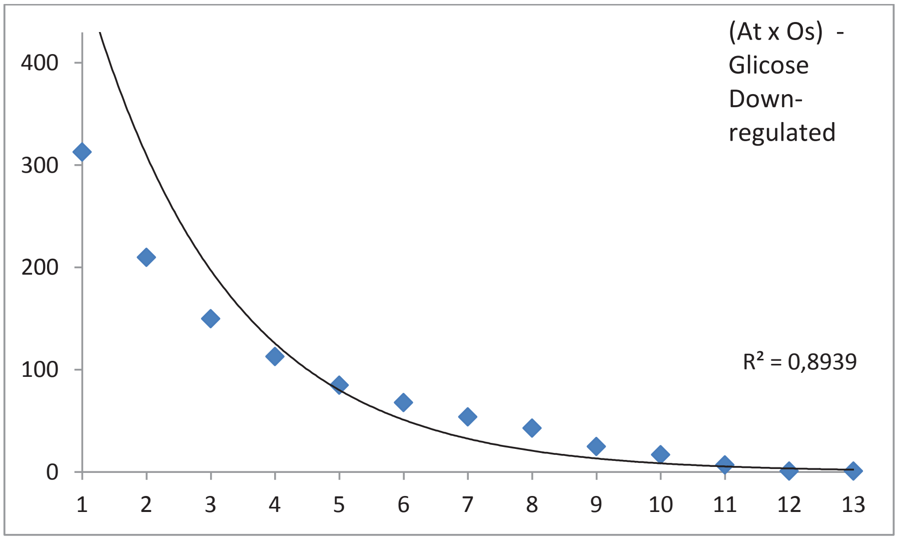

600
540
480
420
360
300
240
180
120
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
(Os x At) - ABA Up-regulated
R² = 0,9191

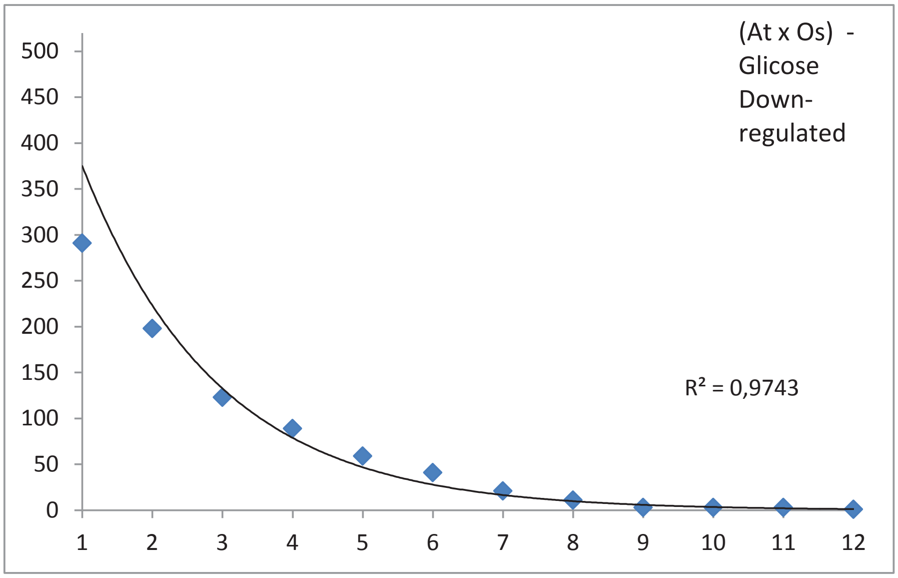

180
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
(Sb x At) - ABA Up-regulated
R² = 0,9809

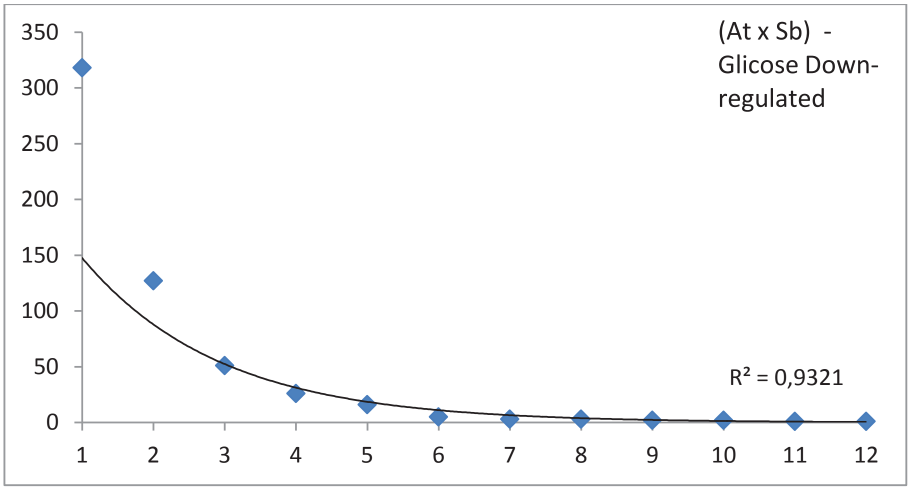

(At x Sb) - ABA Up-regulated
R² = 0,9723
300
250
200
150
100
50
0
1
2
3
4
5
6

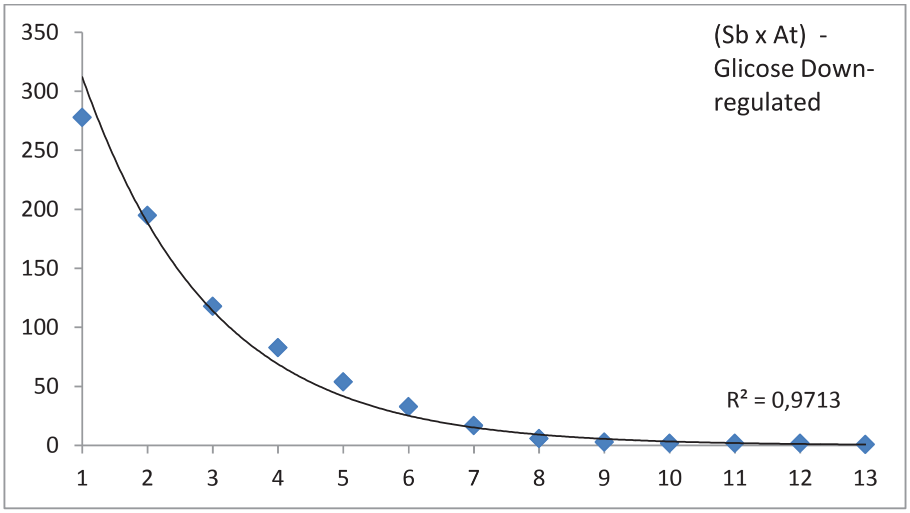

(Sb x Os) - ABA Up-regulated
R² = 0,9433
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

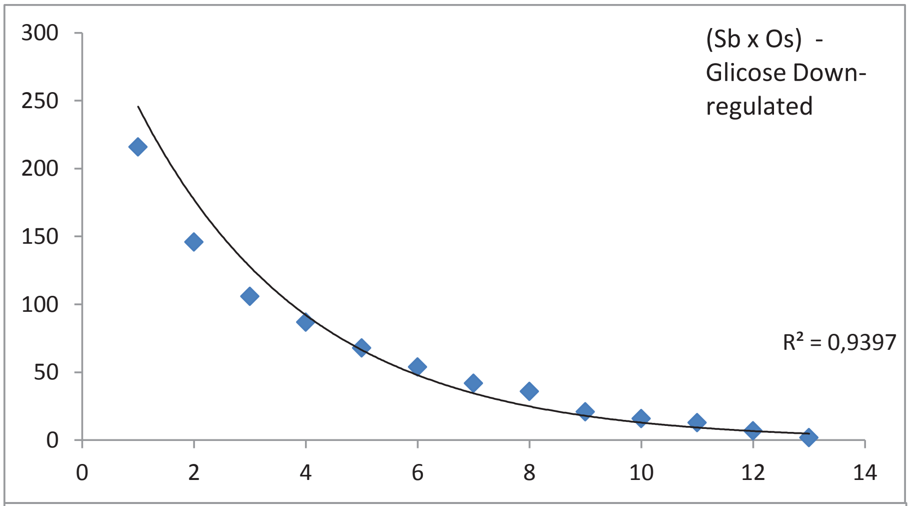

(Os x Sb) - ABA Up-regulated
R² = 0,9547
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7

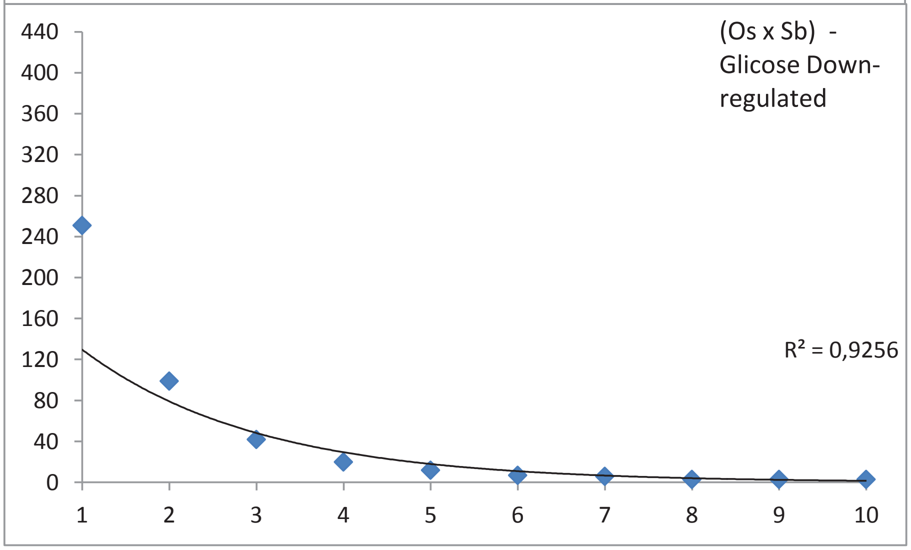

(At x Os) - Glicose Down-regulated
R² = 0,8939
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13

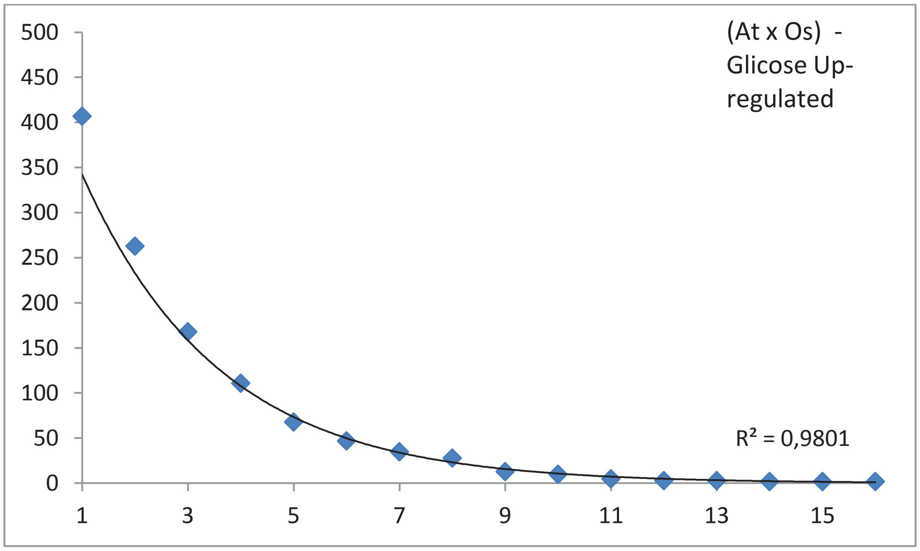

(At x Os) - Glicose Down-regulated
R² = 0,9743
500
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12

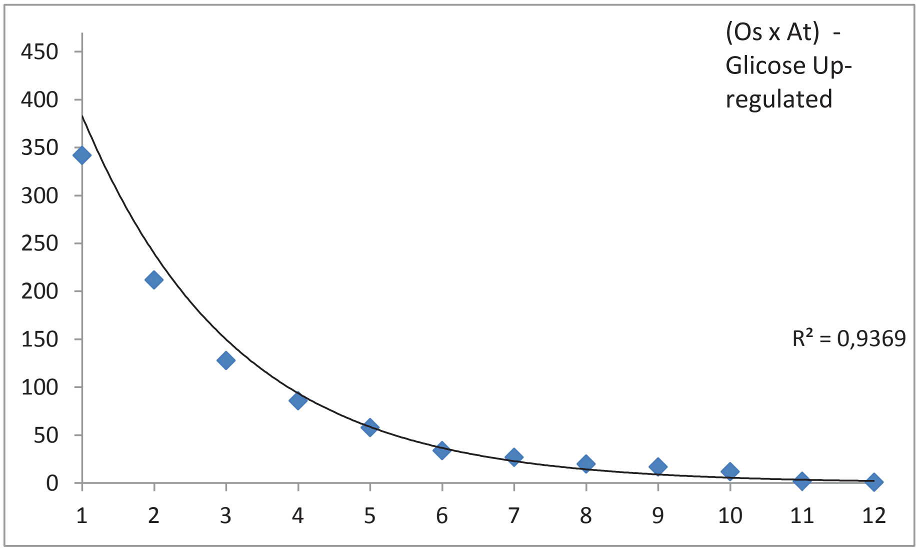

(At x Sb) - Glicose Down-regulated
R² = 0,9321
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12

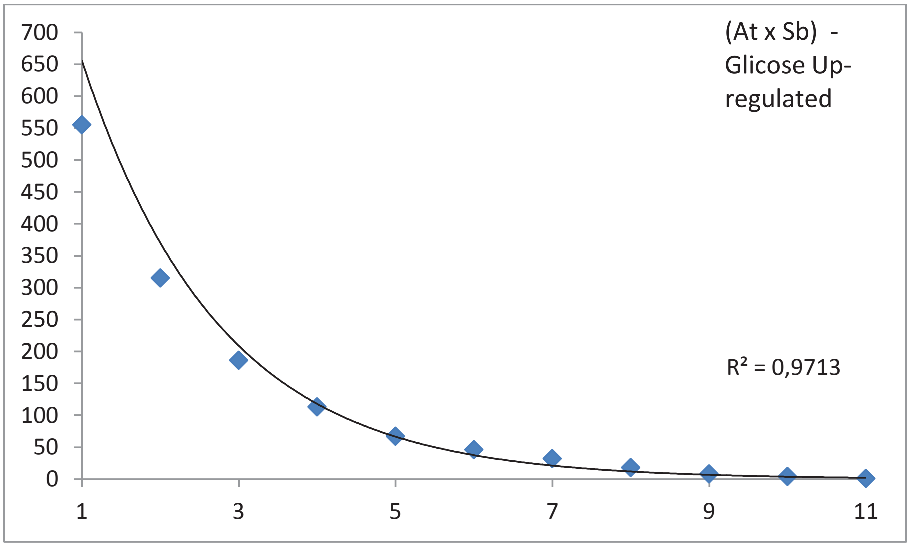

350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
(Sb x At) - Glicose Down-regulated
R² = 0,9713

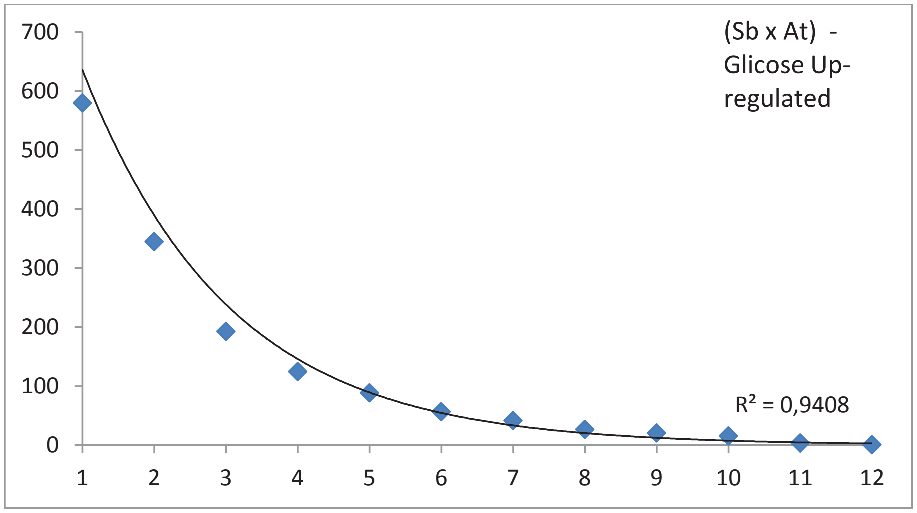

300
250
200
150
100
50
0
0
2
4
6
8
10
12
14
(Sb x Os) - Glicose Down-regulated
R² = 0,9397

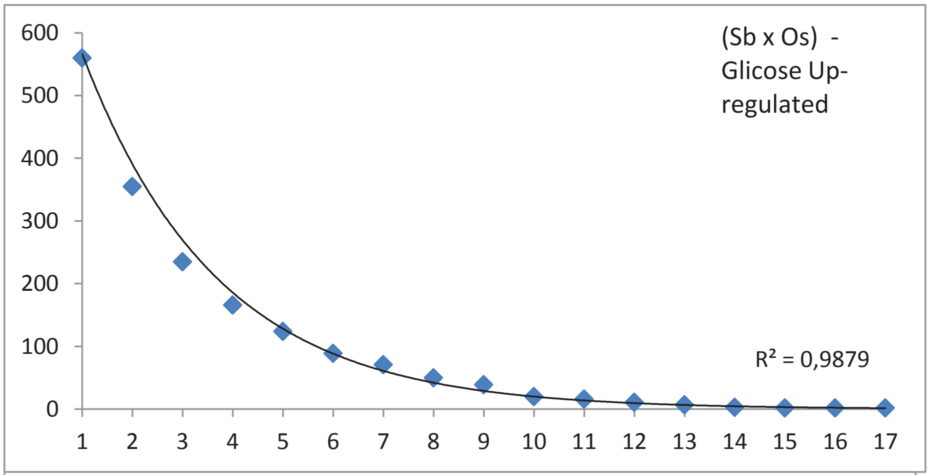

440
400
360
320
280
240
200
160
120
80
40
0
1
2
3
4
5
6
7
8
9
10
(Os x Sb) - Glicose Down-regulated
R² = 0,9256

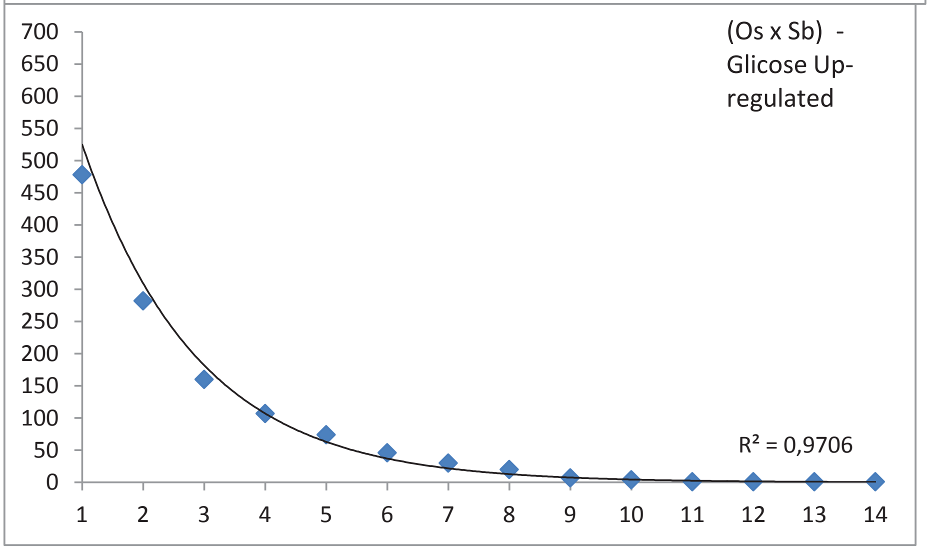

(At x Os) - Glicose Up-regulated
R² = 0,9801
500
450
400
350
300
250
200
150
100
50
0
1
3
5
7
9
11
13
15

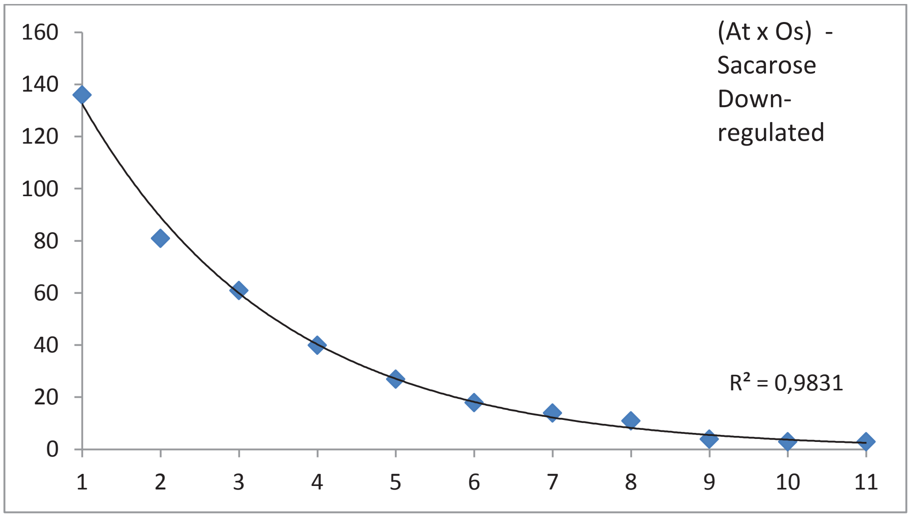

(Os x At) - Glicose Up-regulated
R² = 0,9369
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12

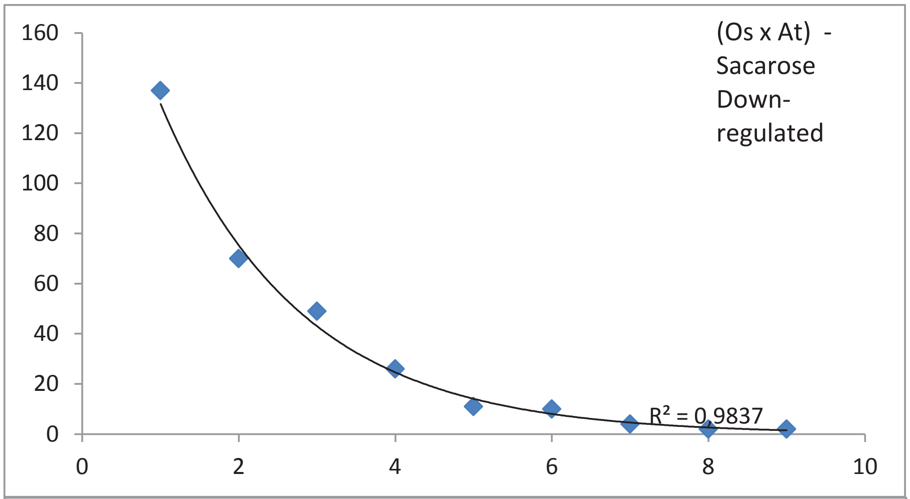

(At x Sb) - Glicose Up-regulated
R² = 0,9713
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
1
3
5
7
9
11

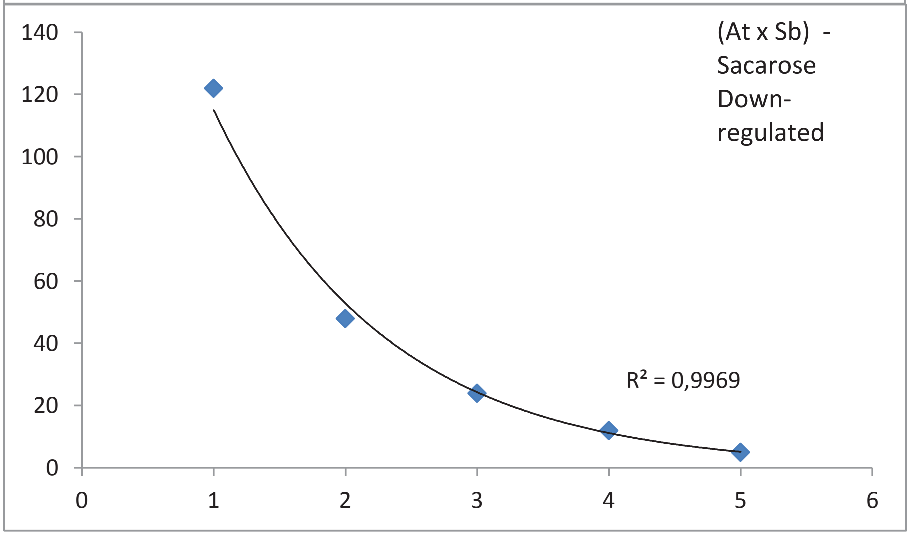

700
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
(Sb x At) - Glicose Up-regulated
R² = 0,9408

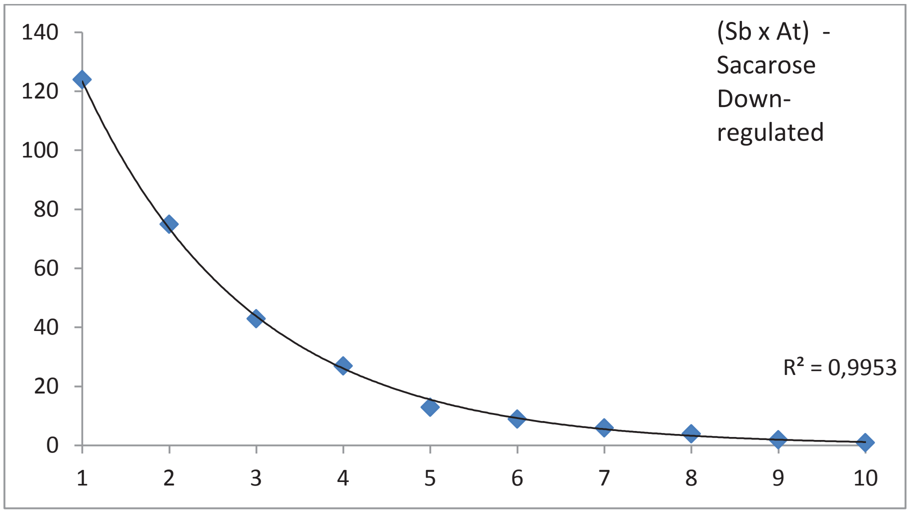

600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
(Sb x Os) - Glicose Up-regulated
R² = 0,9879

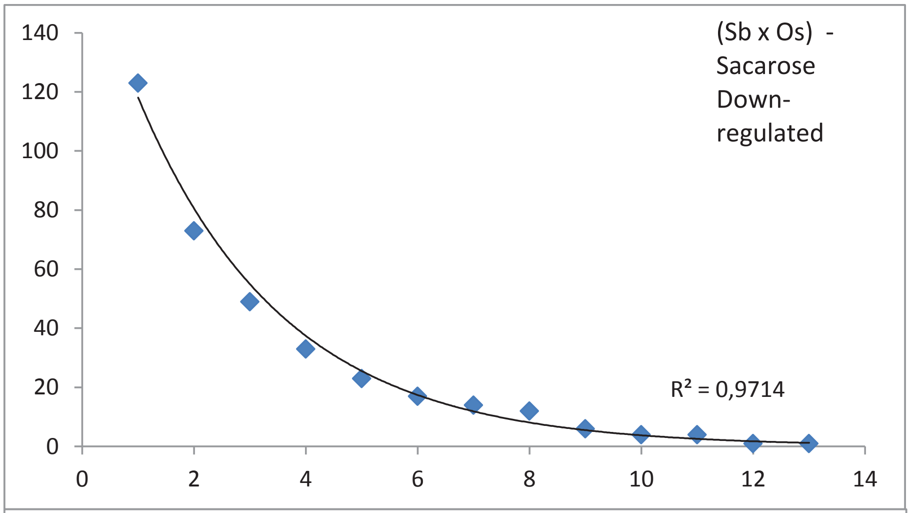

700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
(Os x Sb) - Glicose Up-regulated
R² = 0,9706

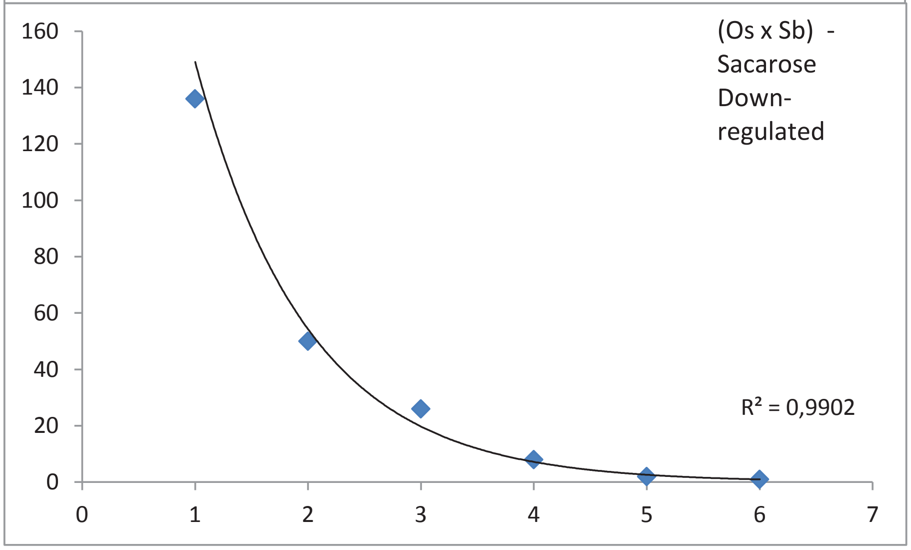

160
140
120
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
(At x Os) - Sacarose Down-regulated
R² = 0,9831

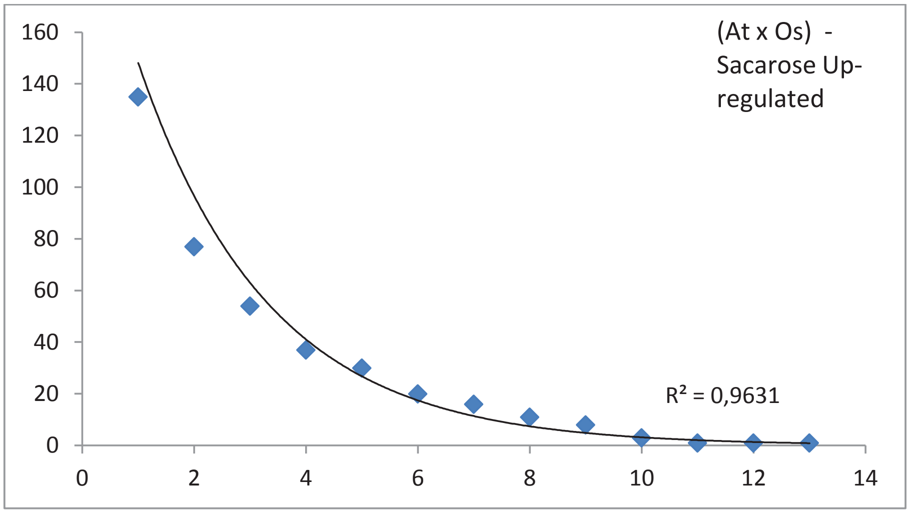

160
140
120
100
80
60
40
20
0
0
2
4
6
8
10
(Os x At) - Sacarose Down-regulated
R² = 0,9837


140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
(At x Sb) - Sacarose Down-regulated
R² = 0,9969

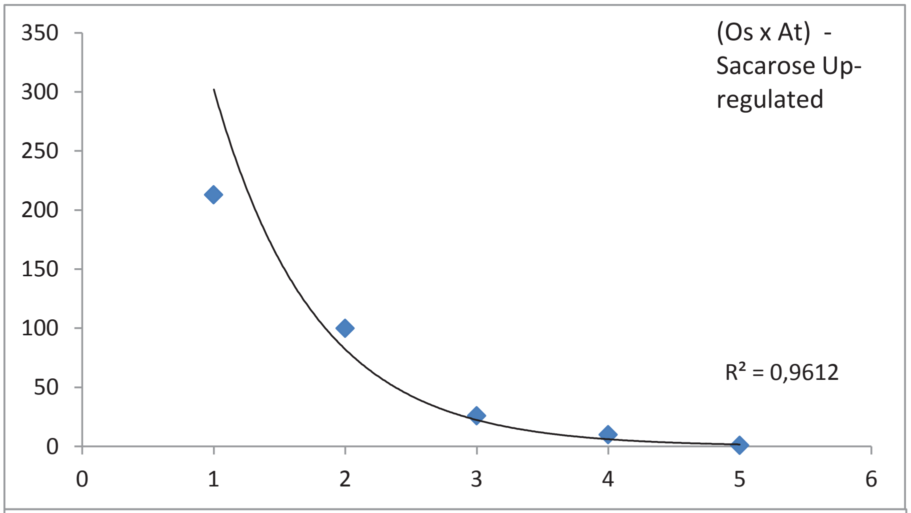

(Sb x At) - Sacarose Down-regulated
R² = 0,9953
140
120
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10


(Sb x Os) - Sacarose Down-regulated
R² = 0,9714
140
120
100
80
60
40
20
0
0
2
4
6
8
10
12
14

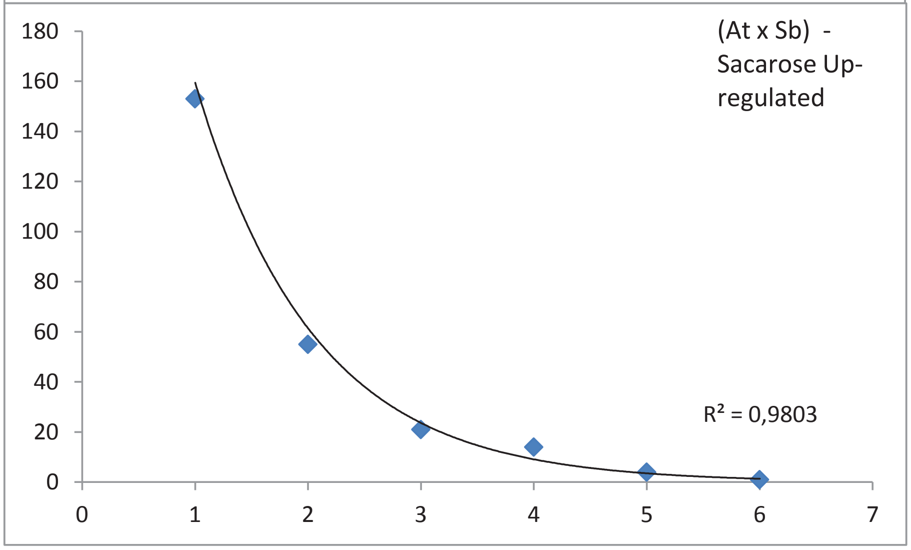

(Os x Sb) - Sacarose Down-regulated
R² = 0,9902
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7



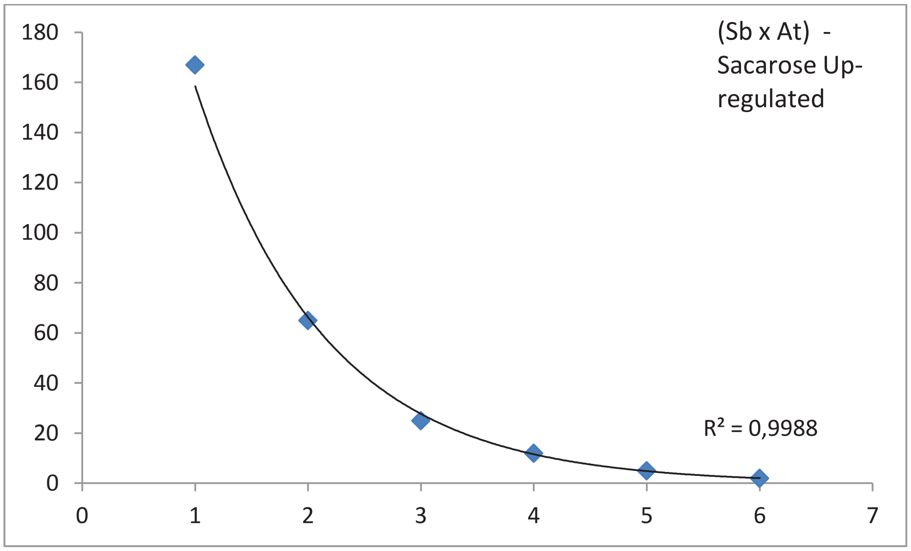
(At x Os) - Sacarose Up-regulated
160
140
120
100
80
60
40
20
0
0
2
4
6
8
10
12
14
R² = 0,9631

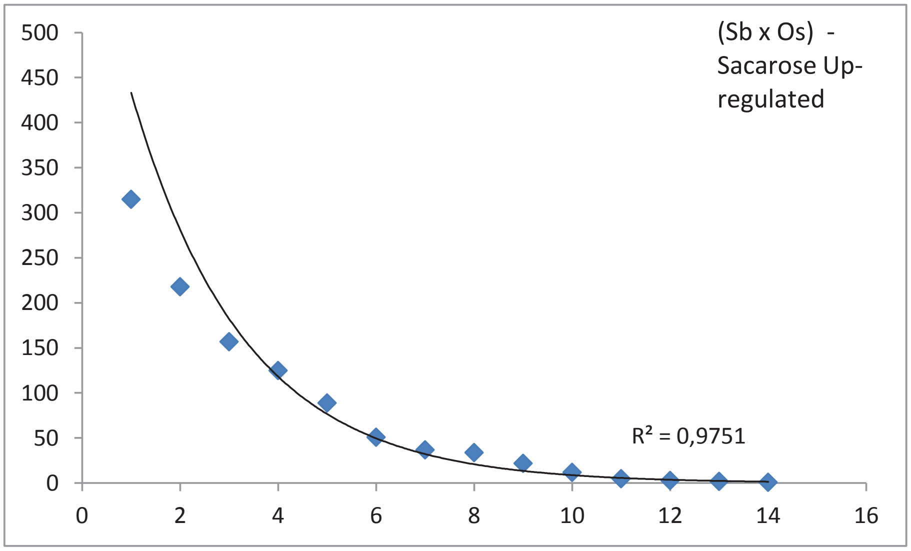


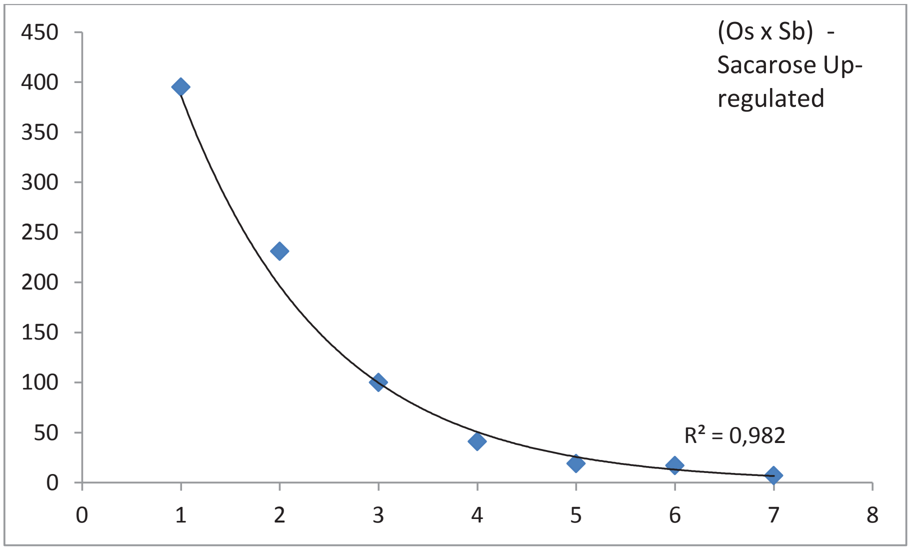
(Os x At) - Sacarose Up-regulated
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
R² = 0,9612



(At x Sb) - Sacarose Up-regulated
180
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
R² = 0,9803


(Sb x At) - Sacarose Up-regulated
R² = 0,9988
180
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7


(Sb x Os) - Sacarose Up-regulated
R² = 0,9751
500
450
400
350
300
250
200
150
100
50
0
0
2
4
6
8
10
12
14
16


(Os x Sb) - Sacarose Up-regulated
R² = 0,982
450
400
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8

Journal of Experimental Botany, Vol. 60, No. 4, pp. 1191–1206, 2009
doi:10.1093/jxb/erp014 Advance Access publication 16 February, 2009
This paper is available online free of all access charges (see http://jxb.oxfordjournals.org/open_access.html for further details)

Journal of Experimental Botany
www.jxb.oxfordjournals.org

RESEARCH PAPER

Expression pattern of four storage xyloglucan mobilization-related genes during seedling development of the rain forest tree *Hymenaea courbaril* L.

A. D. Brandão[1,3], L. E. V. Del Bem[1,2], M. Vincentz[1,2] and M. S. Buckeridge[3,*]

[1] Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil
[2] Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil
[3] Departamento de Botânica, Universidade de São Paulo, Instituto de Biociências São Paulo, SP, Brazil

Received 6 October 2008; Revised 21 December 2008; Accepted 22 December 2008

Abstract

During seedling establishment, cotyledons of the rain forest tree *Hymenaea courbaril* mobilize storage cell wall xyloglucan to sustain growth. The polysaccharide is degraded and its products are transported to growing sink tissues. Auxin from the shoot controls the level of xyloglucan hydrolytic enzymes. It is not yet known how important the expression of these genes is for the control of storage xyloglucan degradation. In this work, partial cDNAs of the genes xyloglucan transglycosylase hydrolase (HcXTH1) and β-galactosidase (HcBGAL1), both related to xyloglucan degradation, and two other genes related to sucrose metabolism [alkaline invertase (HcAlkIN1) and sucrose synthase (HcSUS1)], were isolated. The partial sequences were characterized by comparison with sequences available in the literature, and phylogenetic trees were assembled. Gene expression was evaluated at intervals of 6 h during 24 h in cotyledons, hypocotyl, roots, and leaves, using 45-d-old plantlets. HcXTH1 and HcBGAL1 were correlated to xyloglucan degradation and responded to auxin and light, being down-regulated when transport of auxin was prevented by *N*-1-naphthylphthalamic acid (NPA) and stimulated by constant light. Genes related to sucrose metabolism, HcAlkIN1 and HcSUS1, responded to inhibition of auxin transport in consonance with storage mobilization in the cotyledons. A model is proposed suggesting that auxin and light are involved in the control of the expression of genes related to storage xyloglucan mobilization in seedlings of *H. courbaril*. It is concluded that gene expression plays a role in the control of the intercommunication system of the source–sink relationship during seeding growth, favouring its establishment in the shaded environment of the rain forest understorey.

Key words: Auxin, cell wall, β-galactosidase, *Hymenaea*, invertase, storage mobilization, XTH, xyloglucan.

Introduction

The Leguminosae is among the most important families in the Neotropics. Worldwide, around 18 000 species belong to this family and, from studies of the chemical composition of their seeds, it has been estimated that about half of the species contain galactomannan, an endospermic cell wall storage polysaccharide (Buckeridge *et al.*, 2000). However, some species from the Leguminosae have specialized in a different way, relying on the cell wall polysaccharide xyloglucan. Species from the two genera *Copaifera* and *Hymenaea* use xyloglucan as the carbon source to establish initial growth, and this feature has been thought to be correlated with the success in the colonization of biomes, such as the Amazon and the savannahs in South America, by these genera (Buckeridge *et al.*, 2000).

The genus *Hymenaea* is thought to have originated in Africa and subsequently spread and adapted very well in the Neotropical regions of South America, generating many different species (Lee and Langenheim, 1975). Within the genus, the species *Hymenaea courbaril* is considered as one of the most successful mainly because of its shade and

[*] To whom correspondence should be addressed. E-mail: msbuck@usp.br

drought tolerance (Gerhardt, 1993; Souza and Valio, 1999; Santos and Buckeridge, 2004). The species stores ~45% of the seed dry mass in the cotyledons as xyloglucan (Buckeridge and Dietrich, 1990), whose function has been shown to be the support for initial seedling growth and development until autotrophic growth is established (Santos and Buckeridge, 2004).

Xyloglucan, besides being a source of carbon, also plays an important role in defining the structural properties of plant cell walls and the regulation of plant growth and development (Levy *et al.*, 1997). This polymer is composed of a cellulose-like (1→4)-linked β-D-glucan main chain, which is partially substituted by α-D-Xyl*p* side chains at O-6. Depending on the source, the side chain can be β-D-Gal*p*-(1→2)-α-D-Xyl*p*, α-L-Fuc*p*-(1→2)-β-D-Gal*p*-(1→2)-α-Xyl*p*, or even more complex groups (Buckeridge *et al.*, 2000). Xyloglucans are polysaccharides with remarkable structural regularity. Under hydrolysis with cellulase, a persistent pattern of oligosaccharides is produced. These are usually based on a cellotetraose chain in which three out of every four glucoses in the main chain is branched with xylose, forming a unit named XXXG (for the nomenclature of xyloglucans see Fry *et al.*, 1993). In contrast to other plant tissues, in seeds the storage xyloglucan is composed of galactose branches that form XLXG, XXLG, and XLLG, and no fucosylated oligosaccharides are found. Comparative studies of the fine structure of seed storage xyloglucans have shown the similar structural pattern of subunits, but the *H. courbaril* xyloglucan displays unique structural features, with ~50% of it being composed of a family of oligosaccharides based on five instead of four glucoses in the main chain (XXXXG) (Buckeridge *et al.*, 1992, 1997; Tiné *et al.*, 2006).

During early seedling growth, xyloglucan mobilization proceeds, leading to sucrose production in the cotyledons. Sucrose is transported to the developing sink tissues (shoot and root) (Santos and Buckeridge, 2004). It has been observed for cotyledons of *Tropaelum majus*, *Copaifera langsdorffii*, and *H. courbaril* that the polysaccharide is attacked by three hydrolases and one transglycosylase (Buckeridge *et al.*, 2000). In all xyloglucan mobilization systems studied to date, it has been shown that polysaccharide degradation is achieved by the coordinated action of at least four cell wall hydrolases. The main chain of the polymer is attacked by xyloglucan endo-transglycosylase hydrolase (XTH) which, depending on the species, may be of the hydrolytic type (XET; xyloglucan endo-transglycosylase) or the transglycosylase type (XTH). The oligosaccharides produced are attacked by β-galactosidose, α-xylosidase, and β-glucosidase, so that free glucose and xylose are produced (Buckeridge *et al.*, 2000).

In *H. courbaril* as well as in *T. majus*, auxin is thought to be an important positive regulator of cell wall xyloglucan storage mobilization. It has been proposed that xyloglucan degradation is dependent on the polar transport of auxin in *H. courbaril* (Santos *et al.*, 2004).

In addition, light also seems to influence directly the rate of xyloglucan mobilization in cotyledons of developing seedlings of *H. courbaril* by modulating the production of xyloglucan hydrolases (Santos and Buckeridge, 2004; Santos *et al.*, 2004). The mechanisms involved in this light-mediated xyloglucan hydrolysis are uncertain but are possibly related to the photomorphogenetic growth pattern.

This intercommunication system between light, hormones, and sugar metabolism highlights a very efficient mechanism used to store and use carbon during seedling establishment. The possible control related to a switch from skotomorphogenesis and photomorphogenesis programmes might explain the successful development of the seedlings of *H. courbaril* in the very low light intensity conditions found in the understorey of the rain forests. Lower light intensities lead seedlings to etiolate (skotomorphogenesis) until they cross the litter layer of the forest, after which the photomorphogenesis rules prevail.

Light has also been proposed to stimulate the production of auxin that would be transported to the cotyledons where it would signal for xyloglucan degradation to take place, which would result in sucrose production. The resulting sucrose would then be translocated to the growing seedling and, when the number of leaves is such that net photosynthesis can support autotrophic growth, the shrunken cotyledons would have already degraded most of their reserves and would fall, leaving a young plant ready to face life in the understory of the forest.

It seems, therefore, that the interplay of auxin and light signals plays important roles in the control of storage mobilization in the *H. courbaril* cotyledons. Mobilization means here degradation of xyloglucan to maintain the synthesis of sucrose that will be translocated to sustain growth of developing sink organs.

In order to obtain further insight into xyloglucan mobilization the question was asked of whether auxin and light act by regulating the expression of key genes involved in the mobilization process. In this work, partial cDNA for XTH and β-galactosidase (xyloglucan degradation), sucrose synthase, and alkaline invertase (sucrose metabolism) were cloned and the expression of the corresponding genes under inhibition of polar transport of auxin and the presence or absence of light during the mobilization process were analysed in different organs and during circadian light changes. The results suggest that regulation of mRNA accumulation is important for xyloglucan mobilization in *H. courbaril*.

Materials and methods

Plant material and experimental design

Seeds of *H. courbaril* L. were obtained from trees growing in a gallery forest at Sao João da Boa Vista county (22°00′S; 47°18′W), São Paulo, Brazil. Seeds were scarified manually with sandpaper and allowed to germinate in distilled water in Petri dishes at 25 °C for 15 days, when radicle protrusion occurred. After germination, the seedlings were transferred to pots (0.5 l) containing vermiculite and were allowed to grow in a greenhouse under a 12 h photoperiod at ambient temperature.

Thirty-day-old plants were subjected to shoot excision, and auxin transport inhibition in these plants was induced by application of 200 mM *N*-1-naphthylphthalamic acid (NPA) in lanolin as described previously (Santos *et al.*, 2004). For light/darkness experiments, developing seedlings were transferred to continuous light (50 μmol of photons m^2 s^{-1}) at 28 °C or to darkness at 28 °C for 20 d.

After these treatments, seedlings were allowed to grow until the 45th day, when samples of cotyledons, leaves, hypocotyls, and roots from 15 plants were harvested (Fig. 1). The 45th day was chosen due to the fact that xyloglucan mobilization is approximately half way to being complete and at this point the cotyledons display high level of activities of xyloglucan hydrolases (Tiné *et al.*, 2000; Santos and Buckeridge, 2004).

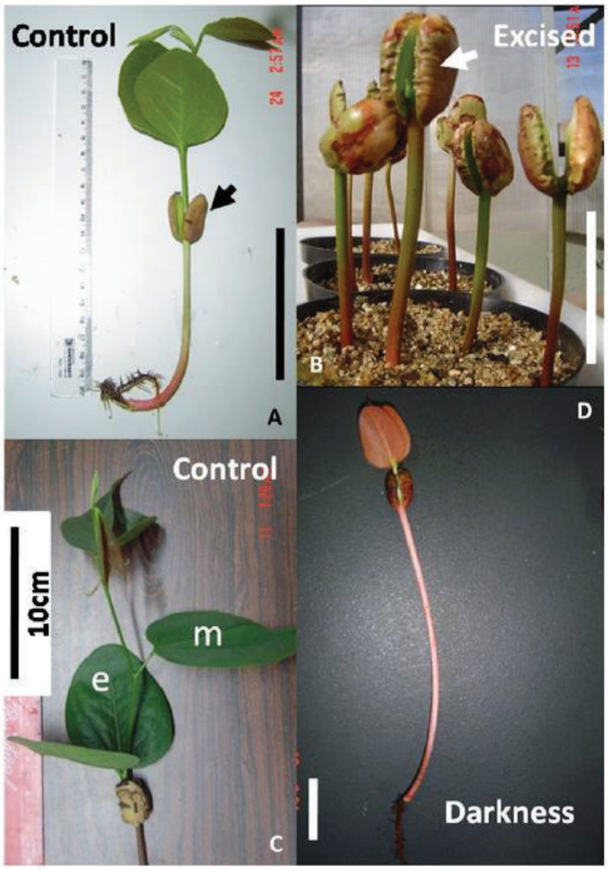


Fig. 1. Aspect of seedlings of *Hymenaea courbaril* on the 45th day after the beginning of imbibition. (A) A typical specimen of *H. courbaril*. Note the investment in the shoot (arrow=cotyledon). (B) Seedlings that had their shoot excised 35 d after imbibition. These will develop new branches in ~30 d (see Santos *et al.*, 2004) (arrow=cotyledon). (C) Closer view of the shoot, showing the eophylls (e) (first leaf) and metaphylls (m) (second leaf). (D) An etiolated seedling of *H. courbaril* after 20 days in darkness. Note that the hypocotyl is ~5 times longer in the dark and eophylls did not develop, characterizing the programme of skotomorphogenesis. Bars represent 10 cm.

Sampling during the 24 h day–night cycle was obtained every 6 h (0, 6, 12, 18, and 24 h). The sampled material was immediately frozen in liquid nitrogen and stored, at most for 3 months, before RNA extraction. The choice to evaluate expression during a whole day was made on the basis of previous observations that xyloglucan hydrolases presented diurnal variations (A Santos and MS Buckeridge, unpublished results). The time in all experiments was recorded in days after the beginning of imbibition. All assays and time points were conducted in triplicate.

Total RNA extraction, primer design, cDNA synthesis and amplification, cloning, and sequencing of PCR products

RNA extraction was performed with different protocols according to the part of the seedling being analysed, i.e. cotyledon, hypocotyl, leaves, or root. For roots and leaves, the method used was the same as that described by Gesteira *et al.* (2003), and for hypocotyls a Concert Total RNA extraction solution (Invitrogen) was used following the manufacturer's instructions. For cotyledons, a procedure described by Daohong *et al.* (2004) was used with the following modifications. A 1 g aliquot of cotyledons was ground into a fine powder in liquid nitrogen and then 10 ml of buffer solution (0.5 M TRIS-HCl pH 8, 0.01 M EDTA, 0.25 M LiCl, 2.5% SDS, and 0.1% β-mercaptoethanol) was added. The mixture was vigorously stirred for 1 min and phenol:chloroform:isoamylic alcohol (25:24:1, v/v/v) was added and stirred again for 1 min. The material was centrifuged at 4 °C, 13 000 *g*, for 15 min. This step was repeated twice. The RNA was then precipitated twice. The first precipitation was performed by adding 0.5 ml of 12 M LiCl and kept for 2 h at –80 °C. After 30 min of centrifugation the pellet was washed with 6 M LiCl. After a further 15 min of centrifugation under the same conditions, the pellet was resuspended in 0.3 ml of diethylpyrocarbonate (DEPC)-treated water and a second precipitation was performed using 1/10 volume of 3 M sodium acetate, pH 5.2 and a 0.5 vol. of absolute ethanol. The mixture was stirred and kept overnight at –80 °C. After 30 min of centrifugation, the pellet was washed with 70% ethanol, resuspended in DEPC-treated water, and stored at –80 °C. RNA integrity was checked by 1% agarose gel electrophoresis with 6% formaldehyde in MOPS buffer (20 M MOPS, 0.6 M sodium acetate, 0.01 M EDTA, pH 8). The RNAs were stored at –80 °C until used for cDNA synthesis.

The reverse transcription of total RNAs to cDNA was performed using Ready-To-Go RT-PCR beads (Amersham Biosciences). The reaction was carried out following the manufacturer's instructions, using 3.5 μg of total RNA samples and 1 mM primer Oligo dT12-18 (Invitrogen). The cDNA was stored at –20 °C until use.

Coding sequences for actin, xyloglucan endo-transglycosylase, β-galactosidase, alkaline invertase, and sucrose synthase from eudicotyledons were obtained from GenBank

(NCBI: http://www.ncbi.nlm.nih.gov/Genbank/index.html). Amino acid sequence alignments were performed with Clustal W (http://www.ebi.ac.uk/clustalw) using default parameters. Gaps were removed for phylogenetic analyses which were performed using the Prodist program to calculate distances among sequences, and the Neighbor–Joining tree constructing algorithm (Saitou and Nei, 1987) was used to establish the evolutionary relationships among the homologous sequences.

To amplify partial cDNAs, pairs of nested and degenerated primers were designed from conserved protein domains detected in the sequences from organisms phylogenetically closely related to *Hymenaea*, i.e. the Leguminosae. The primers synthesized had 17–24 nucleotides (Table S2).

For PCR amplification of cDNAs, the following solution in a total volume of 50 μl was used: 10 μl of cDNA reaction, 5 μl of 10× PCR buffer, 1.5 μM $MgCl_2$ (50 mM), 2.5 μl of dNTP mix (10 μM), 1 mM for the degenerated 3′ primer and 1 mM for the degenerated 5′ primer, and 0.3 μl of *Taq* DNA polymerase (5 U μl^{-1};– Invitrogen). The PCR conditions for β-galactosidase, xyloglucan transglycosylase hydrolase, alkaline invertase, and sucrose synthase were: initial denaturation for 2 min at 95 °C followed by 5 min at 72 °C and 36 cycles (94 °C for 5 s, 55 °C for 1 h 30 min, and 72 °C for 45 s), and a final extension of 10 min at 72 °C. Actin was used as an internal control in semi-quantitative RT-PCR experiments.

The amplification products of the expected size were purified after separation by 2% agarose gel electrophoresis using the kit Wizard SV Gel and PCR Clean-Up System (Promega) according to the manufacturer's instructions.

PCR products were cloned into the TA-Cloning vector (Invitrogen) following the manufacturer's instructions. Sequencing was performed using a DNA sequencing Big Dye Kit™ (Applied Biosystems) following the manufacturer's instructions, and analysed with an automatic sequencer ABI PRISM™ 377 (Perkin Elmer).

The sequences that were obtained were then compared with GenBank sequences using the blastn tool (NCBI: http://www.ncbi.nlm.nih.gov/GenBank/index.html) to confirm their identities.

Semi-quantitative RT-PCR and densitometric quantification

The semi-quantitative RT-PCR experiments were performed using pairs of specific primers which were designed from the partial cDNA sequences of the *H. courbaril* enzymes which were obtained as described above. The primers promoted the amplification of a 360 bp fragment for actin, a 400 bp fragment for β-galactosidase (BGAL), a 330 bp fragment for xyloglucan transglycosylase hydrolase (XTH), a 440 bp fragment for alkaline invertase (AlkIN), and a 410 bp fragment for sucrose synthase (SUS) (Table S2). The number of cycles for PCR amplification was 36 for all genes and 25 for actin. The optical densities of cDNA amplification products were obtained using eagle eye II (Stratagene). All assays were conducted in triplicate, and for each experiment three technical replicates were used. The densitometry analyses of the gels were performed with Gel-Pro Analyser software (V. 3.1 Gel-Pro; Media Cybernetics, Inc.) (data no shown).

Construction of phylogenetic trees

Arabidopsis and rice xyloglucan transglycosylase hydrolase, β-galactosidase, alkaline invertase, and sucrose synthase were identified by using the blast program and the query sequences at TAIR (The *Arabidopsis* Information Resource) in order to identify genes that encode in the *Arabidopsis* genome. Orthologues in rice (*Oryza sativa*) were also searched for using tbastn. Alignment was performed using Clustal W.

Phylogenetic reconstruction was performed from distances calculated from the aligned amino acid sequences using the JTT matrix (Jones *et al.*, 1992), and the tree was inferred by the Neighbor–Joining method (Saitou and Nei, 1987). The numbers of amino acid positions used were 145, 111, 273, and 382 for β-galactosidase, xyloglucan transglycosylase hydrolase, alkaline invertase, and sucrose synthase, respectively. The analyses were carried out using the software MEGA4 (Tamura *et al.*, 2007).

In order to classify the cDNAs obtained from cotyledons of *H. courbaril* β-galactosidase (HcBGAL1; EU370969), xyloglucan transglycosylase hydrolases (HcXTH1; EU370971), alkaline invertase (HcAlkIN1; EU370968), and sucrose synthase (HcSUS1; EU370970), trees were generated using the representatives of the Leguminosae (see Supplementary Table S1 available at *JXB* online for accession numbers).

Results

Phylogenetic characterization of partial cDNA for H. courbaril β-galactosidase, xyloglucan transglycosilase hydrolase, alkaline invertase, sucrose synthase, and actin

The strategy to obtain partial cDNA sequences encoding *H. courbaril* relied basically on the use of two pairs of nested degenerated primers corresponding to conserved sequences. The amplification products with the expected length (Table S2) corresponding to 452 bp for actin, 435 bp for β-galactosidase, 333 bp for XTH, 819 bp for alkaline invertase, and 1146 bp for sucrose synthase were cloned and sequenced. Their identity was confirmed according to their similarity to known sequences. The *H. courbaril* (Hc) amino acid sequences were further characterized by a phylogenetic analysis. As expected, the partial sequences for HcBGAL1, HcXTH1, HcAlkIN1, and HcSUS1 were found to group more strongly with homologous sequences from Leguminosae. This analysis also allowed the more precise classification of the *Hymenaea* sequences. It was found that HcBGAL1 is more closely related to group I β-galactosidases; HcXTH1 to group I XTHs; HcAlkIN1 to group β alkaline invertase, and HcSUS1 to the SUS1 group of sucrose synthase (Supplementary Table S1 at *JXB* online). A comparison of HcBGAL1 with other plant β-galactosidases is shown in Fig. 2. This gene is inserted in a group that contains TBG4, an exogalactanase from tomato, and also genes from mung bean and chickpea. HcBGAL1 separated completely from a group that contains several genes from *Arabidopsis* and rice, except for the genes AtBGAL2, AtBGAL12, and AtBGAL4.

HcXTH1 appears closer to the genes VaXTH1 and VaXTH2 (Fig. 3), both coding for XTHs supposedly related

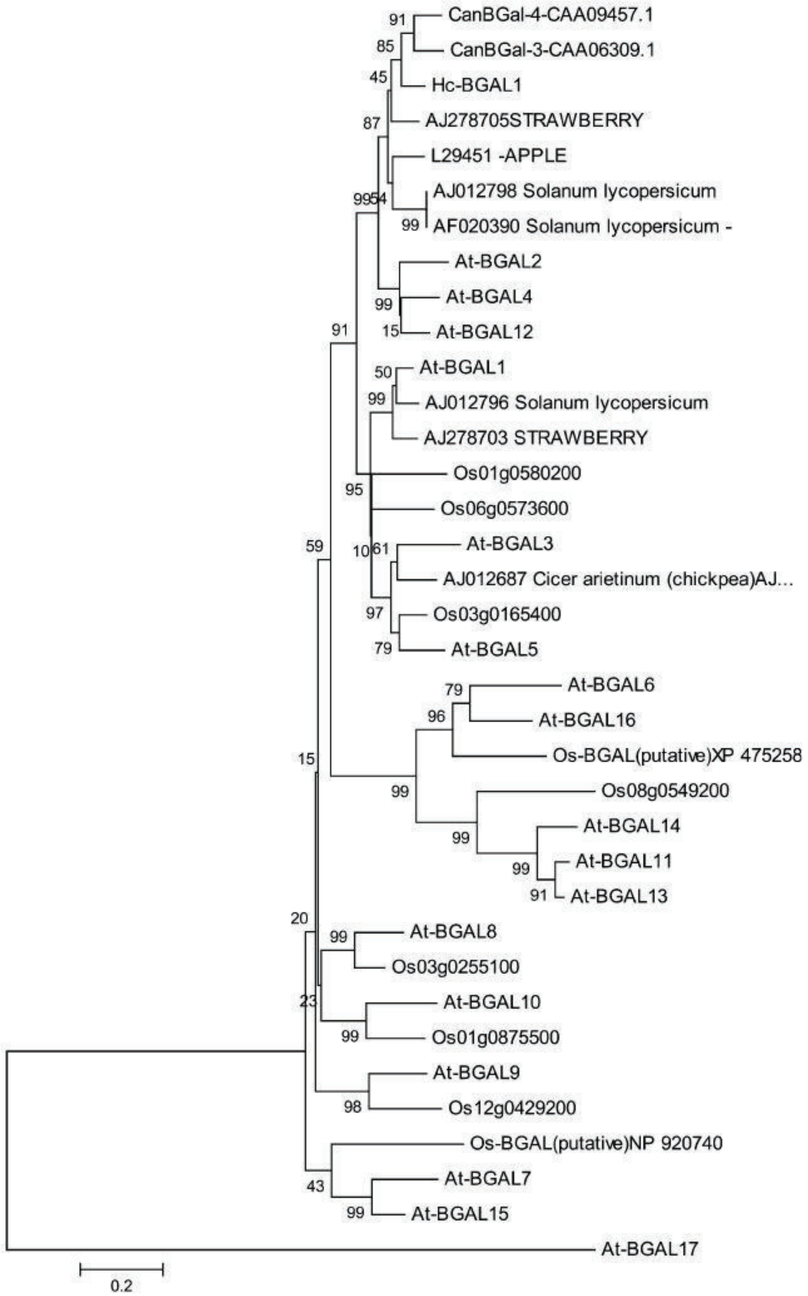


Fig. 2. Phylogenetic relationship of β-galactosidases of Angiosperms. Distances were calculated on the basis of positions 181–332 in *A. thaliana* (At-BGAL4). The nucleotide sequence can be accessed at NCBI: EU370969. Accession numbers are as follows: *A. thaliana* (At-GAL1, AT3G13750; At-GAL2, AT3G52840; At-GAL3, AT4G36360; At-BGAL4, AT5G56870; At-GAL5, AT1G45130; At-BGAL6, AT5G63800; At-GAL7, AT5G20710; At-BGAL8, AT2G28470; At-BGAL9, AT2G32810; At-BGAL10, AT5G63810; At-BGAL11, AT4G35010; At-GAL12, AT4G26140; At-BGAL13, AT2G16730; At-BGAL14, AT4G38590; At-GAL15, AT1G31740, At-BGAL16, AT1G77410; At-BGAL17, AT1G72990; *Oryza sativa* (10 members), Leguminosae (CanBGAL-4, CAA09457; CanBGAL-3, CAA06309; chickpea, AJ012687), tomato (AJ012796; AF020390; AJ012798), apple (L29451), and strawberry (AJ278705; AJ278703). The values presented represent the bootstraps.

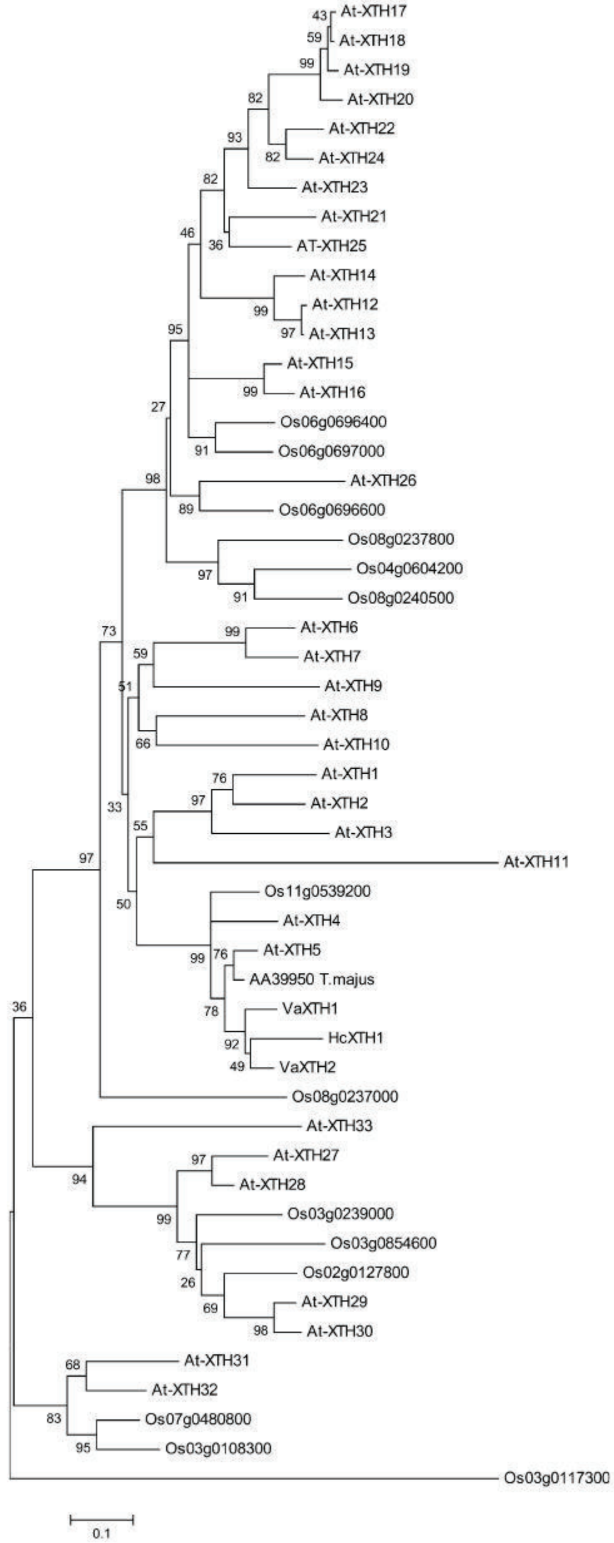


Fig. 3. Phylogenetic relationship of xyloglucan transglycosylase hydrolase (XTH) of Angiosperms. Distances were calculated on the basis of positions 87–197 of *A. thaliana* (At-XTH5). The nucleotide sequence can be accessed at NCBI: EU370971. Accession numbers are as follows: *A. thaliana* (At-XTH1, AT4G13080; At-XTH2, AT4G13090; At-XTH3, AT3G25050; At-XTH4, AT2G06850; At-XTH5, AT5G13870; At-XTH6, AT5G65730; At-XTH7, AT4G37800; At-XTH8, AT1G11545; At-XTH9, AT4G03210; At-XTH10, AT2G14620; At-XTH11, AT3G48580; At-XTH12, AT5G57530; At-XTH13, AT5G57540; At-XTH14, AT4G25820; At-XTH15, AT4G14130; At-XTH16, AT3G23730; At-XTH17, AT1G65310; At-XTH18, AT4G30280; At-XTH19, AT4G30290; At-XTH20, AT5G48070; At-XTH21, AT2G18800; At-XTH22, AT5G57560; At-XTH23, AT4G25810; At-XTH24, AT4G30270; At-XTH25, AT5G57550; At-XTH26, AT4G28850; At-XTH27, AT2G01850; At-XTH28, AT1G14720; At-XTH29, AT4G18990; At-XTH30, AT1G32170; At-XTH31, AT3G44990; At-XTH32, AT2G36870), *Oryza sativa* (14 members), Leguminosae (Va-XTH2, AB086396; VaXTH1, AB086395), and*Tropaeolum majus* (AA39950). The values presented represent the bootstraps.

to cell wall expansion in *Vigna angulares.* It is also noticeable that HcXTH1 belongs to the same possible group of orthologues of an XTH from nasturtium (*T. majus*) that is the only XTH clearly related to storage xyloglucan metabolism (see below).

Relatively little is known about the features of genes related to alkaline invertases. HcAlkIN1 was found to be close to a gene from the legume *Lotus corniculatus* (Lcinv1) (Fig. 4). This class of genes is referred to as neutral alkaline invertase in the literature and its function is poorly understood.

Figure 5 shows the phylogenetic relationship between HcSUS1 and other plants SUS sequences. This gene belongs to a possible group of orthologues including two *Arabidopsis* (SUS1 and SUS4) and three rice genes (Os06g0194900, Os03g0401300, and Os07g0616800). HcSUS1 clustered together with several legumes sequences, among which the most closely related seems to be one from *Pisum sativum* (SUS isoform 3).

Expression patterns of the four genes in response to auxin polar transport inhibition and different light treatments

On the basis of the fact that the presence of gene expression and enzyme activities of the four genes cloned in this work is associated with most plant organs and different isoforms of the proteins, it was decided to evaluate the levels of expression of these genes in roots, cotyledons, leaves, and stem of developing seedlings of *H. courbaril.* The results of the experiments with excision of the top shoot and NPA treatment are shown in Fig. 6 and the experiment with continuous light and darkness is shown in Fig. 7.

Cotyledons: Figure 6A presents the patterns of expression of HcBGAL1 (BG), HcXTH1 (XTH), HcAlkIN1 (AlkN), and HcSUS1 (SUS) in cotyledons of 45-d-old seedlings of *H. courbaril.* Both HcBGAL1 and HcXTH1 presented a peak of mRNA production at midday. HcBGAL1 was expressed in control seedlings with a peak of transcript at 12 h. Transcripts of this gene were found neither under excised conditions (top shoot of the plant excised) nor in

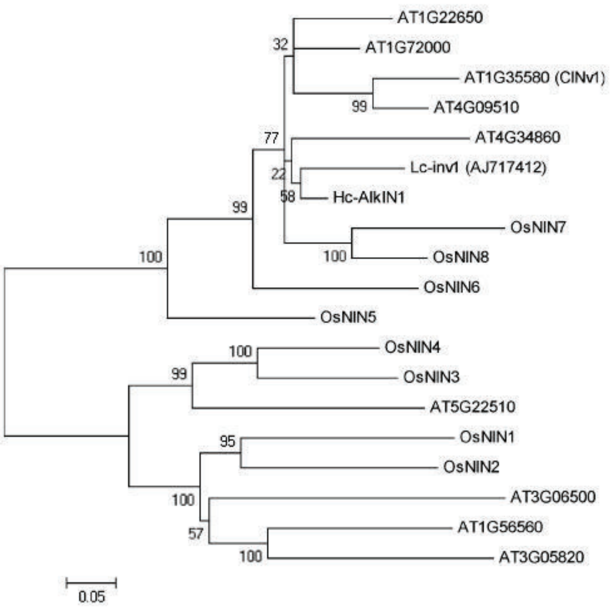


Fig. 4. Phylogenetic relationship of alkaline/neutral invertase of Angiosperms. Distances were calculated on the basis of positions 278–556 of *A. thaliana* (AT4G34860). The nucleotide sequence can be accessed at NCBI: EU370968. Accession numbers are as follows: *A. thaliana* (nine members), *Oryza sativa* (OsNIN1, AY575558; OsNIN2, AY575559; OsNIN3, AY575560; OsNIN4, AY575561; OsNIN5, AY575562; OsNIN6, AY575563; OsNIN7, AY575564; OsNIN8, AY575565), and *Lotus corniculatus* (Lc-inv1, AJ717412). The values presented represent the bootstraps.

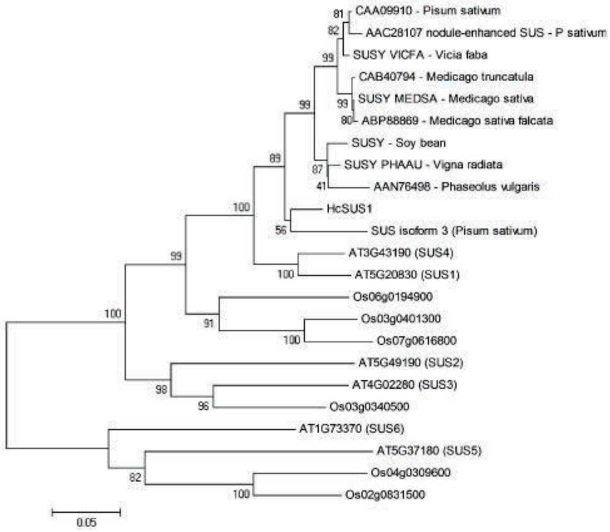


Fig. 5. Phylogenetic relationship of sucrose synthase of Angiosperms. Distances were calculated on the basis of positions 429–810 of *A. thaliana* (SUS4). The nucleotide sequence can be accessed at NCBI: EU370970. Accession numbers are as follows: *A. thaliana* (At-SUS1, AT5G20830; At-SUS2, AT5G49190; At-SUS3, AT4G02280; At-SUS4, AT3G43190; At-SUS5, AT5G37180; At-SUS6, AT1G73370), *Oryza sativa* (six members), and legumes (SUSY VICFA, P31926; SUSY MEDSA, O65026; SUSY PHAAU, Q01390; SUS isoform 3, CAC32462). The values presented represent the bootstraps. Dicot=eudicotyledons.

NPA-treated seedlings (Fig. 6A). HcXTH1 displayed a very similar pattern of expression to HcBGAL1 (Fig. 6A). Under excision treatment, HcAlkIN1was up-regulated and under NPA treatment it was down-regulated (Fig. 6A). HcSUS1 was expressed during the light period, and under excision treatment no transcript was observed. Under NPA treatment, the HcSUS1 gene was expressed similarly to control seedlings, with a peak of transcript at 18 h (Fig. 6A).

Figure 7A shows the expression patterns of the four cloned genes in cotyledons of *H. courbaril* under different light conditions. Under constant light, HcBGAL1 was expressed only at 18 h and in darkness it was expressed only at 0 h and 24 h. Under constant light, no transcripts of HcXTH1 were observed and under darkness the gene was up-regulated (Fig. 7A). HcAlkIN1 was strongly up-regulated in constant light whereas in the darkness it was down-regulated (Fig. 7A). Regarding HcSUS1, in control seedlings a high level of mRNA was observed at 12 h. Cotyledons from seedlings growing under constant light for 20 d presented a peak of expression at 6 h. The level of expression of HcSUS1 remained low and constant in seedlings growing in darkness (Fig. 7A).

Hypocotyls: HcBGAL1 was expressed in control seedlings with a peak of transcript at 6 h, and under excision and NPA treatments the gene was down-regulated (Fig. 6B). HcXTH1 presented peaks of transcripts every 12 h. Under excision and NPA treatments, the gene was down-regulated. HcAlkIN1 showed a peak of transcript at 0 h and 24 h, and under excision and NPA treatments the gene was down-regulated (Fig. 6B). HcSUS1 presented a peak of transcription at 0 h and 24 h, and under excision treatment the gene was constitutively expressed. Under treatment with NPA, no transcript was observed (Fig. 6B).

Under constant light, the HcBGAL1 gene was up-regulated and under darkness it was expressed similarly to control seedlings, with peaks of transcripts every 12 h (Fig. 7B). HcXTH1 was down-regulated under constant light and darkness, being constitutively expressed (Fig. 7B). Under constant light, the HcAlkIN1 gene was strongly up-regulated whereas under darkness the gene was down-regulated. HcSUS1 presented a similar pattern to the control hypocotyls, whereas it was completely inhibited in the dark (Fig. 7B).

Roots: HcBGAL1 was expressed in control seedlings with a peak of transcript at 0 h and 24 h (Fig. 6C). Under excised conditions, the gene was down-regulated and under NPA treatment the gene was expressed similarly to control seedlings, with peaks of transcripts at 0 h and 24 h (Fig. 6C). HcXTH1 displayed a peak of transcription in the middle of the night in roots. It was down-regulated (showing a constitutive behaviour) under excised conditions and it was completely inhibited by NPA treatment (Fig. 6C). HcAlkIN1 was expressed in control seedlings with a peak of transcript at 12 h. Under excised conditions,

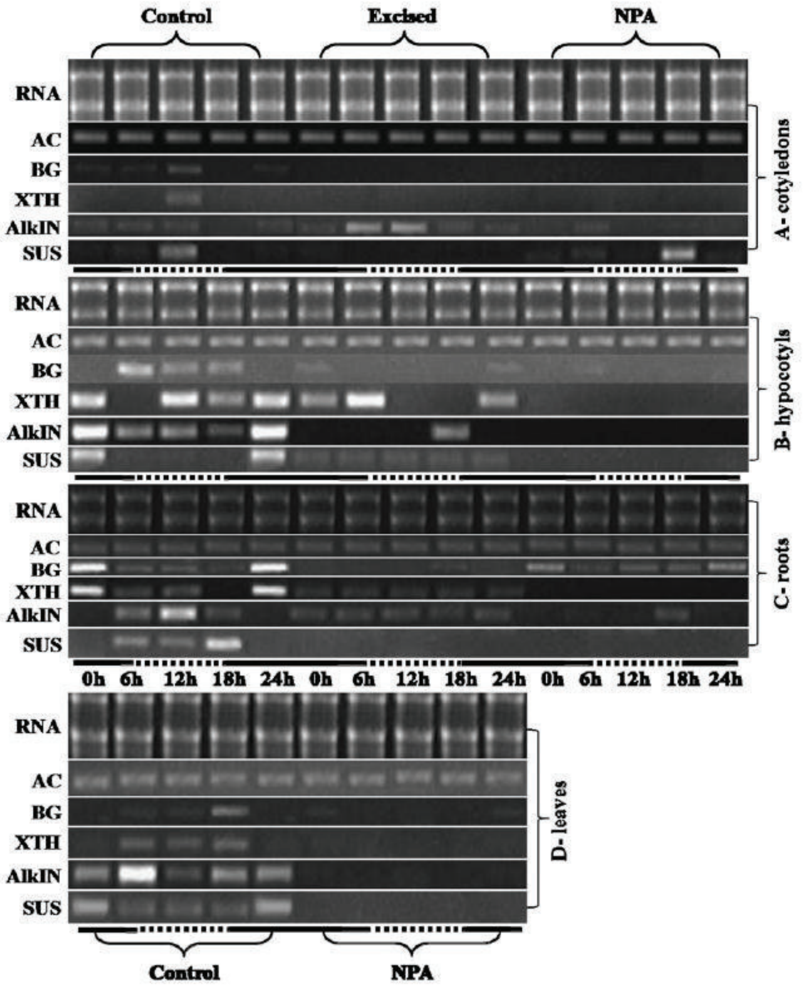


Fig. 6. Patterns of gene expression of actin (AC), HcBGAL1 (BG), HcXTH1 (XTH), HcAlkIN1 (AlkN), and HcSUS1 (SUS) in different organs of 45-d-old seedlings of *H. courbaril*. Gene expression was followed using semi-quantitative RT-PCR during 24 h by sampling every 6 h. Solid white bars represent the dark periods and dashed lines the light periods during the experiment. Treatments were excision of the top shoot and inhibition of auxin polar transport by application of NPA. Densitometry was performed for each gel and essentially confirmed the patterns observed in this figure.

the gene was constitutively expressed and under NPA treatment it has been found to be down-regulated. HcSUS1 was expressed during both light and dark periods, with a peak of transcript at 18 h, and under excision and NPA treatment no transcripts were observed (Fig. 6C).

HcBGAL1 was up-regulated under constant light and down-regulated in the darkness (Fig. 7C). Under constant light, HcXTH1 showed a similar pattern to the control and was down-regulated in darkness (Fig. 7C). HcAlkIN1 was up-regulated in constant light and down-regulated in the dark. Under constant light, HcSUS1 was up-regulated, displaying constitutive expression throughout the day. Under darkness the gene was constitutively expressed from 0 h to 12 h. At 18 h it was down-regulated and transcripts appeared again at 24 h (Fig. 7C).

Leaves: HcBGAL1 was expressed during the light and dark periods, with a peak of transcription at 18 h in control seedlings. Under NPA treatment, the gene was down-regulated. HcXTH1 was expressed with a constitutive peak of transcript at 6, 12, and 18 h. HcAlkIN1 presented a peak of transcription at 6 h, and a peak of transcription of HcSUS1 was observed at 0 h and 24 h. NPA down-regulated all the four genes studied (Fig. 6D).

Under constant light, the HcBGAL1 gene was up-regulated, with peaks of transcripts every 12 h. In darkness, the gene was down-regulated. Under constant light, HcXTH1 of leaves was down-regulated during the day and under darkness the gene was strongly up-regulated (Fig. 7D). Under constant light, HcAlkIN1 behaved similarly to the control, whereas in darkness the gene was down-regulated (Fig. 7D). In constant light HcSUS1 was up-regulated and under darkness the gene was down-regulated (Fig. 7B).

All gels were analysed by densitometry (not shown), confirming the patterns described above.

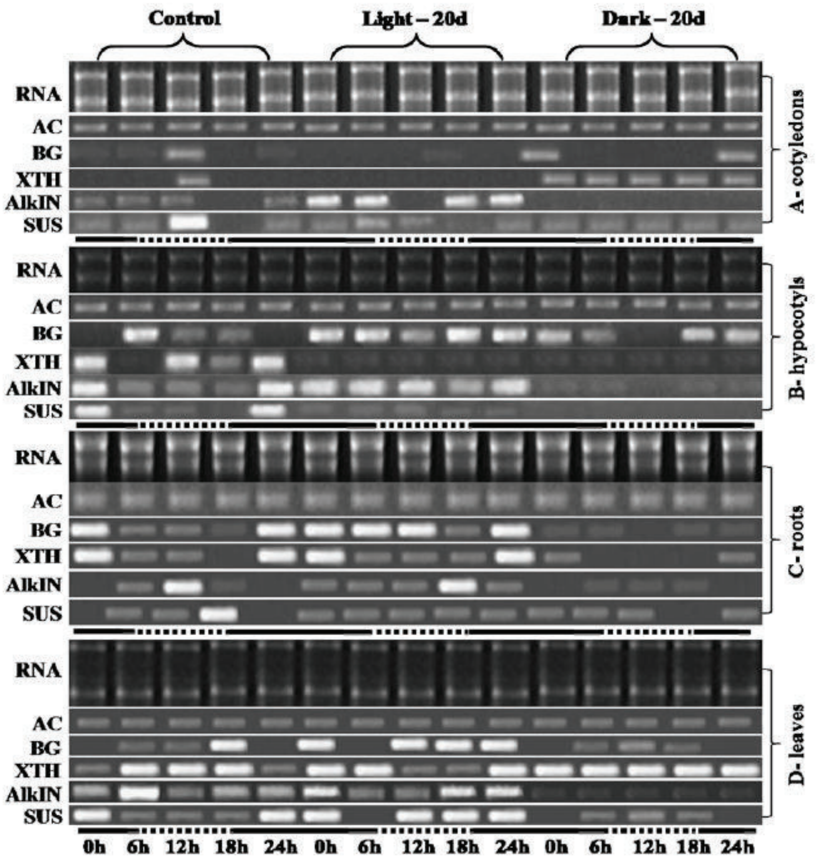


Fig. 7. Patterns of gene expression of actin (AC), HcBGAL1 (BG), HcXTH1 (XTH), HcAlkIN1 (AlkN), and HcSUS1 (SUS) in different organs of 45-d-old seedlings of *H. courbaril.* Gene expression was followed using semi-quantitative RT-PCR during 24 h by sampling every 6 h. White solid bars represent the dark periods and dashed lines the light periods during the experiment. After germination, developing seedlings were kept under continuous light or darkness for 20 d. Densitometry was performed for each gel and essentially confirmed the patterns observed in this figure.

Discussion

Phylogeny and putative biological roles for the four partial sequences cloned from H. courbaril

The cloning of the partial cDNA sequences of the β-galactosidase, XTH, invertase, and sucrose synthase genes from *H. courbaril* permitted phylogenetic analyses to be performed with other known genes (Figs 2–5, respectively).

β-Galactosidase: HcBGAL1 is close to TBG4, a gene belonging to the tomato gene family (Smith and Gross, 2000), and also to CanBGal3 from chickpea (Esteban *et al.*, 2003). TBG4 has the same sequence as the lupin exo-galactanase gene, which has been functionally well characterized as a cell wall gene directly related to cell wall storage mobilization (Buckeridge and Reid, 1994). The same is the case for CanBGal3, which has been experimentally associated with pectin degradation. In the case of HcBGAL1, it remains to be elucidated why this association with pectin-related genes occurs if the main storage polysaccharide is xyloglucan. A possible explanation is that this enzyme might be associated with granting access to xyloglucan-degrading enzymes during mobilization. Amaral (2005) found that the primary walls of the cotyledons of *H. courbaril* are rich in pectins and these walls might have to be modified during mobilization in order to grant access to storage xyloglucan. Alternatively, the preservation of sequence similarity might not be related to enzyme specificity, as in this work the full sequence was not cloned. Indeed, Alcântara *et al.* (1999, 2006) found unique β-galactosidases in xyloglucan-storing cotyledons of *C. langsdorffii* and *H. courbaril* (this enzyme was named *hcbetagal*), respectively. Sequencing of these proteins along with cloning of the full sequence of HcBGAL1 will probably provide the answer to whether or not *hcbetagal* (see also Fig. 8) and HcBGAL1 are the same entity.

XTH: The HcXTH1 gene displays closer similarity to nasturtium NXET1, a gene associated with endotransglycosylation, than to the NXG1, one of the first XET genes cloned from cotyledons of *T. majus* (De Silva *et al.*, 1993) that has been associated with a xyloglucan-storing system analogous to that of *H. courbaril.* Indeed, later, Rose *et al.* (1998) found that XET1 was expressed in all vegetative

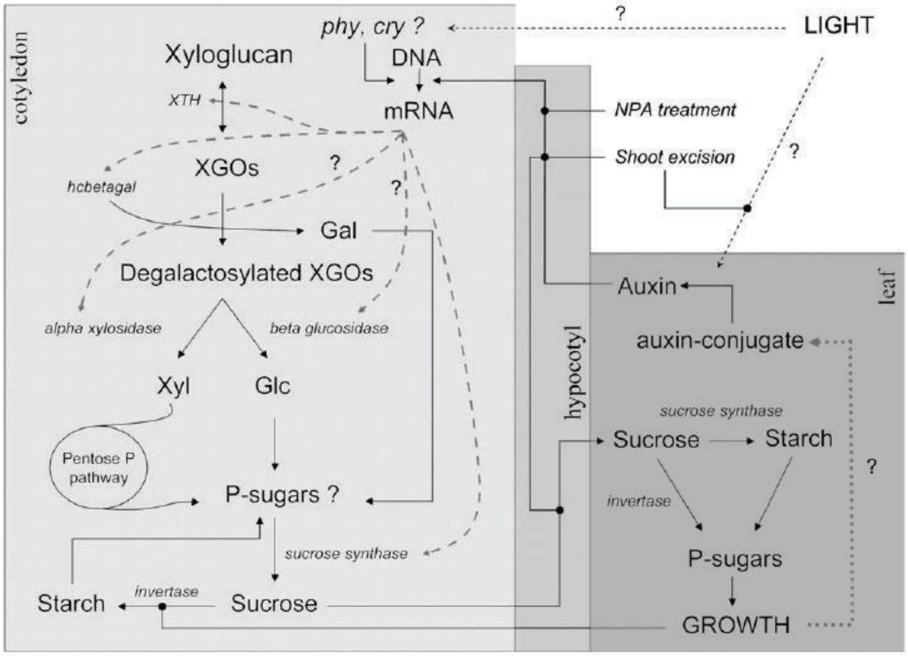


Fig. 8. Pathways of storage xyloglucan catabolism in the cotyledon of a developing seedling of *Hymenaea courbaril*. These steps are considered together as 'xyloglucan degradation'. The figure also shows some of the events that occur in the hypocotyl and developing leaves. The entire process that includes xyloglucan degradation, sucrose production, and transport through the hypocotyl towards the leaves is termed xyloglucan mobilization. The names of the enzymes are shown in italics. Hypothetically the following correlations between enzymes and genes exist: *XTH*=HcXTH1, *hcbetagal*=HcBGAL1, *invertase*=HcAlkIN1, and *sucrose synthase*=HcSUS1. Note that there are several points at which feedback control is present. The figure illustrates the points which are thought to control xyloglucan mobilization according to the experiments with application of NPA, top shoot excision, and presence or absence of light. Question marks were added to indicate hypothetical pathways that need further experiments. *phy*, phytochrome; *cry*, cryptochrome; Gal, galactose; Glc, glucose; Xyl, xylose.

tissues (root, epicotyl, stem, and leaf) except for germinating cotyledons of *T. majus*. Although the xyloglucan degradation system present in cotyledons of legumes (especially *H. courbaril*) has been associated with that present in nasturtium (Buckeridge *et al.*, 2000), it is becoming clear that many of the features of the former systems are quite different. Tiné *et al.* (2000) have already demonstrated that the principal activity of XET in *H. courbaril* needs a supply of oligosaccharides in order to hydrolyse its own storage xyloglucan, therefore being a true XTH. Alcântara (2000) isolated an XTH from cotyledons of *H. courbaril* and found that it is indeed an enzyme of the transglycosylase type and not an XET of the hydrolytic type as found by Fanutti *et al.* (1993) for *T. majus*. Thus, the finding of closer similarity of HcXTH1 to genes such as VaXTH1 and VaXTH2 is not surprising, especially if we consider that their expression is associated with the action of auxin, as demonstrated in this work. The detection of HcXTH1 in several organs of the seedling of *H. courbaril* cannot be considered as specifically marking a single gene, as the primers for PCR most probably allow amplification of cDNA of different closely related genes. The sequence DEIDFEFLGNRTG, according to Okazawa *et al.* (1993), is a conserved sequence present in most XETs found in nature. Therefore, the present detection of this conserved sequence in HcXTH1, along with expression of the gene in several organs of *H. courbaril*, suggests that the primer used is apparently not specific, but reflects the occurrence of expression of several isoforms of legume-specific XTHs. On the other hand, due to the fact that the main event taking place in the cotyledons of *H. courbaril* at the 45th day during seedling development is xyloglucan degradation (Santos and Buckeridge, 2004; Santos *et al.*, 2004), one can expect that the expression patterns observed, at least in this organ, reflect important events related to xyloglucan mobilization.

Alkaline invertase: In cotyledons of *H. courbaril*, relatively high concentrations of sucrose, glucose, and fructose were detected by Santos and Buckeridge (2004) during xyloglucan degradation. A search in the databases using HcAlkIN1 resulted in many sequences with considerable similarity, but there were only a few publications that could indicate a possible physiological role for this invertase. Because it is an alkaline/neutral invertase, it is possible that the product of the gene is active in the cytosol. In this context, its biochemical role might indeed be associated with sucrose degradation. However, it is not possible at this point to understand the role played by HcAlkIN1 in the process of storage mobilization, and further experiments will have to be performed to define the precise function of HcAlkIN1.

Sucrose synthase: One of the genes found to be closely related to HcSUS1 is MtSucS1 which, in *Medicago truncatula*, was found to be associated with the vascular systems in several parts of the plant as well as in developing seedlings (Hohnjec *et al.*, 1999). These authors conclude that the pattern of gene expression observed indicated an involvement of MtSucS1 with the generation of sink strength. This adds support to the hypothesis that HcSUS1 might be associated with sink strength in developing seedlings of *H. courbaril*.

Expression of sucrose synthase genes has been shown to be cell specific, developmentally regulated, or regulated by tissue carbohydrate status (Koch *et al.*, 1992; Ruan *et al.*, 1997). Most of the carbon produced during photosynthesis is channelled through the synthesis of sucrose, which is central to plant growth and development. Synthesis of sucrose certainly occurs in cotyledons of *H. courbaril*, as the products of xyloglucan degradation are thought to be rapidly metabolized into sucrose (Santos and Buckeridge, 2004).

Besides its role in determining accumulation of sucrose in certain tissues, sucrose synthase has also been related to synthesis of cell wall polysaccharides (Amor *et al.*, 1995; Buckeridge *et al.*, 1999). In these cases, the enzyme is thought to be associated with plasma or Golgi membranes and to control the production of UDP-glucose that is used as substrate for synthesis of cellulose, callose, and the mixed linkage β-glucan in grasses. In the present experiments involving excision of the shoot or application of NPA, in all cases of sink or transport organs (leaves, roots, and hypocotyl, respectively) inhibition of expression of HcSUS1 was observed. Thus, it is possible that the lower level of expression of this gene is associated with lower use of sugars to produce cell walls, as the growth rate decreases due to the lack of sugars coming from the storage mobilization organ (i.e. the cotyledon).

Auxin-induced expression patterns for cell wall-related genes

The importance of auxin in plant development has been highlighted by several authors (Romano *et al.*, 1995; Gray *et al.*, 1998; Reed, 2001; Rampey *et al.*, 2004). Auxin signalling has also been implicated as the principal regulatory factor for xyloglucan metabolism during cell wall loosening following development in several plant species (see, for example, Cosgrove, 1993; Catalá *et al.*, 1997; Nishitani, 1997). The regulation of gene expression by auxin has been extensively studied for decades (Theologis and Ray, 1982) and several families of auxin-regulated genes [e.g. members of the SAUR (small auxin-up-regulated) genes, the Aux/IAA gene, and GH3 families] (Catalá *et al.*, 2000; Nakazawa *et al.*, 2001; Reed, 2001; Liscum and Reed, 2002) have been identified. These rapidly induced, auxin-responsive mRNAs are under transcriptional control involving regulatory sequences in the promoter region (Abel and Theologis, 1996; Reed, 2001). Some of these genes controlled by auxin are expressed in distinct spatial and temporal patterns, thus underscoring the diversity of auxin responses in different plant tissues and organs (Abel and Theologies, 1996). However, the patterns of gene expression related to cell wall storage polysaccharides have not been described to date.

Genes that encode XTHs have been implicated by several studies in auxin-mediated regulation of xyloglucan metabolism (Nishitani and Masuda, 1981; Nishitani, 1995; Schindler *et al.*, 1995; Wu *et al.*, 1996; Xu *et al.*, 1996; Catalá *et al.*, 1997, 2001; Davies *et al.*, 1997; Akamatsu *et al.*, 1999; Yokoyama and Nishitani, 2000, 2001; Hyodo *et al.*, 2003; Nakamura *et al.*, 2003; Goda *et al.*, 2004; Yokoyama *et al.*, 2004). The application of indole acetic acid (IAA) on intact seedlings of *Arabidopsis thaliana* up-regulated expression of XTH members (Xu *et al.*, 1995, 1996; Yokoyama and Nishitani, 2001; Goda *et al.*, 2004; Vissenberg *et al.*, 2005). Catalá *et al.* (2000) characterized regulation of Cel7 by auxin during fruit development in tomato. Nakamura *et al.* (2003) analysed two azuki bean XTH genes and concluded that they are up-regulated by auxin but with different responses. Cui *et al.* (2005) observed up-regulation of XET in rice leaf. Osato *et al.* (2006) analysed some members of XTH in roots of *A. thaliana* and concluded that AtXTH19 was up-regulated by auxin.

Considering these observations, gene expression of two of the hydrolases (HcBGAL1 and HcXTH1) which are directly associated with xyloglucan storage degradation were analysed. Expression of genes related to sucrose metabolism in seedlings, i.e. invertase (HcAlkIN1) and sucrose synthase (HcSUS1), was also followed. Due to the observation of changes in the concentrations of sucrose, fructose, and glucose in xyloglucan-degrading cotyledons (Santos and Buckeridge, 2004), these two enzymes can be thought of as key for sucrose metabolism on the 45th day of development and consequently have important roles in the establishment of the seedlings of *H. courbaril*.

The pattern of expression observed in the control treatments of cotyledons suggests that the genes related to xyloglucan degradation (HcBGAL1 and HcXTH1) in seedlings of *H. courbaril* are dominantly expressed during the day time (peaks observed at 12 h) and sucrose metabolism genes (HcAlkIN1 and HcSUS1) are expressed during both the light and dark periods, but with peaks observed in the middle of the day (Figs 6A, 7A).

The fact that activity and protein related to β-galactosidase are found during the night (LIV Amaral and MS Buckeridge, unpublished) suggests that the gap between mRNA accumulation and enzyme production/activity in cotyledons of *H. courbaril* is of ~6 h. Auxin transport is apparently crucial for the xyloglucan degradation process in the cotyledons of *H. courbaril* as the present observations showed that no expression is detectable in these organs when auxin polar transport is inhibited by NPA and also when the sink organs (developing leaves) are excised (Fig. 6). Because these new leaves are thought to produce the auxin that is transported to the cotyledon and induces the activity of the enzymes (Santos *et al.*, 2004), it cannot be concluded from the shoot excision treatment that absence of sink strength inhibits expression of HcBGAL1 and HcXTH1. Thus, it remains to be seen whether sink strength or auxin production (or both) are the main factors that control HcXTH1 and HcBGAL1 expression.

In leaves of *H. courbaril*, HcSUS1 presented a constant level of expression during the whole day (Figs 6D, 7D). The observation that NPA inhibits expression of all genes is intriguing. It might possibly be attributed to the fact that there is no production and transport of sucrose from cotyledons to developing leaves, i.e. the maintenance of the source–sink relationship between the two organs is important to supply sucrose for developing organs.

The hypocotyl can be viewed as the means through which cotyledons (the source) transport synthesized sucrose towards the developing shoot (the main sink). The hypocotyl is also the way through which auxin produced in the top shoot reaches the storage-mobilizing cotyledons. In hypocotyls, HcBGAL1 and HcXTH1 are genes probably associated with cell elongation, one of the main events taking place in this organ. It may be speculated that the sucrose metabolism-related genes HcAlkIN1 and HcSUS1 are probably associated with the supply of energy for hypocotyl tissues. The expression levels of these two genes might also be associated (or at least correlated) with the intensity of transport of sucrose. Therefore, they may be used as an indication of the status of the source–sink relationship between growing tissues and storage degradation. The fact that all four genes had their expression levels severely decreased under the effect of NPA adds a new dimension to the interpretation of the effect observed by Santos *et al.* (2004). Auxin, produced in the shoot and transported to the cotyledons, not only seems to control expression of the genes related to xyloglucan degradation, but also seems to keep the sucrose metabolism in hypocotyls functional. A similar result was obtained when the top shoot was excised. However, HcXTH1 still displayed a high expression level after excision, suggesting that there are independent wall-related mechanisms different from the control by auxin.

Roots seem to be the organ with the lowest correlation with storage mobilization in seedlings of *H. courbaril*, as a lower proportion (~30%) of the xyloglucan degradation products end up in this organ during the whole period of development (Santos and Buckeridge, 2004). Even so, some strong effects of inhibition of polar transport of auxin were observed. Expression of HcBGAL1 was shut down after excision of the shoot, but not under treatment with NPA (Fig. 6C). In the case of HcXTH1, the reverse was observed, with maintenance of expression after excision of the shoots, but complete inhibition under treatment with NPA. These results reflect changes in the wall of roots that are apparently related to the shoot growth. As according to Santos and Buckeridge (2004) the developing leaves display a limited capacity to carry out photosynthesis at this stage of development, it is possible that in the present experiments roots were already becoming dependent on photosynthesis. This hypothesis is supported by the observation that HcSUS1, a gene that is possibly related to sucrose synthesis, was inhibited by both shoot excision and NPA treatment. Thus, the absence of auxin coming from the shoot eliminated expression of an important gene (sucrose synthase) thought to be related to the sink function of the organ (Fig. 6C).

Gene expression and light

Plants are sensitive to a number of abiotic environmental stimuli, including light, and related genes change their expression level in response to light. The existence of distinct gene sets that respond to different stimuli suggests that specific receptors and signal transduction pathways are utilized in response to alterations in light to drive distinct gene expression changes (Iliev *et al.*, 2002).

When a seed germinates in the forest understorey and a seedling starts to develop, a genetic programme named skotomorphogenesis is thought to be active (Alabbadí *et al.*, 2004). This genetic programme maintains hypocotyl elongation and inhibits expression of genes related to photosynthesis. When this seedling reaches a higher light intensity, photosynthesis establishes concomitantly with a rapid increase in leaf area, and hypocotyls cease to elongate.

Seedlings of *H. courbaril* can face microenvironmental conditions in the understorey of the tropical rain forests that are compatible with the maintenance of skotomorphogenesis. Seedlings of *H. courbaril* are usually quite large (30–40 cm tall; Fig. 1) but the light conditions in the understorey where they naturally develop can be as low as 22 μmol photons m^2 s^{-1} (Santos and Buckeridge, 2004). According to these authors, the lower the average light intensity, the slower will be the rate of storage mobilization and the corresponding relative growth rate. Furthermore, these authors reported that cotyledons are proportionally more important when light intensity decreases. Yet an important observation regarding the control of storage mobilization in seedlings of *H. courbaril* by light is that the importance of reserves in the less illuminated environment is directly related to the capacity of the newly developed leaves to establish photosynthesis, i.e. when in higher light intensity, seedlings become autotrophic earlier, becoming independent of the cotyledons sooner, which may induce an earlier fall of cotyledons.

That light interferes with xyloglucan mobilization in cotyledons of *H. courbaril* during seedling development is certain (Santos and Buckeridge, 2004). However, it is not yet known what type of light is involved and by what mechanism it regulates xyloglucan degradation. Although Santos and Buckeridge (2004) observed that a very low red/far-red ratio (0.5) did not increase seedling height significantly in relation to other red/far-red ratios (1.2 and 1.4), in a separate experiment, Santos *et al.* (2004) observed that when cotyledons were kept in the darkness (covered with aluminium foil during mobilization), xyloglucan hydrolase activities were significantly inhibited. These observations raise the hypothesis that light may have a somewhat direct control on xyloglucan mobilization at the cotyledon level, but at the same time is indirectly controlled by light through auxin production and transport in leaves.

To evaluate this hypothesis, in the present work, seedlings were kept in constant light and darkness during the storage mobilization period (from the 30th to the 60th day after the beginning of imbibition), and on the 45th day (the maximal rate of degradation of xyloglucan) gene expression was followed for 24 h.

In the literature, physiological experiments indicate that auxin is the major plant hormone closely connected with light transduction signals (Neff *et al.*, 1999; Steindler *et al.*, 1999). The expression of the AtHB2 gene is regulated by phytochrome, and this photoreceptor may regulate the transport of auxin through AtHB2 gene expression (Carabelli *et al.*, 1996). Reed (2001) suggests that light can also regulate Aux/IAA proteins by stabilizing them in *A. thaliana*. Vandenbussche *et al.* (2003) observed the induction of auxin-up-regulated genes in plants growing in lower light intensities. Braam and Davis (1990) and Xu *et al.* (1995) also observed that TCH4 encodes an XET that is also up-regulated by darkness and auxin. How these several stimuli lead to the common molecular response of TCH4 regulation is unknown.

The present experiments were not designed to test the involvement of phytochrome in the control of xyloglucan degradation, but the observations, together with previous observations by Santos and Buckeridge (2004) and Santos *et al.* (2004), suggest that this is a hypothesis worthwhile testing in future experiments. In general, the present results show that constant light or darkness had a profound effect on the expression of the four genes related to storage xyloglucan mobilization in cotyledons of *H. courbaril*. In cotyledons, where xyloglucan degradation takes place (Tiné *et al.*, 2000), expression of HcBGAL1 was inhibited by light, whereas the expression of HcXTH1 was induced in the dark. The latter observation is consistent with the literature mentioned above. On the other hand, in control seedlings, which were kept under a 12 h photoperiod, the expression of these two genes was apparently synchronized, with both peaking in the middle of the day. When constant light or darkness was imposed on seedlings during the period of storage mobilization, the expression of HcBGAL1 and HcXTH1 lost synchrony. The observation that the expression of these two genes becomes out of phase in constant light or darkness, according to the model proposed by Tiné *et al.* (2000), would disconnect the processes of degalactosylation from the process of transglycosylation, therefore halting the production of oligosaccharides that are the substrates for α-xylosidase and β-glucosidase. The disruption of the chain of biochemical events would then halt xyloglucan degradation. Indeed, cotyledons of seedlings grown in the dark were not capable of hydrolysing reserves, as they did not present the characteristic symptoms (i.e. shrunken cotyledons). Although the effect of different light intensities on gene expression was not tested, it is reasonable to speculate that the desynchronization between galactosylation and transglycosylation is likely to be at least part of the explanation for the delay in xyloglucan degradation in low light intensities observed by Santos and Buckeridge (2004). The lower production of auxin due to lower light intensity (LIV Amaral, HP Santos, and MS Buckeridge, unpublished results) might be the other part of the explanation for these observations. However, more experiments taking into consideration the addition of exogenous auxin to the excised seedling to replace the lack of the top shoot are necessary to clarify how different control systems interact.

Santos *et al.* (2004) proposed a model for storage xyloglucan mobilization in seedlings of *H. courbaril* in which a complex cross-talking network would in fact modulate the source–sink relationship between cotyledons and parts of the developing seedling. The present report extends this model for genes related to xyloglucan mobilization and sucrose metabolism (Fig. 8), showing for the first time that control at the transcription level is also important for xyloglucan storage mobilization. Auxin produced in leaves induces expression of hydrolases in the cotyledons. Hypothetically, auxin production could be related to growth as well as hydrolysis of its conjugated form (Woodward and Bartel, 2005). Regarding the effect of light, it is not possible to know for sure exactly where signalling occurs. Two hypotheses are possible, one being through transformation of conjugated to free IAA and the other related to action through phytochromes and/or cryptocrome. Further investigation will be necessary to test these hypotheses.

In conclusion, at least the four genes studied in this work have their expression controlled by auxin and light, which are both integrated in a communication network that seedlings of this important rain forest tree use to maximize physiological performance (i.e. carbon use efficiency) in the understorey of the very competitive rain forests such as the Atlantic and the Amazon forests.

Supplementary data

Supplementary data are available at *JXB* online.

Acknowledgements

The authors acknowledge financial support from the Ministry of Science and Technology of Brazil (project MCT-FUNDEPAG 2005.0188), FAPESP (grants 2004/10159-8 and 2007/59708-1) and CNPq, to ADB and TMQ for help in harvesting seedlings. AD Brandão was a doctoral student at Cell Biology Department, UNICAMP.

References

Abel S, Theologis A. 1996. Early genes and auxin action. *Plant Physiology* **111,** 9–17.

Akamatsu T, Hanzawa Y, Ohtake Y, Takahashi T, Nishitani K, Komeda YY. 1999. Expression of endoxyloglucan transferase genes in acaulis mutants of *Arabidopsis*. *Plant Physiology* **121,** 715–722.

Alabbadí D, Gil J, Blázquez MA, García-Martínez JL. 2004. Gibberellins repress photomorphogenesis in darkness. *Plant Physiology* **134,** 1050–1057.

Alcântara PHN. 2000. Isolamento e caracterização das enzimas xiloglucano endotransglicosilase e β-galactosidase do catabolismo do xiloglucano de reserva dos cotilédones de *Hymenaea courbaril* L. (Leguminosae-Caesalpinioideae). PhD thesis. Federal University of São Paulo (UNIFESP), São Paulo, Brazil.

Alcântara PHN, Dietrich SMC, Buckeridge MS. 1999. Xyloglucan mobilisation and purification of a (XLLG/XLXG) specific β-galactosidase from cotyledons of *Copaifera langsdorffii* Desf. (Leguminosae). *Plant Physiology and Biochemistry* **37,** 1–11.

Alcântara PHN, Martin L, Silva CO, Dietrich SMC, Buckeridge MS. 2006. Purification of a β-galactosidase from cotyledons of *Hymenaea courbaril* L. (Leguminosae). Enzyme properties and biological function. *Plant Physiology and Biochemistry* **44,** 619–627.

Amaral LIV. 2005. Metabolismo de carboidratos estruturais e de reserva em cotilédones de Hymenaea courbaril L var. stilbocarpa (Jatobá). PhD Thesis University of São Paulo, Department of Botany, São Paulo, Brazil.

Amor Y, Haigler CH, Johnson S, Wainscott M, Delmer DP. 1995. A membrane-associated form of SuSy and its potential role in synthesis of cellulose and callose in plants. *Proceedings of the National Academy of Sciences, USA* **92,** 9353–9357.

Braam J, Davis RW. 1990. Rain-, wind-, and touch-induced expression of calmodulin and calmodulin-related genes in *Arabidopsis*. *Cell* **60,** 357–364.

Buckeridge MS, Carpita NC, Vergara C. 1999. The mechanism of synthesis of a cereal mixed-linkage (1,3),(1,4)-β-D-glucan: evidence for multiple sites of glucosyl transfer in the synthase complex. *Plant Physiology* **120,** 1105–1116.

Buckeridge MS, Crombie HJ, Mendes CJM, Reid JSG, Gidley MJ, Vieira CJ. 1997. A new family of xyloglucan oligosaccharides from cotyledons of *Hymenaea courbaril*: structure determination of the oligosaccharide XXXXG by enzymatic sequencing. *Carbohydrate Research* **303,** 233–237.

Buckeridge MS, Dietrich SMC. 1990. Galactomannans from Brazilian legume seeds. *Revista Brasileira de Botanica* **13,** 109–112.

Buckeridge MS, Reid JSG. 1994. Purification and properties of a novel β-galactosidase or exo-β-(1,4)-galactanase from the cotyledons of germinated *Lupinus angustifolius* L. seeds. *Planta* **192,** 502–511.

Buckeridge MS, Rocha DC, Reid JSG, Dietrich SMC. 1992. Xyloglucan structure and post-germinative metabolism in seeds of *Copaifera langsdorffii* from savannah and forest populations. *Physiologia Plantarum* **86,** 145–151.

Buckeridge MS, Tiné MAS, Lima DU, Santos HP. 2000. Mobilisation of storage cell wall polysaccharides in seeds. *Plant Physiology and Biochemistry* **38,** 141–156.

Carabelli M, Morelli G, Whitelam G, Ruberti I. 1996. Twilight-zone and canopy shade induction of the ATHB-2 homeobox gene in green plants. *Proceedings of the National Academy of Sciences, USA* **93,** 3530–3535.

Catalá C, Rose JKC, Bennett AB. 1997. Auxin regulation and spatial localization of an endo-1,4-β-D-glucanase and a xyloglucan endotransglycosylase in expanding tomato hypocotyls. *The Plant Journal* **12,** 417–426.

Catalá C, Rose JKC, Bennett AB. 2000. Auxin-regulated genes encoding cell wall modifying proteins are expressed during early tomato fruit growth. *Plant Physiology* **122,** 527–534.

Catalá C, Rose JKC, York WS, Albersheim P, Darvill AG, Bennett AB. 2001. Characterization of a tomato xyloglucan endotransglycosylase gene that is down-regulated by auxin in etiolated hypocotyls. *Plant Physiology* **127,** 1180–1192.

Cosgrove DJ. 1993. Wall extensibility: its nature, measurement and relationship to plant cell growth. *New Phytologist* **124,** 1–23.

Cui D, Neill SJ, Tang Z, Cai W. 2005. Gibberellin-regulated XET is differentially induced by auxin in rice leaf sheath bases during gravitropic bending. *Journal of Experimental Botany* **56,** 1327–1334.

Daohong W, Bochu W, Biao L, Chuanren D, Jin Z. 2004. Extraction of total RNA from Chrysanthemum containing hight levels of phenolic and carbohydrates. *Colloids and Surfaces. B, Biosurfaces* **36,** 111–114.

Davies C, Boss PK, Robinson SP. 1997. Treatment of grape berries, a nonclimacteric fruit, with a synthetic auxin retards ripening and alters the expression of developmentally regulated genes. *Plant Physiology* **115,** 1155–1161.

De Silva J, Jarman CD, Arrowsmith DA, Stronach MS, Sidebottom C, Reid JSG. 1993. Molecular characterisation of a xyloglucan specific endo 1,4-β-D glucanase (xyloglucan endo-transglycosylase) from nasturtium seeds. *The Plant Journal* **3,** 701–711.

Esteban R, Dopico B, Muñoz FJ, Romo S, Martín I, Labrador E. 2003. Cloning of a *Cicer arietinum* β-galactosidase with pectin-degrading function. *Plant and Cell Physiology* **44,** 718–725.

Fanutti C, Gidley MJ, Reid JSG. 1993. Action of a pure xyloglucan endo-transglycosylase (formerly called xyloglucan-specific endo-(1,4)-β-D-glucanase) from the cotyledons of germinated nasturtion seeds. *The Plant Journal* **3,** 691–700.

Fry SC, York WS, Albersheim P, *et al*. 1993. An unambiguous nomenclature for xyloglucan-derived oligosaccharides. *Plant Physiology* **89,** 1–3.

Gerhardt K. 1993. Tree seedling development in tropical dry abandoned pasture and secondary forest in Costa Rica. *Journal of Vegetation Science* **4,** 95–102.

Gesteira AS, Micheli F, Ferreira CF, Cascardo CM. 2003. Isolation and purification of functional total RNA from different organs of cacao tree during its interaction with the pathogen *Crinipellis perniciosa*. *Biotechnology* **35,** 494–500.

Goda H, Sawa S, Asami T, Fujioka S, Shimada Y, Yoshida S. 2004. Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in *Arabidopsis*. *Plant Physiology* **134,** 1–19.

Gray WM, Ostin A, Sandberg G, Romano CP, Estelle M. 1998. High temperature promotes auxin-mediated hypocotyl elongation in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* **95,** 7197–7202.

Hohnjec N, Becker JD, Puhler A, Perlick AM, Kuster H. 1999. Genomic organization and expression properties of the MtSucS1 gene, which encodes a nodule-enhanced sucrose synthase in the model legume *Medicago truncatula*. *Molecular Genetics and Genomics* **261,** 514–522.

Hyodo H, Yamakawa S, Takeda Y, Tsuduki M, Yokota A, Nishitani K, Kohchi T. 2003. Active gene expression of a xyloglucan endotransglucosylase/hydrolase gene, XTH9, in inflorescence apices is related to cell elongation in *Arabidopsis thaliana*. *Plant Molecular Biology* **54,** 473–482.

Iliev EA, Xu W, Polisensky DH, OH M, Torisky RS, Clouse SD, Braam J. 2002. Transcriptional and posttranscriptional regulation of *Arabidopsis* TCH4 expression by diverse stimuli. Roles of cis regions and brassinosteroids. *Plant Physiology* **130,** 1–14.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8,** 275–282.

Koch KE, Nolte KD, Duke ER, McCarty DR, Avigne WT. 1992. Sugar levels modulate differential expression of maize sucrose synthase genes. *The Plant Cell* **4,** 59–69.

Lee YT, Langenheim JH. 1975. A systematic revision of the genus *Hymenaea* (Leguminosae; Caesalpinioideae; Detarieae). *University of California Publications in Botany* **69,** 1–109.

Levy S, Maclachlan G, Staehelin LA. 1997. Xyloglucan side chains modulate binding to cellulose during *in vitro* binding assays as predicted by conformational dynamics simulations. *The Plant Journal* **11,** 373–386.

Liscum E, Reed JW. 2002. Genetics of Aux/IAA and ARF action in plant growth and development. *Plant Molecular Biology* **49,** 387–400.

Nakamura T, Yokoyama R, Tomita E, Nishitani K. 2003. Two azuki bean XTH genes, VaXTH1 and VaXTH2, with similar tissue-specific expression profiles are differently regulated by auxin. *Plant and Cell Physiology* **44,** 16–24.

Nakazawa M, Yabe N, Ichikawa T, Yamamoto YY, Yoshizumi T, Hasunuma K, Matsui M. 2001. DFL1, an auxin-responsive GH3 gene homologue, negatively regulates shoot cell elongation and lateral root formation, and positively regulates the light response of hypocotyl length. *The Plant Journal* **25,** 213–221.

Neff MM, Nguyen SM, Malancharuvil EJ, *et al*. 1999. BAS1: a gene regulating brassinosteroid levels and light responsiveness in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* **96,** 15316–15323.

Nishitani K. 1997. The role of endoxyloglucan transferase in the organization of plant cell walls. *International Review of Cytology* **173,** 157–206.

Nishitani K. 1995. Endo-xyloglucan transferase, a new class of transferase involved in cell wall construction. *Journal of Plant Research* **108,** 137–148.

Nishitani K, Masuda Y. 1981. Auxin-induced changes in the cell wall structure: changes in the sugar compositions, intrinsic viscosity and molecular weight distributions of matrix polysaccharides of the epicotyl cell wall of *Vigna angularis*. *Physiologia Plantarum* **52,** 482–494.

Okazawa K, Sato Y, Nakagawa T, Asada K, Kato I, Tomita E, Nishitani K. 1993. Molecular cloning and cDNA sequencing of endoxyloglucan transferase, a novel class of glycosyltransferase that mediates molecular grafting between matrix polysaccharides in plant cell walls. *Journal of Biological Chemistey* **268,** 25364–25368.

Osato Y, Yokoyama R, Nishitani K. 2006. A principal role for AtXTH18 in Arabidopsis thaliana root growth: a functional analysis using RNAi plants. *Journal of Plant Research* **119,** 153–162.

Rampey RA, LeClere S, Kowalczyk M, Ljung K, Sandberg G, Bartel B. 2004. A family of auxin-conjugate hydrolases that contributes to free indole-3-acetic acid levels during *Arabidopsis* germination. *Plant Physiology* **135,** 978–988.

Reed JW. 2001. Roles and activities of Aux/IAA proteins in *Arabidopsis*. *Trends in Plant Science* **6,** 420–425.

Romano CP, Robson PRH, Smith H, Estelle M, Klee H. 1995. Transgene-mediated auxin overproduction in *Arabidopsis*: hypocotyl elongation phenotype and interactions with the hy6-1 hypocotyl elongation and axr1 auxin-resistant mutants. *Plant Molecular Biology* **27,** 1071–1083.

Rose JK, Brummell DA, Bennett AB. 1998. Two divergent xyloglucan endotransglycosylases exhibit mutually exclusive patterns of expression in nasturtium. *Plant Physiology* **110,** 493–499.

Ruan YL, Chourey PS, Delmer DP, Perez-Grau L. 1997. The differential expression of sucrose synthase in relation to diverse patterns of carbon partitioning in developing cotton seed. *Plant Physiology* **115,** 375–385.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4,** 406–425.

Santos HP, Buckeridge MS. 2004. The role of the storage carbon of cotyledons in the establishment of seedlings of *Hymenaea courbaril* L. under different light conditions. *Annals of Botany* **94,** 819–830.

Santos HP, Purgatto E, Mercier H, Buckeridge MS. 2004. The control of storage xyloglucan mobilisation in cotyledons of *Hymenaea courbaril* L. *Plant Physiology* **135,** 287–299.

Schindler T, Bergfeld R, Schopfer P. 1995. Arabinogalactan proteins in maize coleoptiles: developmental relationship to cell death during xylem differentiation but not to extension growth. *The Plant Journal* **7,** 25–36.

Smith DL, Gross KC. 2000. A family of at least seven β-galactosidase genes is expressed during tomato fruit development. *Plant Physiology* **123,** 1173–1183.

Souza RP, Valio IFM. 1999. Carbon translocation as affected by shade in saplings of shade tolerant and intolerant species. *Biologia Plantarum* **42,** 631–636.

Steindler C, Matteucci A, Sessa G, Weimar T, Ohgishi M, Aoyama T, Morelli G, Ruberti I. 1999. Shade avoidance responses are mediated by the ATHB-2 HD-Zip protein, a negative regulator of gene expression. *Development* **126,** 4235–4245.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24,** 1496–1599.

Theologis A, Ray P. 1982. Early auxin-regulated polyadenylylated mRNA sequences in pea stem tissue. *Proceedings of the National Academy Sciences, USA* **79,** 418–421.

Tiné MAS, Cortelazzo AL, Buckeridge MS. 2000. Xyloglucan mobilisation in cotyledons of developing plantlets of *Hymenaea* *courbaril* L. (Leguminosae-Caesalpinoideae). *Plant Science* **154,** 117–126.

Tiné MAS, Silva CO, Lima DU, Carpita NC, Buckeridge MS. 2006. Fine structure of a mixed-oligomer storage xyloglucan from seeds of *Hymenaea courbaril*. *Carbohydrate Polymers* **66,** 444–454.

Vandenbussche F, Vriezen WH, Smalle J, Laarhoven LJJ, Harren FJM, Straeten DVD. 2003. Ethylene and auxin control the *Arabidopsis* response to decreased light intensity. *Plant Physiology* **133,** 517–527.

Vissenberg K, Oyama M, Osato Y, Yokoyama R, Verbelen JP, Nishitani K. 2005. Differential expression of AtXTH17, AtXTH18, AtXTH19 and AtXTH20 genes in *Arabidopsis* roots. Physiological roles in specification in cell wall construction. *Plant and Cell Physiology* **46,** 192–200.

Woodward A, Bartel B. 2005. Auxin: regulation, action and interaction. *Annals of Botany* **95,** 707–735.

Wu SC, Blumer JM, Darvill AG, Albersheim P. 1996. Characterization of an endo-1,4-glucanase gene induced by auxin in elongating pea epicotyls. *Plant Physiology* **110,** 163–170.

Xu W, Campbell P, Vargheese AK, Braam J. 1996. The *Arabidopsis* XET-related gene family: environmental and hormonal regulation of expression. *The Plant Journal* **9,** 879–889.

Xu W, Purugganan MM, Polisensky DH, Antosiewicz DM, Fry SC, Braam J. 1995. *Arabidopsis* TCH4, regulated by hormones and the environment, encodes a xyloglucan endotransglycosylase. *The Plant Cell* **7,** 1555–1567.

Yokoyama R, Nishitani K. 2000. Functional diversity of xyloglucan-related proteins and its implications in the cell wall dynamics in plants. *Plant Biology* **2,** 598–604.

Yokoyama R, Nishitani K. 2001. A comprehensive expression analysis of all members of a gene family encoding cell-wall enzymes allowed us to predict cis-regulatory regions involved in cell-wall construction in specific organs of *Arabidopsis*. *Plant and Cell Physiology* **42,** 1025–1033.

Yokoyama R, Rose JKC, Nishitani K. 2004. A surprising diversity and abundance of XTHs (xyloglucan endotransglucosylase/hydrolases) in rice, classification and expression analysis. *Plant Physiology* **134,** 1088–1099.

BMC Genomics

Research article

Sugarcane genes associated with sucrose content

Flávia S Papini-Terzi[1], Flávia R Rocha[1], Ricardo ZN Vêncio[2], Juliana M Felix[3,5], Diana S Branco[3], Alessandro J Waclawovsky[1], Luiz EV Del Bem[3], Carolina G Lembke[1], Maximiller DL Costa[1], Milton Y Nishiyama Jr[1], Renato Vicentini[4,5], Michel GA Vincentz[3,4], Eugênio C Ulian[5,6], Marcelo Menossi[4] and Glaucia M Souza*[1]

Address: [1]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brazil, [2]BIOINFO-USP Núcleo de Pesquisas em Bioinformática, Universidade de São Paulo, São Paulo, SP, Brazil, [3]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas, SP, Brazil, [4]Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil, [5]Centro de Tecnologia Canavieira, Piracicaba, São Paulo, SP, Brazil and [6]Monsanto do Brasil Ltda, São Paulo, SP, Brazil

Email: Flávia S Papini-Terzi - flastal@yahoo.com.br; Flávia R Rocha - flavia.r.rocha@gmail.com; Ricardo ZN Vêncio - rvencio@gmail.com; Juliana M Felix - felix@ctc.com.br; Diana S Branco - dianabranco@gmail.com; Alessandro J Waclawovsky - ajw.bio@gmail.com; Luiz EV Del Bem - dudelben@gmail.com; Carolina G Lembke - carolina.lembke@gmail.com; Maximiller DL Costa - maximiller@gmail.com; Milton Y Nishiyama - yutakajr@iq.usp.br; Renato Vicentini - rvicentini@ctc.com.br; Michel GA Vincentz - mgavince@unicamp.br; Eugênio C Ulian - eugenio.ulian@gmail.com; Marcelo Menossi - menossi@unicamp.br; Glaucia M Souza* - glmsouza@iq.usp.br

* Corresponding author

Published: 21 March 2009

BMC Genomics 2009, **10**:120 doi:10.1186/1471-2164-10-120

Received: 14 November 2008
Accepted: 21 March 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/120

Abstract

Background -: Sucrose content is a highly desirable trait in sugarcane as the worldwide demand for cost-effective biofuels surges. Sugarcane cultivars differ in their capacity to accumulate sucrose and breeding programs routinely perform crosses to identify genotypes able to produce more sucrose. Sucrose content in the mature internodes reach around 20% of the culms dry weight. Genotypes in the populations reflect their genetic program and may display contrasting growth, development, and physiology, all of which affect carbohydrate metabolism. Few studies have profiled gene expression related to sugarcane's sugar content. The identification of signal transduction components and transcription factors that might regulate sugar accumulation is highly desirable if we are to improve this characteristic of sugarcane plants.

Results -: We have evaluated thirty genotypes that have different Brix (sugar) levels and identified genes differentially expressed in internodes using cDNA microarrays. These genes were compared to existing gene expression data for sugarcane plants subjected to diverse stress and hormone treatments. The comparisons revealed a strong overlap between the drought and sucrose-content datasets and a limited overlap with ABA signaling. Genes associated with sucrose content were extensively validated by qRT-PCR, which highlighted several protein kinases and transcription factors that are likely to be regulators of sucrose accumulation. The data also indicate that aquaporins, as well as lignin biosynthesis and cell wall metabolism genes, are strongly related to sucrose accumulation. Moreover, sucrose-associated genes were shown to be directly responsive to short term sucrose stimuli, confirming their role in sugar-related pathways.

Conclusion -: Gene expression analysis of sugarcane populations contrasting for sucrose content indicated a possible overlap with drought and cell wall metabolism processes and suggested signaling and transcriptional regulators to be used as molecular markers in breeding programs. Transgenic research is necessary to further clarify the role of the genes and define targets useful for sugarcane improvement programs based on transgenic plants.

Background

The importance of bioenergy-generating crops such as sugarcane is increasing rapidly and is likely to play an increasing role given the environmental and economical challenges of fossil fuel usage. Sugarcane belongs to the *Saccharum* L. genus, which derives from crosses of the domesticated species *S. officinarum* (a group that has sweet canes with thick and juicy culms), natural hybrids (*S. sinense* and *S. barberi*) and *S. spontaneum* (a wild species with no sugar and thin culms). All modern cultivars are derived from a few intercrossings of these hybrids [1-5]. Sucrose content is a phenotypic characteristic selected over centuries by breeding programs. Sugarcane cultivars differ in both maximum sucrose accumulation capacity and accumulation dynamics during growth [6]. Breeding programs routinely perform crosses to identify genotypes able to produce more sucrose early in the crop season to allow for continuous sugar production throughout the year. The internodes mature progressively towards the base of the culms with an increasing concentration of sucrose at the base. Sucrose content in the mature internodes can reach around 20% of the culms dry weight while lower sucrose levels are observed in younger internodes where glucose and fructose are predominant.

The improvement of modern cultivars could be achieved by identifying genes associated with important agronomic traits, such as sucrose content. These genes can then be used to generate transgenic plants or can serve as molecular markers for map-assisted breeding [7]. Internodes have been expression-profiled during culm development [8-12], but differences between cultivars that contrast for sucrose content have not been extensively reported. Understanding differences in the regulation of genes related directly or indirectly to sucrose accumulation in different cultivars is an important step if we want to aid breeding for sugar yield improvement. It is also important to understand the impact of environmental stresses on sucrose accumulation and the role of hormones in integrating stress signaling and developmental cues. Water stress, for example, reduces yield drastically and therefore, drought-tolerant sugarcane cultivars might be critically important in a scenario of cultivation expansion since much of the land available for sugarcane cultivation is located in regions subjected to drought. Drought responses include immediate protective measures and long term growth alterations [13]. Modulation of gene expression under this stress [14-19] involves ABA-dependent and independent pathways [13]. Carbohydrate metabolism is also related to abiotic stress responses since some aspects of the regulation of sugar metabolism are mediated by ABA and fructose, raffinose and trehalose act as osmoprotectants [20]. It is important to emphasize that some sugars (such as glucose, trehalose and sucrose) are important signaling molecules that affect plant growth and development including germination, early vegetative growth and flowering, as well as a variety of physiological processes such as photosynthesis, resource partition and defense responses [21-26]. The pathways activated by sugars cross-talk with other pathways, including those related to hormonal, cell cycle control and nitrogen responses [27-30]. ABA and sucrose were shown to be involved in the control of sucrose levels in plant cells [21] but the underlying mechanisms are still unknown.

We previously used cDNA microarrays to identify sugarcane genes that are responsive to drought and ABA [31]. The cDNAs are derived from a collection of 237,954 ESTs developed by the SUCEST sugarcane EST project [32] which were assembled into 43,141 putative, unique sugarcane transcripts that are referred to as Sugarcane Assembled Sequences (SAS). In this report we present the results of a large-scale analysis of the transcriptome of thirty genotypes grown in the field. cDNA microarrays were used to compare high- and low-Brix individuals and a comparison was made to reveal gene expression patterns that correlate with sucrose content, culm development, sugar treatments, drought and ABA treatment. We performed an extensive validation of cDNA microarray data using pooled plants, as well as individual genotypes. The results indicate a close relationship between sucrose content and drought signaling.

Results

cDNA microarrays were used to identify genes that were differentially expressed in genotypes contrasting for sucrose content. The arrays preparation, validation and analysis were done as previously described [31]. Multiple crossings were performed for twelve years among *S. officinarum* and *S. spontaneum* (Population 1) and between commercial varieties SP80-180 and SP80-4966 (Population 2) to generate genotypes with extreme values of sugar content. The simplest way to access phenotypic differences with a high degree of confidence is to measure sucrose in the culm juice. This can be done in the field using a simple refractometer that evaluates Brix (soluble solids content). In sugarcane most of the soluble solids in the juice (70 to 91%) correspond to sucrose. Using this approach, thousands of genotypes can be phenotyped and contrasting individuals among the populations can be selected for further agronomic evaluation. Brix measurements were taken from 500 individuals of each population and the extreme clones in this population were selected and evaluated for sucrose content (see Additional file 1). To evaluate gene expression samples were collected from single individuals as well as from pools of seven or eight plants grown for seven, ten and elevenmonths.

Two experimental designs were used to perform transcriptome comparisons: (I), internodes 1, 5 and 9 from high

Brix plants were compared to the same internodes from low Brix plants (HB vs LB) in both populations or (II), mature internodes 9 were compared to immature internodes 1 from plants with high or low Brix in population 2 [33]. Twenty six hybridizations were performed revealing 239 genes associated with sucrose content and regulated during culm development (see Additional file 2 and Figure 1).

A total of 117 genes were found to be differentially expressed in at least one comparison between high and low Brix genotypes (internodes 1, 5 or 9), and ten genes (SCCCLR1048F03.g, SCCCLR2003E10.g, SCCCRZ1001 F02.g, SCCCRZ1001H05.g, SCCCRZ1002E08.g, SCEZS T3147A10.g, SCJFRZ2007F10.g, SCAGLR1043E04.g, SCS BHR1050B11.g and SCVPCL6041F01.g) were found to be differentially expressed in both populations analyzed (see Additional file 2). Among these SAS, we found three transcription factors, two aquaporins and two transcripts related to development. The gene expression comparison between mature and immature internodes showed a total of 173 differentially expressed genes (see Additional file 2 and Figure 1).

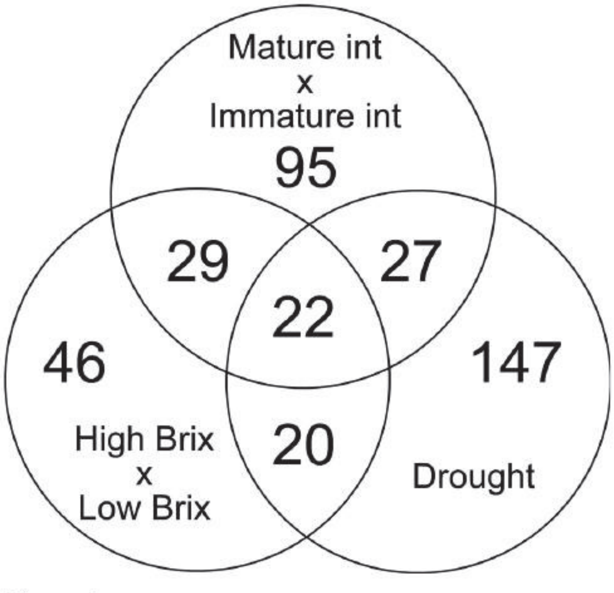


Figure 1

Comparison of differential gene expression associated with sucrose content, culm development and drought responses in sugarcane. Genes were identified as associated with sucrose content if they were differentially expressed when high Brix or low Brix pools of plants were compared. Genes regulated during culm development were identified by comparing Mature and Immature Internodes. The drought-responsive genes were found to be induced or repressed by drought after 24, 72 or 120 h of water deficit. The figure represents a Venn diagram of the three differential expression data sets. Technical replicates range from 2 to 16 since genes are spotted several times in the same array. The credibility level used to define outliers was 0.96 in all three data sets.

Table 1 lists a selection of the differentially expressed genes along with the number of biological samples that displayed altered expression when high and low Brix pools of plants were compared (HB vs LB) and when mature and immature internodes were compared (MI vs II). The expression data sets were compared to those obtained for plants exposed to drought conditions or ABA treatment [31] (see Additional file 2). Comparison to ABA treated plants yielded eleven differentially expressed genes in common, including the *ScPKABA1-3* (SCRFL R1034G06.g) and the *ScMAPK-4* (SCSBAM1084E01.g), which were both more expressed in high Brix and repressed by ABA, and a PP2C-like protein phosphatase (SCEPRZ1010E06.g) which showed the opposite profile. Comparison to drought-regulated genes showed an extensive overlap in differential expression between the two datasets. Between 117 and 173 genes associated with high sucrose content and internode development, respectively, 43.6% and 28.3% were previously shown to be altered by drought while twenty-two genes were altered in all conditions analyzed (Figure 1).

Expression data of forty-two genes was also obtained using qRT-PCR. We determined gene expression differences for pools of extreme individuals from both populations (Figure 2), in mature and immature internodes (Figure 3) and in response to drought and ABA treatment (Figure 4). The significance of the data obtained by qRT-PCR was inferred statistically by calculating values of P for expression differences against the reference sample (see Methods for details). Overall gene expression data obtained using cDNA microarrays was confirmed in qRT-PCR experiments for over 80% of the genes tested, even when the target RNA derived from a distinct biological replicate. We also investigated, using qRT-PCR, how the expression levels varied among the individual genotypes from Population 1 (Figure 5). In this case, the value of P was calculated against the average expression level across genotypes and the validation rate was around 58%. Additional file 3 lists all the values of P for the validated genes.

In order to unravel signaling aspects of sucrose accumulation, we asked whether genes differentially expressed in contrasting Brix genotypes or in mature-versus-immature internodes could represent direct sucrose- and/or glucose-regulated genes and, therefore, be part of the sucrose- and glucose-response pathways. To this end, sugarcane seedlings were treated with 3% sucrose or 3% glucose for 4 h and the expression of thirty-four genes was analyzed by qRT-PCR. The expression of thirty of these genes was

Table 1: Selection of SAS showing differential expression when high and low Brix plants were compared or when mature and immature internodes were compared.

SAS	category	sub category 1	sub category 2	HB vs LB	MI vs II	Drought	ABA	Suc	Gluc
SCCCLR1022D05.g	adapter	14-3-3 protein	GF14		↓				
SCCCRZ1001D02.g	adapter	14-3-3 protein	GF14		↓↓↓↓				
SCEQRT1031D02.g	adapter	14-3-3 protein	GF14		↓↓				
SCEQRT1025D06.g	adapter	14-3-3 protein	GF14		↓	↓↓			
SCVPLR1049C09.g	calcium metabolism	calmodulin-binding protein	AAA family ATPase (cell division cycle protein 48 sub-family)		↓				
SCCCRZ1C01H06.g	calcium metabolism	calmodulin-binding protein	Apyrase (Nucleoside diphosphatase)		↓↓↓↓	↓↓			
SCJLLR1108H07.g	calcium metabolism	calmodulin-binding protein	Ca(2+)-ATPase		↓				
SCEZLB1012F10.g	calcium metabolism	calmodulin-binding protein	Cyclic nucleotide-gated calmodulin-binding ion channel	↑					
SCCCAM1001A03.g	calcium metabolism	calmodulin-binding protein	Multidrug resistant (MDR) ABC transporter	↑				↑↑	
SCRFLR2037F09.g	calcium metabolism	calreticulin	CRT2 Calreticulin 2	↓	↓↓	↑↑		↑↑↑	↑↑↑
SCCCLR2C02A05.g	cell wall metabolism	expansin	EXPA11	↓	↓↓				
SCQGRT1040G03.g	cell wall metabolism	expansin	OsEXPA23	↑		↓↓			
SCACSB1037A07.g	cell wall metabolism	cytochrome P450	P-coumaroyl shikimate 3'-hydroxylase	↓↓					
SCEZHR1087F06.g	cell wall metabolism	cytochrome P450	Ferulate-5-hydroxylase	↓	↑↑				
SCSGFL4193B05.g	cell wall metabolism	cytochrome P450	Cinnamic acid 4-hydroxylase	↓					
SCRFLR1012F12.g	cell wall metabolism	.	Caffeic acid 3-O-methyltransferase	↑↑	↑↑				
SCBFLR1039B05.g	cell wall metabolism	polysaccharide metabolism	Xyloglucan endotransglycosylase		↓↓↓↓				
SCCCLR1048D07.g	cell wall metabolism	lignin	Phenylalanine ammonia-lyase	↑		↓↓			
SCEQRT1024E12.g	cell wall metabolism	lignin	Phenylalanine ammonia-lyase	↑	↓	↓↓	↑↑	↑↑↑	↑↑↑
SCSGAM1094D05.g	cell wall metabolism	lignin	Phenylalanine ammonia-lyase	↓	↓				
SCCCCL6002B05.g	hormone biosynthesis	auxin	Nitrilase	↑		↑↑			
SCEQRT1028H06.g	hormone biosynthesis	auxin	Nitrilase		↓↓	↑↑			
SCRFLR1012D12.g	hormone biosynthesis	auxin	Nitrilase	↑	↓	↑↑			
SCVPLR2012A10.g	hormone biosynthesis	ethylene	ACC oxidase	↑	↓↓				
SCCCRT1001E01.g	hormone biosynthesis	jasmonic acid	Lipoxygenase	↓	↓↓↓↓	↓↓		↓↓↓	↓↓↓
SCJFRT1007H07.g	hormone biosynthesis	jasmonic acid	Lipoxygenase	↓	↓				
SCCCLR1C03G01.g	hormone biosynthesis	jasmonic acid	Omega-6 fatty acid desaturase	↓	↓	↑↑	↑↑		
SCCCAM2004G02.g	hormone-related	auxin	Auxin transport/ auxin eflux carrier (OsPIN1c)	↓					
SCCCLR2002F08.g	hormone-related	auxin	dormancy/auxin associated family (auxin-repressed)	↓	↑↑				

Table 1: Selection of SAS showing differential expression when high and low Brix plants were compared or when mature and immature internodes were compared. *(Continued)*

SCBGLR1023D05.g	pathogenicity	R-gene transduction	Zinc finger protein (LSD1)	↑	↓↓↓	↓↓		↑↑↑	
SCAGLR1043F02.g	protein metabolism	calmodulin-binding protein	HSP70 (heat shock)	↑↑	↓	↑↑			
SCCCCL3120G07.g	protein metabolism	calmodulin-binding protein	HSP70 (heat shock)	↑		↑↑			
SCCCRZ1003A03.g	protein metabolism	calmodulin-binding protein	HSP70 (heat shock)		↑				
SCEQRT2099H01.g	protein kinase	calcium-dependent	ScCDPK-27		↓				
SCVPAM1055A12.g	protein kinase	casein kinase	ScCKI-11	↑	↓	↑↑			
SCCCLR1C04G08.g	protein kinase	casein kinase	ScCKI-3	↑					
SCCCLR1022H07.g	protein kinase	cell cycle-related	ScCDK-11		↓				
SCBGLR1096C08.g	protein kinase	cell cycle-related	ScCDK-18		↓				
SCVPRT2081G05.g	protein kinase	cell cycle-related	ScCDK-3		↓				
SCRLFL1012B10.g	protein kinase	cell cycle-related	ScCDK-6		↓				
SCSBAM1084E01.g	protein kinase	MAPK/MAPKK/ MAPKKK	ScMAPK-4	↑	↑↑		↓↓		
SCEPAM1020A03.g	protein kinase	other	ScATN1-2	↓					
SCVPCL6042B07.g	protein kinase	other	ScCyclin G-associated kinase-like protein-1		↓				
SCJFRZ2032C08.g	protein kinase	SNF-like kinase	ScCIPK-14		↑	↑↑			
SCBFSB1046D04.g	protein kinase	SNF-like kinase	ScCIPK-16	↑				↓↓↓	
SCMCRT2103B04.g	protein kinase	SNF-like kinase	ScCIPK-21	↑↑	↑	↓↓		↓↓	
SCCCLR1C05B07.g	protein kinase	SNF-like kinase	ScCIPK-3	↑		↑↑		↓↓↓	↓↓↓
SCJLRZ1023H04.g	protein kinase	SNF-like kinase	ScCIPK-9		↓↓	↓↓			
SCEPRZ1009C10.g	protein kinase	SNF-like kinase	ScOSA PK-1		↓↓			↓↓	↓↓
SCCCST1004A07.g	protein kinase	SNF-like kinase	ScOSA PK-7		↓				
SCACLR2007G02.g	protein kinase	SNF-like kinase	ScPKABA1-1	↑↑				↓↓	↓↓↓
SCRFLR1034G06.g	protein kinase	SNF-like kinase	ScPKABA1-3	↑			↓↓	↓↓↓	↓↓↓
SCJFRZ2032G01.g	protein kinase	SNF-like kinase	ScSnRK1-2		↓↓	↓↓		↑	
SCCCCL5002B10.g	protein kinase	undefined	ScPK-BI2		↓↓↓				
SCJLLR1054C03.g	protein kinase	undefined	ScPK-BIII7	↑					
SCMCSD2061D05.g	protein kinase	undefined unclassified	ScUPK-46 (CIPK)	↓					
SCCCLB1001D03.g	protein phosphatase	serine/threonine PPM family	PP2A/Catalytic Subunit		↓				
SCEZLR1052F07.g	protein phosphatase	serine/threonine PPM family	PP2A/Subunit A		↓				
SCEPRZ1010E06.g	protein phosphatase	serine/threonine PPM family	PP2C-like	↓↓	↓	↑↑	↑↑	↓↓↓	↓↓↓
SCEZHR1088E02.g	protein phosphatase	tyrosine phosphatase	Dual Specificity Protein Phosphatases (DSPP)	↑↑↑	↓	↑↑		↓↓↓	↓↓↓
SCMCST1051F08.g	protein phosphatase	tyrosine phosphatase	Tyrosine Specific Protein Phosphatases (PTP)			↓↓			
SCSBHR1056H08.g	receptor	ethylene	EIN2		↑				
SCUTLR2023D06.g	transcription factor	CCAAT	ScCA2P5	↑					
SCCCLR1066G08.g	transcription factor	HGM (high mobility group protein)		↑		↓↓			
SCBFAD1046D01.g	transcription factor	HLH (helix-loop-helix)	ScbHLH1		↓↓				
SCCCRZ1001H05.g	transcription factor	HLH (helix-loop-helix)	ScbHLH7	↑↑		↓↓		↓↓↓	↓↓↓
SCAGLR1021G10.g	transcription factor	homeobox	ScHB2		↓↓			↓↓↓	↓↓↓
SCRLAM1010D08.g	transcription factor	homeobox	ScHB41		↓↓				
SCEZLB1010E10.g	transcription factor	hormone-related/ auxin	ScABI40	↓					

Table 1: Selection of SAS showing differential expression when high and low Brix plants were compared or when mature and immature internodes were compared. *(Continued)*

SCCCLR1024F10.g	transcription factor	hormone-related/ auxin	ScARF46		↓			
SCCCRZ1001G10.g	transcription factor	hormone-related/ Aux/IAA	ScAUX1134		↓↓↓↓	↓↓	↓↓↓	↓↓↓
SCVPLR2005H03.g	transcription factor	hormone-related/ Aux/IAA			↓↓			
SCJFRZ2009F04.g	factor transcription	hormone-related/ Aux/IAA			↓			
SCJLLR1054C09.g	transcription factor	hormone-related/ Aux/IAA			↓↓			
SCUTST3086B02.g	transcription factor	hormone-related/ ethylene/AP2/EREBP	ScEREB59		↓			
SCCCLR1001D10.g	transcription factor	hormone-related/ ethylene/AP2/EREBP	DRE binding factor 2		↑	↑↑		
SCBGFL4052C11.g	transcription factor	hormone-related/ ethylene	ScEIL1		↓			
SCCCRZ1004H12.g	transcription factor	hormone-related/ ethylene	ScEIL2	↓	↓			
SCCCRZ2C03D11.g	transcription factor	hormone-related/ gibberellin	ScGRAS71		↓↓			
SCEPRZ1008F02.g	transcription factor	LIM (protein-protein interaction)		↓	↓↓			
SCQGLR1085G10.g	transcription factor	MADS	ScMADS17		↑	↓↓		
SCSFAD1124E07.g	transcription factor	MYB	ScMYB70		↑			
SCRURT2010A10.g	transcription factor	MYB	ScMYB120		↓			
SCCCLR2003E10.g	transcription factor	NAM (no apical meristem)	ScNAC27	↓↓			↓↓↓	↓↓↓
SCRUAD1132D09.g	transcription factor	NAM (no apical meristem)	ScNAC51					
SCACLR1130H08.g	transcription factor	zinc finger protein	ScYAB16		↓			
SCEZST3147A10.g	transcription factor	zinc finger protein	ScC3H84	↓	↓			
SCCCCL4003D08.g	transcription factor	zinc finger protein	ScC3H95		↓			
SCQGRZ3011D06.g	transcription factor	zinc finger protein/ alfin-like	ScALF9		↓			
SCCCRZ1002E08.g	stress	drought and cold response	Aquaporin (plasma membrane)	↓	↓↓			
SCCCST3001H12.g	stress	drought and cold response	Aquaporin (plasma membrane)	↑	↓↓			
SCEQRT2100B02.g	stress	drought and cold response	Aquaporin (plasma membrane)	↑	↓↓			
SCCCLR1024C03.g	stress	drought and cold response	Aquaporin (tonoplast intrinsic protein)	↓	↓			
SCCCRZ1001F02.g	stress	drought and cold response	Aquaporin (tonoplast intrinsic protein)	↓	↓			
SCQGLR1085F11.g	stress	drought-induced	Dehydrin	↓	↓↓↓	↑↑	↓↓↓	↓↓↓
SCCCLR2C01F06.g	stress	wound-induced	wound-responsive family protein	↑↑↑	↑		↑↑↑	↑↑↑

The table also shows differential expression of the same SAS as seen in [31] for plants submitted to drought and ABA treatment. Differential expression refers to cDNA microarray analysis except for the last two columns, which refer to qRT-PCR data obtained in samples of plantlets treated with sucrose or glucose. The table lists a selection of SAS whose expression was enriched or decreased in two technical replicates for each biological sample. For a complete list see additional file 2. The up arrow indicates that the SAS is more expressed, the down arrow indicates that the SAS is less expressed. The number of arrows indicates the number of hybridizations.

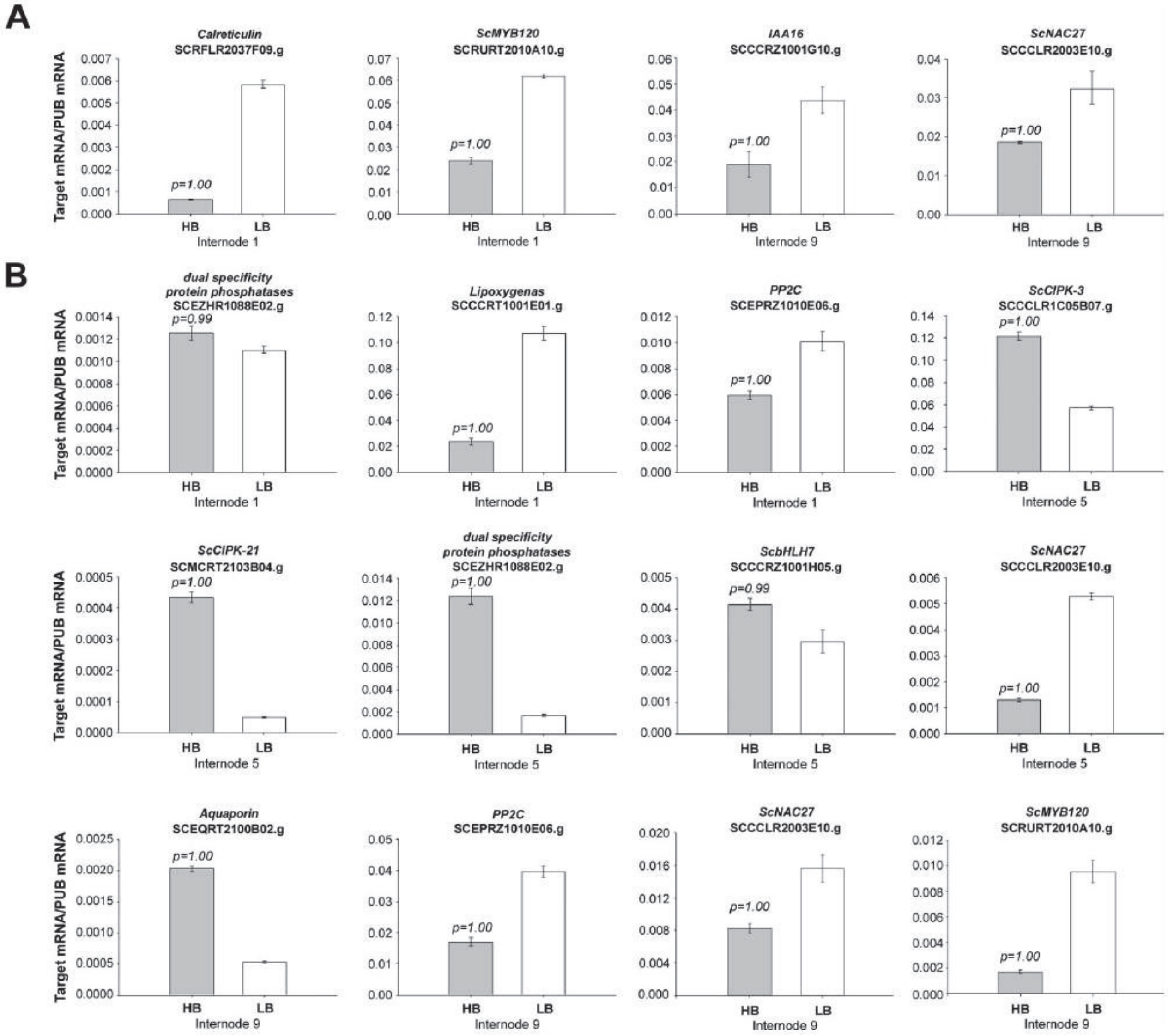


Figure 2

Real Time PCR (qRT-PCR) analysis of Populations gene expression. The y axis refers to the relative expression ratio between target mRNA versus the reference mRNA (polyubiquitin-PUB SCCCST2001G02.g). The relative expression levels were determined in Internode 1, 5 and 9 tissues from a pool of the eight individuals with the highest Brix measures (HB) and the eight individuals with the lowest Brix measures (LB) from Population 1 (A) and from a pool of the seven individuals with the highest Brix measures (HB) and the seven individuals with the lowest Brix measures (LB) from Population 2 (B). The reactions for the target mRNA and reference mRNA were carried out in parallel and each reaction was performed in triplicates. Error bars were calculated as described previously [31]. The transcript levels for the reference genes were verified not to vary in response to the treatments. The values of P correspond to the probability Pr(HB>LB) and Pr(HB<LB) for up- and down-regulated genes, respectively. The SAS was considered differentially expressed when P ≥ 0.95.

affected by sucrose, of which six were also found to be regulated by 3% manitol (osmotic control) and thus, were not considered as true sucrose-responsive genes (see Additional file 3). Figure 6 shows the expression pattern of fifteen of these genes. Among the twenty-four sucrose-regulated genes, nineteen were also found to respond to glucose, indicating a significant overlap between these two signaling pathways (see Additional file 3 and Figure 6). This is not unexpected since sucrose can be readily converted to glucose and sucrose-specific responsive pathways have been identified previously. The five genes, identified here as genuine sucrose-regulated genes, include

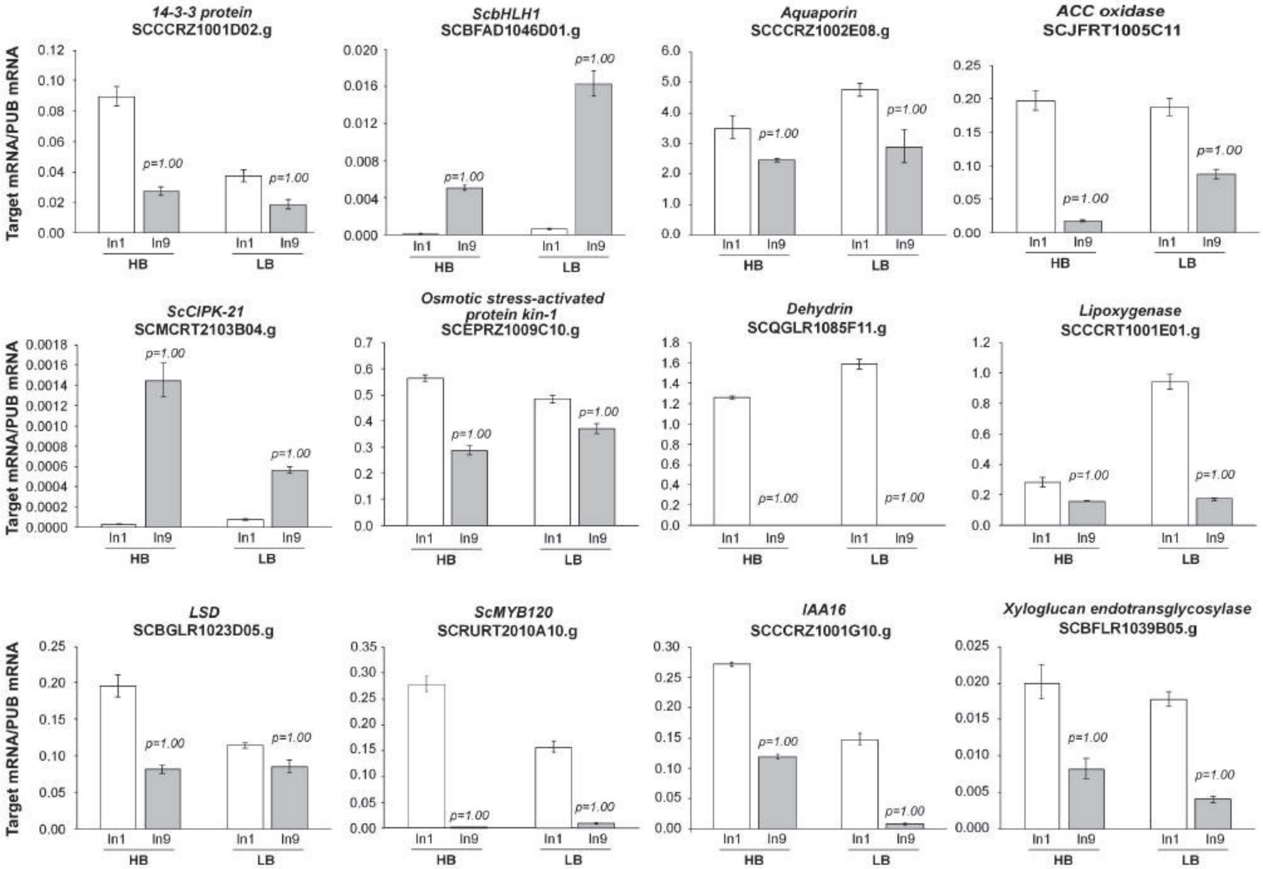


Figure 3

Real Time PCR (qRT-PCR) analysis of internode developmental gene expression. The y axis refers to the relative expression ratio between target mRNA versus the reference mRNA (polyubiquitin SCCCST2001G02.g). The relative expression levels were determined in Internode 1 and 9 tissues from a pool of the seven individuals with the highest Brix measures (HB) and the seven individuals with the lowest Brix measures (LB) of Population 2. All reactions were carried out in parallel and each reaction was performed in triplicates. Error bars were calculated as described previously [31]. The transcript levels for the reference genes were verified to not vary in response to the treatments. The P values correspond to the probability Pr(MI>II) and Pr(MI<II) for up- and down-regulated genes, respectively when In9 and In1 samples were compared. The values of P were calculated for the HB and LB pools of plants independently. The SAS was considered differentially expressed when P ≥ 0.95.

three SNF1-like kinases, a pathogen-response related protein and a multidrug resistance ABC transporter (see Additional file 3). A weak overlap with ABA signaling was detected, since only three sucrose/glucose-regulated genes were also modulated by ABA (Table 1). Finally, we noticed that thirteen of the twenty-four genes exhibited opposite regulatory responses in high Brix genotypes and/or mature internodes as compared to the short-term sugar-induced regulation in seedlings (data not shown). Together, these data establish the existence of a correlation between high sucrose content and early sucrose and/or glucose-responsive genes, some of which may be relays of signal transduction pathways triggered by these sugars.

In addition, we sought to obtain some insight into the extent to which the short term sucrose and/or glucose regulatory cascade is conserved between sugarcane, a monocot and *Arabidopsis thaliana* (Arabidopsis), a model eudicot organism. Therefore, we compared the data obtained in this study on sugarcane seedlings with results described for *Arabidopsis* seedlings under similar experimental conditions (3% glucose [30] or 0,5% sucrose [34]). Among the twenty-four sugar-regulated sugarcane genes, six of them, along with their eight orthologues in *Arabidopsis* (forming five groups of orthologues) were found to be similarly regulated by glucose and/or sucrose (see Additional file 4 and Additional file 5). The groups of

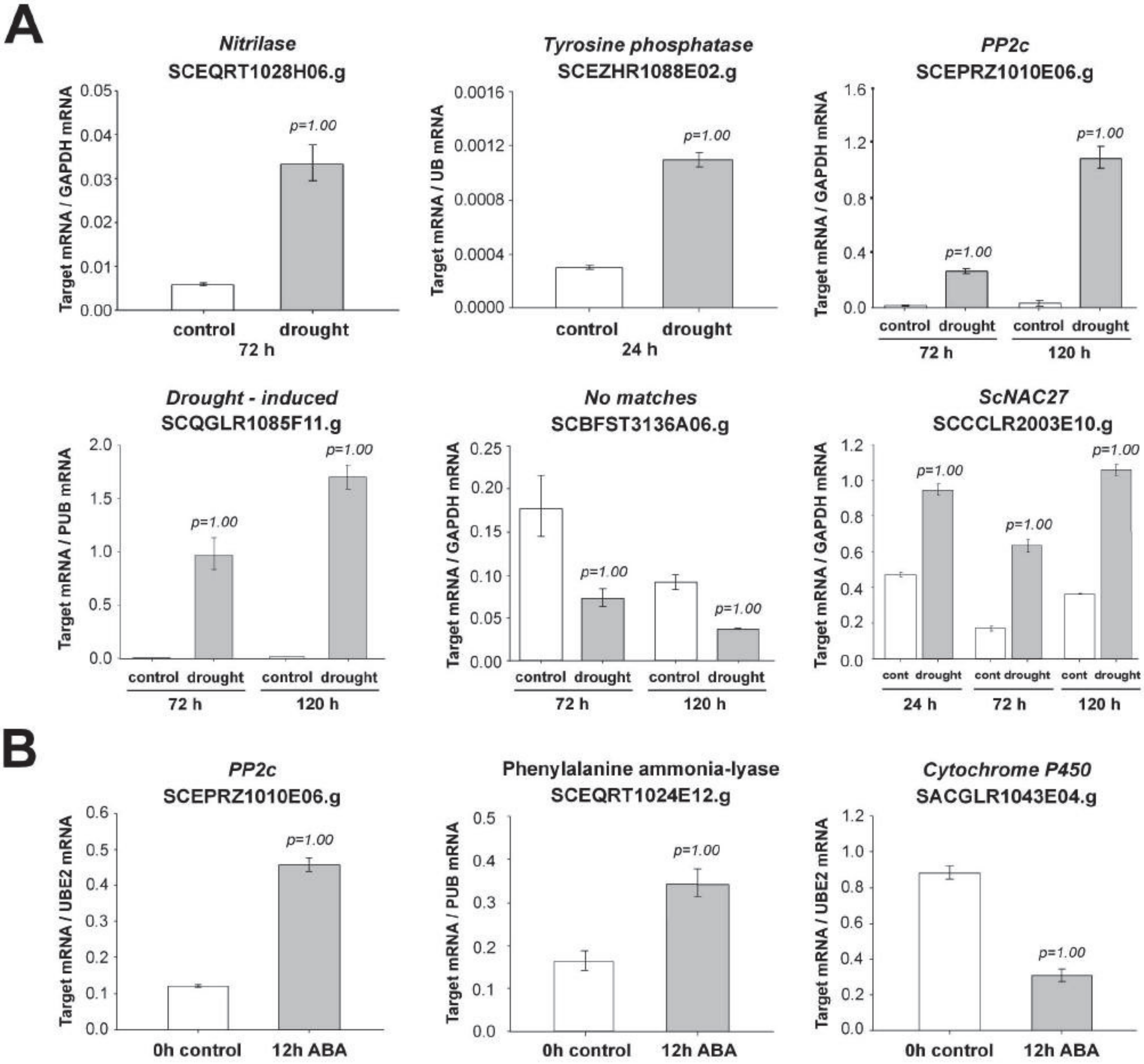


Figure 4

Real Time PCR (qRT-PCR) analysis of drought and ABA-responsive gene expression. The y axis refers to the relative expression ratio between target mRNA versus the reference mRNA (polyubiquitin SCCCST2001G02.g; GAPDH Gene ID: 542367; UBE2 SCBGLR1002D06.g) in sugarcane plants treated with ABA for 12 h or drought conditions for 24, 72 or 120 h. The reactions for the target mRNA and reference mRNA were carried out in parallel and each reaction was performed in triplicates. Error bars were calculated as described previously [31]. The transcript levels for the reference genes were verified to not vary in response to the treatments. The values of P correspond to the probability Pr (Treated>Control) and Pr (Treated<Control) for up- and down-regulated genes, respectively. The SAS was considered differentially expressed when P ≥ 0.95.

orthologues correspond to SNF1-like kinases (SCRFLR1034G06.g and SCACLR2007G02.g – *At1g 78290*), two calreticulin genes (SCRFLR2037F09.g – *At1g56340* and *At1g09210*), an auxin/IAA transcription factor gene (SCCCRZ1001G10.g – *At3g04730*), a defense and cell wall-related gene encoding a phenyl ammonia-lyase (SCEQRT1024E12.g – *At2g37040, At3g53260, At3g10340*) and a dehydrin gene (SCQGLR1085F11.g –

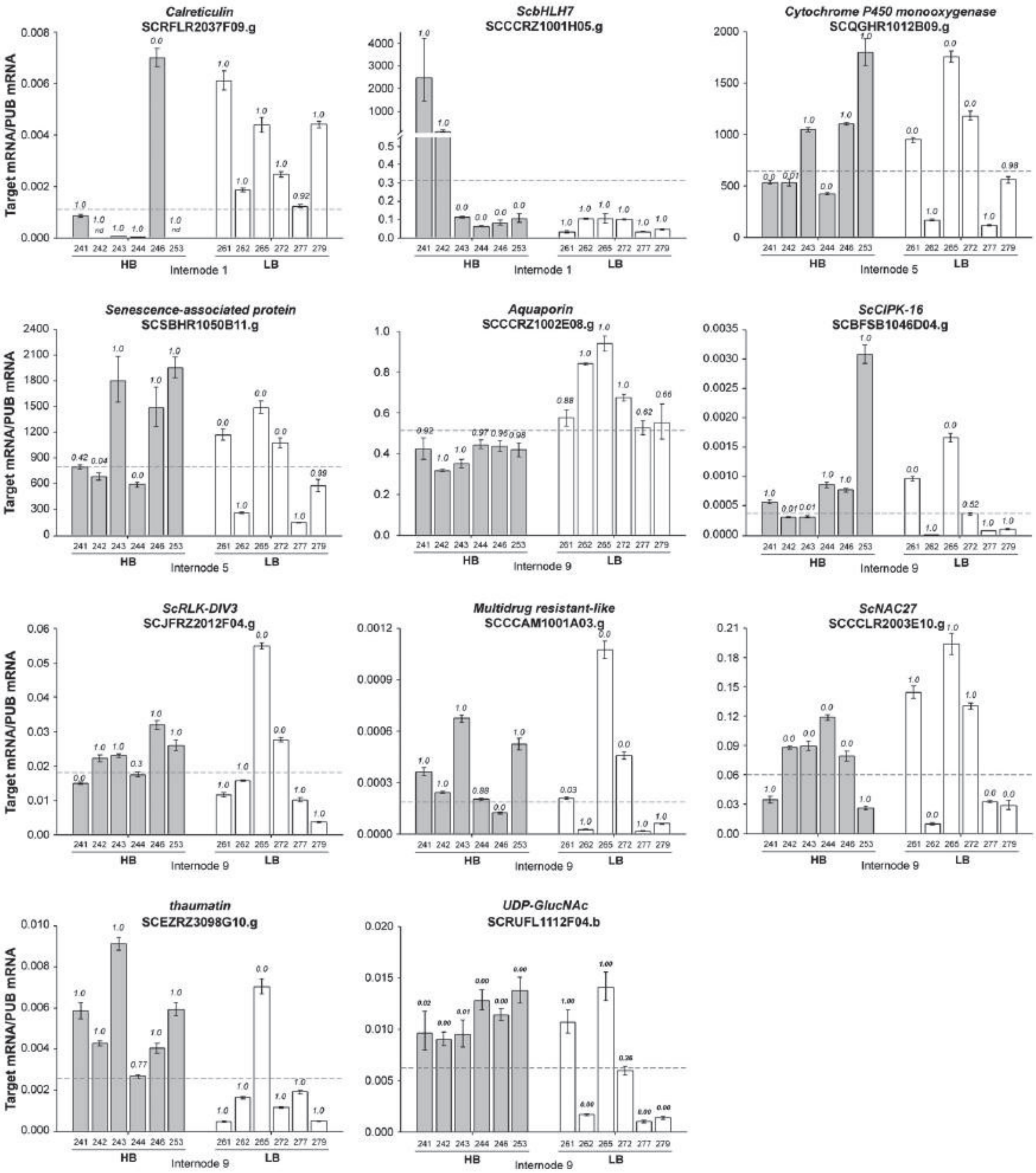


Figure 5

Real Time PCR (qRT-PCR) analysis of individual genotypes gene expression. The y axis refers to the relative expression ratio between target mRNA versus the reference mRNA (polyubiquitin SCCCST2001G02.g). The relative expression levels were determined in Internode 1, 5 and 9 tissues from six individuals with the highest Brix measures (CTC98-241, CTC98-242, CTC98-243, CTC98-244, CTC98-246 and CTC98-253) and six individuals with the lowest Brix measures (CTC98-261, CTC98-262, CTC98-265, CTC98-272, CTC98-277 and CTC98-279) of Population 2. All reactions were carried out in parallel and each reaction was performed in triplicates. Error bars were calculated as described previously [31]. The transcript levels for the reference genes were verified to not vary in response to the treatments. The significance of differential gene expression was determined considering normal distributions for each tested condition and comparing them to the average expression for all samples (dotted line). The values of P correspond to the probability Pr (GenotypeX>average) and Pr (GenotypeX<average) for up- and down-regulated genes. P values were calculated for each genotype independently. The SAS was considered differentially expressed when P ≥ 0.95.

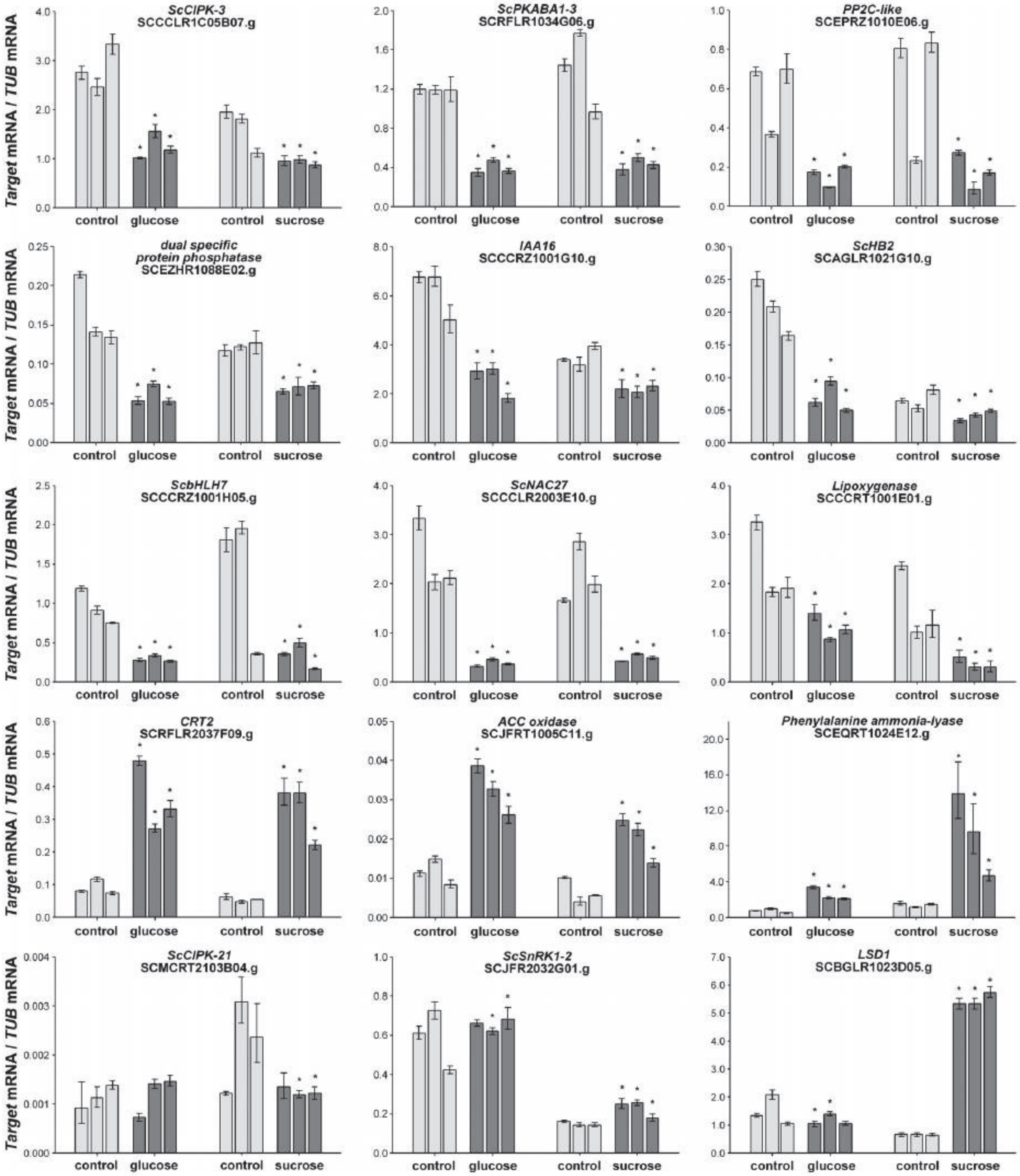


Figure 6

Quantitative PCR (qRT-PCR) analysis of sucrose and glucose responsive genes. The y axis refers to the relative expression ratio between target mRNA versus the reference mRNA (tubulin SCCCRZ1002H03.g) for 3 different experiments in sugarcane thirteen-old day seedlings treated with 3% glucose and 3% sucrose for 4 h. R1, R2 and R3 refers to three control and three sucrose and glucose independent treatments. Error bars were calculated as described previously [31]. The transcript levels for the reference genes were verified to not vary in response to the treatments. The values of P correspond to the probability Pr (Treated>Control) and Pr (Treated<Control) for up- and down-regulated genes, respectively. The SAS was considered differentially expressed when P ≥ 0.95.

At3g50980) (see Additional file 4). Furthermore, two Arabidopsis genes, the CUC1/NAC-type transcription factor (*At3g1550*) and a wound-responsive gene (*At4g10270*) and their closely related respective sugarcane homologues (SCCCLR2003E10.g and SCCCLR2C01F06.g) were found to be similarly regulated by sugars (Table 1).

Discussion

Sugarcane partitions carbon into sucrose that can accumulate to 0.7 M in culms [35]. This unique characteristic has been exploited and improved by humans through breeding. Studies that shed light on the molecular mechanisms behind this feature include gene expression and signaling studies on sink and source regulation [36], QTL studies for sucrose accumulation [37] and gene expression profiling during internode maturation [10-12,38]. Such studies indicated that genes associated with sucrose metabolism are not abundantly expressed in culm tissues while genes related to synthesis and catalysis of sucrose are turned off during internode maturation. Genes involved in cellulose synthesis, cell wall metabolism and lignification are also regulated during this process. The activity of genes associated with internode development was evaluated in genotypes of *S. robustum* (which does not accumulate sucrose to high levels), *S. officinarum* and in a hybrid [39]. Mature internodes of all three genotypes showed decreased expression of cell wall metabolism-associated genes and increased expression of genes related to sucrose metabolism. While the general conclusion of these studies does not appear to be in agreement, it is important to note that the genotypes, environment and age of plants used were different and that a larger sampling may be necessary to define gene profiles in sugarcane.

In this work, we evaluated mature and immature internodes of thirty genotypes using cDNA microrrays and qRT-PCR. Genes associated with sucrose content were defined through the analysis of segregating populations selected for one or three generations [40]. Internodes 1, 5 and 9 (In1, 5 and 9) were collected from plants grown in the field. Among the genes found to be differentially expressed were those related to hormone signaling (auxin, ethylene, jasmonates), stress responses (drought, cold, oxidative), cell wall metabolism, calcium metabolism, protein kinases, protein phosphatases and transcription factors. We compared high Brix plants against low Brix plants by hybridizing pairwise the In1, In5 and In9 tissues directly (HB vs LB hybridizations) or by hybridizing mature against immature internodes (MI vs II). We validated gene expression by qRT-PCR in pools of clones and many individual genotypes. We also investigated if genes associated with sucrose content were responsive to sucrose or glucose treatments. Many of the sucrose-associated genes that are regulated during development are associated with drought responses or are modulated by ABA or sugars, as discussed below (see Additional file 2 and Table 1).

Protein kinases and calcium signaling

Protein phosphorylation appears to play a predominant role in sucrose accumulation and culm development. We have previously categorized sugarcane proteins with a PKinase domain using a phylogenetic approach and named sugarcane protein kinases (PKs) according to the groups obtained, similarity to other kinases and additional domains observed [31]. We now add evidence that several of these genes are regulated during culm development.

A total of fifty-four genes corresponding to PKs, protein phosphatases (PPases) or receptor-like kinases (RLKs) were differentially expressed in high Brix plants or during culm maturation (see Additional file 2). *ScMAPK-4* (SCSBAM1084E01.g) was more highly expressed in high Brix and mature internodes (Table 1). A MAPK kinase was reported to be involved in the regulation of source metabolism by glucose and stress, which is an indication that *ScMAPK-4* might be important in establishing sink-source relationships in sugarcane [36,41]. The most predominant PK category altered is the SNF1-like kinase family of proteins. In yeast, SNF1 regulates the expression of genes coding for carbohydrate metabolism and other metabolic enzymes [42]. In plants, SNF1-related kinases have been named SnRK1 [43] and comprise three distinct sub-families (SnRK1, SnRK2 and SnRK3). In sugarcane, we have identified members of all three sub-families [31]. Analogous to SNF1, plant SnRK1s also regulate carbon metabolism at the level of gene expression. At least three important biosynthetic enzymes have been identified as biological substrates of SnRK1s: hydroxymethylglutaryl-CoA reductase (HMG-CoA reductase) [44]; sucrose-phosphate synthase [45] and nitrate reductase [46]. It is possible to make a direct parallel between sucrose accumulation and the gene expression levels for an ScSnRK1 (SCJFRZ2032G01.g). *ScSnRK1-2* and four 14-3-3 proteins of the GF14 type (SCCCLR1022D05.g, SCCCRZ1001D02.g, SCEQRT1031D02.g and SCEQRT1025D06.g) were expressed at lower levels in mature internodes (Table 1). 14-3-3 proteins, together with a SnRK1, phosphorylate and inhibit the enzyme sucrose phosphate synthase (SPS) *in vitro* [45,47]. Our findings suggest that the decrease in the expression of these genes in the mature internodes may allow for increased sucrose accumulation. We also observed that *ScSnRK1-2* was induced by sucrose treatment, while most of the *ScCIPKs* and *ScPKABA* and *ScOSAPK* genes were repressed (Table 1). This is an interesting finding that may functionally distinguish the pathways triggered by these kinases in response to sucrose and stress.

Members of the SnRK2 and SnRK3 sub-family including two Osmotic Stress-Activated Kinases – OSA-PK (SCEPRZ1009C10.g and SCCCST1004A07.g) and three CBL-interacting Protein Kinases – CIPK (SCJFRZ 2032C08.g, SCMCRT2103B04.g and SCJLRZ1023H04.g) were identified as developmentally regulated during culm maturation (Table 1). Most importantly, three CIPKs (SCBFSB1046D04.g, SCMC RT2103B04.g, SCCC LR1C05B07.g) were more highly expressed in high Brix plants. CBL are regulatory subunits similar to calcineurin that bind to and respond to calcium signals [48]. It has been shown that OSA-PKs and CIPKs mediate drought, osmotic, saline and cold stresses in response to ABA and calcium [49]. Among our differentially expressed genes we found nine genes associated with calcium signaling (SCVPLR1049C09.g, SCCCRZ1C01H06.g, SCJL LR1108 H07.g, SCEZLB1012F10.g, SCCCAM1001A03.g, SCAGL R1043F02.g, SCCCCL3120G07.g, SCCC RZ1003A03.g, SCRFLR2037F09.g) and a calcium-dependent protein kinase (SCEQRT2099H01.g – *ScCDPK-27*) that also indicates a role for this second messenger in sucrose accumulation in sugarcane (Table 1). Sucrose synthesis control depends on the activity of the sucrose phosphate synthase, which catalyses the synthesis of sucrose 6-phosphate from UDP-glucose and fructose 6-phosphate. Sucrose breakdown depends on the activity of invertase, which breaks down sucrose into glucose and fructose, and on the activity of sucrose synthase, that converts sucrose in fructose and UDP-glucose in the presence of UDP [35]. Several studies have shown that some CDPKs phosphorylate and regulate sucrose synthase [50-53]. Studies on the maize sucrose synthase showed that phosphorylation of this enzyme on the Ser-15 by CDPKs stimulates the sucrose breakdown activity of this enzyme [50,52]. Besides, CDPKs can phosphorylate residue Ser-170 of this enzyme directing it to the degradation pathway via proteosome 26S [52,54]. The decrease in expression of *ScCDPK-27* in the mature internode correlates well with increased sucrose in this organ. The activity of sucrose synthase modulates the source-drain relationship [55,56], which eventually determines sucrose content in sugarcane internodes. Additionally, some CDPKs can phosphorylate and inactivate the enzyme sucrose phosphate synthase [57,58], which might contribute to lower sucrose in culms when this enzyme is expressed in high levels, such as seen in low Brix genotypes. Since sucrose biosynthesis is a process regulated by calcium, CDPKs and SnRKs, the genes differentially expressed observed in the high Brix genotypes may all contribute and act as critical control points in sucrose accumulation in this grass.

Drought signaling

We found a prevalence of gene families regulated by ABA, drought and other stresses among the genes associated with sucrose content [33]. Sixty-nine genes associated with sucrose content were identified to be regulated in response to drought and eleven to ABA (see Additional file 2). This is a strong indication that some of the pathways associated with sucrose content and culm development may overlap with stress signaling pathways. A similar finding was described by Casu and colleagues that found many stress-related genes expressed in culms [11]. Overall, around 32% of the genes previously found to be responsive to drought are associated with sucrose content (Figure 1). It is generally known that sugarcane plants maturation is favoured by the exposure to a period of low water precipitation. It is possible that increased sucrose content is dependent on a drought season and that plants converge the drought and sucrose signaling pathways to sustain growth even during the stress season. Sugars that transduce stress signals or act as osmoprotectants, like fructose, raffinose and trehalose [20] could be central players during this process. A PP from the PP2C family (SCEPRZ1010E06.g) similar to a PPase that transduces the ABA signal was identified as associated with Brix, drought and ABA (Table 1). This PPase was less expressed in high Brix, reduced in the mature internodes and induced by drought and ABA. A similarity search showed that this *PP2C* is very similar to *ABI1* and *ABI2* from *Arabidopsis thaliana*. *PP2C*s that are part of the ABA signaling pathway, such as *ABI1*, *ABI2*, *AtPP2CA* and *AtP2CHA*, have their transcript levels increased by this phytohormone [59,60]. Among the processes regulated by ABI1 and ABI2 is stomatal closure, which is certainly one of the first protective measures during drought signaling. Moreover, some of the ABA biosynthesis enzymes are induced by drought and osmotic stress [61]. The fact that sugarcane genes associated with ABA and drought signaling are differentially regulated in plants with distinct sucrose accumulation capacities is an indicative that the role of ABA is well worth being further verified.

Drought responses vary depending on the duration and intensity of the stimulus and comprehend protective alterations and long term growth changes [13]. Many genes responsive to drought have been catalogued [14-18]. The drought stimulus lead to transient calcium fluxes, the activation of calcium sensors, the accumulation of reactive oxygen species, the activation of the MAPK pathway and the induction of several transcription factors including DREB2A, DREB2B [62] and NAC [63]. We have found, among the genes associated with sucrose content, many components of the gene families mentioned above. A MAPK was more expressed in high Brix and in mature internodes and repressed by ABA (SCSBAM1084E01.g), a *DREB* was induced during culm maturation (SCCCLR1001D10.g) (Table 1) and over forty stress responsive genes were identified (Additional file 2). A *DEHYDRIN TYPE 1* (SCQGLR1085F11.g) regulated by the DREB signaling cascade [64] was dramatically repressed in mature internodes of high and low Brix plants and induced after 72 h and 120 h drought treatment (Table 1). A barley dehydrin gene, *DHN1*, was highly expressed in cells cultured at 25°C and 2°C in media containing high sucrose but our data indicated the dehydrin expression may not be a consequence to high sucrose since immature internodes do not have high levels of this sugar [65]. Overexpression of *DREB2A* in *Arabidopsis thaliana* led to the generation of transgenic plants more tolerant to drought [66,67]. It is possible that some of the genotypes may indeed be more resistant to drought and thus able to grow and accumulate more sucrose, but additional experiments are necessary to verify this hypothesis.

Transcription factors and hormone signaling

We have recently integrated and evaluated the SUCEST and PlantGDB http://planttfdb.cbi.pku.edu.cn/web/index.php?sp=so EST databases for putative Transcription Factors and identified 2,406 candidate TFs. These were classified into families and can be found at http://grassius.org[68]. We found twenty-one transcription factors (TFs) regulated during culm development (Table 1). The great majority was more expressed in the immature internodes, including two members of the Homeobox Knotted1-homeodomain (SCAGLR1021G10.g and SCRLAM1010D08.g), which have been shown to be involved in developmental processes in maize [69]. Developmentally regulated genes include a homolog (SCBFAD1046D01.g) to anthocyanin regulatory R-S protein containing a helix-loop-helix (HLH) domain, that controls tissue-specific synthesis of anthocyanin pigments [70]. Nine transcription factors were identified as differentially expressed when high Brix and low Brix genotypes were compared including an *ARF6* (*AUXIN RESPONSE FACTOR6a*) (SCEZLB1010E10.g), a *NAM* (*NO APICAL MERISTEM*) (SCCCLR2003E10.g) and an *EIL* (*ETHYLENE INSENSITIVE3-LIKE*) (SCCCRZ1004H12.g) (Table 1). The *NAM* transcript was less abundant in both populations analysed, negatively regulated by sucrose and glucose treatment and induced by drought (Table 1 and Figure 4). NAM transcription factors in *Oryza sativa* have been described as important regulators of drought tolerance [63]. This may indicate a connection between these signaling pathways [71], possible co-regulation associated with sucrose content and cross-talks or signaling overlaps between sugar sensing, sugar mobilization and drought responses.

Among the genes more expressed in the immature internodes, we also found several genes similar to auxin, ethylene and giberellin-responsive TFs (Table 1). We found a second TF related to *ARF6* (SCCCLR1024F10.g) and four *AUXIN RESPONSE PROTEINS* (SCCCRZ1001G10.g, SCVPLR2005H03.g, SCJFRZ2009F04.g, SCJLLR1054C09.g) more expressed in immature internodes. Signaling by auxins during culm development was also pointed out by the altered expression of two nitrilase genes (SCEQRT1028H06.g and SCRFLR1012D12.g) with a putative role in this hormone's biosynthesis, an auxin efflux carrier (SCCCAM2004G02.g) and a *AUXIN REPRESSED PROTEIN* (SCCCLR2002F08.g) which is upregulated in mature internodes. One of the nitrilases (SCEQRT1028H06.g) was induced after 72h of drought [31] and its expression level was confirmed by qRT-PCR (Figure 4). Another drought-induced nitrilase (SCCCCL6002B05.g) was found more expressed in high Brix plants. This gene is highly similar to the maize *ZmNIT2* gene, which converts indole-3-acetonitrile to indole-3-acetic acid [72]. Overall, differential expression of auxin signaling genes during internode development and/or association with sucrose content was observed in twenty different biological samples.

Ethylene was observed as a putative modulator of this process (Table 1). One *EIL* was found less expressed in high Brix plants (SCCCRZ1004H12.g) and two were less expressed in mature internodes (SCBGFL4052C11.g and SCCCRZ1004H12.g). Besides, one ACC oxidase (SCVPLR2012A10.g) was found to be more expressed in high Brix plants and less expressed in mature internodes.

Jasmonic Acid biosynthesis also seems to have a relevant role in culm development since several enzymes envolved in methyl jasmonate biosynthesis were found to be more expressed in immature internodes, two lipoxygenases (SCCCRT1001E01.g and SCJFRT1007H07.g) and an Omega-6 fatty acid desaturase (SCCCLR1C03G01.g) (Table 1).

Cell wall biosynthesis

There were several genes with a putative function in cell wall metabolism that were among the differentially expressed genes, such as the expansins SCQGRT1040G03.g and SCCCLR2C02A05.g (Table 1). Similar genes were found to be expressed in two-day-old rice seedlings, a stage where rapid cell elongation occurs accompanied by cell division [73]. The authors believe EXP activity may be required for cell expansion. Expansins may act in the relaxation of the cell wall, possibly by breaking the bonds between cellulose microfibrils and matrix polysaccharides [74,75] allowing for cell expansion. Our data indicates that, in sugarcane, a gene similar to *EXPA23* (SCQGRT1040G03.g) is more expressed in 7-month old high Brix plants as compared to low-Brix plants, and the *EXPA11* (SCCCLR2C02A05.g) in turn, is more expressed in low Brix plants (after 11 months of planting) (Table 1). As mentioned above, we have evidence that auxin signaling is highly active in immature internodes. Auxin signaling is associated with plant cell expansion [76], which may be an additional evidence that the high Brix plants selected have cell expansion alterations that might confer higher sucrose accumulation capacity. These observations are corroborated by the identification of a *XYLOGLUCAN ENDO-B-1,4 GLUCANASE* (*XTH*) (SCBFLR1039B05.g) that is more expressed in immature internodes (Table 1). XTHs can hydrolyse xyloglucans, major components of plant cell walls, and transglycosylate residues into growing xyloglucan chains, that may be important during tissue expansion [77]. Our data is in agreement with previous findings from Casu and colleagues [12] that identified five *XTHs* and four β-expansins less expressed in mature internodes, as well as two caffeoyl-CoA O-methyltransferases induced in mature internodes.

We found five genes of the lignin biosynthesis pathway associated with sucrose content (Table 1). The first step in lignin biosynthesis in plants is the deamination of L-phenylalanine by Phenylalanine Ammonia-Lyase (PAL) to cinnamic Acid. PAL is the first enzyme of the phenylpropanoid pathway [78]. It converts L-phenylalanine into *trans*-cinnamic acid (*t*-CA), which is further transformed in plants into many phenylpropanoid compounds, such as lignins, antioxidants, anthocyanins and flavonoid nodulation factors. We found three *PAL* genes associated with sucrose content and more expressed in immature internodes (SCCCLR1048D07.g, SCEQRT1024E12.g and SCSGAM1094D05.g). SCEQRT1024E12.g was induced after ABA treatment and repressed after 72 and 120 h drought stress. The second step in lignin biosynthesis is catalyzed by a cinnamate 4-hydroxylase (C4H) [79]. We found one SAS less expressed in high Brix similar to a *C4H* (SCSGFL4193B05.g). Down the pathway *p*-Coumaroyl is transformed into Caffeoyl CoA by a *p*-coumaroyl shikimate 3'-hydroxylase (C3H). One SAS similar to a *C3H* (SCACSB1037A07.g) was found less expressed in high Brix plants. We also observed a *FERULATE 5*-HYDROXYLASE (*F5H*) (SCEZHR1087F06.g) and a *CAFFEIC ACID 3-O-METHYLTRANSFERASE* (*COMT*) (SCRFLR1012F12.g) more expressed in mature internodes. *F5H* was less expressed in high Brix while *COMT* was induced (Table 1). While all the above mentioned genes may have a role in cell wall metabolism, it is important to note that *trans*-cinnamic acid can also be converted into salicylic acid and anthocyanins [80] and, until the activity of these enzymes is verified, the data can only indicate a putative alteration in cell wall biosynthesis and modification in the accumulation of sucrose in culms.

Cell wall biosynthesis can reduce sucrose accumulation since carbon fluxes directed to plant growth and cell wall expansion may alter carbon partitioning into sucrose. It is also possible that sucrose accumulation may trigger increased lignification. One of the PAL enzymes was induced by sucrose treatment up to 14-fold indicating that this enzyme is highly responsive to sucrose. An induction of a *COMT* gene has already been described during culm maturation [12] but this is the first report implicating a PAL, C4H, C3H, F5H and COMT in sucrose content. It is possible that some of the genotypes analyzed also differ in biomass content and a continued agronomic evaluation is necessary to assess how gene expression in the selected genotypes is related to other characteristics, such as cell wall composition, growth rates, internode size and width, number of internodes and drought tolerance, for instance. Many parameters besides Brix may differ among the genotypes and have not yet been evaluated. It is possible also that the high Brix genotypes may be more amenable for acid and enzymatic hydrolysis and cellulosic ethanol production. Silencing of lignin biosynthesis genes has been shown to benefit sugar release for lignocellulosic biomass fermentation [81]; it will be thus interesting to test if altered biomass has been selected for during the breeding process. At any rate, the alteration of these cell wall biosynthesis genes in association to Brix content is an interesting indication of a correlation between these processes.

A relationship between high sugar content and sugar signaling

We showed that a set of genes associated with sucrose content is also early sugar-responsive. Since most of these genes are related to signal transduction (kinases, phosphatases, transcription factors, hormone synthesis) they are likely to constitute upstream components of the sugar regulatory cascade. These findings raise the interesting possibility that sugar signaling may somehow influence sugar accumulation capacity in sugarcane. How these genes may influence sucrose accumulation is an open question. Interestingly, between sucrose/glucose-treated young seedlings and high Brix genotypes, contrasting expression patterns were found for 15 genes, while the remaining 9 genes presented similar regulatory trends (Table 1 and data not shown). This latter set of genes may be related to higher sugar fluxes and/or higher sugar sensitivity in high Brix genotypes. Opposite regulatory patterns between young seedlings and internodes of high Brix genotypes are more difficult to explain, but could reflect differential developmental-dependent controls. A comparative analysis with *Arabidopsis* showed that among the twenty-four sugarcane genes tested, five *Arabidopsis*-sugarcane probable groups of orthologues and two pairs of *Arabidopsis*-sugarcane close homologues (sister clades) were apparently regulated similarly by sugars in seedlings (see Additional file 4). Orthologous genes of the signal transduction-class, such as those encoding transcription factors or kinases with conserved regulatory features, are likely to represent important players in the sugar signal transduction pathways and this can now be tested. Within this framework, it should also be interesting to further analyze the integration/interaction of the *Arabidopsis* CUC1/NAC-type transcription factor (At3g1550), which controls shoot apical meristem formation [82] in the sugar regulatory network.

Data validation across genotypes

To confirm gene expression and evaluate transcript levels we performed qRT-PCR reactions for forty-two genes. With a probability value higher than 0.95 we observed that 80% of the gene expression data obtained using cDNA microarrays were compatible with the qRT-PCR data.

Validation of developmental regulation was elucidative of differences among high Brix and low Brix populations. *ScCIPK-21* for instance, a gene more expressed in high Brix and in mature internodes, was found to be much more induced during culm development in the high Brix plants, what may be an indicative that induction of this gene may lead to higher sucrose levels. A category that was consistently more expressed in immature internodes and high Brix and that has been seen to be responsive to drought in other plants is the aquaporin family of proteins. We wanted to verify if genes of this family could be useful expression markers of sucrose content. Five aquaporins, from both the MIP and PIP sub-families (SCCC RZ1002E08.g, SCCCST3001H12.g, SCEQRT2100B02.g, SCCCLR1024C03.g, and SCCCRZ1001F02.g) were regulated during culm development and two of them found to be associated with high Brix (SCCCST3001H12.g and SCEQRT2100B02.g) in population 2 (Table 1). This large family of proteins is primarily involved in the regulation of water movement between cells and cell compartments, although many of them also facilitate the passage of small solutes [83,84]. The accumulation of sucrose in such high concentrations as seen in sugarcane cells certainly represents an osmotic challenge that demands efficient control of solute compartmentalization and water transport out of the vacuoles. As key players in the equilibration of water potentials via regulation of membrane permeability, aquaporins may have a fundamental role in the process of sugar storage in sugarcane vacuoles. It has been observed in *Arabidopsis* that loss of the aquaporin TIP1.1 severely affects carbohydrate metabolism and transport [85], and the authors postulate that this aquaporin could be involved in a vesicle-based routing of carbohydrates towards the central vacuole. In our study, expression of one aquaporin correlated to lower sucrose content in both populations analyzed. Since differential expression was determined in pools of seven or eight individuals we decided to verify how many of the genotypes in those pools presented the observed expression patterns. Expression data was obtained using qRT-PCR for twelve of the sixteen extreme individuals of Population 1 (Figure 5). We calculated the average expression levels across all twelve genotypes for the Aquaporin gene (SCCCRZ1002E08.g) and observed that the gene was less expressed in all high Brix genotypes and more expressed in half of the low Brix genotypes. This is a strong indicative that low expression of this gene has been segregated and selected by the breeding process and is strongly associated with high sucrose content. To verify if this would be the case for other differentially expressed genes we evaluated the expression profiles of eleven genes in the twelve extreme genotypes of Population 1. For some genes, such as the HLH TF (SCCCRZ1001H05.g), expression was consistently lower than average in all low Brix genotypes, but found to be higher in only two of the high Brix genotypes (Figure 5). In this case, the two genotypes had almost 10,000 fold increased expression for the genes, which may account for the differential expression observed in the pool used for cDNA microarray analysis. This is an extreme case and, in general, most of the data shows around 50% of validation in the individual genotype validations.

Conclusion

Gene expression analysis of sugarcane populations contrasting for sucrose content indicated a possible overlap of sugar, drought and cell wall metabolism processes and suggested signaling and transcriptional regulators that might be useful as molecular markers in breeding programs or as primary targets in a sugarcane improvement program based on transgenic plants. This work is a first survey on gene expression related to sucrose content and some similarities point to conservation between monocot and dicot sucrose responses. This observation can help to point the important players in the sugar signal transduction pathways. Due to the diversity of roles described for the identified genes, additional experiments will be necessary to elucidate their possible roles in the sugarcane sucrose accumulation process. Our group is currently generating transgenic plants with modified expression levels for these genes to confirm the hypothesis raised for their function.

Methods

Biological samples

Population 1 was derived from two intra-specific polycrosses, one among 21 *Saccharum officinarum* genotypes (Caiana Fita, IK76108, Lahaina, MZ151, MZ151 roxa, Sabura, Salangor, Sinimbu, NG213, Fiji 47, Hinahina 18, Manjri Red, Muntok Java, NG77142, Soff 8268, SS601, Sylva, NG2880, Vae Vae Ula, IJ76315 and IN8425) and the other combining 13 *Saccharum spontaneum* genotypes (IN8458, IN8488, Krakatau, SES 147b, US56158, US7440, US851008, UM721, UM691, SES 194, IK7686, US56193 and US571723). The individuals of these polycrosses were crossed amongst themselves and for each generation, 500 individuals were sampled for soluble solids (Brix degree). The extreme segregants of the F3 hybrid individuals were planted in a field in single rows of 5 m using standard sugarcane cultivation practices. Tissue samples were collected in March of the following year, when plants were 10 months old. The Brix degree content of the 4th–5th internodes of each sugarcane stalk was measured with a portable refractometer (N1 model, ATAGO, Japan). Additional file 1 lists Brix measurements for the extreme individuals of both populations [33] and the corresponding sucrose concentration. The average Brix value for high sugar individuals was 18.10 +/- 1.44 and for low sugar individuals was 6.70 +/- 0.96 for Population 1. Sucrose content was 9.2% in high Brix individuals and 1.1% in low Brix individuals for this population. Sugars were determined as described [86,87].

Population 2 was derived from a cross between two commercial varieties (SP80-180 × SP80-4966). Five hundred sugarcane F1 plants were field-grown. Seven plants with extreme Brix values were selected. Population 2 was less contrasting than Population 1, with an average high Brix of 18.47 +/- 1.41 and average low Brix of 13.65 +/- 1.27.

Sucrose accumulating tissues (sink tissues, herein internodes) were collected from field grown plants. We have previously determined using Pair-wise Pearson correlation calculations a high correlation of gene expression between individuals collected at the same time or within a short interval of time (0.84 to 0.88), and a lower correlation between individuals collected in different years (0.61 to 0.64) [88]. Mature (In9), intermediate (In5) and immature internodes (In1) were then collected from four selected plants of each genotype at 7, 10 and 11 months after planting. Tissue collected from the four plants was pooled, therefore each biological sample corresponds to a certain tissue derived from four plants, and the gene expression data reflects the average expression of the pooled plant tissue. A total of 132 biological samples were selected for gene expression studies from both populations. RNA was extracted from tissues of individuals or pools of eight individuals as described [31].

The cultivar SP90-1638 (Internal Technical Report, CTC, 2002), sensitive to drought, was used for the water deprivation experiments. The experiments were previously described [31]. Briefly, plants were transferred to pots containing moist sand under greenhouse conditions. Regular watering was maintained for 90 days, being suppressed after this period for the experimental group. Aerial parts of six plants for each experimental point were collected 24, 72 and 120 h after the onset of drought for the control and experimental groups.

For ABA treatment, plants derived from shoot apex of 2-month-old sugarcane plants were *in vitro* cultivated for approximately three months in a growth chamber as described [31]. ABA (Sigma Chem. Co) was added to the culture medium to a final concentration of a 100 μmol.L $^{-1}$ whereas control plants were treated with distilled water. Leaves were collected after 0.5h, 1h, 6h and 12 h of exposure to ABA. Six plantlets were sampled for each time point.

For sucrose treatment, seeds obtained from a crossing between SP891046 and IAC912195 varieties were imbibed in water, incubated for 10 min at 52°C to open the panicle and sterilized by a 5-min treatment in 70% ethanol followed by 20-min in 2.5% sodium hypoclorite. Seeds were then washed 5 times in sterile water and then transferred to a Musharige and Skoog half-strength solid growing media [89] containing 0.5% of sucrose. Plates were incubated in continuous light for 12 days at 28°C. Subsequently, _seedlings were transferred to liquid MS/2 growth medium without any sugar and further grown for 24 h under weak agitation (60 rpm) and constant light before being treated with 3% sucrose or 3% glucose or 3% mannitol (stock solution of 30% in water) or just with water as control for 4 h. RNA was extracted using Concert® (Invitrogen, USA) according to the manufacturer's recommendations.

Gene expression data

cDNA microarray experiments were conducted and data extracted as described previously [31]. SUCEST SAS consensus sequences can be found at http://sucestfun.iq.usp.br/sucestfun/database/Sugarcane-Clustering.fasta. The corresponding Sugarcane Gene Index contigs http://compbio.dfci.harvard.edu/tgi/plant.html can be searched at http://sucestfun.iq.usp.br/cgi-bin/pub_sucest/clustering_comparison.cgi and downloaded at http://sucest-fun.org/pub/brix/data/SAS_vs_SGI.blastn.csv and http://sucest-fun.org/pub/brix/data/SGI_vs_SAS.blastn

.csv. The designed microarray contains 1830 genes which yielded 1545 good-quality PCR fragments. Reverse transcription, labeling and hybridizations were done using the reagents provided with the CyScribe Post-Labeling kit (GE Healthcare) or SuperScript™ Plus Indirect cDNA Labeling System (Invitrogen, USA). The microarrays were scanned according to the manufacturer's instructions using the Generation III System (Molecular Dynamics). Hybridizations were carried out as described [33].

Two technical replicates were obtained for each microarray experiment. Data were collected using the ArrayVision (Imaging Research Inc.) software. The fluorescence ratios were normalized in the MxS space, where M is the base 2 logarithm of the intensities ratio and S is the base 2 logarithm of the average intensity of each spot. The M values were normalized to account for systematic errors using the LOWESS fitting. The identification of differentially expressed genes was performed using a local implementation of the Outliers Search method [31]. The SAS presenting more than 70% of its replicates outside fold-change cut-off curves were defined as differentially expressed. Raw data can be found at https://sucestfun.iq.usp.br/pub/brix/index.html.

Validation of microarray results by real-time PCR (qRT-PCR)

Real-time PCR reactions were done essentially as described [31]. The ratio between the relative amounts of the target gene and the endogenous control gene in the qRT-PCR reactions was determined based on the Ct method [90] with modifications. The normalized expression level was calculated as $L = 2^{-\Delta Ct}$ and $\Delta CT = C_{T,\ target} - C_{T,\ reference}$. A polyubiquitin (PUB) gene (SCCCST2001G02.g) was used as an endogenous reference in the qRT-PCR reactions of high Brix and low Brix samples after verification that its mRNA levels were similar in the populations and individual tissues. This PUB gene was also used for the sucrose-responsive gene expression validation. Drought samples were normalized using a GAPDH (CA254672.1 [91]), PUB (SCCCST2001G02.g) or Ubiquitin (SCCCLR1048F12.g) genes and ABA samples using a UBE2 (*ubiquitin conjugating enzyme E2*) (SCBGLR1002D06.g) or PUB gene (SCCCST2001G02.g).

To access the statistical significance of expression ratios, we assumed a log-normal model and calculated the probability P = Pr(sample>reference) and P = Pr(sample<reference) for up- and down-regulated genes, respectively. The expression profile was considered validated when $P \geq 0.95$. For validation of gene expression differences among all different genotypes the probability value P of being greater or smaller than the average expression across all individuals was calculated depending on whether the condition was respectively up- or down-regulated according to the microarray data.

Comparative sequence analysis

Comparative analysis of sugarcane sugar-responsive genes was done by constructing phylogenetic trees containing the corresponding most similar plant sequences. A tblastx search [92] with the sugar-regulated SAS against a green plants protein data set including 365,187 proteins sequences obtained from several genomes (*Arabidopsis thaliana*, version 7.0 – http://www.arabidopsis.org; *Populus trichocarpa*, version 1.1 – http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html; *Glycine max*, version 0.1 – http://www.phytozome.net/soybean.php; *Oryza sativa*, version 5.0 – http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml; *Sorghum bicolor*, version 1.4 – http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html; *Selaginella moellendorffii*, version 1.0 – http://genome.jgi-psf.org/Selmo1/Selmo1.home.html; *Physcomitrella patens patens*, version 1.1 – http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html; *Volvox carteri*, version 1.0 – http://genome.jgi-psf.org/Volca1/Volca1.home.html; Chlamydomonas reinhardtii, version 3.0 – http://genome.jgi-psf.org/chlre3/chlre3.home.html; *Ostreococcus lucimarinus*, version 2.0 – http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.html; *Ostreococcus tauri*, version 2.0 – http://genome.jgi-psf.org/Ostta4/Ostta4.home.html; *Micromonas pusilla CCMP1545*, version 2.0 – http://genome.jgi-psf.org/MicpuC2/MicpuC2.home.html; *Micromonas strain RCC299*, version 2.0 – http://genome.jgi-psf.org/MicpuN2/MicpuN2.home.html) was performed. For each SAS, the first 40 best matches, or all matches obtained if this number were lower than 40, were selected for further analysis. The conserved domains found among the 40 selected sequences were aligned using ClustalW [93] to produce ungapped alignments. The phylogenetic relationship of these aligned sequences was then constructed using the Neighbor-Joining method [94] using p-distance. All analysis were conducted in MEGA4 software [95]. This process allowed identifying the most probable orthologues sequences of the SAS. The Arabidopsis orthologues and the Arabidopsis more closely related homologues (usually one sister clade which may include at least two Arabidopsis sequences) were compared with the set of Arabidopsis genes regulated by glucose [30] and/or sucrose [34] using VennMaster 0.37.3 http://www.informatik.uni-ulm.de/ni/staff/HKestler/vennm/doc.html.

Authors' contributions

FSPT, FRR, AJW, CGL, MDLC and DB conducted cDNA microarray and qRT-PCR experiments. LEVDB conducted the phylogenetic analysis. ECU was responsible for sugarcane cultivation and germplasm sample collection. MYNJ, RZNV and RV were responsible for bioinformatic analysis and database development. MV, MM and GMS are group leaders, designed the experiments, analysed the data and had intelectual input in all activities listed above. All authors contributed to datamining, read the manuscript and approved it.

Note

Additional files can be found at http://sucest-fun.org/pub/brix

Raw data has been uploaded to GEO Database (Series GSE14732)

Additional material

Additional file 1

Brix degree and sugar content of populations. *Brix degree, sucrose, glucose and fructose were determined from 10-month old plants of Population 1 and 11-month old plants of Population 2. The measurements were made from juice extracted from the 9th internode. Brix measurements of these populations have been previously described [33]*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-120-S1.xls]

Additional file 2

SAS showing differential expression when high and low Brix plants were compared or when mature and immature internodes were compared using cDNA microarrays. *The table also shows differential expression of the same SAS as seen in [31] for plants submitted to drought and ABA treatment. The table lists a SAS whose expression was enriched or decreased as determined by the Outliers Search Method in two technical replicates for each biological sample. The expression ratio for each technical replicate is in brackets.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-120-S2.xls]

Additional file 3

P value of qRT-PCR. *Genes associated with sucrose content, drought, ABA and sugars were validated by qRT-PCR. The tables indicate all the genes evaluated and the values of P for differential expression.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-120-S3.doc]

Additional file 4

Sugarcane and Arabidopsis orthologues similarly regulated by sucrose and glucose. *Orthologies between Sugarcane and Arabidopsis were assigned using the Neighbor-Joining method [94]. The Arabidopsis orthologues were compared with the set of Arabidopsis genes regulated by glucose [30] and/or sucrose [34].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-120-S4.xls]

Additional file 5

Inferred phylogenetic relationships among tblastx hits using the sugarcane SAS as queries. *The amino acid alignments were performed with ClustalX. The distances were obtained by p-distance and topography inferred with Neighbor-Joining (NJ) using only the aligned blocks (complete deletion). Analysis were conducted in MEGA4. The continuous blocks show regulation by sucrose and the pointed blocks show regulation by glucose (in both cases red for induction and green for repression). **A** – SCRFLR2037F09.g (Calreticulin 2); **B** – SCEQRT1024E12.g (Phenylalanine ammonia-lyase); **C** – SCCCRZ1001G10.g (IAA16); **D** – SCACLR2007G02.g and SCRFLR1034G06.g (canePKABA1-1 and canePKABA1-3); **E** – SCQGLR1085F11.g (Dehydrin). The sequences names correspond to those present in the protein data sets showed in Material & Methods: **AT** –* Arabidopsis thaliana*; **Gm** –* Glycine max *(soybean); **jgi|Poptr1** –* Populus trichocarpa*; **LOC Os** –* Oryza sativa *(rice); **Sb** –* Sorghum bicolor *(sorghum); **jgi|Selmo1** –* Selaginella moellendorffii*; **jgi|Phypa1_1** –* Physcomitrella patens patens*; **jgi|MicpuC2** –* Micromonas pusilla CCMP1545*, **jgi|MicpuN2** –* Micromonas strain RCC299*; **jgi|Volca1** –* Volvox carteri*; **jgi|Chlre3** –* Chlamydomonas reinhardtii.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-120-S5.pdf]

Acknowledgements

We thank Dr. Carlos Hotta and Dr. Michael Lawton for a critical review of the manuscript. We are indebted to Maria Cristina Falco, Sabrina Moutinho Chabregas, Paulo Gouvêa e William Burniquist for their technical support in the development of this work. This work was funded by Fundação de Àmparo à Pesquisa do Estado de São Paulo (FAPESP). G. M. S., M. M. are recipients of productivity fellowships from CNPq.

References

1. Daniels J, Roach BT: **Taxonomy and evolution in sugarcane.** In *Sugarcane improvement through breeding* Edited by: Heinz D. Amsterdam: Elsevier Press; 1987:7-84.
2. Daniels J, Daniels C: **Geographical, historical and cultural aspect of the origin of the Indian and Chinese sugarcanes S. barberi and S. sinense.** *Sugarcane Breeding newsletter* 1975, **36:**4-23.
3. Roach BT: **Nobilisation of sugarcane.** *Proc Int Soc Sugar Cane Technol* 1972, **14:**206-216.
4. Arceneaux G: **Cultivated sugarcanes of the world and their botanical derivation.** *Proc Int Soc Sugar Cane Technol* 1967, **12:**844-854.
5. Price S: **Interspecific hybridization in sugarcane breeding.** *Proc Int Soc Sugar Cane Technol* 1965, **12:**1021-1026.
6. Lingle SE: **Seasonal internode development and sugar metabolism in sugarcane.** *Crop Science* 1997, **37(4):**844-854.
7. Menossi M, Silva-Filho MC, Vincentz M, Van-Sluys MA, Souza GM: **Sugarcane Functional Genomics: gene discovery for agronomic trait development.** *Int J Plant Genomics* 2008:1-11.
8. Carson DL, Botha FC: **Genes expressed in sugarcane maturing internodal tissue.** *Plant cell reports* 2002, **20:**1075-1081.
9. Carson DL, Huckett BI, Botha FC: **Sugarcane ESTs differentially expressed in immature and maturing internodal tissue.** *Plant Sci* 2002, **162:**289-300.
10. Casu RE, Grof CP, Rae AL, McIntyre CL, Dimmock CM, Manners JM: **Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis.** *Plant molecular biology* 2003, **52(2):**371-386.

11. Casu RE, Dimmock CM, Chapman SC, Grof CP, McIntyre CL, Bonnett GD, Manners JM: **Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling.** *Plant Mol Biol* 2004, **54(4):**503-517.
12. Casu RE, Jarmey JM, Bonnett GD, Manners JM: **Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by Affymetrix GeneChip Sugarcane Genome Array expression profiling.** *Functional & integrative genomics* 2007, **7(2):**153-167.
13. Riera M, Valon C, Fenzi F, Giraudat J, Leung J: **The genetics of adaptive responses to drought stress: abscisic acid-dependent and abscisic acid-independent signalling components.** *Physiologia Plantarum* 2005, **123:**111-119.
14. Bray EA: **Molecular Responses to Water Deficit.** *Plant Physiol* 1993, **103(4):**1035-1040.
15. Bohnert HJ, Nelson DE, Jensen RG: **Adaptations to Environmental Stresses.** *The Plant cell* 1995, **7(7):**1099-1111.
16. Ingram J, Bartels D: **The Molecular Basis of Dehydration Tolerance in Plants.** *Annu Rev Plant Physiol Plant Mol Biol* 1996, **47:**377-403.
17. Seki M, Narusaka M, Abe H, Kasuga M, Yamaguchi-Shinozaki K, Carninci P, Hayashizaki Y, Shinozaki K: **Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray.** *The Plant cell* 2001, **13(1):**61-72.
18. Oono Y, Seki M, Nanjo T, Narusaka M, Fujita M, Satoh R, Satou M, Sakurai T, Ishida J, Akiyama K, *et al.*: **Monitoring expression profiles of Arabidopsis gene expression during rehydration process after dehydration using ca 7000 full-length cDNA microarray.** *Plant J* 2003, **34(6):**868-887.
19. Yamaguchi-Shinozaki K, Shinozaki K: **Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses.** *Annual review of plant biology* 2006, **57:**781-803.
20. Gupta AK, Kaur N: **Sugar signalling and gene expression in relation to carbohydrate metabolism under abiotic stresses in plants.** *J Biosci* 2005, **30(5):**761-776.
21. Smeekens S: **Sugar induced signal transduction in plants.** *Annu Rev Plant Physiol Plant Mol Biol* 2000, **51:**49-81.
22. Rolland F, Winderickx J, Thevelein JM: **Glucose-sensing and -signalling mechanisms in yeast.** *FEMS yeast research* 2002, **2(2):**183-201.
23. Koch K: **Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development.** *Current opinion in plant biology* 2004, **7:**235-246.
24. Gibson SI: **Control of Plant development and gene expression by sugar signaling.** *Current opinion in plant biology* 2005, **8:**93-102.
25. Rolland F, Baena-Gonzalez E, Sheen J: **Sugar sensing and signaling in plants: conserved and novel mechanisms.** *Annual review of plant biology* 2006, **57:**675-709.
26. Moore B, Zhou L, Rolland F, Hall Q, Cheng W-H, Liu Y-X, Hwang I, Jones T, Sheen J: **Role of the Arabidopsis glucose sensor HXK1 in nutrient, light and hormonal signaling.** *Science* 2003, **3000:**332-336.
27. Halford NG, Paul MJ: **Carbon metabolite sensing and signalling.** *Plant biotechnology journal* 2003, **1(6):**381-398.
28. Zhou L, Jang J-C, Jones TL, Sheen J: **Glucose and ethylene signal transduction crosstalk revealed by an Arabidopsis glucose-insensitive mutant.** *Proc Nat Acad SciUSA* 1998, **95:**10294-10299.
29. Cheng WH, Endo A, Zhou L, Penney J, Chen HC, Arroyo A, Leon P, Nambara E, Asami T, Seo M, *et al.*: **A unique short-chain dehydrogenase/reductase in Arabidopsis glucose signaling and abscisic acid biosyntesis and functions.** *Plant Cell* 2002, **14(11):**2723-2743.
30. Li Y, Lee KK, Walsh S, Smith C, Hadingham S, Sorefan K, Cawley G, Bevan MW: **Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine.** *Genome Res* 2006, **16:**414-427.
31. Rocha FR, Papini-Terzi FS, Nishiyama MY Jr, Vencio RZ, Vicentini R, Duarte RD, de Rosa VE Jr, Vinagre F, Barsalobres C, Medeiros AH, *et al.*: **Signal transduction-related responses to phytohormones and environmental challenges in sugarcane.** *BMC Genomics* 2007, **8:**71.
32. Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Giglioti EA, Lemos MV, Coutinho LL, *et al.*: **Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane.** *Genome Res* 2003, **13(12):**2725-2735.
33. Papini-Terzi FS, Felix JM, Rocha FR, Waclawovsky AJ, Ulian EC, Chabregas S, Falco MC, Nishiyama-Jr MY, Vêncio RZN, Vicentini R, *et al.*: **The SUCEST-FUN Project: identifying genes that regulate sucrose content in sugarcane plants.** *Proc Int Soc Sugar Cane Technol* 2007, **26(6):**25-29.
34. Osuna D, Usadel B, Morcuende R, Gibon Y, Bläsing OE, Höhne M, Günter M, Kamlage B, Trethewey R, Scheible WR, *et al.*: **Temporal responses of transcripts, enzyme activities and metabolites after adding sucrose to carbon-deprived Arabidopsis seedlings.** *Plant J* 2007, **49(3):**463-491.
35. Moore PH: **Temporal and spatial regulation of sucrose accumulation in the sugarcane stem.** *Austr J Plant Physiol* 1995, **22:**661-679.
36. McCormick AJ, Cramer MD, Watt DA: **Changes in photosynthetic rates and gene expression of leaves during a source-sink perturbation in sugarcane.** *Annals of Botany* 2008, **101:**89-102.
37. Ming R, Liu SC, Moore PH, Irvine JE, Paterson AH: **QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane.** *Genome Res* 2001, **11(12):**2075-2084.
38. Casu R, Dimmock C, Thomas M, Bower N, Knight D: **Genetic and expression profiling in sugarcane.** *Proc Int Soc Sugar Cane Technol* 2001, **24:**542-546.
39. Watt D, McCormics A, Govender C, Crame M, Huckett B: **Increasing the utility of genomics in unraveling sucrose accumulation.** *Field Crop Res* 2005, **92:**149-158.
40. Jansen RC, Nap JP: **Genetical genomics: the added value from segregation.** *Trends Genet* 2001, **17(7):**388-391.
41. Ehness R, Ecker M, Godt D, Roitsch T: **Glucose and stress independently regulate source/sink relations and defense mechanisms via signal transduction pathways involving protein phosphorylation.** *The Plant cell* 1997, **9:**1825-1841.
42. Woods A, Munday MR, Scott J, Yang X, Carlson M, Carling D: **Yeast SNF1 is functionally related to mammalian AMP-activated protein kinase and regulates acetyl-CoA carboxylase in vivo.** *The Journal of biological chemistry* 1994, **269(30):**19509-19515.
43. Halford NG, Hardie DG: **SNF1-related protein kinases: global regulators of carbon metabolism in plants?** *Plant Mol Biol* 1998, **37(5):**735-748.
44. Barker JH, Slocombe SP, Ball KL, Hardie DG, Shewry PR, Halford NG: **Evidence that barley 3-hydroxy-3-methylglutaryl-coenzyme a reductase kinase is a member of the sucrose nonfermenting-1-related protein kinase family.** *Plant Physiol* 1996, **112(3):**1141-1149.
45. Sugden C, Donaghy PG, Halford NG, Hardie DG: **Two SNF1-related protein kinases from spinach leaf phosphorylate and inactivate 3-hydroxy-3-methylglutaryl-coenzyme A reductase, nitrate reductase, and sucrose phosphate synthase in vitro.** *Plant Physiol* 1999, **120(1):**257-274.
46. Douglas P, Pigaglio E, Ferrer A, Halfords NG, MacKintosh C: **Three spinach leaf nitrate reductase-3-hydroxy-3-methylglutaryl-CoA reductase kinases that are required by reversible phosphorylation and/or Ca2+ ions.** *The Biochemical journal* 1997, **325(Pt 1):**101-109.
47. Toroser D, Athwal GS, Huber SC: **Site-specific regulatory interaction between spinach leaf sucrose-phosphate synthase and 14-3-3 proteins.** *FEBS letters* 1998, **435(1):**110-114.
48. Sanders D, Pelloux J, Brownlee C, Harper JF: **Calcium at the crossroads of signaling.** *The Plant cell* 2002, **14(Suppl):**S401-417.
49. Boudsocq M, Lauriere C: **Osmotic signaling in plants: multiple pathways mediated by emerging kinase families.** *Plant Physiol* 2005, **138(3):**1185-1194.
50. Huber SC, Huber JL, Liao PC, Gage DA, McMichael RW Jr, Chourey PS, Hannah LC, Koch K: **Phosphorylation of serine-15 of maize leaf sucrose synthase. Occurrence in vivo and possible regulatory significance.** *Plant Physiol* 1996, **112(2):**793-802.
51. Zhang XQ, Lund AA, Sarath G, Cerny RL, Roberts DM, Chollet R: **Soybean nodule sucrose synthase (nodulin-100): further analysis of its phosphorylation using recombinant and authentic root-nodule enzymes.** *Archives of biochemistry and biophysics* 1999, **371(1):**70-82.
52. Hardin SC, Tang GQ, Scholz A, Holtgraewe D, Winter H, Huber SC: **Phosphorylation of sucrose synthase at serine 170: occur-**

rence and possible role as a signal for proteolysis. *Plant J* 2003, **35(5)**:588-603.

53. Hardin SC, Winter H, Huber SC: **Phosphorylation of the amino terminus of maize sucrose synthase in relation to membrane association and enzyme activity.** *Plant Physiol* 2004, **134(4)**:1427-1438.
54. Hardin SC, Huber SC: **Proteasome activity and the post-translational control of sucrose synthase stability in maize leaves.** *Plant Physiol Biochem* 2004, **42(3)**:197-208.
55. Zrenner R, Salanoubat M, Willmitzer L, Sonnewald U: **Evidence of the crucial role of sucrose synthase for sink strength using transgenic potato plants (Solanum tuberosum L.).** *Plant J* 1995, **7(1)**:97-107.
56. Hanggi E, Fleming AJ: **Sucrose synthase expression pattern in young maize leaves: implications for phloem transport.** *Planta* 2001, **214(2)**:326-329.
57. McMichael RW Jr, Bachmann M, Huber SC: **Spinach Leaf Sucrose-Phosphate Synthase and Nitrate Reductase Are Phosphorylated/Inactivated by Multiple Protein Kinases in Vitro.** *Plant Physiol* 1995, **108(3)**:1077-1082.
58. Pagnussat GC, Fiol DF, Salerno GL: **A CDPK type protein kinase is involved in rice SPS light modulation.** *Physiol Plant* 2002, **115(2)**:183-189.
59. Merlot S, Gosti F, Guerrier D, Vavasseur A, Giraudat J: **The ABI1 and ABI2 protein phosphatases 2C act in a negative feedback regulatory loop of the abscisic acid signalling pathway.** *Plant J* 2001, **25(3)**:295-303.
60. Tahtiharju S, Palva T: **Antisense inhibition of protein phosphatase 2C accelerates cold acclimation in Arabidopsis thaliana.** *Plant J* 2001, **26(4)**:461-470.
61. Nambara E, Marion-Poll A: **Abscisic acid biosynthesis and catabolism.** *Annual review of plant biology* 2005, **56**:165-185.
62. Xiong L, Schumaker KS, Zhu JK: **Cell signaling during cold, drought, and salt stress.** *The Plant cell* 2002, **14(Suppl)**:S165-183.
63. Hu H, Dai M, Yao J, Xiao B, Li X, Zhang Q, Xiong L: **Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice.** *Proc Natl Acad Sci USA* 2006, **103(35)**:12987-12992.
64. Lee SC, Lee MY, Kim SJ, Jun SH, An G, Kim SR: **Characterization of an abiotic stress-inducible dehydrin gene, OsDhn1, in rice (Oryza sativa L.).** *Mol Cells* 2005, **19(2)**:212-218.
65. Tabaei-Aghdaei SR, Pearce RS, Harrison P: **Sugars regulate cold-induced gene expression and freezing-tolerance in barley cell cultures.** *J Exp Bot* 2003, **54(387)**:1565-1575.
66. Sakuma Y, Maruyama K, Osakabe Y, Qin F, Seki M, Shinozaki K, Yamaguchi-Shinozaki K: **Functional analysis of an Arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression.** *The Plant cell* 2006, **18(5)**:1292-1309.
67. Sakuma Y, Maruyama K, Qin F, Osakabe Y, Shinozaki K, Yamaguchi-Shinozaki K: **Dual function of an Arabidopsis transcription factor DREB2A in water-stress-responsive and heat-stress-responsive gene expression.** *Proc Natl Acad Sci USA* 2006, **103(49)**:18822-18827.
68. Yilmaz A, Nishiyama MYJ, Fuentes BG, Souza GM, Janies D, Gray J, Grotewold E: **GRASSIUS: A Platform for Comparative Regulatory Genomics across the Grasses.** *Plant Physiol* 2009, **149(1)**:171-180.
69. Vollbrecht E, Veit B, Sinha N, Hake S: **The developmental gene Knotted-1 is a member of a maize homeobox gene family.** *Nature* 1991, **350(3615)**:241-243.
70. Ludwig SR, Habera LF, Dellaporta SL, Wessler SR: **Lc, a member of the maize R gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the myc-homology region.** *Proc Natl Acad Sci USA* 1989, **86(18)**:7092-7096.
71. Aoki K, Ogata Y, Shibata D: **Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology.** *Plant Cell Physiol* 2007, **48(3)**:381-390.
72. Park WJ, Kriechbaumer V, Möller A, Piotrowski M, Meeley RB, A G, Glawischnig E: **The Nitrilase ZmNIT2 converts indole-3-acetonitrile to indole-3-acetic acid.** *Plant Physiol* 2003, **133(2)**:794-802.
73. Shin JH, Jeong DH, Park MC, An G: **Characterization and transcriptional expression of the alpha-expansin gene family in rice.** *Mol Cells* 2005, **20(2)**:210-218.
74. Cosgrove DJ, Li LC, Cho HT, Hoffmann-Benning S, Moore RC, Blecker D: **The growing world of expansins.** *Plant Cell Physiol* 2002, **43(12)**:1436-1444.
75. McQueen-Mason S, Cosgrove DJ: **Disruption of hydrogen bonding between plant cell wall polymers by proteins that induce wall extension.** *Proc Natl Acad Sci USA* 1994, **91(14)**:6574-6578.
76. Mockaitis K, Estelle M: **Integrating transcriptional controls for plant cell expansion.** *Genome Biol* 2004, **5(11)**:245.
77. Farrokhi N, Burton RA, Brownfield L, Hrmova M, Wilson SM, Bacic A, Fincher GB: **Plant cell wall biosynthesis: genetic, biochemical and functional genomics approaches to the identification of key genes.** *Plant biotechnology journal* 2006, **4(2)**:145-167.
78. MacDonald MJ, D'Cunha GB: **A modern view of phenylalanine ammonia lyase.** *Biochem Cell Biol* 2007, **85(3)**:273-282.
79. Boudet AM, Kajita S, Grima-Pettenati J, Goffner D: **Lignins and lignocellulosics: a better control of synthesis for new and improved uses.** *Trends Plant Sci* 2003, **8(12)**:576-581.
80. Lee HI, León J, Raskin I: **Biosynthesis and metabolism of salicylic acid.** *Proc Natl Acad Sci USA* 1995, **92(10)**:4076-4079.
81. Chen F, Dixon RA: **Lignin modification improves fermentable sugar yields for biofuel production.** *Nature biotechnology* 2007, **25**:759-761.
82. Takada S, Hibara K, Ishida T, Tasaka M: **The CUP-SHAPED COTYLEDON1 gene of Arabidopsis regulates shoot apical meristem formation.** *Development* 2001, **128(7)**:1127-1135.
83. Maurel C, Chrispeels MJ: **Aquaporins. A Molecular Entry into Plant Water Relations.** *Plant Physiol* 2001, **125**:135-138.
84. Chaumont F, Moshelion M, Daniels MJ: **Regulation of plant aquaporin activity.** *Biol Cell* 2005, **97(10)**:749-764.
85. Ma S, Quist TM, Ulanov A, Joly R, Bohnert HJ: **Loss of TIP1;1 aquaporin in Arabidopsis leads to cell and plant death.** *Plant J* 2004, **40(6)**:845-859.
86. DiCesare JL: **The analysis of carbohydrates on high eficiency columns using an aqueous fase.** *Chromatography Newsletter* 1980, **8(2)**:52-56.
87. **Instant coffee – Determination of free and total carbohydrate contents – Method using high-performance anion-exchange chromatography.** Standardization IOf. ISO 11292 Geneva: ISO; 1995.
88. Papini-Terzi FS, Rocha FR, Vencio RZ, Oliveira KC, Felix Jde M, Vicentini R, Rocha Cde S, Simoes AC, Ulian EC, di Mauro SM, *et al.*: **Transcription profiling of signal transduction-related genes in sugarcane tissues.** *DNA Res* 2005, **12(1)**:27-38.
89. Murashige T, Skoog F: **A revised medium for rapid growth and bioassys with tobacco tissue culture.** *Physiol Plant* 1962, **15**:473-497.
90. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.** *Methods* 2001, **25(4)**:402-408.
91. Iskandar HM, Simpson RS, Casu RE, Bonnett GD, MacLean DJ, Manners JM: **Comparison of Reference Genes for Quantitative Real-Time Polymerase Chain Reaction Analysis of Gene Expression in Sugarcane.** *Plant Molecular Biology Reporter* 2004, **22**:325-337.
92. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
93. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
94. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4(4)**:406-425.
95. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24(8)**:1596-1599.

DECLARAÇÃO

Declaro para os devidos fins que o conteúdo de minha dissertação de Mestrado/tese de Doutorado intitulada "Evolução de famílias multigênicas e redes de regulação em plantas":

(   ) não se enquadra no § 3º do Artigo 1º da Informação CCPG 01/08, referente a bioética e biossegurança.

Tem autorização da(s) seguinte(s) Comissão(ões):

(✗) CIBio – Comissão Interna de Biossegurança , projeto No. 2008/01, Instituição: IB - Unicamp.

(   ) CEUA – Comissão de Ética no Uso de Animais , projeto No. ______________, Instituição: ______________________________.

(   ) CEP - Comissão de Ética em Pesquisa, protocolo No. ______________, Instituição: ______________________________

** Caso a Comissão seja externa ao IB/UNICAMP, anexar o comprovante de autorização dada ao trabalho. Se a autorização não tiver sido dada diretamente ao trabalho de tese ou dissertação, deverá ser anexado também um comprovante do vínculo do trabalho do aluno com o que constar no documento de autorização apresentado.*

____
Aluno: Luis Eduardo Vieira Del Bem

____
Orientador: Michel Georges Albert Vincentz

Para uso da Comissão ou Comitê pertinente:
(✓) Deferido (   ) Indeferido

____
Carimbo e assinatura

Prof. Dr. MARCELO LANCELLOTTI
Presidente da Comissão Interna de Biossegurança
Instituto de Biologia - UNICAMP

Para uso da Comissão ou Comitê pertinente:
(   ) Deferido (   ) Indeferido

____
Carimbo e assinatura