ANÁLISE EVOLUTIVA E FUNCIONAL DOS GENES DA FAMÍLA F-BOX EM LEGUMINOSAS (FABACEAE)

DANIEL BELLIENY RABELO

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO

CAMPOS DOS GOYTACAZES - RJ.

ABRIL, 2013.

ANÁLISE EVOLUTIVA E FUNCIONAL DOS GENES DA FAMÍLA F-BOX EM LEGUMINOSAS (FABACEAE)

DANIEL BELLIENY RABELO

Dissertação apresentada à Universidade Estadual do Norte Fluminense Darcy Ribeiro para obtenção do título de Mestre em Biociências e Biotecnologia.

ORIENTADOR: Dr. THIAGO MOTTA VENANCIO

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO

CAMPOS DOS GOYTACAZES - RJ.

ABRIL, 2013.

ANÁLISE EVOLUTIVA E FUNCIONAL DOS GENES DA FAMÍLA F-BOX EM LEGUMINOSAS (FABACEAE)

DANIEL BELLIENY RABELO

Dissertação apresentada à Universidade Estadual do Norte Fluminense Darcy Ribeiro para obtenção do título de Mestre em Biociências e Biotecnologia.

ORIENTADOR: Dr. THIAGO MOTTA VENANCIO

Aprovada em 1 de abril de 2013

COMISSÃO EXAMINADORA

_____

Prof. Dr. Thiago Motta Venancio (PhD-UENF)

_____

Prof. Dr. Jorge Hernandez Fernandez (PhD-UENF)

_____

Prof. Dr. Vanildo Silveira (PhD-UENF)

_____

Prof.ª Dr.ª Adriana Silva Hemerly (PhD-UFRJ)

Este trabalho é dedicado a Elzi Bellieny (*in memorian*)

Agradecimentos

À Regina Bellieny pelos valores ensinados. Pela dedicação, e amor que que provou inúmeras vezes serem infindáveis. Nada do que conquistei teria sido possível sem sua mão terna a me guiar.

Aos padrinhos Carlos Manoel Bellieny e Fátima Bellieny que sempre forneceram todo apoio e atenção em todas as fases deste percurso.

Ao orientador Dr. Thiago Venancio pelo olhar atento e perfeccionismo incansáveis que contribuem de forma inestimável para o crescimento de seus atentos *padawans*.

Aos colegas de laboratório, presentes ou que passaram, pelo convívio harmonioso e tranquilo com leves toques de diversão: Newton, Ana Laura, Cris, Isabela, Esther. Em especial ao amigo da vizinhança, Gustavo Lemos.

À professora Drª Elenir Oliveira, a quem sempre serei grato por ter participado na minha iniciação científica como orientadora. E por ter incentivado esse trabalho de mestrado.

Aos amigos que estão sempre presentes e são extensões da família: Estevão, Marco, Raphael, Victor, Nicholas, Letícia, Ygor e Lyncoln.

À namorada Andara Gualberto pela participação fundamental nessa caminhada. Pelo carinho, paciência, apoio e pela dedicação imensa. Também por todos os ensaios de apresentação que assistiu (e ainda vai assistir).

RESUMO

A modificação covalente é uma estratégia utilizada frequentemente na regulação de uma série de processos biológicos. Dentre os modificadores polipeptídicos conhecidos, a ubiquitina é o mais utilizado, promovendo tipicamente a degradação de seus alvos. O processo de ubiquitinação depende primordialmente de três enzimas: uma enzima de ativação da ubiquitina (E1), uma enzima de conjugação à ubiquitina (E2), e uma ubiquitina-ligase (E3), que é responsável pelo reconhecimento das proteínas-alvo. A ativação da ubiquitina ocorre por meio da ação da E1. E2 e E3 cooperam para montagem de uma cadeia multiubiquitina na proteína alvo. A ligação de moléculas de ubiquitina orienta a degradação das proteínas pelo complexo 26S proteassomo. A poliubiquitinação está associada a processos fisiológicos diversos, desde regulação do ciclo celular e desenvolvimento até resposta a estímulos ambientais. O complexo SCF (Skp1-Cullin1-F-box) é uma das ubiquitina-ligases mais bem descritas na literatura atualmente. Proteínas da família F-box são responsáveis pelo reconhecimento de substratos para atuação do complexo na via de ubiquitinação. As expansões ocorridas no grupo das F-box as coloca entre as maiores famílias gênicas de plantas. O presente trabalho apresenta o impacto diferencial causado por eventos de duplicação sobre a expansão e conservação da família F-box em diferentes espécies de leguminosas (*Glycine max* (soja) e *Medicago truncatula*). Nossos resultados mostram grandes diferenças tanto no conteúdo de genes F-box entre espécies próximas, quanto no meio pelo qual esta família se prolifera. Enquanto na soja os genes F-box tendem a ser mais antigos e dispostos em regiões sintênicas com outras dicotiledôneas, em *Medicago* o repertório desta família evoluiu preponderantemente por duplicações locais, ou em tandem. Reportamos ainda a divergência transcricional em vários destes genes oriundos de duplicações recentes, bem como a rápida evolução de transcrição tecido-específica e recrutamento para processos biológicos distintos como nodulação e desenvolvimento da semente, colocando-os como potenciais reguladores de processos adaptativos.

ABSTRACT

Covalent modification of proteins is often used to regulate a number of biological processes. Among the polypeptide modifiers, ubiquitin is the most used, typically inducing protein degradation. Ubiquitination primarily depends on three enzymes: ubiquitin activating enzyme (E1), ubiquitin conjugating enzyme to (E2), and ubiquitin ligase (E3), being the latter responsible for substrate specificity. Ubiquitination occurs through the concerted action of E1s, E2s, E3s and other less-conserved accessory proteins to assemble a polyubiquitin structure in the target protein. The attachment of ubiquitin molecules typically drive protein degradation by the 26S proteasome complex. Polyubiquitination is associated with various physiological processes, from cell cycle regulation and development to stress response. The SCF complex (Skp1-Cullin1-F-box) is one of the best described ubiquitin ligases and the F-box proteins are the subunit responsible for substrate specificity. Moreover, due to several lineage-specific expansions, the F-box family is arguably the largest gene families in plants. This work presents the differential impact caused by large-scale and tandem duplication events on the expansion and maintenance of F-box family in two closely-related legumes (*Glycine max* (soybean) and *Medicago truncatula*). Our results show large differences in both, content and origins of the F-box family members. While soybean F-box genes tend to be older and arranged in synthenic regions with other dicots, the *M. truncatula* F-box repertoire mainly evolved by local (tandem). Several of these young F-box genes evolved divergent transcriptional patterns and recruitment to various biological processes (e.g. nodulation and seed development), indicating that they are potential regulators of adaptive processes.

ABREVIAÇÕES

NBS-LRR – *nucleotide-binding-site leucine-rich-repeat*

HECT - *homologous to the E6-AP carboxyl terminus*

RING – *really interesting new gene*

Cul1 – *cullin 1*

FBX – F-box

TD-FBX – F-box duplicado localmente (tandem)

TD-F-box - F-box duplicado localmente (tandem)

***HAI_IN_RH –** *hours after inoculation _ inoculated _ root hair*

***HAI_UN_RH -** *hours after inoculation _ mock-inoculated _ root hair*

Índice

Introdução

A relevância da soja no âmbito econômico

Sementes são cultivadas para diversas finalidades, dentre elas a produção de biocombustíveis e alimentação humana e animal. Sementes acumulam proteínas, carboidratos e lipídeos, que são consumidos para sustentar as etapas iniciais do desenvolvimento pós-germinativo (Schmidt et al., 2011). Durante o desenvolvimento da semente, o acúmulo de proteínas de reserva é regulado por uma complexa rede de componentes genéticos, fisiológicos e ambientais (Golombek et al., 2001; BrocardGifford et al., 2003; Elke et al., 2005; Fait et al., 2006; Weber et al.,2005). A soja (*Glycine max* Merr.) é a leguminosa mais cultivada no mundo, constituindo uma importante fonte de proteína e óleo. A produção mundial de soja e o rendimento por hectare aumentou de forma constante ao longo do século passado em decorrência do desenvolvimento tecnológico e lançamento de cultivares adaptadas a diferentes ambientes. A soja é um componente chave da alimentação humana e animal, sendo responsável por mais da metade da produção de oleaginosas do mundo [United States Department of Agriculture (USDA) Serviço AGRÍCOLAS Exteriores do banco de dados de produção, abastecimento e distribuição - http://www.usdabrazil.org.br]. Em termos de massa de sementes produzidas, a soja é o quarto cultivo mais importante do mundo e ocupa nos Estados Unidos o segundo lugar em área plantada (FAOSTAT 2010; http://faostat.fao.org/default.aspx). Em 1990, as importações líquidas chinesas de soja foram de um 1 teragrama (Tg). Esse número aumentou para 33 Tg até 2007, acompanhando o crescimento exponencial da economia daquele país (FAOSTAT 2010; http://faostat.fao.org/).

A produção atual mundial de soja é de mais de 255.000 Tg, dos quais mais de 95% são provenientes de apenas sete países (Ainsworth et al., 2011). Embora a soja tenha sido inicialmente domesticada na China e permanecido ausente do Novo Mundo até a colonização europeia, atualmente 80% da produção mundial é proveniente de Estados Unidos, Brasil e Argentina. O rendimento da soja nesses três países têm aumentado constantemente ao longo das últimas duas décadas (Ainsworth et al., 2011), sendo esta crescente aliada ao dramático aumento na área plantada praticamente dobraram a produção mundial da cultura desde 1990

(Ainsworth et al., 2011). A produção comercial de soja no Brasil começou na década de 1960 e superou 40 milhões de toneladas em 2002. O Brasil atualmente é o segundo maior produtor de soja do mundo ao lado dos EUA (FAOSTAT 2010: http://faostat.fao.org/). O rendimento médio da soja no Brasil é de quase 3 toneladas por hectare, podendo ser considerado altamente produtivo e comparável ao de outros grandes produtores das Américas do Norte e do Sul (FAOSTAT 2010; http://faostat.fao.org/).

Nodulação e fixação biológica do nitrogênio

Uma característica importante das leguminosas é a sua capacidade de estabelecer simbiose com bactérias fixadoras de nitrogênio (família Rhizobiaceae) (Schmutz et al., 2010), num processo denominado nodulação. O desenvolvimento de nódulos depende de complexas interações entre os simbiontes, sendo sua morfologia frequentemente distinta em diferentes leguminosas (Jones et al., 2007; Maunoury et. al., 2008). Na interação entre *Medicago truncatula* e *Sinorhizobium meliloti*, os primeiros processos de sinalização envolvem flavonóides derivados de plantas e fatores Nod derivados de *Rhizobium*, induzindo a divisão celular no córtex radicular interno e levando à formação do primórdio nodular (Maunoury et al., 2010). Tem sido evidenciado nos últimos anos que a assimilação de nitrogênio e o metabolismo de carbono são bioquimicamente coordenados por uma complexa rede transcricional e enzimática (revisado por Nunes-Nesi, 2010). Esta simbiose baseia-se na alocação de carbono para o simbionte *Bradyrhizobium japonicum* em troca de nitrogênio fixado. A partir de ortólogos conhecidos em outras espécies, foram identificadas em soja 28 nodulinas e 24 genes reguladores da nodulação (Schmutz et al., 2010).

Dentre os estresses abióticos, a seca é o fator que causa mais prejuízos, prejudicando o desenvolvimento da planta e resultando em perdas de aproximadamente 40% (Ladrera et al., 2007). A seca também limita significativamente a fixação biológica de nitrogênio em nódulos, pois afeta a disponibilidade de oxigênio e carbono para a fixação de nitrogênio pelo bacteróide, além de induzir o acúmulo de ureídes nos nódulos (Ladrera et al 2007; Marino et

al.2007). Várias centenas de genes que são especificamente induzidos ou reprimidos durante a nodulação tem sido identificados, sendo sua maioria ativada em estágios tardios da organogênese (El Yahyaoui et al., 2004). Estudos sobre os perfis transcricionais relacionados a interação simbióticas entre *M. truncatula* e *S.meliloti* documentaram 13 possíveis fatores de transcrição (TFs) regulados positivamente e 11 regulados negativamente em nódulos. Nodulinas correguladas junto a estes TFs também foram encontradas, sugerindo fortemente que elas tenham sua transcrição controlada por estes fatores.

Características do genoma da soja

A recente publicação do genoma da soja (Schmutz et al., 2010) abriu grandes perspectivas para pesquisas bioquímicas e fisiológicas. Dentre estas iniciativas, estudos em larga escala vem sendo conduzidos para estudar o transcriptoma, proteoma e metaboloma de diferentes tecidos e condições fisiológicas da espécie (Libault et al., 2010). O genoma da soja está entre os maiores genomas vegetais sequenciados até o momento e apresenta boa qualidade se comparado a outros drafts genômicos de alta qualidade (Schmutz et al., 2010). O genoma de soja tem aproximadamente 950 megabases, sendo 57% das bases contidas em regiões heterocromáticas, ricas em repetições e próximas aos centrômeros. Dos 46.430 genes preditos com alta confiança, 34.073 (73%) apresentaram ortólogos em outras angiospermas, sendo classificados em 12.523 famílias (Schmutz et al., 2010). Expansões em tandem de famílias de genes são comuns na cultura da soja e incluem NBS-LRR, F-box, proteínas de resposta a auxina, e outros domínios comumente encontrados em grandes famílias em plantas (Schmutz et al., 2010).

A maioria das espécies de angiospermas e vertebrados descendem de ancestrais que sofreram duplicações de genoma, seja através de autopoliploidia ou alopoliploidia. Evidências citogenéticas, morfológicas e genômicas sugerem que a maioria das angiospermas (60-70%) sejam derivadas de ancestrais que sofreram algum evento de poliploidização. É normalmente esperado que alterações de ploidia sejam deletérias em decorrência do desbalanço genético e metabólico (Otto e Whitton, 2000), apesar de as evidências de duplicação genômica serem claras em

muitos dos genomas sequenciados, com surpreendente prevalência em plantas (Sémon e Wolfe, 2007).

Dado o alto número de duplicações genômicas, alguns cientistas argumentam que a poliploidização ofereça grande potencial adaptativo. Crow e Wagner (2006) sugeriram que a duplicação do genoma poderia reduzir o risco de extinção através de redundância funcional, robustez mutacional, e taxas aumentadas de evolução e adaptação (Van de Peer et al., 2009). Contudo, é importante notar que em 500-600 milhões de anos de evolução dos vertebrados, não mais que duas (ou três em teleósteos) duplicações genômicas completas foram detectadas até o momento (Van de Peer et al., 2009), evidenciando que a prevalência dos eventos de poliploidização observada em plantas não se repete em vertebrados.

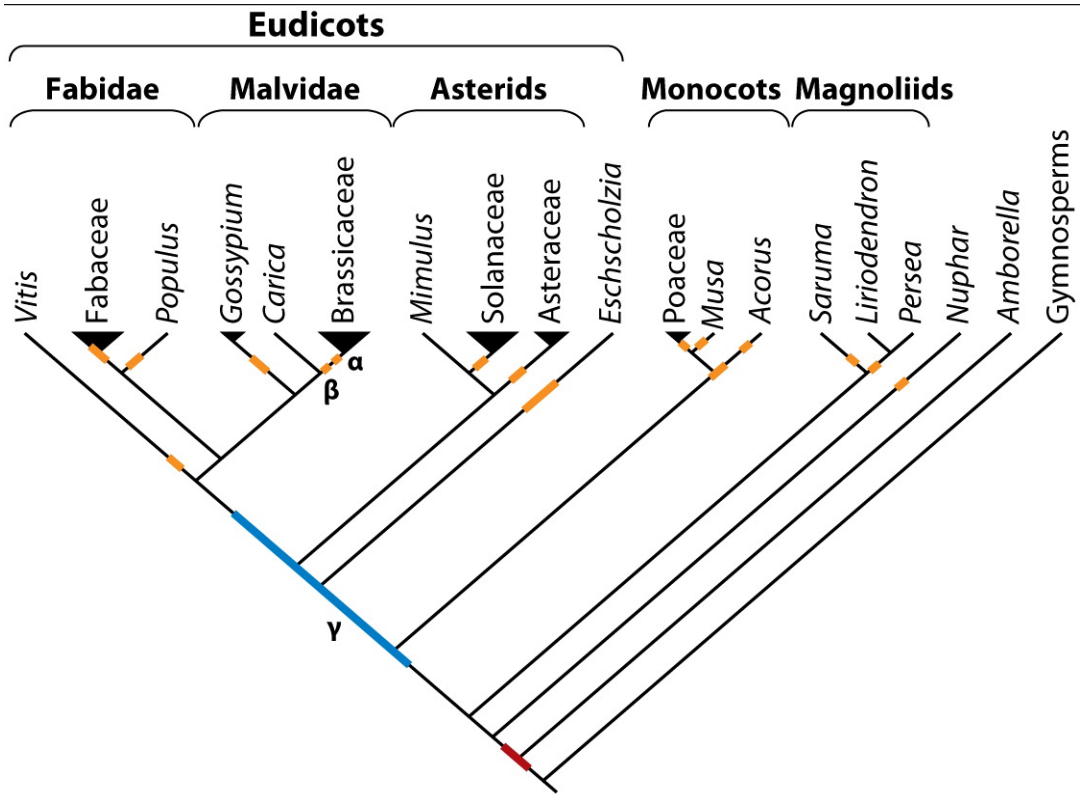

Figura 1: Árvore filogenética representada de forma simplificada. As barras coloridas representam eventos de duplicação de genomas. Resultado de poliploidias ocorridas nas linhagens. Fonte: Soltis e Soltis, 2009.

Efeitos da abordagem biotecnológica sobre a produção da soja

Melhoramento genético, lançamento de novos cultivares e avanço das técnicas agrícolas têm contribuído para aumentos substanciais na produção de soja ao longo dos anos (Specht et al., 1999). Apesar do sucesso dos programas de melhoramento genético clássico, é imperativa a necessidade de se compreender as bases genéticas e bioquímicas de fenótipos de interesse a fim de identificar potenciais alvos para futuras melhorias e extrapolação para outras espécies (Ainsworth et al., 2011).

Análises históricas de cultivares lançadas ao longo dos últimos 90 anos indicam que um dos principais fatores do aumento do rendimento é o número de sementes por planta (Morrison et al., 2000; Jin et al 2010). Por outro lado, patógenos e doenças impõem severas limitações à produção agrícola mundial. Enquanto as plantas são naturalmente imunes à grande maioria dos potenciais patógenos de bactérias e fungos do ambiente, eventualmente patógenos e pragas conseguem se estabelecer, resultando em enormes prejuízos econômicos. O controle de patógenos tem sido essencialmente feito por programas de melhoramento genético e/ou utilização de pesticidas, sendo a primeira estratégia extremamente eficaz na geração de plantas altamente resistentes a patógenos específicos (Wally e Punja, 2010).

A correlação histórica positiva entre fotossíntese e rendimento de soja sugere que esta via poderia ser um alvo promissor para ganhos de rendimento adicional (Ainsworth et al., 2011). Outra importante função que tem sido constante alvo de investigação é o armazenamento de lipídios. *Arabidopsis thaliana* e soja possuem números comparáveis de genes envolvidos na síntese de lipídios de reserva, alongamento de ácidos graxos e produção de cera/cutina (Schmutz et al., 2010). Curiosamente, o número de genes envolvidos na sinalização lipídica, degradação de lipídios de reserva, e síntese de lipídios de membrana é de duas a três vezes maior em soja que em *A. thaliana*, indicando uma maior complexidade regulatória no metabolismo de lipídeos na leguminosa (Schmutz et al., 2010).

A via de ubiquitinação

A modificação de proteínas por polipeptídeos pode alterar vários aspectos bioquímicos, tanto em células eucarióticas quanto procarióticas. Embora o sistema de ubiquinação seja exclusivo de eucariotos, diversas proteínas da via apresentam homologia com proteínas procarióticas (Burroughs *et al.,* 2009*;* Iyer *et al.,* 2006). Dentre esses peptídeos modificadores, o mais conhecido é a ubiquitina, um peptídeo altamente conservado em eucariotos (Hochstrasser, 2000; Hershko e Ciechanover, 1998; Pickart, 2001; Weissman, 2001; Glickman e Ciechanover, 2002). Após processamento da pré-ubiquitina para a sua forma ativa monomérica, a ubiquitina pode ser covalentemente ligada a outras proteínas por um conjunto de processos bioquímicos coletivamente chamados de ubiquitinação ou ubiquitilação.

A ubiquitinação tipicamente induz a degradação específica de seus alvos. A degradação de proteínas regulatórias mediada por ubiquitinação desempenha uma importante função no controle de diversos processos como a progressão do ciclo celular, transdução de sinal, regulação transcricional e endocitose (Fang e Weissman, 2004 ; Hershko e Ciechanover, 1998). A ubiquitina possui sete resíduos de lisina (K27, K29, K33, K48 e K63) cujos vários tipos de ligações são associadas a diferentes funções celulares. Por exemplo, as cadeias poliubiquitina ligados por K48, estão envolvidos em degradação proteossomal, enquanto a ubiquitinação por K63 é um sítio de acoplamento para mediação de internalização de proteínas de membrana, interação proteína-proteína ou mudanças conformacionais (Mallette e Richard, 2012). Em geral, o esqueleto da via ubiquitinação é composto de por três enzimas: a enzima de ativação da ubiquitina, ou E1; E2, ou enzima de conjugação da ubiquitina; e E3, ou ubiquitina-ligase (Figura 2). E1, a primeira enzima da via da ubiquitinação, forma uma ligação tiol-éster entre a cisteína do seu sítio ativo e a glicina carboxi-terminal da ubiquitina. Cada molécula de E1 totalmente carregada leva duas ubiquitinas ativadas. A ubiquitina ativada na E1 é subsequentemente transferida ao sítio ativo de uma E2 por transesterificação. E3 liga-se à E2, já combinada à ubiquitina, e ao substrato (proteína-alvo) facilitando a formação de um isopeptídeo ligado pela glicina carboxi-terminal da ubiquitina e um resíduo de lisina do substrato ou à uma outra ubiquitina previamente aderida a esse substrato

(Hershko e Ciechanover, 1998; Pickart, 2001). Alternativamente, a E3 pode ter a ubiquitina ligada covalentemente a sua estrutura e transferi-la posteriormente ao substrato, de forma E2-independente. A organização da cascata enzimática de conjugação é hierárquica: Há uma E1; um número limitado de E2's, cada uma das quais pode servir várias E3's; presentes em dezenas ou até centenas (muitas delas ainda não identificadas) (Venancio *et al.,* 2009).

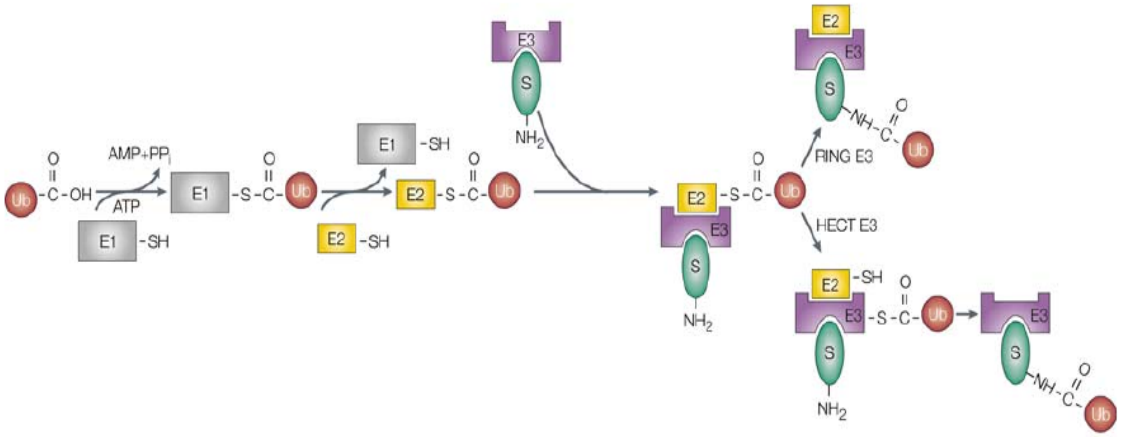

Figura 2: A via de ubiquitinação. Ubiquitina livre (Ub) é ativada de modo dependente de ATP com a formação de ligação tiol-éster entre E1 e a porção C-terminal da ubiquitina. A ubiquitina é transferida para uma ou várias E2's diferentes. E2 associa-se a E3, que já deverá possuir um substrato a ela ligado. A ubiquitina é então transferida diretamente de E2 ao substrato. Fonte: Fang e Weissman, 2004.

A reação da E3 envolve pelo menos duas etapas distintas: ligação da E3 ao substrato através do sinal de ubiquitinação, e a ligação covalente de uma ou mais ubiquitinas ao substrato. Além disso, os substratos que serão marcados para reconhecimento pelo proteassomo normalmente possuem uma cadeia poliubiquitina ligada a K48 estendida de uma das ubiquitinas já ligadas e ele (Pickart, 2001). As ubiquitina-ligases (E3) pertencem tipicamente à famílias gênicas com o domínio HECT, que transfere a ubiquitina diretamente do substrato ligado a um domínio não catalítico. Proteínas HECT são encontradas em diversas linhagens de eucariotos e são definidas por um domínio HECT de aproximadamente 350 aminoácidos (Scheffner, 1998). São proteínas, em geral, de grande massa molecular (90 a 200kDa) com domínios N-terminais longos. Nessa classe de ubiquitina-ligases, o domínio N-terminal é aparentemente responsável pela ligação ao substrato

enquanto o domínio C-terminal HECT atua na transferência direta da ubiquitina de uma ligação tiol-éster para o substrato.

Por outro lado, os membros da outra classe catalítica, como SCF e complexos APC, utilizam um domínio RING-*finger* para facilitar a ubiquitinação (Jackson *et al*., 2000). O APC (*anaphase-promoting complex*) foi a primeira ubiquitina-ligase multicomponente descrita, e é necessária para a degradação de substratos que controlam a transição da metáfase para anáfase e para a degradação da ciclina B para encerramento da mitose (Peters, 1998; Wolf e Jackson, 1998; Zachariae e Nasmyth, 1999). De maneira similar ao complexo SCF, o APC contém um homólogo da culina (Apc2), e uma proteína RING-H2-*finger* similar a Roc1/Rbx1, chamada Apc11.

O complexo SCF (Skp1-*Cullin*1-F-box) contém pelo menos quatro proteínas: Skp1, Cul1, Roc1/Rbx1/Hrt1, e uma proteína F-box. Os substratos do SCF são ligados diretamente por adaptadores chamados de proteínas F-box. Estes contém um motivo de aproximadamente 45 aminoácidos (domínio F-box) e ligam-se aos substratos por uma interação proteína-proteína (Bai *et al*., 1996). O domínio F-box é responsável pela ligação com Skp1, que por sua vez associa-se com a porção N-terminal de Cul1 (Jackson *et al.,* 2000). Proteínas F-box direcionam a ubiquitinação de substratos específicos (Bai *et al.,* 1996; Skowyra *et al.,* 1997), mas a gama de F-boxes e substratos é bastante ampla. Há pelo menos 19 proteínas F-box em *S. cerevisiae*, mais de 100 em *C. elegans*, e aproximadamente 50 descritas em vertebrados (Cenciarelli *et al.,* 1999; Regan-Reimann *et al.,* 1999; Winston *et al.,* 1999). Assim, como muitas dessas proteínas F-box constituem ubiquitina-ligases SCF ativas, o número de substratos de complexos SCF pode ser bastante variado (centenas ou milhares). Em complexos SCF bem estudados como SCF(Cdc4), SCF(Grr1), SCF(Skp2) e SCF(β-TrCP), a fosforilação parece ser necessária para ligação da proteína F-box. No complexo SCF (Cdc4), cada um dos substratos conhecidos necessita estar fosforilado para que ocorra a ubiquitinação (Deshaies, 1999). Cul1 é um membro da família das culinas, que, em humanos, incluem pelo menos Cul1-Cul5, e Apc2. As culinas organizam e ativam o complexo E3 e colaboram no recrutamento de E2. Roc1/Rbx1 contém domínio RING-H2-*finger*, e promove a associação da proteína Cul1 com a enzima de

conjugação à ubiquitina (E2), além de aumentar a atividade da própria ubiquitina ligase (Kamura *et al.,* 1999; Ohta *et al.,* 1999; Seol *et al.,*1999; Skowyra *et al.,* 1999). A RING-H2 pertence a classe das RING-*finger* (*Really Interesting New Protein colocar o significado a primeira vez que aparece a sigla RING*) que contém um octeto de cisteína e resíduos de histidina que participam da ligação com E2. A reconstituição bioquímica de complexos SCF sugere que esses quatro componentes são suficientes para ubiquitinar substratos específicos (Skowyra *et al.,* 1997; Feldman *et al.,* 1997).

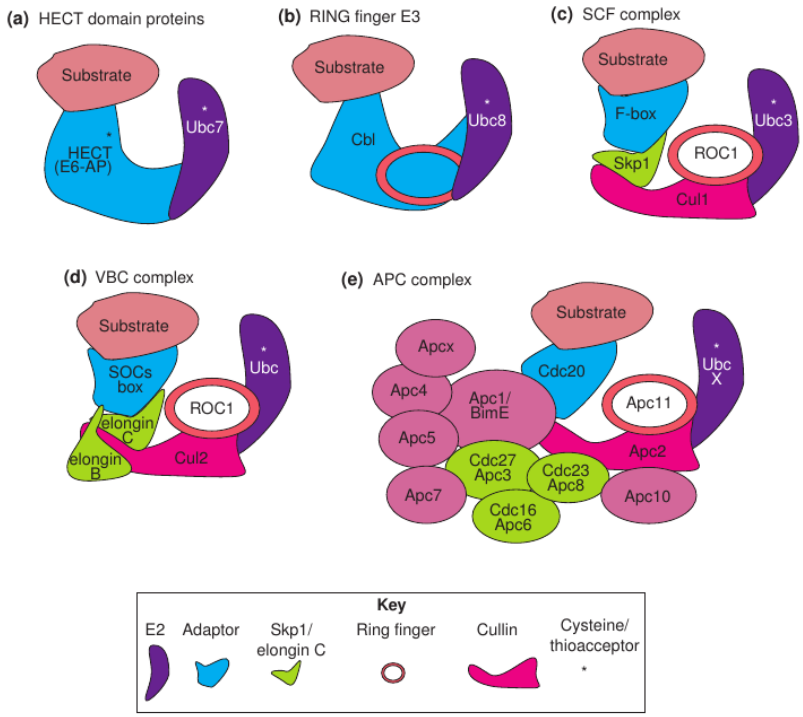

Figura 3: Comparação entre algumas das principais ubiquitina-ligases. (a) Proteína de domínio HECT. (b) Proteína de domínio RING-finger. (c) Complexo SCF. (d) Complexo VBC. (e) Complexo APC. Fonte: Jackson et al., 2000

São conhecidas diversas proteínas cujo mecanismo de ação é semelhante ao da ubiquitina. Pelo menos 70 famílias de Ubl's (*ubiquitin-like modifiers*) estão distribuídas entre os eucariotos, das quais aproximadamente 20 estavam presentes no ancestral comum todos os organismos do clado (Burroughs *et al.,* 2012). A diversificação inicial das Ubls em eucariotos teve um importante papel na emergência de subestruturas celulares características dos eucariotos, bem como de

sistemas referentes a compartimentalização núcleo-citoplasma, tráfego vesicular, direcionamento lisossomal, processamento protéico no retículo endoplasmático e dinâmica da cromatina (Burroughs *et al.,* 2012; Venancio *et al.,* 2009). Dentre todas essas, a SUMO (*small-ubiquitin-related modifier*) e a RUB1 (*related-to-ubiquitin* 1) tem recebido a maior atenção. O primeiro alvo de SUMO a ser identificado foi a proteína de transporte nucleoplasmático RunGAP1. Sua atuação na importação de algumas proteínas através do poro nuclear depende do sistema de conjugação de SUMO. Nesse caso específico, SUMO liga-se a um resíduo específico de lisina da proteína RunGap1 (Haas e Siepmann, 1997). Em contraste do grande número de proteínas modificadas por SUMO, um número menor de proteínas modificadas por RUB1 é conhecido até o momento. Todas as proteínas modificadas por RUB1 conhecidas até o momento pertencem à família das culinas (Hochstrasser, 1998; Read *et al.,* 2000). Uma E3 conhecida como SCF (β-TrCP), responsável pela ubiquitinação de proteínas IκB, necessita ter sua Cul1 rubilada para realizar a ligação da ubiquitina à proteína-alvo citada (Read *et al.,* 2000). É possível que a via de RUB1 tenha evoluído como um regulador do sistema de ubiquitinação, para limitar eventualmente a auto-ubiquitinação das subunidades de E3, o que pode regular negativamente a atividade de E3. Ou ainda, para acionar uma mudança conformacional em um complexo ubiquitina-E2-E3 subsequentemente formado que estimula a poliubiquitinação do substrato (Hochstrasser, 2000).

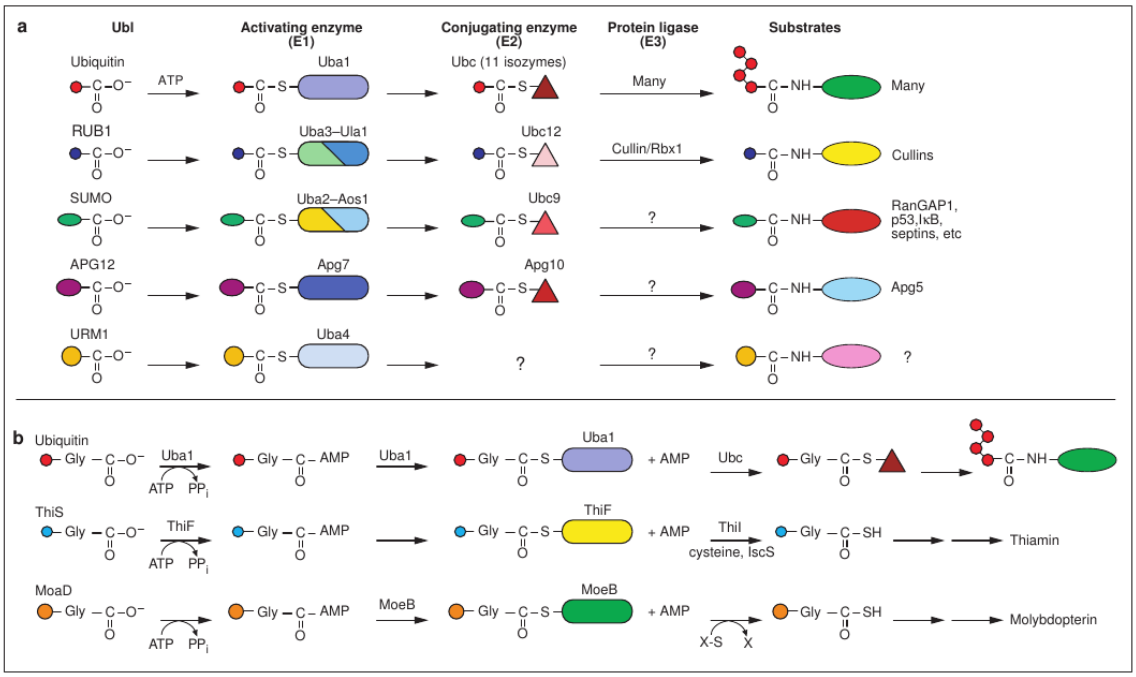

Figura 4: A família ubiquitina de proteínas modificadoras. (a) Via de conjugação à ubiquitina das proteínas Ubls (ubiquitin-like) incluindo as respectivas E1s, E2s, E3s e os seus substratos. (b) Comparação entre ativação de Ub/Ubl e vias biossintéticas de cofatores enzimáticos. Fonte: Hochstrasser, 2000.

A família multigênica F-box

A expansão de famílias gênicas (gerando múltiplos parálogos) tem sido proposta como uma das principais fontes de matéria-prima para a adaptação de organismos a diferentes ambientes (Li, 1983; Nei e Rooney, 2005; Xu et al., 2009). Genes parálogos podem dividir funções (subfuncionalização) ou adquirir novas funções (neofuncionalização). Alternativamente, genes duplicados podem ainda sofrer mutações, sendo inativados e deletados ou permanecendo no genoma na forma de pseudogenes (Hua et al., 2011). A família F-box é uma das maiores famílias multigênicas em plantas, estando alguns de seus membros envolvidos na regulação do desenvolvimento e respostas a estresse (Yan et al., 2011). Xu *et al.*, 2009 reportou 692, 337 e 779 genes F-box em *Arabidopsis thaliana*, *Populus trichocarpa,* e *Oryza sativa* respectivamente, ilustrando a grande representatividade da família no grupo. Proteínas F-box recrutam especificamente substratos do complexo ubiquitina ligase (E3) conhecido como SCF (Skp1-Cullin-F-box), sendo assim responsável pela especificidade do reconhecimento (Lechner et al., 2006). De

forma mais ampla, sabe-se que a degradação proteica pelo proteossomo (um processo relativamente conservado durante o curso da evolução) requer a ligação de várias moléculas de ubiquitina às proteínas alvo. E1 e E2 são pouco específicas, enquanto diferentes E3s reconhecem conjuntos de substratos distintos para ubiquitinação. A ubiquitina ligase (E3) mais bem caracterizada é o complexo SCF (SKP1-CUL1-F-box) (Cardozo et al., 2004). Proteínas F-box são definidas por uma assinatura peculiar de aproximadamente 45-60 aminoácidos na porção N-terminal (Kipreos e Pagano, 2000) que determinam o domínio F-box propriamente dito. Enquanto a extremidade oposta (C-terminal) compreende um módulo que recruta o substrato do complexo. Alguns dos módulos de recrutamento mais bem caracterizados são: LRR (leucine-rich repeat), Kelch, WD-40, Tubby e armadillo (Gagne et al., 2002; Xu et al., 2009).

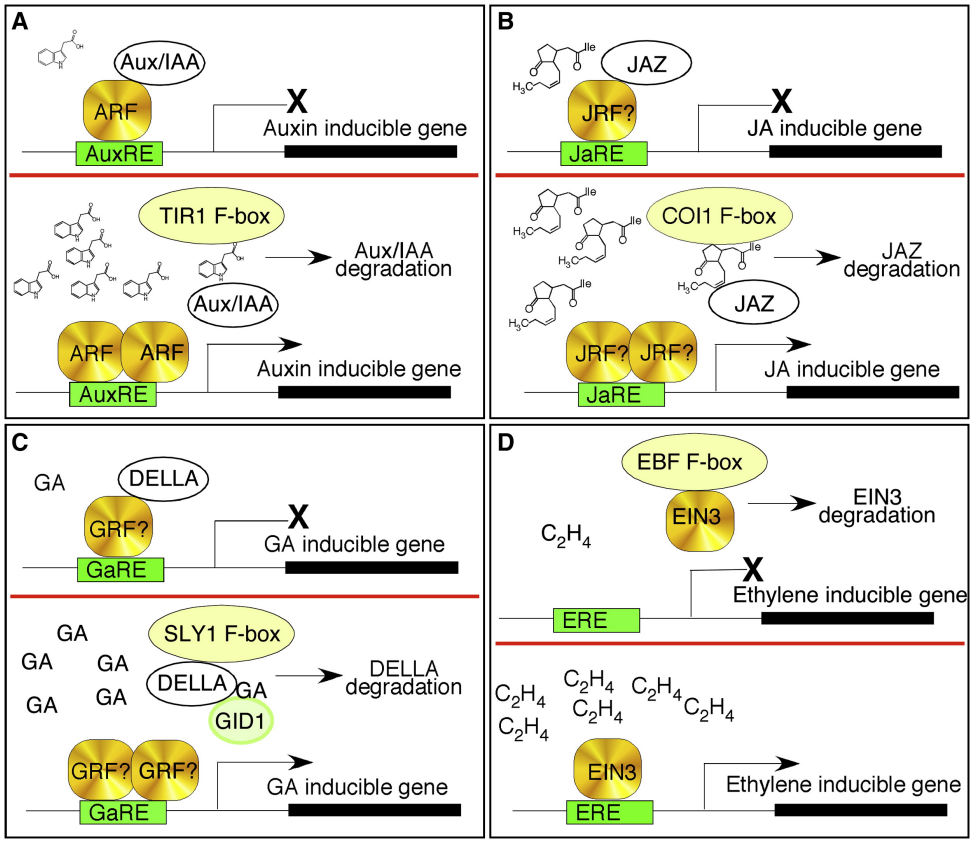

Figura 5: Participação crucial da via de ubiquitinação na sinalização por diversos fitormônios.A) Auxina; B) Jasmonato; C) Giberelina; D) Etileno. A ubiquitinação leva normalmente a degradaçãode repressores transcricionais, ativando programas de expressão gênica específicos. Adaptado de McSteen e Zhao (2008).

Um dado importante a respeito dos F-box, é a grande variação quantitativa com que se apresentam nos genomas das diferentes espécies (Bai et al., 1994; Gagne et al., 2002). Em *Saccharomyces cerevisiae*, *Caenorhabidtis elegans*, *Drosophila melanogaster* e *Homo sapiens* há, respectivamente: 14, 337, 24 e 38 genes F-box (Kipreos e Pagano, 2000). Especificamente em plantas, Hua *et al.,* 2011 confirmaram a variação substancial do número de F-boxes em espécies vegetais sem correlação com o tamanho dos genomas. Este mesmo estudo confirmou ainda que espécies próximas filogeneticamente podem apresentar conjuntos bastante divergentes de F-boxes devido a ganhos e perdas de genes ocorridos nas linhagens. Provavelmente devido a seu papel na ubiquitinação e degradação, genes F-box tem sido associados ao controle genético de vários processos essenciais tais como: desenvolvimento de plântulas, respostas a diferentes estresses, embriogênese, sinalização hormonal, organogênese floral, senescência e resistência a patógenos (Lechner et al., 2006). Os alvos mais notórios dos complexos SCF em plantas são fatores de transcrição pertencentes às principais vias de sinalização mediadas pelos principais fitormônios (Figura 2). Ótimos exemplos de tais F-boxes são Tir1 e Afb1-5 (auxina); Coi1 (jasmonato); Sly1 e SNE (giberelina); Etp1-2 (etileno) e Aip2 e KEG (ácido abscísico) (Santner e Estelle, 2010).

Eventos de duplicação em genomas vegetais

O aumento recente no volume de dados anotados de genomas de plantas tem fomentado diversas hipóteses a respeito da biologia e história evolutiva destas espécies. Mais especificamente, os genomas recentemente anotados de *G. max* (soja) (Schmutz et al., 2010) e de *M. truncatula* (alfafa) (Young et al., 2011) contribuíram substancialmente para o entendimento de inúmeras características no grupo das Fabaceae, dentre as quais destaca-se a formação de nódulos nas raízes para fixação de nitrogênio. Particularmente importante do ponto de vista biológico e econômico, o mecanismo de fixação de nitrogênio (pela simbiose com bactérias *Rhizobium*) tem tido sua origem associada a eventos de duplicação de genoma inteiro (WGD, whole-genome duplications) ocorridos na linhagem das leguminosas

(Young et al., 2011), clado ao qual pertencem espécies com genomas sequenciados, como: *G. max, M. truncatula, e Phaseolus vulgaris.* Ainda que o conhecimento a respeito dos eventos de duplicação nos genomas, bem como suas implicações na biologia dos organismos, tenha avançado consideravelmente nos últimos anos, pouco se explorou as diferenças nos padrões de conservação/deleção dos alelos duplicados entre os organismos. O emprego de análises de sintenia em genomas tem se mostrado uma estratégia valiosa na descoberta de regiões de conservação nos genomas de espécies vegetais utilizando métodos de genômica comparativa.

O termo sintenia (do Grego; syn = junto, taenia = faixa) é usado em genética para indicar a presença de dois ou mais *loci* no mesmo cromossomo. As análises de sintenia foram elaboradas antes do surgimento dos genomas completamente sequenciados. Atualmente, o conceito de sintenia tem sido expandido para abordar questões de homeologia (homologia residual de cromossomos originalmente completamente homólogos). Os primeiros mapas comparativos entre genomas inteiros entre espécies foram desenvolvidos entre membros da família das Solanaceae (McCouch, 2001). Desvendar a relação entre segmentos cromossomais conservados e a relação funcional dos elementos contidos nesses segmentos é uma importante questão no âmbito da genômica computacional. Frequentemente sugere-se que regiões apresentando conteúdo gênico similar em diferentes espécies são evidências de relação filogenética e traçam, através do curso evolutivo, a herança da função a partir de um ancestral comum. Em um determinado genoma, a presença de regiões extensas de duplicação indicam uma duplicação ancestral de larga-escala; ou ainda uma duplicação de genoma inteiro. Por outro lado, segmentos contendo genes homólogos em vários genomas indicam a conservação da estrutura gênica no último ancestral comum das espécies em questão (Sakar *et al.,* 2011). Estudos clássicos demonstraram que regiões multigênicas homólogas entre várias espécies estariam sob seleção negativa e por isso teriam sua organização estrutural preservada, sendo os operons os melhores exemplos deste fenômeno (Ermolaeva, 2005). Por outro lado, estudos em eucariotos mostram que, salvo algumas exceções (e.g. genes Hox), estes genes encontram-se em regiões sintênicas simplesmente por causa do curto tempo de divergência entre as espécies em questão. No primeiro caso sugerido, o exemplo mais conhecido é o dos operons presentes em genomas procarióticos.

Objetivos

- Determinar, com base em estudos de sintenia, os padrões de conservação dos genes da família F-box nas espécies analisadas;
- Investigar o impacto dos eventos de duplicação locais (tandem) e de segmento (incluindo duplicações de genoma inteiro) sobre os conjuntos de genes F-box nas diferentes espécies estudadas;
- Avaliar os padrões teciduais de expressão de genes F-box duplicados localmente em *G. max* e *M. truncatula*.

Metodologia

Aquisição de dados

A escolha adequada das ferramentas e dos repositórios de dados é essencial para o sucesso de qualquer projeto de bioinformática. Os dados genômicos foram inicialmente obtidos do Phytozome (Department of Energy's Joint Genome Institute and the Center for Integrative Genomics; University of California Regents; http://www.phytozome.net).

Análise de domínios

Todas as sequências de proteínas foram obtidas e submetidas e análise com o algoritmo foi utilizado o algorítimo HMMER3 (Finn et al., 2011) para detectar *de novo* o domínio F-box (PF00646). Visando, desta forma, uniformizar e otimizar os parâmetros de busca para todos os genomas estudados.

Manipulação dos dados

Os dados obtidos foram processados usando *scripts* escritos em linguagem Perl (www.perl.org), visando a consolidação de informações de diferentes fontes em um formato único a ser utilizado nas etapas subsequentes do projeto. Todos os

passos do trabalho envolvendo comparação de listas; análise, normalização e integração de dados foram realizados através de scripts escritos em perl durante o desenvolvimento do projeto.

Análise de sintenia

Utilizou-se BLASTp (Altschul et al., 1997) para comparar par-a-par os conjuntos de proteínas preditas para cada espécie estudada utilizando o corte de E-value ≤ 0.01. Utilizamos então estes resultados para alimentar o software DAGchainer (Haas et al., 2004), que é capaz de identificar as regiões de sintenia nos genomas dos organismos, bem como identificar regiões duplicadas nos genomas. Foram aplicados os parametros padrão do DAGchainer, exceto pela exigência de pelo menos quatro genes alinhados. Após a identificação das regiões de sintenia/duplicação nos genomas, foram gerados ideogramas circulares utilizando o software Circos (Krzywinski et al., 2009) .

No sentido de avaliar a possibilidade dos genes F-box estarem preferencialmente localizados dentro ou fora das regiões de sintenia foi realizada uma simulação computacional onde as regiões sintênicas detectadas entre dois dados genomas tiveram os identificadores de seus genes embaralhados. Este processo foi repetido no sentido de gerar 10.000 conjuntos de regiões sintênicas, que foram então usados para comparar as frequências esperada e observada de F-boxes em regiões sintênicas.

Identificação de ortólogos e reconstrução filogenética

A detecção de similaridade entre as proteínas analisadas foi realizada usando BLAST (Altschul et al., 1997). Alinhamento múltiplo das proteínas homólogas foi realizado utilizando o software MUSCLE (Edgar 2004) e visualizado com o programa Jalview (Clamp *et al.,* 2004). Reconstruções filogenéticas serão realizadas

com a utilização do software RAxML (Stamatakis et al., 2008), que utiliza máxima-verossimilhança para predição da filogenia baseado nas sequencias fornecidas.

Especificidade da expressão gênica

Baseado no algorítmo desenvolvido por Yanai *et al.,* 2004, foi desenvolvido um programa em linguagem de programação Perl capaz de fornecer um índice de especificidade da expressão tecidual de cada gene presente no transcriptoma. Com base nos dados de expressão de cada gene, o algorítmo retorna um valor entre 0 (perfil de expressão mais constitutivo) e 1 (perfil de expressão mais tecido-específico).

Resultados

Simulação de regiões de sintenia

Com o objetivo de investigar o grau de conservação dos genes da família F-box em espécies distintas de leguminosas (*G. max* e *M. truncatula*) foi realizada uma rotina de simulações computacionais que randomizou as regiões de sintenia entre os genomas das espécies (Figura 6). Para essa análise foram utilizados dois genomas de espécies externas ao grupo Fabaceae: *A. thaliana* e *Vitis vinifera*. Os genes presentes nas regiões sintênicas entre as espécies foram substituídos por genes aleatórios recuperados do conjunto completo dos genomas das espécies. As proporções de representatividade dos genes encontradas nas regiões sintênicas originais foram mantidas nas randomizações. Observou-se um padrão bastante distinto de conservação da família F-box entre as espécies. Ao passo que *G. max* e

V. vinifera tendem a apresentar maior retenção de genes F-box ancestrais em seus genomas, *M. truncatula* e *A.thaliana* possuem maior grau de reposição de genes dessa família. Essa diferença deve-se à extensão de regiões de duplicação tandem presentes em *M. truncatula* e *A. thaliana*.

	Glycine max	*Medicago truncatula*	*Arabidopsis thaliana*	*Vitis vinifera*
Glycine max		202(179.2 ± 10.5)	196(181.1 ± 10.5)	186(179.4 ± 10.5)
Medicago truncatula	152(217.7 ± 12.7)		104(156.6 ± 11.2)	82(156.4 ± 11.4)
Arabidopsis thaliana	124(272.6 ± 12.6)	110(206.7 ± 11.9)		92(224.4 ± 12.3)
Vitis vinifera	74(51.7 ± 5.8)	55(40.9 ± 5.4)	62(43.6 ± 5.5)	

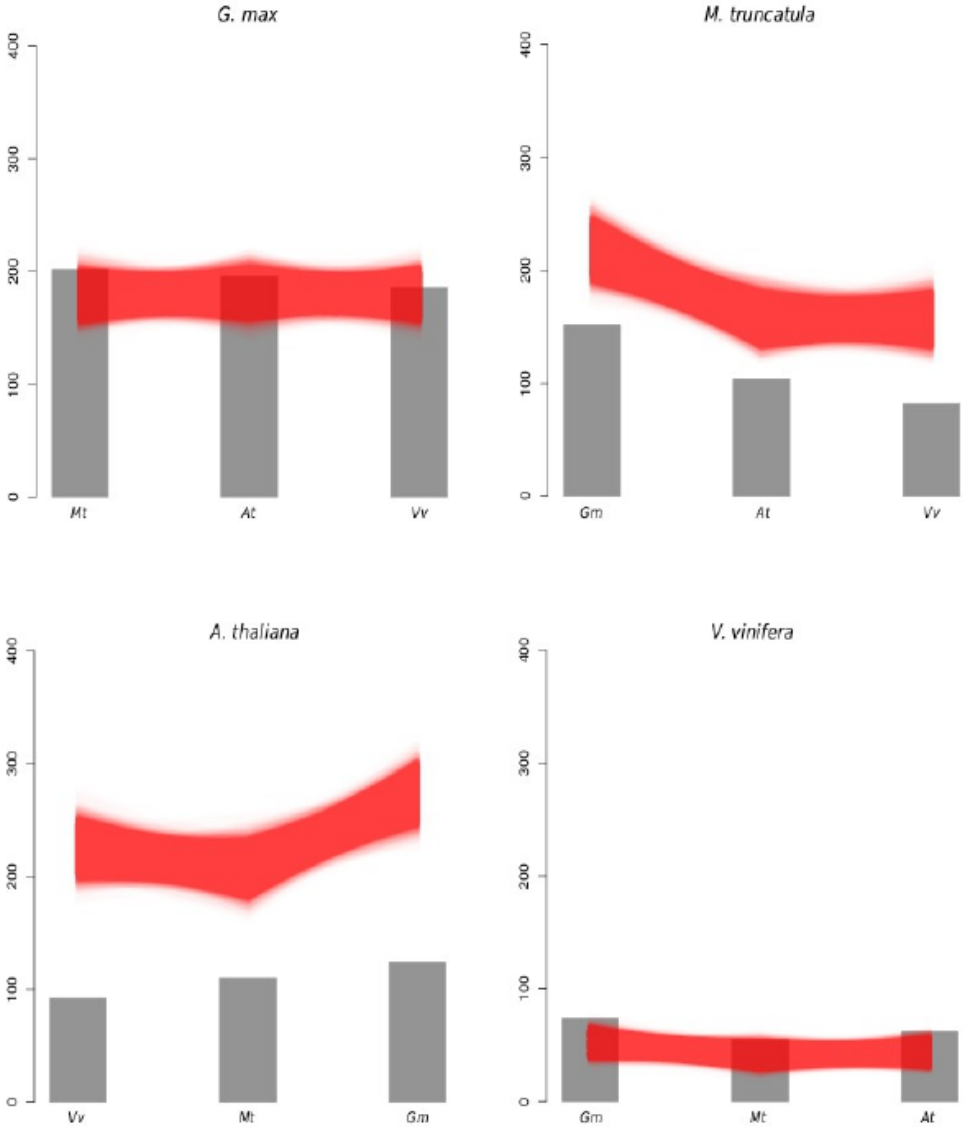

Figura 6: Os gráficos de barra mostram as quantidades encontradas de genes F-box nas regiões de sintenia (par-a-par) entre as espécies comparadas. As nuvens ilustram as quantidades de FBX encontradas nas 10.000 regiões sintênicas randomizadas computacionalmente. Na tabela, os números de FBX em cada comparação de sintenia com os respectivos valores de desvio padrão encontrados nas simulações.

Análise dos perfia de expressão em blocos de duplicação local (tandem)

Os blocos duplicados localmente (tandem) em *M. truncatula* mostraram-se mais extensos que nas demais espécies analisadas. A Tabela 1 lista os segmentos de duplicação no genoma da espécie que contém genes da família F-box.

Destacados estão os segmentos com maior enriquecimento em termos de proporção (total de F-box/total de genes no segmento) de genes da família. Os perfis transcricionais dos genes F-box oriundos de eventos de duplicação local (tandem) foram analisados nas espécies de leguminosas abordadas no trabalho: *G. max* (Libault *et al.,* 2010; Severin *et al.,* 2010) e *M. truncatula* (Benedito *et al.,* 2008). Neste resultado (Figura 7), é possível observar alguns clusters de expressão tecidual, de forma mais evidente, em *M. truncatula* (mais conspícuos em sementes). O resultado evidencia ainda que genes F-box recentemente duplicados, como os originados de duplicações locais, estão atendendo a demandas biológicas sendo expressos em diversos tecidos em ambas as espécies.

Tabela 1: A tabela mostra os blocos duplicados em tandem no genoma de M. truncatula. Em amarelo estão marcados as regiões com maior proporção de genes pertencentes à família F-box em relação ao total de genes presentes no bloco.

##Gene ID	Total FBX	DAGChainer_alignment_header
Medtr1g008500	4	## alignment MtChr1 vs. MtChr1 Alignment #1 score = 2544.6 (num aligned pairs: 73):
Medtr1g012890	3	## alignment MtChr1 vs. MtChr1 Alignment #103 score = 304.1 (num aligned pairs: 11):
Medtr1g016490	2	## alignment MtChr1 vs. MtChr1 Alignment #93 score = 320.3 (num aligned pairs: 12):
Medtr1g020700	6	## alignment MtChr1 vs. MtChr1 Alignment #23 score = 641.6 (num aligned pairs: 21):
Medtr1g025860	2	## alignment MtChr1 vs. MtChr1 Alignment #3 score = 1555.4 (num aligned pairs: 40):
Medtr1g028440	1	## alignment MtChr1 vs. MtChr1 Alignment #162 score = 202.4 (num aligned pairs: 8):
Medtr1g043480	4	## alignment MtChr1 vs. MtChr1 Alignment #2 score = 1598.5 (num aligned pairs: 52):
Medtr1g068390	2	## alignment MtChr1 vs. MtChr1 Alignment #43 score = 497.8 (num aligned pairs: 13):
Medtr1g093010	17	## alignment MtChr1 vs. MtChr1 Alignment #5 score = 1307.6 (num aligned pairs: 36):
Medtr1g110860	2	## alignment MtChr1 vs. MtChr1 Alignment #18 score = 754.5 (num aligned pairs: 21):
Medtr1g137960	3	## alignment MtChr1 vs. MtChr1 Alignment #22 score = 647.2 (num aligned pairs: 18):
Medtr1g143190	1	## alignment MtChr1 vs. MtChr1 Alignment #10 score = 899.2 (num aligned pairs: 30):
Medtr2g008430	10	## alignment MtChr2 vs. MtChr2 Alignment #2 score = 2469.6 (num aligned pairs: 66):
Medtr2g011000	3	## alignment MtChr2 vs. MtChr2 Alignment #6 score = 1427.4 (num aligned pairs: 33):
Medtr2g026900	18	## alignment MtChr2 vs. MtChr2 Alignment #15 score = 858.6 (num aligned pairs: 35):
Medtr2g030030	8	## alignment MtChr2 vs. MtChr2 Alignment #18 score = 707.4 (num aligned pairs: 23):
Medtr2g033660	4	## alignment MtChr2 vs. MtChr2 Alignment #16 score = 835.5 (num aligned pairs: 22):
Medtr2g077470	5	## alignment MtChr2 vs. MtChr2 Alignment #40 score = 432.8 (num aligned pairs: 17):
Medtr2g091950	2	## alignment MtChr2 vs. MtChr2 Alignment #3 score = 2165.6 (num aligned pairs: 60):
Medtr2g096920	5	## alignment MtChr2 vs. MtChr2 Alignment #1 score = 2640.7 (num aligned pairs: 70):
Medtr2g123270	6	## alignment MtChr2 vs. MtChr2 Alignment #13 score = 879.5 (num aligned pairs: 29):
Medtr2g126940	2	## alignment MtChr2 vs. MtChr2 Alignment #110 score = 217.0 (num aligned pairs: 10):
Medtr3g006080	6	## alignment MtChr3 vs. MtChr3 Alignment #26 score = 689.9 (num aligned pairs: 22):
Medtr3g009050	9	## alignment MtChr3 vs. MtChr3 Alignment #61 score = 460.7 (num aligned pairs: 15):
Medtr3g013620	4	## alignment MtChr3 vs. MtChr3 Alignment #20 score = 794.4 (num aligned pairs: 19):
Medtr3g016530	15	## alignment MtChr3 vs. MtChr3 Alignment #41 score = 562.3 (num aligned pairs: 18):
Medtr3g020690	4	## alignment MtChr3 vs. MtChr3 Alignment #34 score = 603.9 (num aligned pairs: 17):
Medtr3g026350	4	## alignment MtChr3 vs. MtChr3 Alignment #22 score = 750.8 (num aligned pairs: 24):
Medtr3g029580	30	## alignment MtChr3 vs. MtChr3 Alignment #19 score = 808.5 (num aligned pairs: 26):
Medtr3g040840	5	## alignment MtChr3 vs. MtChr3 Alignment #10 score = 972.4 (num aligned pairs: 33):
Medtr3g044250	7	## alignment MtChr3 vs. MtChr3 Alignment #48 score = 518.6 (num aligned pairs: 15):
Medtr3g087770	3	## alignment MtChr3 vs. MtChr3 Alignment #334 score = 212.4 (num aligned pairs: 7):
Medtr3g104440	5	## alignment MtChr3 vs. MtChr3 Alignment #4 score = 1569.4 (num aligned pairs: 51):
Medtr3g110040	6	## alignment MtChr3 vs. MtChr3 Alignment #5 score = 1224.4 (num aligned pairs: 34):
Medtr3g123450	1	## alignment MtChr3 vs. MtChr3 Alignment #49 score = 511.6 (num aligned pairs: 18):
Medtr3g137080	5	## alignment MtChr3 vs. MtChr3 Alignment #14 score = 914.5 (num aligned pairs: 24):
Medtr3g144040	2	## alignment MtChr3 vs. MtChr3 Alignment #72 score = 427.8 (num aligned pairs: 17):
Medtr3g149860	2	## alignment MtChr3 vs. MtChr3 Alignment #93 score = 377.0 (num aligned pairs: 12):
Medtr3g156220	2	## alignment MtChr3 vs. MtChr3 Alignment #69 score = 434.4 (num aligned pairs: 14):
Medtr3g166440	2	## alignment MtChr3 vs. MtChr3 Alignment #17 score = 848.2 (num aligned pairs: 21):
Medtr4g005380	8	## alignment MtChr4 vs. MtChr4 Alignment #93 score = 454.7 (num aligned pairs: 11):
Medtr4g025790	34	## alignment MtChr4 vs. MtChr4 Alignment #1 score = 4171.0 (num aligned pairs: 121):
Medtr4g033500	8	## alignment MtChr4 vs. MtChr4 Alignment #44 score = 794.5 (num aligned pairs: 30):
Medtr4g036010	2	## alignment MtChr4 vs. MtChr4 Alignment #53 score = 702.7 (num aligned pairs: 19):
Medtr4g073670	3	## alignment MtChr4 vs. MtChr4 Alignment #292 score = 179.1 (num aligned pairs: 8):
Medtr4g110360	4	## alignment MtChr4 vs. MtChr4 Alignment #155 score = 307.8 (num aligned pairs: 11):
Medtr4g125290	11	## alignment MtChr4 vs. MtChr4 Alignment #8 score = 1657.5 (num aligned pairs: 55):
Medtr4g133940	2	## alignment MtChr4 vs. MtChr4 Alignment #4 score = 2031.6 (num aligned pairs: 69):
Medtr4g143470	2	## alignment MtChr4 vs. MtChr4 Alignment #238 score = 219.4 (num aligned pairs: 11):
Medtr4g156560	2	## alignment MtChr4 vs. MtChr4 Alignment #41 score = 836.3 (num aligned pairs: 32):
Medtr4g160910	1	## alignment MtChr4 vs. MtChr4 Alignment #74 score = 574.9 (num aligned pairs: 25):
Medtr5g011800	8	## alignment MtChr5 vs. MtChr5 Alignment #4 score = 3446.6 (num aligned pairs: 99):
Medtr5g016800	1	## alignment MtChr5 vs. MtChr5 Alignment #10 score = 1360.4 (num aligned pairs: 41):
Medtr5g022890	6	## alignment MtChr5 vs. MtChr5 Alignment #7 score = 2259.8 (num aligned pairs: 73):
Medtr5g027510	2	## alignment MtChr5 vs. MtChr5 Alignment #6 score = 2565.1 (num aligned pairs: 69):
Medtr5g040840	8	## alignment MtChr5 vs. MtChr5 Alignment #3 score = 3708.1 (num aligned pairs: 120):
Medtr5g048810	5	## alignment MtChr5 vs. MtChr5 Alignment #14 score = 1116.9 (num aligned pairs: 51):
Medtr5g066660	9	## alignment MtChr5 vs. MtChr5 Alignment #23 score = 843.6 (num aligned pairs: 35):
Medtr5g082790	11	## alignment MtChr5 vs. MtChr5 Alignment #2 score = 4195.5 (num aligned pairs: 122):
Medtr5g087550	2	## alignment MtChr5 vs. MtChr5 Alignment #63 score = 454.0 (num aligned pairs: 24):
Medtr5g105510	15	## alignment MtChr5 vs. MtChr5 Alignment #1 score = 6665.4 (num aligned pairs: 221):
Medtr6g006320	2	## alignment MtChr6 vs. MtChr6 Alignment #20 score = 1031.5 (num aligned pairs: 34):
Medtr6g013700	2	## alignment MtChr6 vs. MtChr6 Alignment #5 score = 1662.0 (num aligned pairs: 48):
Medtr6g017560	3	## alignment MtChr6 vs. MtChr6 Alignment #281 score = 251.1 (num aligned pairs: 9):
Medtr6g036560	2	## alignment MtChr6 vs. MtChr6 Alignment #30 score = 910.9 (num aligned pairs: 24):
Medtr6g081060	1	## alignment MtChr6 vs. MtChr6 Alignment #188 score = 357.3 (num aligned pairs: 10):
Medtr6g092780	3	## alignment MtChr6 vs. MtChr6 Alignment #98 score = 548.4 (num aligned pairs: 20):
Medtr7g022020	8	## alignment MtChr7 vs. MtChr7 Alignment #2 score = 2546.5 (num aligned pairs: 74):
Medtr7g032170	1	## alignment MtChr7 vs. MtChr7 Alignment #37 score = 531.8 (num aligned pairs: 21):
Medtr7g047630	3	## alignment MtChr7 vs. MtChr7 Alignment #8 score = 1121.7 (num aligned pairs: 32):
Medtr7g068230	6	## alignment MtChr7 vs. MtChr7 Alignment #44 score = 485.8 (num aligned pairs: 17):
Medtr7g075920	6	## alignment MtChr7 vs. MtChr7 Alignment #23 score = 659.1 (num aligned pairs: 18):
Medtr7g080880	2	## alignment MtChr7 vs. MtChr7 Alignment #62 score = 395.0 (num aligned pairs: 10):
Medtr7g088650	14	## alignment MtChr7 vs. MtChr7 Alignment #4 score = 2304.5 (num aligned pairs: 78):
Medtr7g107990	16	## alignment MtChr7 vs. MtChr7 Alignment #1 score = 3036.7 (num aligned pairs: 87):
Medtr7g116940	4	## alignment MtChr7 vs. MtChr7 Alignment #15 score = 766.7 (num aligned pairs: 20):
Medtr7g127560	2	## alignment MtChr7 vs. MtChr7 Alignment #96 score = 294.4 (num aligned pairs: 9):
Medtr7g135540	3	## alignment MtChr7 vs. MtChr7 Alignment #35 score = 536.5 (num aligned pairs: 18):
Medtr7g138350	2	## alignment MtChr7 vs. MtChr7 Alignment #26 score = 638.6 (num aligned pairs: 19):
Medtr8g015780	7	## alignment MtChr8 vs. MtChr8 Alignment #24 score = 754.1 (num aligned pairs: 23):
Medtr8g025220	1	## alignment MtChr8 vs. MtChr8 Alignment #7 score = 1225.6 (num aligned pairs: 37):
Medtr8g034220	1	## alignment MtChr8 vs. MtChr8 Alignment #17 score = 910.0 (num aligned pairs: 23):
Medtr8g044370	1	## alignment MtChr8 vs. MtChr8 Alignment #2 score = 1976.3 (num aligned pairs: 57):
Medtr8g074530	8	## alignment MtChr8 vs. MtChr8 Alignment #93 score = 334.7 (num aligned pairs: 14):
Medtr8g076990	1	## alignment MtChr8 vs. MtChr8 Alignment #13 score = 944.2 (num aligned pairs: 36):
Medtr8g076990	2	## alignment MtChr8 vs. MtChr8 Alignment #13 score = 944.2 (num aligned pairs: 36):
Medtr8g088880	4	## alignment MtChr8 vs. MtChr8 Alignment #6 score = 1250.5 (num aligned pairs: 35):
Medtr8g093360	3	## alignment MtChr8 vs. MtChr8 Alignment #94 score = 329.9 (num aligned pairs: 10):
Medtr8g102650	4	## alignment MtChr8 vs. MtChr8 Alignment #77 score = 376.9 (num aligned pairs: 13):
Medtr8g112970	3	## alignment MtChr8 vs. MtChr8 Alignment #125 score = 261.0 (num aligned pairs: 6):
Medtr8g120510	2	## alignment MtChr8 vs. MtChr8 Alignment #40 score = 569.7 (num aligned pairs: 16):
Medtr8g122430	3	## alignment MtChr8 vs. MtChr8 Alignment #85 score = 350.1 (num aligned pairs: 13):
Medtr8g140060	1	## alignment MtChr8 vs. MtChr8 Alignment #44 score = 542.9 (num aligned pairs: 16):
Medtr8g148150	4	## alignment MtChr8 vs. MtChr8 Alignment #3 score = 1637.4 (num aligned pairs: 52):

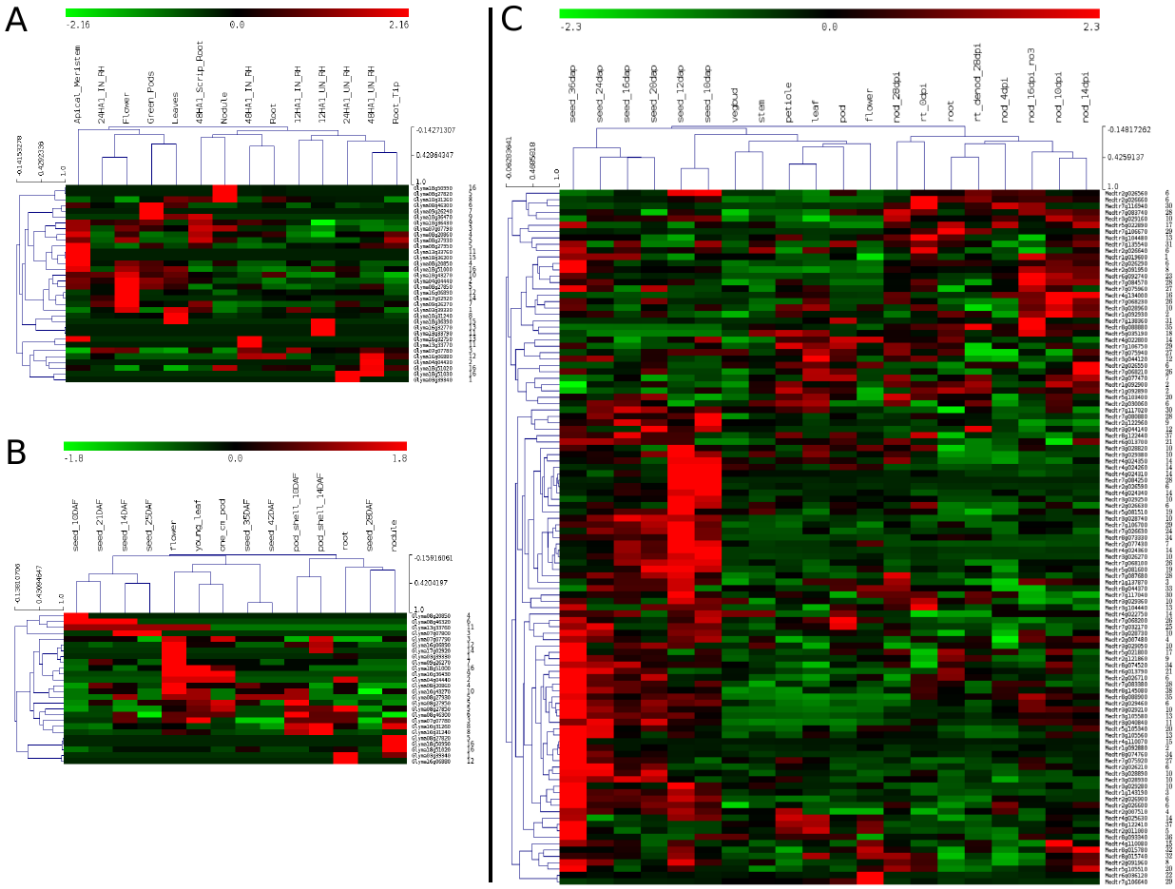

Figura 7: Na figura estão representados os TD-F-boxes e seus respectivos níveis de expressão em diversos tecidos de M. truncatula (esquerda) e G. max (direita) obtidos de 3 fontes distintas. Genes e tecidos estão clusterizados no heatmap por método HCL (Hierarchical Clustering).

Análise das regiões sintenicas, blocos de duplicação tandem e clusters de expressão

As regiões de sintenia entre as duas leguminosas estudadas e *V. vinifera* foram expressas em um ideograma circular que representa a totalidade dos 3 genomas analisados (Figura 8). Escolheu-se *V. vinifera* como elemento de comparação por se tratar de uma rosídea basal que sofreu poucos rearranjos em seu genoma ao longo de sua história evolutiva. Podendo assim oferecer um bom parâmetro comparativo. Na figura 8 estão reprentados ainda os blocos de duplicação local encontrados nos genomas das três espécies. A representação de linhas expressa o número de genes F-box contidos em intervalos de 100 genes nos genomas, evidenciando a intensa representação dos membros dessa família em *M. truncatula* comparado às outras duas. Estão mostrados no ideograma ainda os clusters de expressão tecidual encontrados na análise anterior, incluindo genes TD-F-box. É possível observar clusters gênicos de expressão tecidual mantidos em posições próximas no genoma de *G. max* e *M. truncatula*.

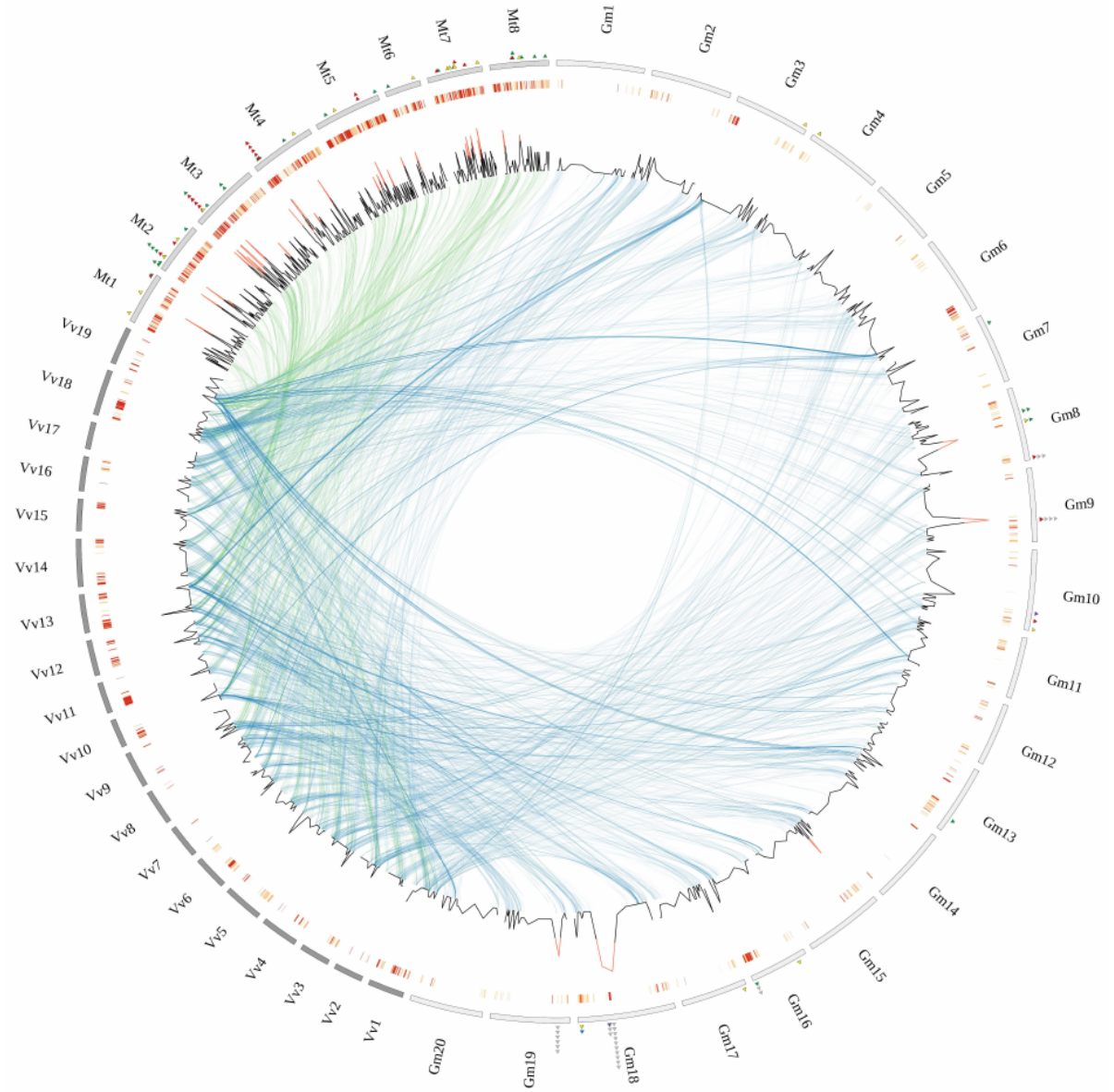

Figura 8: No ideograma estão dispostos os cromossomos de M. truncatula (cinza) e G. max(cinza-claro) e V. vinifera (cinza-escuro) na auréola externa. Acima da circunferência estão representados: TD-Fboxes pertencentes aos clusters de expressão tecidual em M. truncatula (verde: sementes [estágio tardio do desenvolvimento] ; vermelho: sementes [embriogênese tardia]; amarelo: nódulo e em G. max (cinza: não expressos; roxo: meristema apical; amarelo: nódulo; laranja: flor; preto: vagens verdes); Duplicações locais ao longo dos genomas estão representados na auréola seguinte (mais interna): em laranja-claro; regiões com duplicações sobrepostas estão representadas em vermelho; O line-plot sinaliza a quantidade de genes F-box acada 100 genes no genoma; intervalos (de 100 genes) com 5 F-boxes ou mais são realçados em vermelho. Os links (região central do ideograma) conectam genes sintênicos entre as espécies: V. vinifera x M. truncatula(verde), V. vinifera x G. max (azul).

Expressão gênica em M. Truncatula

O gráfico (Figura 9) mostra o log dos níveis de expressão gênica em *M. truncatula* baseados no experimento de Benedito *et al.,* 2008. Nota-se que todos os tecidos/órgãos amostrados no experimento apresentam níveis similares de expressão de genes F-box. Embora a maioria dos genes F-box, duplicados recentemente em tandem ou não, apresentem níveis de expressão relativamente baixos, há representantes da família com níveis de expressão tecidual altos. Dois dos genes mais expressos na família são, de fato, oriundos de duplicação local recente (Medtr4g1340000 e Medtr1g092880).

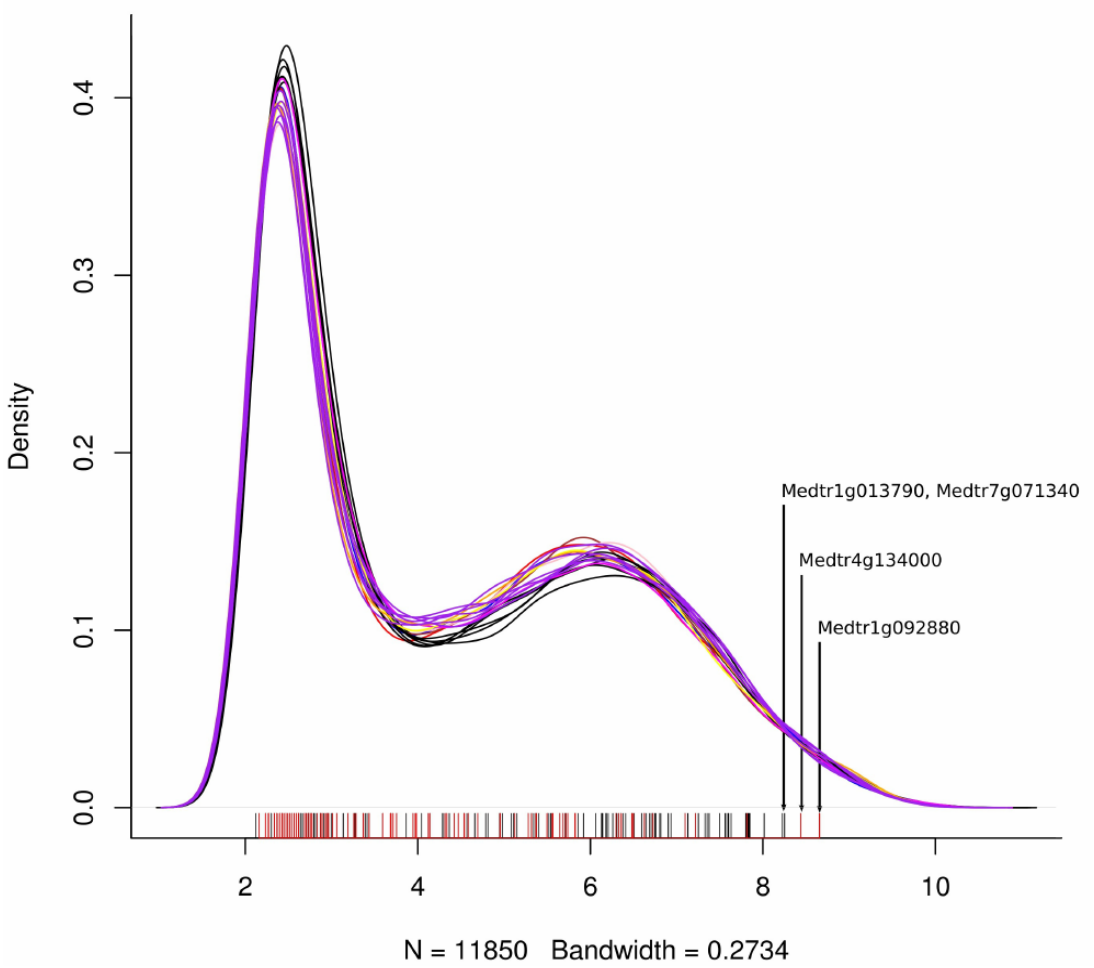

Figura 9: Plot de densidade: log dos dados de expressão gênica total em M. truncatula. As linhas representam os níveis de expressão gênica por tecido/órgão: sementes (preto), pecíolo (azul), caule (vermelho), broto vegetativo (marrom), flor (rosa), folha (magenta), vagem (amarelo), raiz (laranja), nódulos (roxo). As marcações na parte inferior do gráfico representam genes FBX (preto) e TD-FBX (vermelho).

Domínios presentes em proteínas F-box

Com o intuito de investigar a representatividade de domínios específicos em proteínas F-box ao longo do curso evolutivo de organismos vegetais, foi realizada uma busca seguida da quantificação em proteínas traduzidas por genes presentes em regiões sintênicas entre espécies (Figuras 10-13). Alguns domínios C-terminais (responsáveis pelo reconhecimento do substrato a ser recrutado pelo complexo SCF) se destacam, como por exemplo: Kelch, domínios LRR, Tub, e FBA (F-box *associated*). Esse resultado abre questões relativas às funcões específicas de tais proteínas, que poderiam estar envolvidas em processos basais da biologia vegetal. Em seguida investigou-se quais seriam os domínios protéicos mais frequentes em regiões sintênicas entre soja e outras quatro espécies, incluindo duas fixadoras de nitrogênio (*M. truncatula*, e *Prunnus persica*), e duas espécies externas a esse grupo pertencentes ao grupo das rosídeas (*A. thaliana*, e *V. vinifera*) (Figuras 14-17). A prevalência mais evidente encontrada refere-se à família Pkinase. Uma família altamente conservada em diversos organismos que responde por processos celulares variados, como: divisão, proliferação, apoptose e diferenciação. Interessante notar que em regiões sintênicas entre *G. max* e *M. truncatula*, a família F-box aparece na lista das mais conservadas (barra marcada em cor laranja na figura 14). Possivelmente por estar envolvida em processos biológicos estratégicos para o grupo Fabaceae, como a nodulação por exemplo.

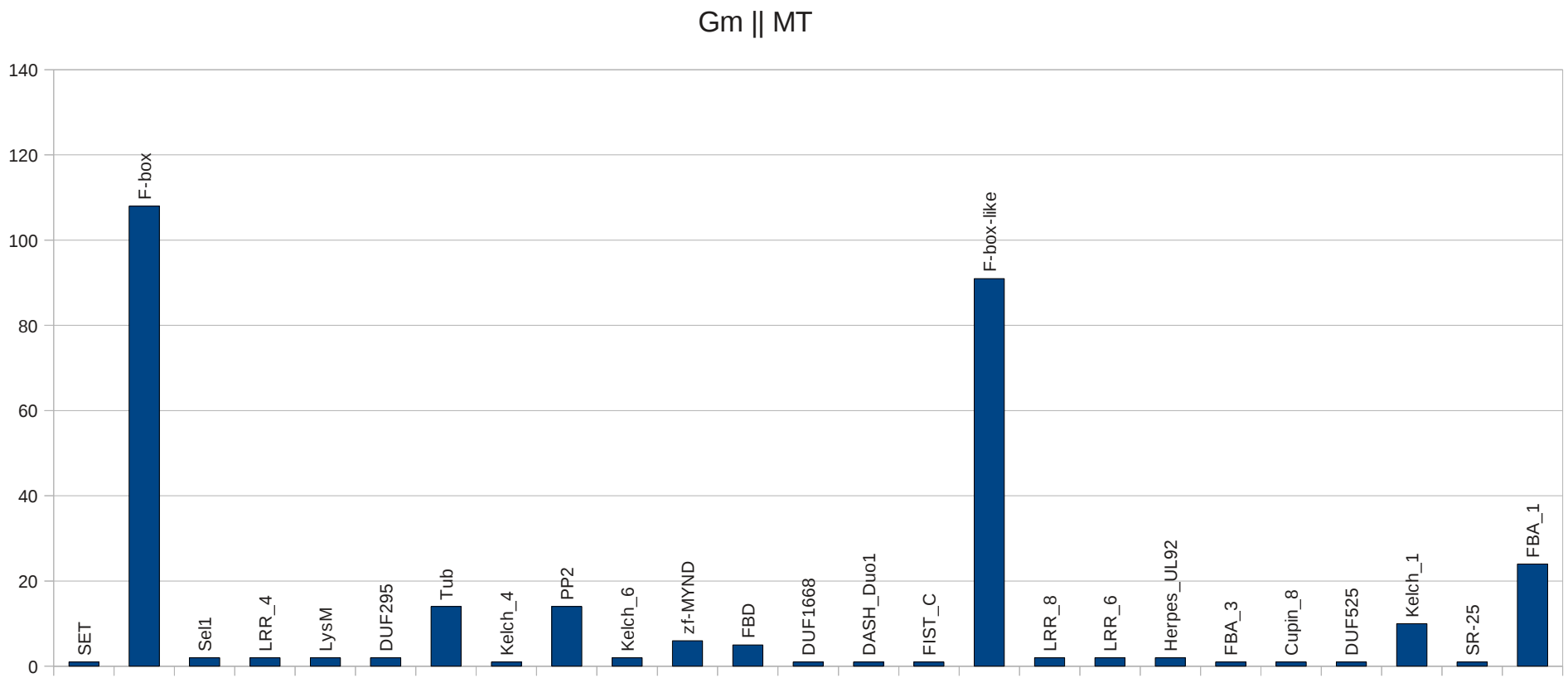

Figura 10: O gráfico mostra os domínios mais comuns encontrados nas proteínas F-box conservadas em regiões sintênicas entre M. truncatula e G. max

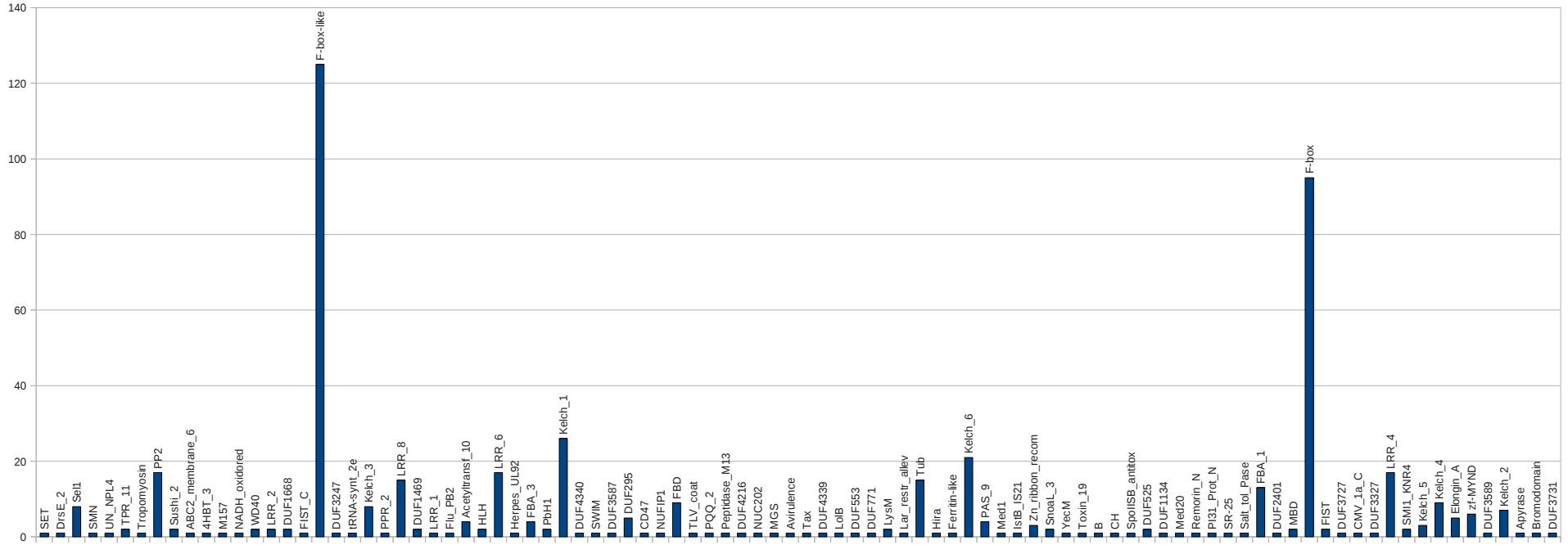

Figura 11: O gráfico mostra os domínios mais comuns encontrados nas proteínas F-box conservadas em regiões sintênicas entre P. persica. e G. max

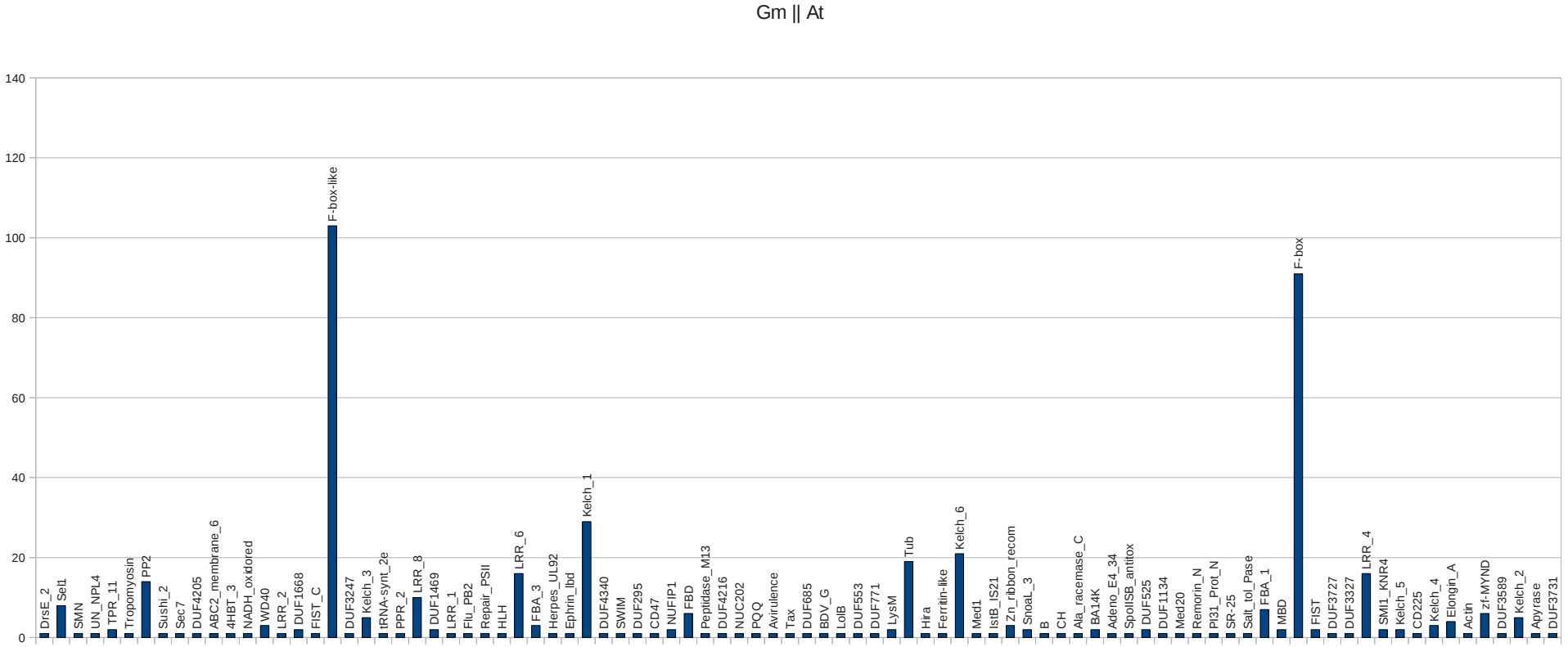

Figura 12: O gráfico mostra os domínios mais comuns encontrados nas proteínas F-box conservadas em regiões sintênicas entre A. thaliana. e G.max

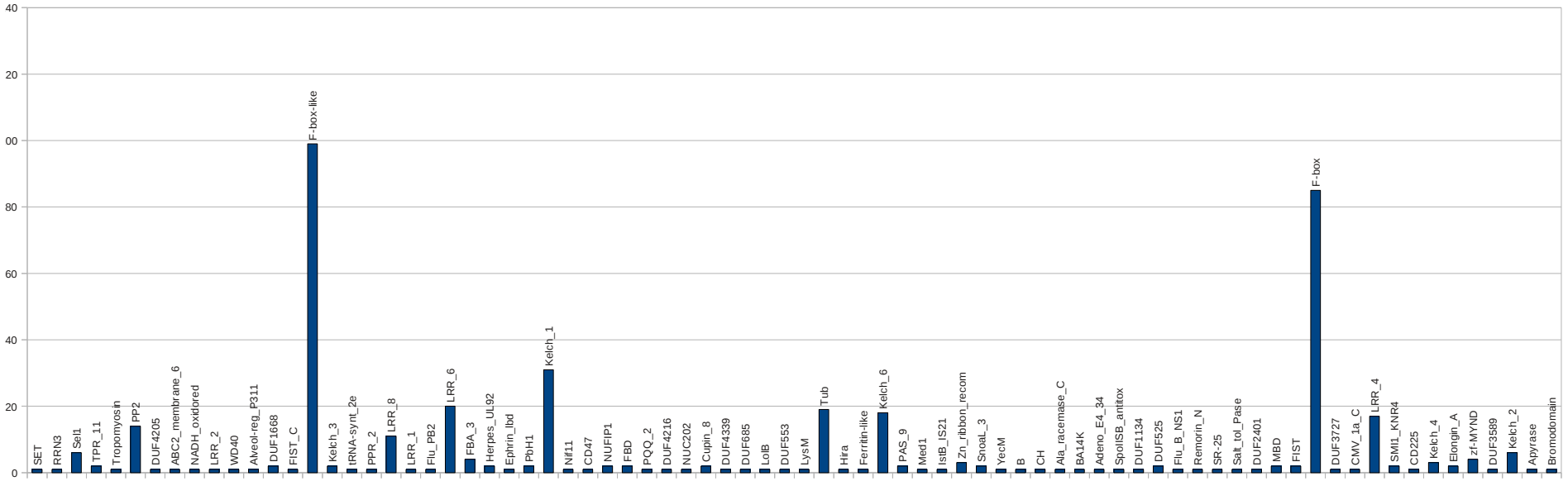

Figura 13: O gráfico mostra os domínios mais comuns encontrados nas proteínas F-box conservadas em regiões sintênicas entre V. vinifera. e G. max

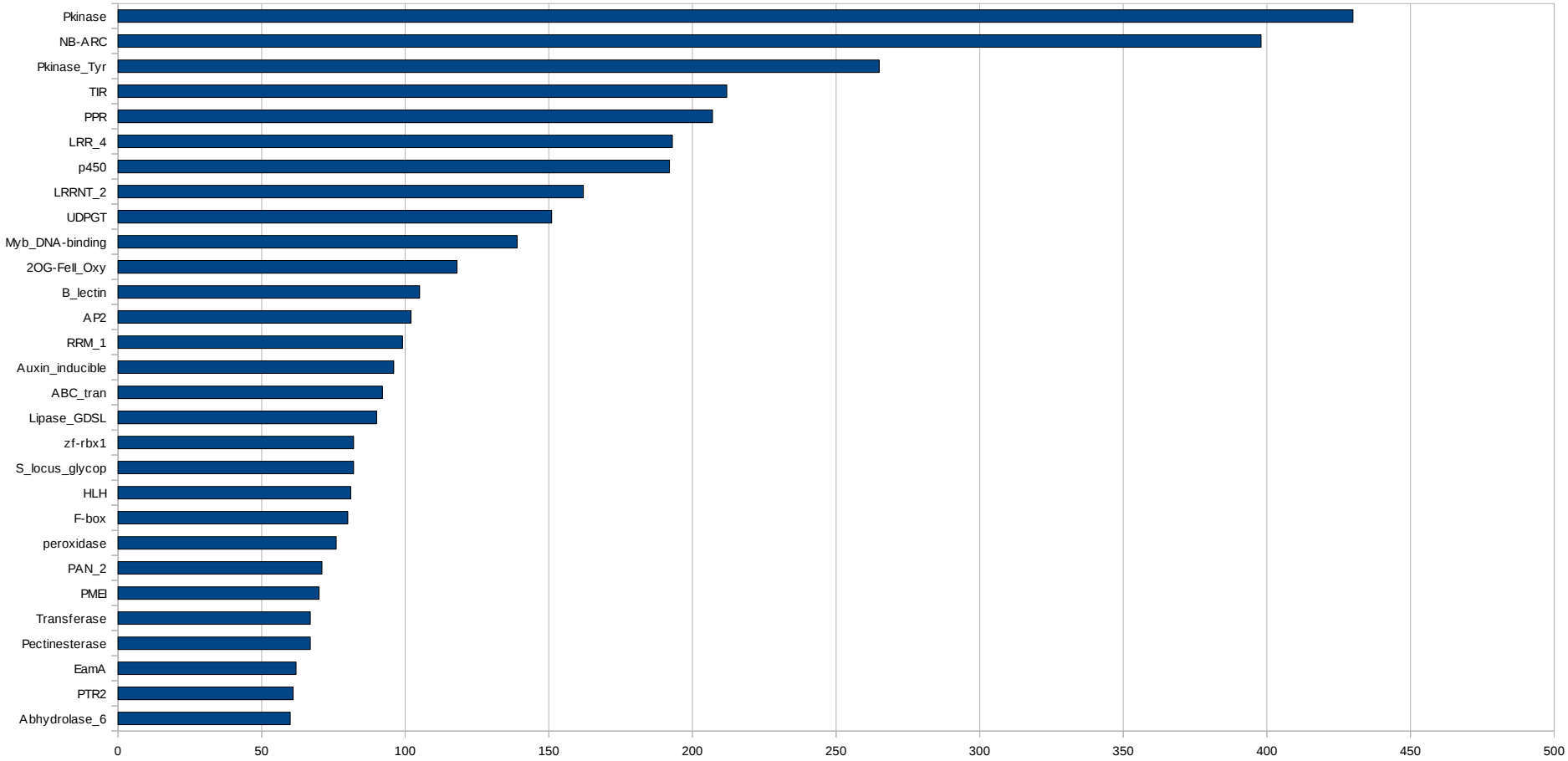

Figura 14: As barras ilustram a quantidade de genes de cada família presente nas regiões sintênicas entre M. truncatula e G. max.

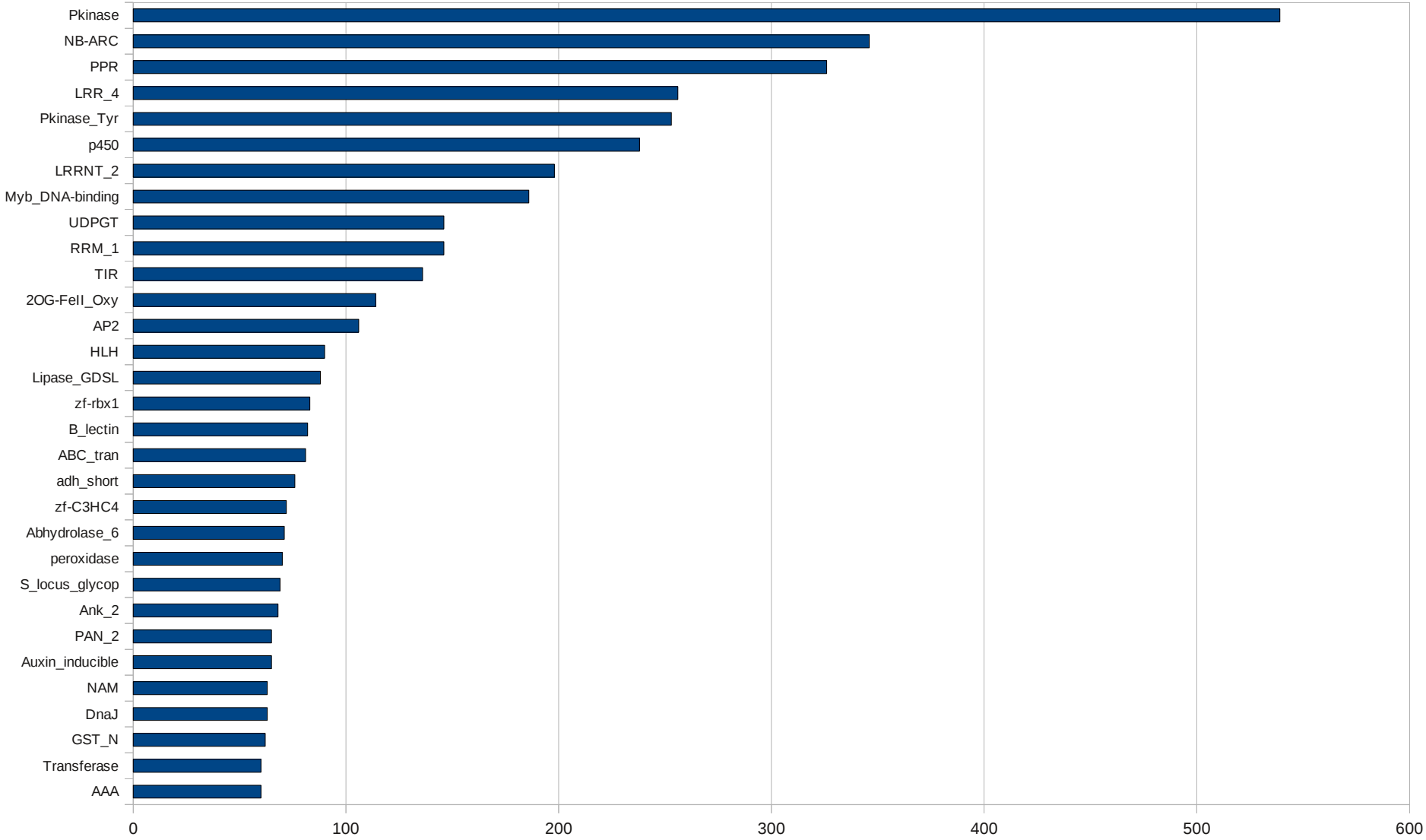

Figura 15: As barras ilustram a quantidade de genes de cada família presente nas regiões sintênicas entre P. persica e G. max.

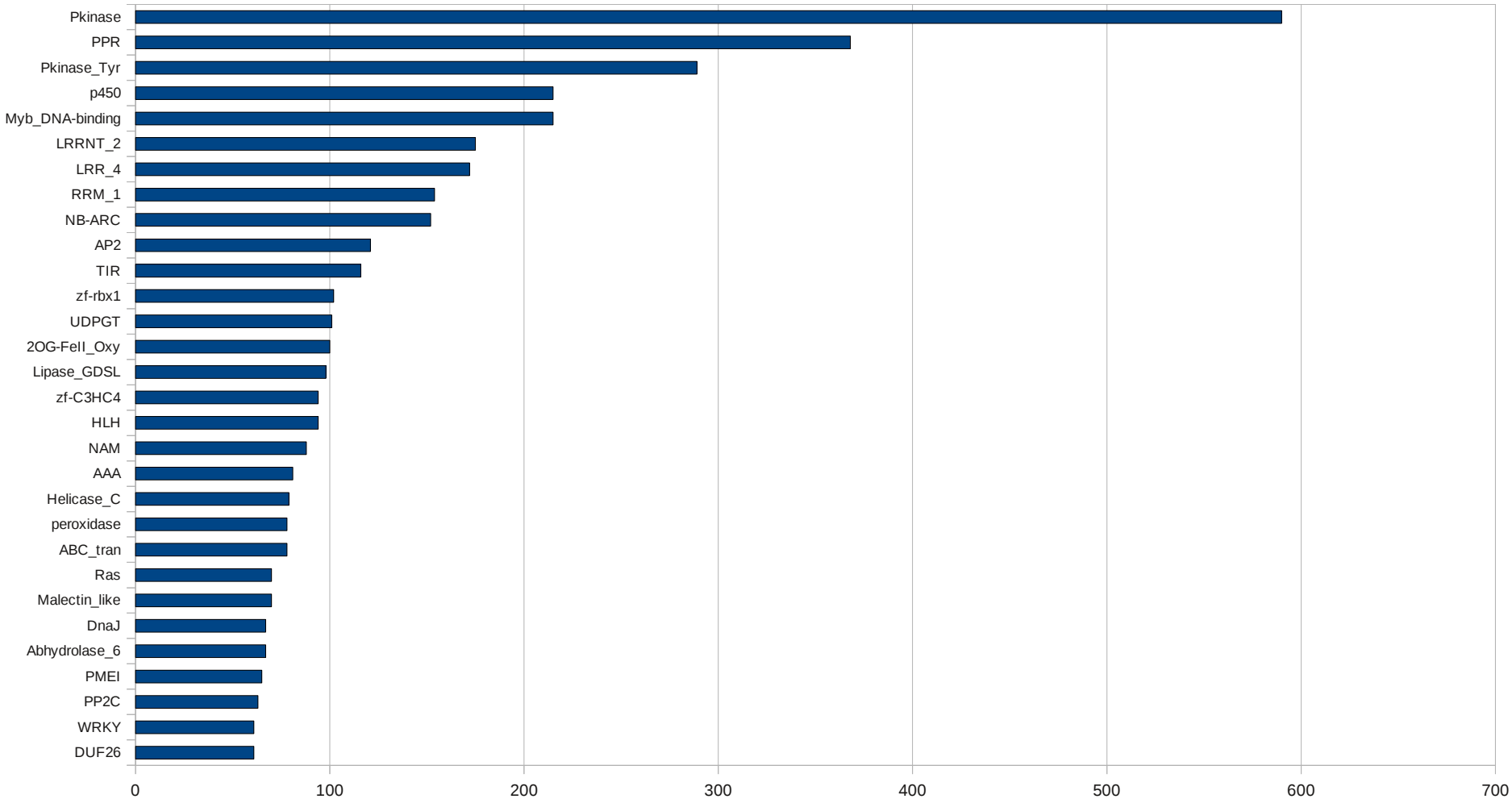

Figura 16: As barras ilustram a quantidade de genes de cada família presente nas regiões sintênicas entre A. thaliana e G. max.

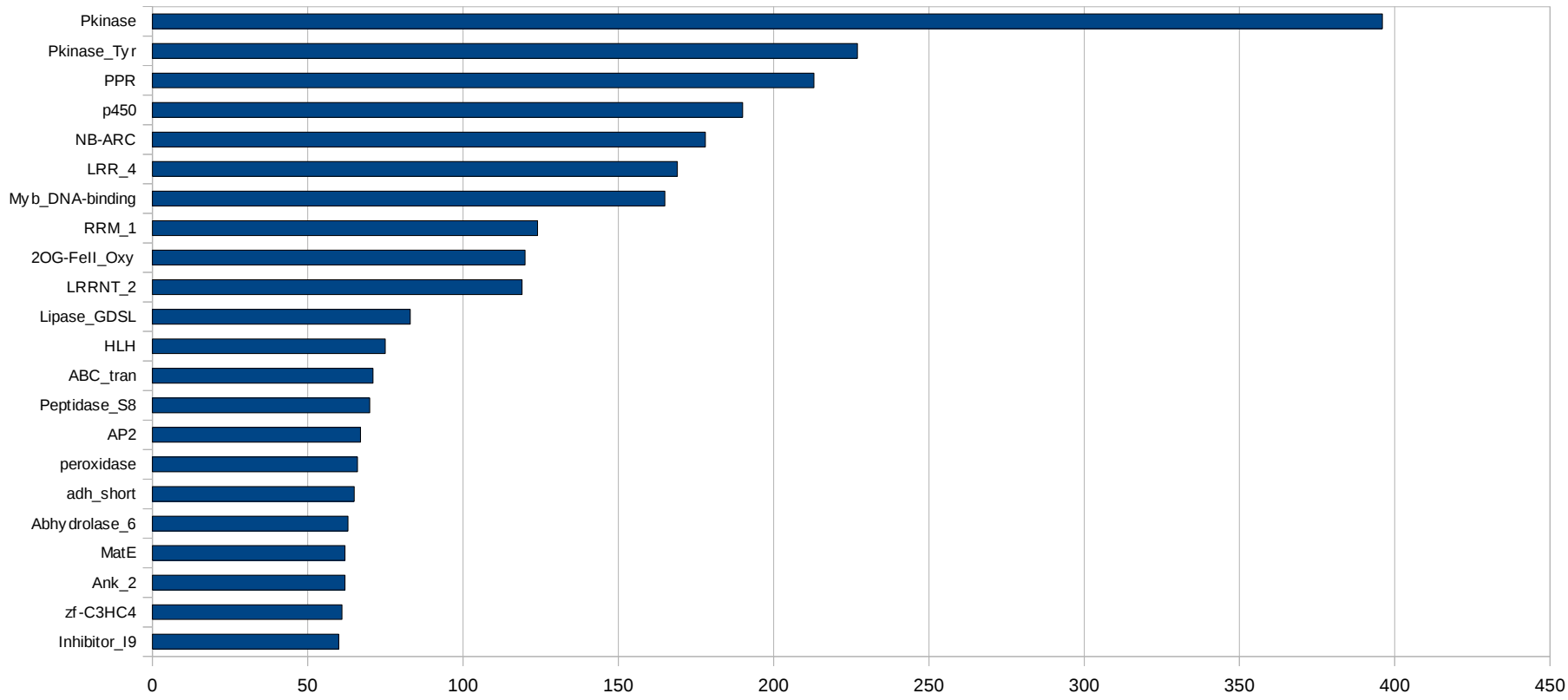

Figura 17: As barras ilustram a quantidade de genes de cada família presente nas regiões sintênicas entre V. vinifera e G. max.

Análise preliminar da expressão gênica em tecidos/órgãos de G. max

A primeira abordagem sobre a expressão tecidual de genes F-box em soja foi realizada em escala global. Foram quantificados os genes F-box expressos em cada tecido amostrado no experimento de Libault *et al.,* 2010. Para traçar um comparativo com esses números e observar se algum viés seria observado nos números, viu-se também a expressão gênica geral em todos os tecidos. A comparação entre as duas quantificações foi realizada por uma razão (genes F-box/Total geral de genes). Embora a razão em nódulos tenha se apresentado maior que a dos demais tecidos analisados, nenhum viés foi observado na expressão de F-boxes nos tecidos estudados (Figuras 18-20). Na Tabela 2 estão dispostos os genes que possuem os mais elevados níveis de expressão para cada tecido e suas respectivas anotações funcionais. Observa-se uma clara prevalência de genes relacionados a combate de patógenos em pêlos radiculares. Um possível resultado da intensa interação deste órgão com microrganismos abundantes no solo. Na Tabela 3 estão listados os identificadores dos genes da família F-box oriundos de duplicações locais expressos em pelo menos um dos dois estudos de transcriptoma utilizados (Libault *et al.,* 2010; Severin *et al.,* 2010), bem como aqueles não expressos em nenhum dos dois citados trabalhos. Foi realizada ainda uma análise para determinação do nível de especificidade de expressão tecidual dos genes encontrados no transcriptoma publicado por Libault *et al.,* 2010 (Figura 21). Pode-se notar uma tendência clara a um perfil intermediário entre a expressão constitutiva (valores próximos de 0) e a tecido-específica (valores próximos a 1). Contudo a frequência com que os genes apresentam o perfil tecido-específico é bem mais alta que a frequência de constitutivos.

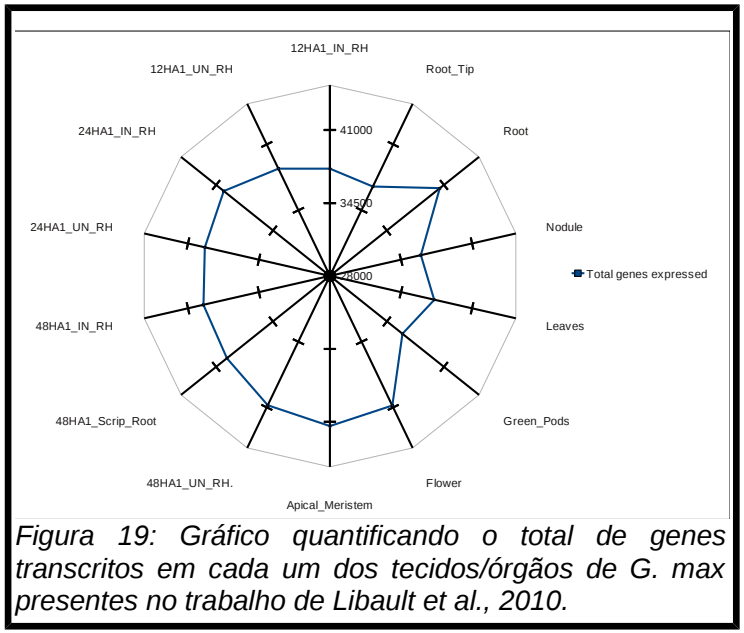

Figura 19: Gráfico quantificando o total de genes transcritos em cada um dos tecidos/órgãos de G. max presentes no trabalho de Libault et al., 2010.

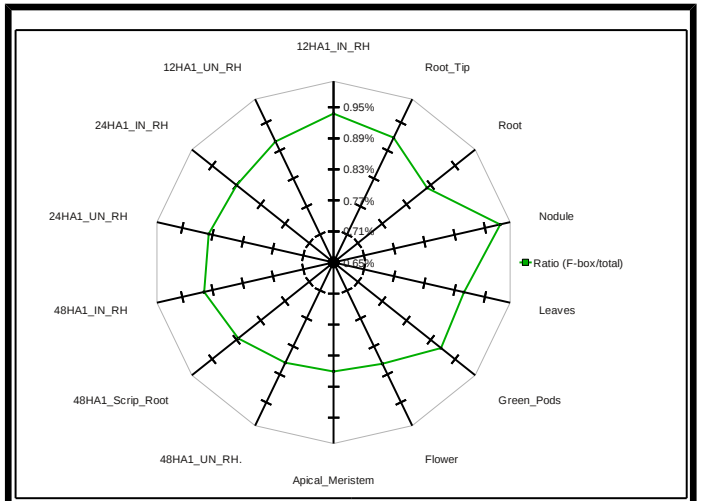

Figura 18: Gráfico ilustrando a razão entre o total de genes F-box sobre o total geral de genes transcritos em cada um dos tecidos/órgãos de G. max presentes no trabalho de Libault et al., 2010.

Tissue	F-box genes expressed	Total genes expressed	Ratio (F-box/total)
12HA1_IN_RH	352	37543	0,94%
12HA1_UN_RH	351	38595	0,91%
24HA1_IN_RH	357	40122	0,89%
24HA1_UN_RH	354	39446	0,90%
48HA1_IN_RH	359	39590	0,91%
48HA1_Scrip_Root	352	39748	0,89%
48HA1_UN_RH.	353	40795	0,87%
Apical_Meristem	356	41364	0,86%
Flower	354	40810	0,87%
Green_Pods	332	36268	0,92%
Leaves	341	37547	0,91%
Nodule	356	36311	0,98%
Root_Tip_Stacey	357	40522	0,88%
Root_Tip	338	36835	0,92%

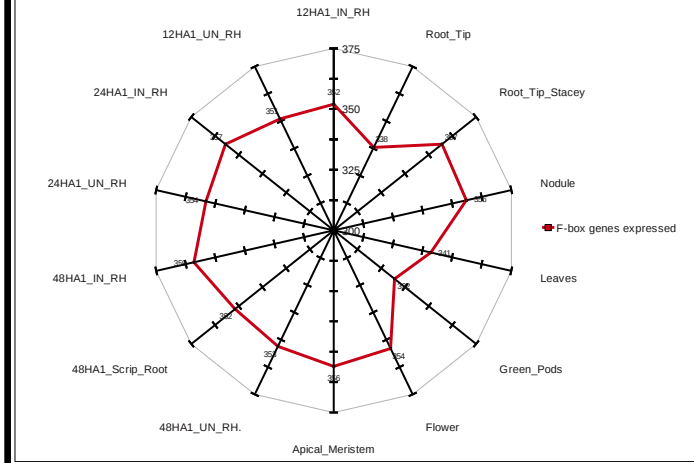

Figura 20: Gráfico quantificando o total de genes F-box transcritos em cada um dos tecidos/órgãos de G. max presentes no trabalho de Libault et al., 2010.

Tabela 2: A tabela mostra os identificadores dos genes mais expressos para cada tecido/órgão amostrado no experimento de Libault et al., 2010. A terceira coluna mostra o valor obtido no experimento. A quarta coluna mostra a anotação funcional (caso haja alguma).

Genes mais expressos nos tecidos de G. max

Tissue	Gene ID	Expression data	Functional annotations
12HAI_IN_RH	Glyma07g37270.1	13964,5762828089	Pfam:00407 Pathogenesis-related protein Bet v I family
12HAI_UN_RH	Glyma07g37270.1	11680,2079120842	Pfam:00407 Pathogenesis-related protein Bet v I family
24HAI_IN_RH	Glyma07g37270.1	9564,0796669732	Pfam:00407 Pathogenesis-related protein Bet v I family
24HAI_UN_RH	Glyma07g37270.1	15820,7850159478	Pfam:00407 Pathogenesis-related protein Bet v I family
48HAI_IN_RH	Glyma03g34310.1	5916,2232060398	Pfam:00230 Major intrinsic protein Panther:19139 AQUAPORIN TRANSPORTER KOG:0223 Aquaporin (major intrinsic protein family) KEGGORTH:09873 aquaporin TIP.
48HAI_Scrip_Root	Glyma09g12250.1 Glyma09g12260.1	16239,7206216676	no annotations..
48HAI_UN_RH	Glyma07g37270.1	5825,3609906294	Pfam:00407 Pathogenesis-related protein Bet v I family
Apical_Meristem_Stacey	Glyma07g01730.1	10473,6072594855	Pfam:03767 HAD superfamily, subfamily IIIB (Acid phosphatase)
Flower_Stacey	Glyma05g25810.1	8650,6601039265	Pfam:00504 Chlorophyll A-B binding protein Panther:21649 CHLOROPHYLL A/B BINDING PROTEIN KEGGORTH:08912 light-harvesting complex II chlorophyll a/b binding protein 1.
Green_Pods_Stacey	Glyma07g01730.1	12642,7631916093	Pfam:03767 HAD superfamily, subfamily IIIB (Acid phosphatase)
Leaves_Stacey	Glyma05g25810.1	10438,4437577971	Pfam:00504 Chlorophyll A-B binding protein Panther:21649 CHLOROPHYLL A/B BINDING PROTEIN KEGGORTH:08912 light-harvesting complex II chlorophyll a/b binding protein 1
Nodule_Stacey	Glyma08g14020.1	25511,9992404484	no annotations.
Root_Stacey	Glyma09g12200.3 : Glyma09g12200.4 : Glyma09g12200.5 : Glyma09g12200.6 : Glyma09g12200.7	13909,5068471547	no annotations.
Root_Tip_Stacey	Glyma17g02260.1	5748,6431879413	Pfam:01179 Copper amine oxidase, enzyme domain Pfam:02727 Copper amine oxidase, N2 domain Pfam:02728 Copper amine oxidase, N3 domain

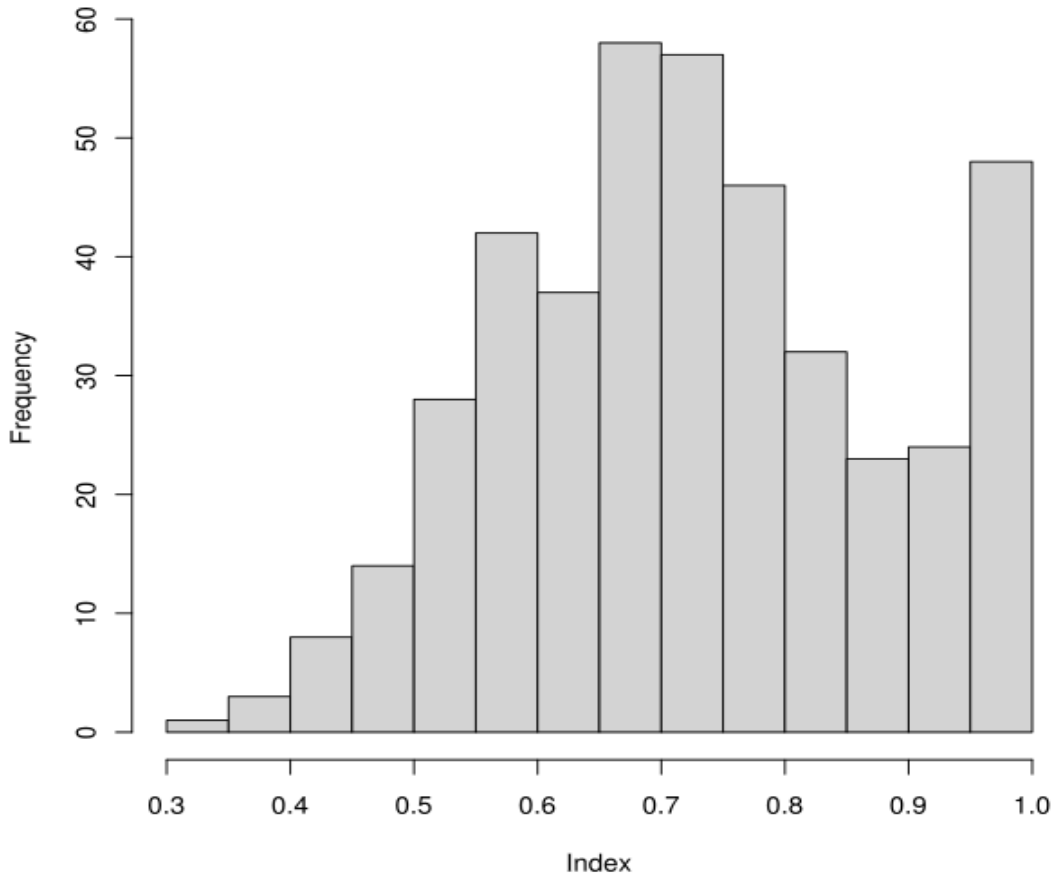

Figura 21: O histograma ilustra as frequências encontradas para os valores de índice de especificidade de expressão tecidual nos genes de G. max. Os valores foram calculados com base nos dados obtidos no experimento de Libault et al., 2010.

Discussão

Inicialmente foram realizadas quantificações nos dados obtidos de estudos de transcriptomas já publicados (Severin *et al.,* 2010; Libault *et al.,* 2010). As análises preliminares mostraram que não há nenhum tecido que apresentasse enriquecimento visível no conteúdo de genes F-box expressos. A quantificação de domínios C-terminais nas proteínas da família F-box também não revelou nenhum viés. A avaliação da especificidade da expressão tecidual dos genes F-box mostrou um comportamento que tende a um perfil intermediário entre o tecido-específico e o constitutivo. Não há deslocamento visível no sentido da expressão constitutiva. Contudo parece haver uma quantidade considerável de genes que tendem a um perfil tecido específico. A reconstrução filogenética dos genes dessa família mostrou-se ineficiente (dado não mostrado), uma vez que as sequências das proteínas divergem de forma a não permitir que o algoritmo utilizado trace relações confiáveis entre os clados.

No sentido de compreender a evolução da família F-box no grupo das leguminosas, foram utilizados os genomas de *G. max* (Gm) e *M. truncatula* (Mt). Também foram inseridos *A. thaliana* (At) (Eurosídeas II) e *V. vinifera* (Vv) (rosídea basal) como membros externos ao grupo das leguminosas. Foram encontrados números bastante distintos de F-boxes entre as espécies analisadas, consequência de ganhos e perdas linhagem-específicos – 480 (Gm), 913 (Mt), 688 (At), 147 (Vv). A alta variação encontrada na quantidade de genes F-box nas espécies estudadas nos levou a investigar a arquitetura genômica da família nestes organismos. Para tanto, a fração de F-boxes presentes em regiões de sintenia entre os genomas constituem um bom indicativo de conservação evolutiva (Figura 5). A significância estatística do resultado é suportada pela proporção de F-boxes nas 10.000 regiões sintênicas simuladas computacionalmente. Novamente, diferenças marcantes foram encontradas em espécies filogeneticamente próximas. Dentre os genes F-box de soja encontrados em sintenia com Vv, 95,7% estão mapeados em regiões de duplicação de segmento. Tal fato aponta para uma grande contribuição dos dois eventos de duplicação de genoma inteiro, ocorridos após a divergência entre as rosídeas basais e as eurosídeas, para o atual arcabouço de F-boxes encontrado no

genoma da soja. Por outro lado, observamos que Mt possui a menor fração de suas F-boxes em regiões de sintenia, resultado do grande número de duplicações locais que formam longos arranjos de genes F-box em tandem. Por exemplo, observa-se uma região densa em F-boxes (30 genes da família ao longo de ~368Kb). Nota-se ainda que muitos destes genes estão transcricionalmente ativos em diferentes tecidos (Figura 4). Apenas 15% (72 de 480) dos F-boxes de Gm são oriundos de duplicações locais recentes, ao contrário dos 53.8% encontrados em Mt (491 de 913) (Figura 3; Figura 5). Uma região densa em genes F-box foi encontrada no cromossomo 18 de Gm. Contendo 16 genes F-box e pelo menos 5 potencialmente inativos (genes que perderam o domínio F-box e ainda assim apresentam semelhança com outros F-boxes) ao longo de ~497Kb. Em nenhum dos transriptomas utilizados (Libault et al., 2010; Severin et al., 2010) no estudo foi detectado qualquer traço de expressão destes genes. Contudo, em uma perspectiva ampla, observamos que 53,57% (30/56) dos demais genes F-box duplicados em tandem em *G. max* são expressos em pelo menos uma condição/tecido. Foram encontrados padrões transcricionais bastante distintos nestes arranjos de F-boxes. Ao passo que alguns genes vizinhos retêm padrões similares, outros são claramente divergentes (Figura 4). No cromossomo 18, por exemplo, Glyma18g51020, Glyma18g50990 e Glyma18g51000 formam uma tríade. Enquanto os dois primeiros são transcritospreferencialmente em nódulos, e estão provavelmente envolvidos em interações com bactérias *Rhizobium*, o terceiro é transcrito principalmente em tecidos aéreos. Este dado sugere uma diversificação funcional recente neste arranjo de tandem-F-boxes. Alguns outros F-boxes presentes em arranjos tandem desenvolveram expressão preferencial em nódulo (Figura 4), sugerindo que a ubiquitinação mediada pelo complexo SCF desempenha papel importante na fixação de nitrogênio em soja. Em *M. truncatula*, o perfil global de transcrição dos F-boxes duplicados localmente pode ser dividido em três secções com base nos perfis de expressão: embriogênese tardia e fase de transição; desenvolvimento tardio de semente; e nódulos (Figura 4). Observou-se também que muitos F-boxes duplicados localmente são transcritos em altos níveis em determinados tecidos (Figura 6). Embora não pareça haver correlação entre a vizinhança dos F-boxes e a transcrição prefencial em nódulos, alguns F-boxes oriundos de duplicações locais (Medtr2g091950, Medtr4g134000 e Medtr5g035190) apresentam altos níveis de

transcrição em nódulos e respondem positivamente ao tratamento com NO3 (Figura 4).

Conclusões

➔ Os eventos de duplicação de genoma inteiro ocorridos após a divergência entre as rosídeas basais e o ancestral das eurosídeas contribuíram marcadamente para o atual conjunto de F-boxes encontradas em soja.

➔ Duplicações locais (tandem) constituem a principal força evolutiva na proliferação da família F-box em *M. truncatula;*

➔ Diversas F-boxes duplicadas recentemente evoluíram perfis transcricionais tecido-específicos, indicando sua participação em processos biológicos importantes, como por exemplo a nodulação;

➔ Os inventários de F-boxes podem variar drasticamente entre espécies, mesmo as evolutivamente próximas.

Referências bibliográficas

Ainsworth, E. A., Yendrek, C. R., Skoneczka, J. A., Long, S. P. (2011), Accelerating yield potential in soybean: potential targets for biotechnological improvement. Plant, Cell & Environment.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25: 3389-3402.

Golombek, S., Rolletschek, H., Wobus, U., Weber, H. (2001) Control of storage protein accumulation during legume seed development. J Plant Physiol 158: 457–464.

Bai, C., Richman, R., Elledge, S.J. (1994) Human cyclin F. EMBO J 13:6087– 6098.

Bai, C., Sen, P., Hofmann, K., Ma L., Goebl, M., Harper, J.W., Elledge, S.J. (1996) SKP1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. Cell 86, 263–274

Brocard-Gifford, I.M., Lynch, T.J., Finkelstein, R.R. (2003) Regulatory networks in seeds integrating developmental, abscisic acid, sugar, and light signaling. Plant Physiol 131: 78–92.

Burroughs, A.M., Iyer, L.M., Aravind, L. (2009) Natural history of the E1-like superfamily: implication for adenylation, sulfur transfer, and ubiquitin conjugation. Proteins 75: 895–910.

Burroughs, A.M., Iyer, L.M., Aravind, L. (2012) Structure and evolution of ubiquitin and ubiquitin-related domains. Methods Mol Biol 832:15-63.

Cardozo, T., Pagano, M. (2004) The SCF ubiquitin ligase: insights into a molecular

machine. Nat Rev Mol Cell Biol 5: 739–751.

Cenciarelli, C., Chiaur, D.S., Guardavaccaro, D., Parks, W., Vida,l M., Pagano, M. (1999) Identification of a family of human F-box proteins. Curr. Biol. 9, 1177–1179

Clamp, M., Cuff, J., Searle, S.M., Barton, G.J. (2004) The Jalview Java alignment editor. Bioinformatics. 12;20(3):426-7.

Crow, K. D. e Wagner, G. P. (2006) What is the role of genome duplication in the evolution of complexity and diversity? Mol. Biol. Evol. 23, 887–892.

Deshaies, R.J. (1999) SCF and Cullin/Ring H2-based ubiquitin ligases. Annu. Rev. Cell Dev. Biol. 15, 435–467

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 19;32(5):1792-7.

Elke, G.R., Blackmore, D., Offler, C.E., Patrick, J.W. (2005) Increased capacity for sucrose uptake leads to earlier onset of protein accumulation in developing pea seeds. Funct Plant Biol 32997–1007.

El Yahyaoui, F., Küster, H., Ben Amor, B., Hohnjec, N., Pühler, A., Becker, A., Gouzy, J., Vernié, T., Gough, C., Niebel, A., Godiard, L., Gamas, P. (2004) Expression profiling in *Medicago truncatula* identifies more than 750 genes differentially expressed during nodulation, including many potential regulators of the symbiotic program. Plant Physiol. 136(2):3159-76.

Ermolaeva, M. (2005) Operon finding in bacteria. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. 2886–2891.

Fait A., Angelovici, R., Less, H., Ohad, I., Urbanczyk-Wochniak, E., Fernie, A.R., Galili, G. (2006) *Arabidopsis* seed development and germination is associated with temporally distinct metabolic switches. Plant Physiol 142: 839–854.

Fang, S., Weissman, A.M. (2004) A field guide to ubiquitylation. Cell Mol Life Sci 61: 1546–1561

Feldman, R.M., Correll, C.C., Kaplan, K.B., Deshaies, R.J. (1997) A complex of Cdc4p, Skp1p and Cdc53p/cullin catalyzes ubiquitination of the phophorylated CDK inhibitor Sic1p. Cell 91, 221–230

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach; J Mol Evol. 17(6):368-76.

Finn, R.D., Clements, J., Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res; 39: 29–37.

Fitch, W.M. (2000) Homology a personal view on some of the problems. Trends Genet; 16:227-31.

Food And Agriculture Organization Of The United Nations (FAOSTAT 2010;http://faostat.fao.org/ default.aspx).

Gagne, J.M., Downes, B.P., Shiu, S.H., Durski, A.M., Vierstra, R.D. (2002) The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in Arabidopsis. Proc Natl Acad Sci USA 99:11519 –11524.

Gonzalez-Pastor, J.E., San Millan, J.L., Castilla, M.A., Moreno, F. (1995) Structure and organization of plasmid genes required to produce the translation inhibitor microcin C7. J. Bacteriol. 177, 7131–7140.

Glickman, M.H., Ciechanover, A. (2002) The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. Physiol. Rev. 82: 373–428

Haas, B.J., Delcher, A.L., Wortman J.R., Salzberg, S.L. (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics, 20(18):3643-3646.

Haas, A.L., Siepmann, T. J. (1997) Pathways of ubiquitin conjugation. FASEB J. 11, 1257–1268.

Hershko, A., Ciechanover, A. (1998) The ubiquitin system. Annu Rev Biochem 67: 425–479.

Hochstrasser, M. (1998) There's the Rub: a novel ubiquitin-like modification involved in cell cycle regulation. Genes Dev. 12, 901–907.

Hochstrasser, M. (2000) Evolution and function of ubiquitin-like protein-conjugation systems. Nat Cell Biol 2: E153–157.

Hua, Z., Zou, C., Shiu, S.H., Vierstra, R.D. (2011) Phylogenetic comparison of F-Box (FBX) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift. PLoS One. 6(1)

Huson, D., Richter, D., Rausch, C., Dezulian, T., Franz, M., Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics; 8: 460.

Iyer, L.M., Burroughs, A.M., Aravind, L. (2006) The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. Genome Biol 7: R60.

Jackson, P.K., Eldridge, A.G., Freed, E., Furstenthal L., Hsu J.Y., Kaiser B.K., Reimann J.D., (2000). The lore of the RINGs: Substrate recognition and catalysis by ubiquitin ligases. Trends Cell Biol. 10: 429–439.

Jones, K.M., Kobayashi, H., Davies, B.W., Taga, M.E., Walker, G.C. (2007) How rhizobial symbionts invade plants: the Sinorhizobium-Medicago model. Nat Rev Microbiol 5: 619–633.

Kamura, T., Koepp, D.M., Conrad, M.N., Skowyra, D., Moreland, R.J., Iliopoulos, O.,

Lane, W.S., Kaelin, W.G. Jr, Elledge, S.J., Conaway, R.C., Harper, J.W., Conaway, J.W. (1999) Rbx1, a component of the VHL tumor suppressor complex and SCF ubiquitin ligase. Science 284, 657–661

Kipreos, E.T., Pagano, M. (2000) The F-box protein family. Genome Biol. 1(5).

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. Genome research, 19(9):1639-1645.

Jin, J., Liu, X., Wang, G., Mi, L., Shen, Z., Chen, X., Herbert, S.J. (2010) Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. Field Crops Research 115, 116–123.

Ladrera, R., Marino, D., Larrainzar, E., González, E.M., Arrese-Igor, C. (2007) Reduced carbon availability to bacteroids and elevated ureides in nodules, but not in shoots, are involved in the nitrogen fixation response to early drought in soybean. Plant Physiol 145:539–546.

Lechner, E., Achard, P., Vansiri, A., Potuschak, T., Genschik, P. (2006) F-box proteins everywhere. Curr Opin Plant Biol, 9(6):631-8.

Li, W. H. (1983) Evolution of duplicate genes and pseudogenes. In: Nei M, Koehm RK, eds. Evolution of genes and proteins. Sunderland, MA: Sinauer Associates, Inc. pp 14–37.

Libault, M., Farmer, A., Brechenmacher, L., May, G.D., Stacey, G. (2010) Soybean root hairs: a valuable system to investigate plant biology at the cellular level. Plant Signal Behav, 5(4):419-21.

Marino, D., Frendo, P., Ladrera, R., Zabalza, A., Puppo, A., Arrese-Igor, C., González, E.M. (2007) Nitrogen fixation control under drought stress. Localized or systemic? Plant Physiol 143:1968–1974.

Maunoury, N., Redondo-Nieto, M., Bourcy, M., Van de Velde, W., Alunni, B., Laporte, P., Durand, P., Agier, N., Marisa, L., Vaubert, D., Delacroix, H., Duc, G., Ratet, P., Aggerbeck, L., Kondorosi, E., Mergaert, P. (2010) Differentiation of symbiotic cells and endosymbionts in *Medicago truncatula* nodulation are coupled to two transcriptome-switches. PLoS One, 4;5(3)

McCouch, S.R. (2001) Genomics and Synteny. Plant Physiol. 125:152–155.

Morrison, M.J., Voldeng, H.D., Cober, E.R. (2000) Agronomic changes from 58 years of genetic improvement of short-season soybean cultivars in Canada. Agronomy Journal 92, 780–784.

Nei, M., Rooney, A.P. (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152.

Nunes-Nesi, A., Fernie, A.R., Stitt, M. (2010) Metabolic and signaling aspects underpinning the regulation of plant carbon nitrogen interactions. Mol Plant, 3(6):973-96.

Oya, T., Nepomuceno, A.L., Neumaier, N., Farias, J.R.B., Tobita, S., Ito, O. (2004) Drought Tolerance Characteristics of Brazilian Soybean Cultivars Evaluation and characterization of drought tolerance of various Brazilian soybean cultivars in the field. Plant Prod. Sci. 7 (2) : 129-137.

Otto, S.P., Whitton, J. (2000) Polyploid incidence and evolution. Annu Rev Genet, 34:401-437.

Pickart, C.M. (2001) Mechanisms underlying ubiquitination. Annu. Rev. Biochem. 70: 503–533

Peters, J.M. (1998) SCF and APC: the Yin and Yang of cell cycle regulated proteolysis. Curr. Opin. Cell Biol. 10 (6), 759–768 Read, M.A., Brownell, J.E.,

Gladysheva, T.B., Hottelet, M., Parent, L.A., Coggins, M.B., Pierce, J.W., Podust, V.N., Luo, R.S., Chau, V., Palombella, V.J. (2000) Nedd8 modification of cul-1 activates SCF(beta(TrCP-dependent ubiquitination of IkappaBalpha. Mol. Cell Biol. 20, 2326–2333.

Saeed, A.I., Bhagabati, N.K., Braisted, J.C., Liang, W., Sharov, V., Howe, E.A., Li, J., Thiagarajan, M., White, J.A., Quackenbush, J. (2006) TM4 microarray software suite. Methods Enzymol; 411:134–193.

Sarkar, A., Soueidan, H., Nikolski, M. (2011) Identification of conserved gene clusters in multiple genomes based on synteny and homology. BMC Bioinformatics. 12(Suppl 9): S18.

Santner, A., Estelle, M. (2010) The ubiquitin-proteasome system regulates plant hormone signaling. Plant J, 61(6):1029-40.

Scheffner, M. (1998) Ubiquitin, E6-AP, and their role in p53 inactivation. Pharmacol. Ther. 78, 129–139

Schmidt, M.A., Barbazuk, W.B., Sandford, M., May, G., Song, Z., Zhou, W., Nikolau, B.J., Herman, E.M. (2011) Silencing of soybean seed storage proteins results in a rebalanced protein composition reserving seed protein content without major collateral changes in the metabolome and ranscriptome. Plant Physiol, 156(1):330-45.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., *et al*. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–183.

Sémon, M., Wolfe, K.H. (2007) Consequences of genome duplication. Curr Opin Genet Dev. 17(6):505-12.

Seol, J.H., Feldman, R.M., Zachariae, W., Shevchenko, A., Correll, C.C., Lyapina, S., Chi, Y., Galova, M., Claypool, J., Sandmeyer, S., Nasmyth, K., Deshaies, R.J.,

Shevchenko, A., Deshaies, R.J. (1999) Cdc53/cullin and the essential Hrt1 RING-H2 subunit of SCF define a ubiquitin ligase module that activates the E2 enzyme Cdc34. Genes Dev. 13, 1614–1626

Skowyra, D., Craig, K.L., Tyers, M., Elledge, S.J., Harper, J.W. (1997) F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. Cell 91, 209–219

Skowyra, D., Koepp, D.M., Kamura, T., Conrad, M.N., Conaway, R.C., Conaway, J.W., Elledge, S.J., Harper, J.W. (1999) Reconstitution of G1 cyclin ubiquitination with complexes containing SCFGrr1 and Rbx1. Science 284, 662–665

Specht, J.E., Hume, D.J., Kumudinia, S.V. (1999) Soybean yield potential – a genetic and physiological perspective. Crop Science 39, 1560–1570.

Stamatakis, A., Hoover, P., Rougemont, J. (2008) A rapid bootstrap algorithm for the RAXML web-servers. Syst Biol. 75:758–771.

United States Department of Agriculture (USDA - http://www.usdabrazil.org.br).

Weissman, A.M. (2001) Themes and variations on ubiquitylation. Nat. Rev. Mol. Cell Biol. 2: 169–178

Van de Peer, Y., Maere, S., Meyer, A. (2009) The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10(10):725-32.

Venancio, T.M., Balaji, S., Iyer, L., Aravind, L. (2009) Reconstructing the ubiquitin network: cross-talk with other systems and identification of novel functions. Genome Biol 10: R33.

Wally, O., Punja, Z.K. (2010) Genetic engineering for increasing fungal and bacterial disease resistance in crop plants. GM Crops. 1(4):199-206.

Weber, H., Borisjuk, L., Wobus, U. (2005) Molecular physiology of legume seed development. Annu Rev Plant Biol 56: 253–279.

Winston, J.T., Koepp, D.M., Zhu, C., Elledge S.J., Harper, J.W., (1999) A family of mammalian F-box proteins. Curr. Biol. 9, 1180–1182

Wolf, D.A., Jackson, P.K. (1998) Cell cycle: oiling the gears of anaphase. Curr. Biol. 8, R636–R639

Xu, G., Ma, H., Nei, M., Kong, H. (2009) Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proc Natl Acad Sci U S A. 20;106(3):835-40.

Yan, Y.S., Chen, X.Y., Yang, K., Sun, Z.X., Fu, Y.P., Zhang, Y.M., Fang, R.X. (2011) Overexpression of an F-box protein gene reduces abiotic stress tolerance and promotes root growth in rice. Mol Plant, 4(1):190-7.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21: 650–659.

Young, N.D., Debelle, F., Oldroyd, G.E., Geurts, R., Cannon, S.B., Udvardi, M.K., Benedito, V.A., Mayer, F., Gouzy, J., Schoof, H. *et al* (2011) The Medicago genome provides insight into the evolution of hizobial symbioses. Nature 480(7378):520-524.

Zachariae, W., Nasmyth, K. (1999) Whose end is destruction: cell division and the anaphase-promoting complex. Genes Dev. 13, 2039–2058