UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

LEANDRO COSTA DO NASCIMENTO

"Análise de expressão gênica diferencial entre diversas bibliotecas de soja"

Este exemplar corresponde à redação final da tese defendida pelo(a) candidato (a) Leandro Costa do Nascimento

e aprovada pela Comissão Julgadora.

Dissertação apresentada ao Instituto de Biologia para obtenção do Título de Mestre em Genética e Biologia Molecular, na área de Bioinformática.

Orientador: Prof. Dr. Gonçalo Amarante Guimarães Pereira

Campinas, 2010

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP

N17a

Nascimento, Leandro Costa do
Análise de expressão gênica diferencial entre diversas bibliotecas de soja / Leandro Costa do Nascimento. – Campinas, SP: [s.n.], 2010.

Orientador: Gonçalo Amarante Guimarães Pereira.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Soja. 2. Transcriptoma. 3. Banco de dados. 4. Integração de dados. I. Pereira, Gonçalo Amarante Guimarães. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

(rcdt/ib)

Título em inglês: Analysis of differential gene expression between different libraries of soybean.
Palavras-chave em inglês: Soybean; Transcriptome; Database; Data integration.
Área de concentração: Bioinformática.
Titulação: Mestre em Genética e Biologia Molecular.
Banca examinadora: Gonçalo Amarante Guimarães Pereira, Michel Eduardo Beleza Yamagishi, Eliseu Binneck.
Data da defesa: 09/12/2010.
Programa de Pós-Graduação: Genética e Biologia Molecular.

Campinas, 09 de dezembro de 2010.

BANCA EXAMINADORA

Prof. Dr. Gonçalo Amarante Guimarães Pereira
(Orientador)

Assinatura

Prof. Dr. Michel Eduardo Beleza Yamagishi

Assinatura

Dr. Eliseu Binneck

Assinatura

Dr. Jorge Maurício Costa Mondego

Assinatura

Dr. Leandro Carrijo Cintra

Assinatura

Dedico este trabalho a toda minha família: meus pais, Pedro e Janeide, e meus irmãos, Bruno e Lucas.

AGRADECIMENTOS

Ao professor Dr. Gonçalo Amarante Guimarães Pereira, pela confiança cedida no início e durante este trabalho.

A todos os membros da equipe de bioinformática do Laboratório de Genômica e Expressão – LGE pelo apoio em pontos fundamentais do trabalho, principalmente ao Marcelo Falsarella Carazzolle que junto comigo tornou possível a construção das ferramentas aqui descritas.

A todos os participantes do projeto GENOSOJA, principalmente aos colegas da Embrapa – Soja, responsáveis pela geração dos dados aqui apresentados, sem os quais, eu não conseguiria desenvolver este trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) pela bolsa concedida durante todo o tempo do mestrado.

Aos meus amigos de Jundiaí (Jhonny, Leonardo e Loammy) e de Campinas (Gleidson, Javier, Roberto, Osvaldo, Lucas e Osmar), companheiros de festas e churrascos durante o período do mestrado.

A toda minha família por me apoiarem em todos os momentos. Vocês são, sem dúvida, pilares fundamentais em toda a minha vida e, mais uma vez, foram durante este trabalho.

SUMÁRIO

LISTA DE ABREVIAÇÕES **8**

LISTA DE FIGURAS **10**

LISTA DE TABELAS **12**

RESUMO **13**

ABSTRACT **14**

INTRODUÇÃO **15**

O Projeto GENOSOJA **16**

CAPÍTULO 1: CONCEITOS BÁSICOS **19**

O código genético **19**

Transcrição e tradução **20**

Estudos de expressão de genes **23**

CAPÍTULO 2: DADOS PÚBLICOS DE SOJA **25**

Genoma da soja **25**

Análise de ESTs **26**

cDNAs Full-Length **37**

Resumo dos dados públicos de soja **38**

CAPÍTULO 3: DADOS GERADOS PELO GENOSOJA 39

Tags de SuperSAGE 39

Bibliotecas subtrativas de cDNA 47

MicroRNAs 55

Resumo dos dados gerados pelo projeto GENOSOJA 59

CAPÍTULO 4: INTEGRAÇÃO DE DADOS 61

CAPÍTULO 5 - CONCLUSÕES E PERSPECTIVAS 69

REFERÊNCIAS BIBLIOGRÁFICAS 80

LISTA DE ABREVIAÇÕES

cDNA – DNA complementar – DNA obtido a partir de transcritos

EST – Expressed Sequence Tag – sequência obtida a partir do sequenciamento de cDNA

ESALQ – Escola de Agricultura Luís de Queiroz

USP – Universidade de São Paulo

ISGC – Consórcio Internacional de estudos genômicos em soja

UNICAMP – Universidade Estadual de Campinas

LGE – Laboratório de Genômica e Expressão

DNA – Ácido desoxirribonucléico

RNA – Ácido ribonucléico

CDS – regiões do RNA maduro que são traduzidas em proteínas, ou seja, regiões codantes

UTR – regiões do RNA maduro que não são traduzidas em proteínas

SAGE – Serial Analysis of Gene Expression – metodologia para quantificação de expressão de genes

JGI – Joint Genome Institute – Instituto de pesquisa dos Estados Unidos

bp – pares de bases

kb – kilobases (um mil pares de bases)

mb – megabases (um milhão de pares de bases)

gb – gigabases (um bilhão de pares de bases)

NCBI – National Center for Biotechnology Information – Banco de dados público de sequências de organismos

RPKM – Reads por Kilobase por Milhão de reads – medida de expressão no caso de transcritos curtos

SAM – Sequences Alignment/Map Format – Formato utilizado por programas de alinhamento

miRNAs – microRNAs

UFRGS – Universidade Federal do Rio Grande do Sul

Gbrowse – Generic Genome Browser – programa para visualização de sequências genômicas

GFF – Generic Feature Format – formato utilizado por programas de alinhamento

LISTA DE FIGURAS

Figura 1: Estrutura de dupla hélice da molécula de DNA **20**

Figura 2: Transcrição em eucariotos **21**

Figura 3: Código genético padrão **22**

Figura 4: Identificação da sequência de um possível gene eucarioto através de um preditor de genes **24**

Figura 5: Distribuição (%) dos ESTs de soja entre os diversos cultivares **28**

Figura 6: Distribuição (%) dos ESTs de soja entre os diversos tecidos **28**

Figura 7: *Reference assembly* **30**

Figura 8: Número de *contigs* e *singlets* de cada montagem **31**

Figura 9: Número de ESTs por contig em cada montagem **32**

Figura 10: Número de bases por contig em cada montagem **32**

Figura 11: Identificando erros nas montagens **33**

Figura 12: Montagens x Genes preditos do genoma **34**

Figura 13: Anotação utilizando o AutoFACT **35**

Figura 14: Expressão gênica diferencial entre bibliotecas de ESTs **36**

Figura 15: Experimento de SAGE **40**

Figura 16: Bancos de dados onde as tags foram alinhadas (%) **43**

Figura 17: Número de alinhamentos por tag..... **44**

Figura 18: Interface de visualização das amostras de SuperSAGE..... **46**

Figura 19: Interface para visualização dos dados de bibliotecas subtrativas..... **52**

Figura 20: Pipeline da análise *ab-initio* das bibliotecas subtrativas..... **53**

Figura 21: Visualização da região de um possível gene **54**

Figura 22: Pipeline utilizado na identificação de microRNAs..... **58**

Figura 23: Número de genes alvo por MicroRNA **59**

Figura 24: SuperSAGE x Bibliotecas subtrativas..... **62**

Figura 25: Modelo relacional resumido do banco de dados **65**

Figura 26: Interface web para integração dos dados..... **67**

Figura 27: Visualização de dados com o Gbrowse..... **69**

LISTA DE TABELAS

Tabela 1: Principais produtores de soja **15**

Tabela 2: Estatística do processo de trimagem **29**

Tabela 3: Parâmetros do CAP3 para as montagens realizadas **30**

Tabela 4: Dados públicos de soja **38**

Tabela 5: Amostras de SuperSAGE do projeto GENOSOJA **42**

Tabela 6: Alinhamentos das tags com bancos de dados de soja **43**

Tabela 7: Número de genes identificados nas amostras **45**

Tabela 8: Tags diferencialmente expressas nas amostras **46**

Tabela 9: Bibliotecas subtrativas do projeto GENOSOJA **48**

Tabela 10 – Número de genes encontrados nas bibliotecas subtrativas **50**

Tabela 11: Número de sequências das bibliotecas de RNAs pequenos **57**

Tabela 12: Número de sequências únicas e diferenciais das bibliotecas **57**

Tabela 13: Resumo dos dados gerados pelo GENOSOJA **60**

RESUMO

A soja é uma das principais *commodities* da economia internacional, sendo sua produção mundial de cerca de 220 milhões de toneladas por safra. Além de ser um alimento rico em proteínas e usado para a fabricação de óleo vegetal, a planta vem ganhando visibilidade devido a possibilidade de ser usada na fabricação de biocombustíveis, principalmente o biodiesel. Para o Brasil, a soja tem grande importância na balança comercial, sendo o país o segundo maior produtor do mundo. Neste contexto, no ano de 2007, o governo brasileiro estabeleceu um consórcio de pesquisas em soja - denominado GENOSOJA - com o objetivo de identificar características genéticas que possam facilitar o processo produtivo da planta, com foco nos diversos estresses que acometem a produção nacional, como a ocorrência de secas, o ataque de pragas e a doença da ferrugem asiática, causada pelo fungo *Phakopsora pachyrhizi*. Este trabalho está inserido no escopo do GENOSOJA, propondo a construção de bancos de dados contendo informações disponíveis nos diversos bancos públicos (sequências genômicas, ESTs e cDNA full-lenght), integrando-as com as informações geradas no decorrer do projeto (tags de SuperSAGE, bibliotecas subtrativas de cDNA e microRNAs). Além disso, foram construídas diversas interfaces web que oferecem aos usuários diversas funcionalidades, incluindo: comparações estatísticas, consultas por palavras-chave, dados sobre anotação e expressão dos genes nas diversas condições e experimentos estudados. Dessa forma, o ferramental de bioinformática aqui apresentado pode facilitar a compreensão de como as diferenças de expressão gênica da planta podem afetar características de importância agronômica.

ABSTRACT

Soybean is one of the main commodities in the international economy, with a world production of about 220 millions of tons per harvest. Besides being a protein rich food and used for vegetable oil production, the plant has been gaining visibility due to the possibility of being to make biofuels, especially biodiesel. The soybean culture is of great importance in the Brazilian economy, being the country the second largest producer in the world. In this context, in 2007, the Brazilian government established a research consortium in soybean – called GENOSOJA - aiming to identify genetic traits that may facilitate the production process of the plant, focusing on the different stresses that affect the national production, as the occurrence of drought, pests' attacks and the asian rust disease, caused by the *Phakopsora pachyrhizi* fungus. This work is inserted in the GENOSOJA, proposing to build a set of databases containing information available in several public databases (genomic sequences, ESTs and full-length cDNA), integrating them with information generated during the project (SuperSAGE tags, cDNA subtractive libraries and miRNAs). Additionally, several web interfaces were built. They offer to users many features, including: statics comparisons, keyword searches, data about annotation and gene expression in different experiments and conditions. Thus, the bioinformatics tools presented here may facilitate the understanding of how the differences in gene expression can affect plant traits with agronomic importance.

INTRODUÇÃO

A soja é uma planta da família das fabáceas (leguminosas) nativa do sudeste da Ásia. Trata-se de um grão com elevado poder alimentício, contendo quantidades significativas de diversos aminoácidos. O farelo da planta, por exemplo, é utilizado por organizações humanitárias, em muitos casos, como único alimento protéico fornecido a populações atingidas pela fome. Além disso, o óleo de soja é um dos tipos de óleo mais consumidos no mundo, sendo utilizado para o preparo de alimentos e em rações animais. A planta também é eficaz na prevenção de doenças, pois contém isoflavonas que ajudam na prevenção de diversos tipos de câncer, como o de pulmão e o de mama (Coward *et al*, 1993).

Por safra, são produzidas em todo o mundo, cerca de 250 milhões de toneladas de soja. O Brasil aparece como 2º maior produtor mundial, sendo responsável por cerca de um quarto da produção mundial. A Tabela 1 apresenta uma comparação entre a produção dos líderes mundiais de soja nas últimas safras.

Tabela 1: Principais produtores de soja - Produção dos líderes mundiais da soja em milhões de toneladas nas quatro últimas safras. Compilado a partir de Embrapa – Soja (http://www.cnpso.embrapa.br)

	2006/2007	2007/2008	2008/2009	2009/2010
EUA	86,77	72,9	80,5	91
Brasil	58,4	60	57,1	68,7
Argentina	45,1	47	32,3	54
Mundo	236,08	220,9	210,6	259,89

A produção brasileira está concentrada principalmente na região Centro-Oeste, na área do cerrado, sendo o estado de Mato Grosso o maior produtor nacional. O modelo de produção é totalmente baseado no agronegócio para exportação e a área plantada teve uma expansão de 25% nas últimas décadas. (Brandão *et al*, 2005).

A soja tem vital importância na balança comercial brasileira, respondendo sozinha por 10% do total de nossas exportações e cerca de 24% de nossas exportações agropecuárias. Além de ser um produto forte para exportação, diversas oportunidades têm surgido no mercado interno, como o uso da soja em biocombustíveis, mais precisamente o biodiesel. Apesar de outras espécies – como a mamona, a canola, o girassol e o dendê – serem mais produtivas para a obtenção do biodiesel (Santos *et al*, 2006), pesquisas realizadas pela Escola de Agricultura Luís de Queiroz (ESALQ), da Universidade de São Paulo (USP), demonstram que a soja é o produto mais viável para utilização imediata na fabricação do combustível no Brasil atualmente, por se tratar da única planta, dentre as possíveis, com estrutura de produção, distribuição e esmagamento de grãos já estabelecidos.

O Projeto GENOSOJA

A estabilidade da cultura brasileira de soja, tanto do ponto de vista econômico quanto ambiental, tem sido constantemente ameaçada devido a ações climáticas e ataques de pragas e doenças que até pouco tempo não causavam danos tão severos a cultura. Como exemplos, podemos citar: a ocorrência de secas, a ocorrência de

pragas, como o ataque de diferentes espécies de nematóides e a doença da ferrugem asiática (FAS) – causada pelo fungo *Phakopsora pachyrhizi* – que gera perdas bilionárias para os produtores de soja de diversos países. Com isso, cresce a importância de programas de melhoramento genético, visando a descoberta de novas técnicas de plantio, irrigação e prevenção de pragras, elevando assim a produção e diminuindo os custos com a mesma.

Por outro lado, nos últimos anos, o crescente desenvolvimento das tecnologias de sequenciamento tem facilitado diversas pesquisas na área genômica, permitindo maior compreensão da funcionalidade dos genes de diversas espécies, incluindo plantas como *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000) e *Oryza sativa* (arroz) (Sequencing Project International Rice Genome, 2005). Devido a imensa quantidade de dados gerados durante tais estudos, o sucesso dos mesmos é altamente dependente de uma análise apropriada de bioinformática, visando disponibilizar ferramentas e bancos de dados adequados para mineração das informações e, dessa forma, possibilitar a identificação de genes relacionados a resposta a determinados estresses. Com esta identificação, é possível, através de técnicas de biologia molecular, isolar tais genes e transferi-los entre cultivares e espécies, criando assim, novas variedades, mais resistentes às principais doenças e pragas.

Neste contexto, o governo brasileiro deu início, no ano de 2007, ao projeto GENOSOJA. Trata-se de um consórcio nacional de estudos em soja envolvendo diversas das principais instituições de pesquisa do Brasil, propondo integrar

informações do genoma da planta com informações sobre a expressão dos genes, dando ênfase a estresses que acometem a produção nacional. Dentre os objetivos do GENOSOJA está a criação de um banco de dados relacional, integrando os resultados obtidos através das diversas estratégias previstas no projeto como SuperSAGE, bibliotecas subtrativas e microRNAs. O projeto também prevê a integração entre os dados brasileiros com os dados gerados pelo Consórcio Internacional de estudos genômicos em soja (*ISGC*), liderado pelos Estados Unidos.

Dentro do escopo do GENOSOJA, o Laboratório de Genômica e Expressão (LGE – http://www.lge.ibi.unicamp.br) do Instituto de Biologia da Universidade Estadual de Campinas (UNICAMP) ficou responsável pela análise de bioinformática dos dados do projeto, incluindo submissão, tratamento e anotação de sequências, bem como armazenamento e gerenciamento do banco de dados. O presente trabalho, descreve as metodologias empregadas nas análises, incluindo a quantificação de dados de expressão de genes dentre as diversas condições estudadas pelo projeto. Também são mostrados os diversos dados públicos de soja obtidos na literatura e um esquema resumido do banco de dados. Além desta introdução, ele está dividido em 5 capítulos sendo: (*i*) definição de alguns conceitos básicos de genética, (*ii*) descrição dos diversos dados públicos de soja utilizados no projeto, (*iii*) descrição dos dados gerados pelo GENOSOJA e das metodologias empregadas na análise dos mesmos, (*iv*) integração dos dados das diversas fontes e resumo do banco de dados e (*v*) conclusão.

CAPÍTULO 1: CONCEITOS BÁSICOS

O objetivo deste capítulo é descrever alguns conceitos básicos de genética, que serão utilizados mais adiante. Outras definições, mais específicas de cada um dos tipos de dados trabalhados serão apresentadas no decorrer do texto.

O código genético

Os seres vivos possuem, no interior de suas células, uma molécula chamada DNA, responsável por armazenar todas as informações genéticas de um organismo, bem como transmiti-las aos seus descendentes (Watson & Berry, 2003). Tal molécula é composta por quatro diferentes nucleotídeos que são formados por uma pentose, um grupo fosfato e uma base nitrogenada que é diferente para cada um dos quatro. O DNA tem uma estrutura em formato de dupla-hélice (Watson & Crick, 1953). A dupla-hélice é composta por duas fitas de nucleotídeos de direções opostas, conforme apresentado na Figura 1. Entre as duas fitas ocorre um pareamento dos nucleotídeos através da formação de pontes de hidrogênio entre as suas bases (adenina – A - com timina – T - e guanina – G - com citosina - C).

É no DNA que se encontra o conjunto de genes de um organismo. Os genes estão relacionados às características de tal organismo, como o sexo, coloração do olho e da pele, mecanismo de defesa contra doenças, entre diversas outras. Apesar de alguns cientistas definirem genes como qualquer região do DNA que é transcrita (participa da transcrição – processo descrito nos próximos parágrafos) (Forsdyke,

2009), a definição mais aceita e que será utilizada neste trabalho é que gene é toda região do DNA codificada em proteína.

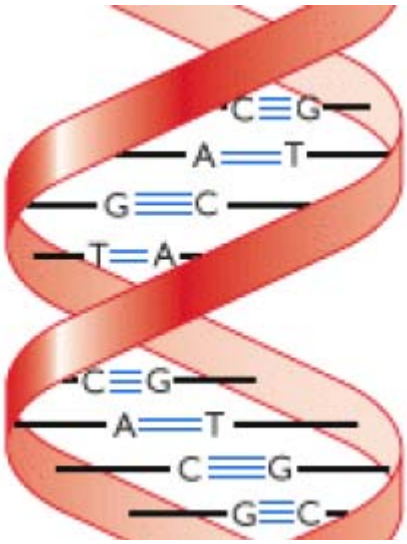


Figura 1: Estrutura de dupla hélice da molécula de DNA.

Transcrição e tradução

A partir da sequência de DNA relativa a um gene, ocorrem importantes processos celulares, como a transcrição e a tradução. A transcrição é a confecção de uma cópia de RNA (chamada de transcrito) a partir da sequência de DNA do gene. Alguns transcritos, por sua vez, serão convertidos em proteínas através do processo de tradução.

O início da transcrição ocorre com a ligação da enzima RNA polimerase nas proximidades da região do gene no genoma (região identificada como promotor de um

gene). No caso dos organismos eucariotos, como a soja, o transcrito inicial ainda é processado através de um mecanismo conhecido como *splicing*, onde ocorre a exclusão dos introns e junção dos exons, conforme apresentado na Figura 2. O transcrito final contém somente a região que pode ser codificada em proteína (todos os exons, ou, CDS) e as UTRs (*Untraslated regions* – regiões que estão nos transcritos, mas não participam da síntese de proteínas). A transcrição pode dar origem a diversos tipos de transcritos, como os RNAs transportadores (tRNA), os RNAs ribossomais (rRNA), os microRNAs (miRNA) e os RNAs mensageiros (mRNA). Estes últimos são os únicos que podem ser convertidos em proteínas.

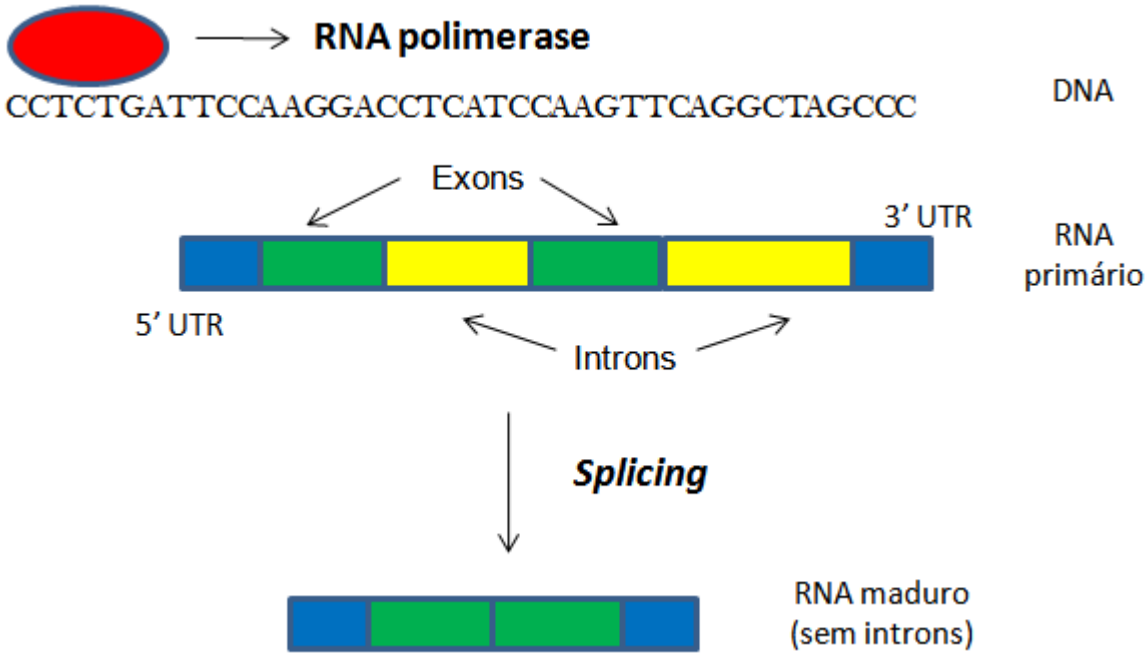


Figura 2: Transcrição em eucariotos – Incluindo o processo de *splicing*, onde ocorre a remoção dos introns (regiões em amarelo) e a junção dos exons (regiões em verde) dando origem ao RNA maduro ou final.

A tradução consiste na síntese de proteínas a partir de moléculas de RNA mensageiro. O processo ocorre com auxílio dos RNAs transportadores, responsáveis pelo transporte dos aminoácidos do citoplasma até o ribossomo. A sequência de uma proteína é baseada no código genético, onde é possível obter a correspondência entre os aminoácidos e os possíveis trios de nucleotídeos (códons). A Figura 3 mostra o código genético padrão, utilizado pela maioria dos organismos.

Visualizando o código genético (Figura 3), observa-se que ele é degenerado, ou seja, um aminoácido pode ser codificado por um ou mais códons (Crick *et al*, 1961). Com isso, mudanças de bases (polimorfismos) na sequência de DNA dos genes não necessariamente ocasionam mudanças na sequência de aminoácidos da proteína codificada.

UUU, UUC: phe UUA, UUG: leu	UCU, UCC, UCA, UCG: ser	UAU, UAC: tyr UAA, UAG: stop	UGU, UGC: cys UGA: stop UGG: trp
CUU, CUC, CUA, CUG: leu	CCU, CCC, CCA, CCG: pro	CAU, CAC: his CAA, CAG: gln	CGU, CGC, CGA, CGG: arg
AUU, AUC, AUA: ile AUG: met	ACU, ACC, ACA, ACG: thr	AAU, AAC: asn AAA, AAG: lys	AGU, AGC: ser AGA, AGG: arg
GUU, GUC, GUA, GUG: val	GCU, GCC, GCA, GCG: ala	GAU, GAC: asp GAA, GAG: glu	GGU, GGC, GGA, GGG: gly

Figura 3: Código genético padrão.

Estudos de expressão de genes

Considerando que os genes estão diretamente ligados a defesa de um organismo contra uma determinada doença, o estudo de expressão de genes tem sido utilizado como referência para a biotecnologia. Sabe-se que, sob a ação de determinadas doenças, um conjunto específico de genes do organismo estará super expresso e outro conjunto reprimido. Entender como a expressão de cada um desses genes contribui na resposta final a doença pode ser um caminho para o desenvolvimento de organismos mais resistentes a ela (Shinozaki *et al*, 2003; Shinozaki & Yamaguchi-Shinozaki, 2007; Van de Mortel *et al*, 2007).

A maioria dos estudos de expressão de genes são baseados em experimentos de transcriptoma (transcritos) como ESTs, SAGE, entre outros. Todos eles utilizam o que é chamado de DNA complementar (cDNA). As sequências de cDNA são produzidas através da ação da enzima transcriptase reversa, que, a partir de um transcrito, gera a sequência de DNA relativa a ele. Com isso, a sequência de cDNA contém somente dados relativos aos transcritos, incluindo as regiões codantes (CDS) e as UTRs, sem bases relativas aos introns.

Em outros casos, porém, não existem dados de transcritos do organismo estudado. Nessa hipótese, é possível utilizar métodos computacionais para a identificação das regiões do genoma que possivelmente codificam proteínas. Tal metodologia é conhecida como predição de genes e tem sido utilizada em diversos estudos de expressão gênica (Libault *et al*, 2010; Severin *et al*, 2010). A função dos preditores de genes é, dada uma sequência de DNA qualquer, identificar onde

começam e terminam os genes, e, no caso dos eucariotos, as fronteiras exon-intron, como demonstra a Figura 4.

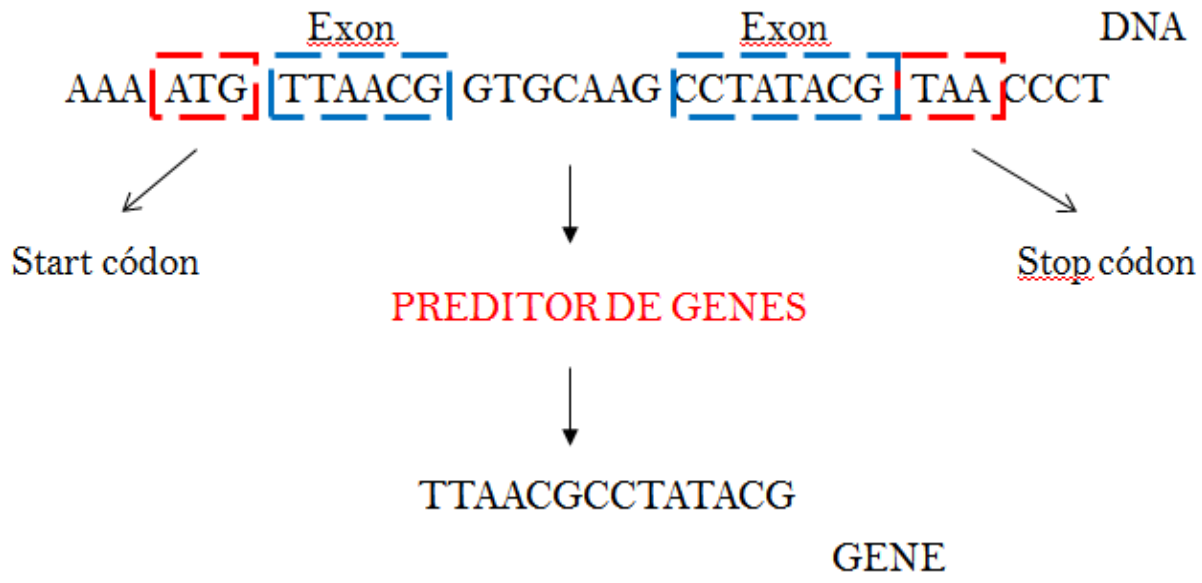


Figura 4: Identificação da sequência de um possível gene eucarioto através de um preditor de genes – Além de reconhecer os códons de início (Start códon) e de término (Stop códon) dos genes, um preditor de genes de eucariotos tem que identificar os introns de uma sequência.

Os preditores de genes podem ser baseados em métodos *ab-initio*, isto é, utilizam somente informação da própria sequência, como conteúdo GC e códon usage; ou comparativos, isto é, utilizam informações de sequências de transcritos ou de proteínas de espécies próximas a de estudo. Outros preditores, geralmente mais utilizados, se baseiam em informações de ambos os tipos, tanto da sequência genômica quanto da comparação com outras sequências.

CAPÍTULO 2: DADOS PÚBLICOS DE SOJA

Para a construção do banco de dados do GENOSOJA, inicialmente foram coletados diversos dados públicos de soja disponíveis na literatura. Este capítulo descreve todos esses dados, os bancos de dados onde foram obtidos e a análise de bioinformática envolvida em cada um deles.

Genoma da soja

A soja apresenta um genoma com tamanho aproximado de 1,1 Gb (Arumuganathan & Earle, 1991) e razoavelmente complexo devido a diversos ciclos de duplicações ocorridos nos últimos 45 milhões de anos, possuindo entre 40 e 60% de sequências repetitivas (Goldberg, 1978; Shoemaker *et al*, 1996; Schmutz *et al*, 2010).

Em janeiro de 2010, o consórcio internacional de pesquisas em soja liderado pelo *JGI* concluiu o sequenciamento do genoma da planta. Os estudos foram baseados no cultivar *Williams 82*, o mais plantado nos Estados Unidos. A montagem final foi realizada com 13 milhões de sequências, com cobertura de aproximadamente 7,2X do genoma e teve como resultado um total de 950 Mb divididos em 20 cromossomos. Além disso, um total de 66.153 genes foram preditos a partir das sequências genômicas, sendo 46.430 deles com alta confiabilidade (Schmutz *et al*, 2010). Os dados estão disponíveis através do link http://www.phytozome.net/soybean e foram os primeiros a serem armazenados no servidor do projeto. As sequências genômicas foram utilizadas como referência para integração dos dados utilizando o programa

Gbrowse, etapa esta que será descrita mais adiante, no capítulo de integração de dados.

Análise de ESTs

ESTs (*expressed sequence tags*) são sequências curtas de mRNA que são selecionadas de forma aleatória a partir de bibliotecas de cDNA. O sequenciamento de ESTs em larga escala tem diversas aplicações, como a descoberta de novos genes (Adams *et al*, 1991), identificação de polimorfismos de nucleotídeo único (SNPs) (Useche *et al*, 2001) e facilitar a análise de proteomas (Jongeneel, 2000). Através de análises comparativas e estatísticas (Audic & Claviere, 1997) é possível obter informações sobre o nível de expressão dos genes entre os diversos tecidos estudados.

Para uma análise de expressão de genes com ESTs é preciso, inicialmente, separar as sequências por espécie ou biblioteca. Após esta etapa é necessário "*trimar*" os fragmentos, ou seja, remover as sequências de baixa qualidade, de vetor, ribossomais e cauda poly A/T. Ao final do processo, são descartadas sequências curtas, geralmente menores que 100 pares de bases (bp). Toda esta etapa é realizada pelo programa bdtrimmer (Baudet & Dias, 2005).

Após a *trimagem*, o processo de montagem – ou clusterização - reduz a redundância de dados e facilita a mineração dos mesmos. A montagem é realizada através da busca de sobreposição entre os ESTs, onde, as sequências formadas pela

sobreposição de vários fragmentos são chamadas de *contigs*. Os ESTs sem sobreposição são chamados de *singlets*. Para esta etapa, podem ser utilizados diversos programas descritos na literatura, como o *TIGR* (Sutton *et al*, 1995), o *Newbler* (Margulies *et al*, 2005) e o CAP3 (Huang & Madan, 1999). Este último foi considerado o mais adequado para montagem de ESTs (Liang *et al*, 2000) e, dessa forma, será utilizado neste trabalho.

A etapa final da análise consiste na identificação de função biológica (anotação) dos unigenes (*contigs* e *singlets*) da montagem. Para isto, são realizadas buscas por similaridade usando informações de genes conhecidos armazenados em bancos de dados públicos utilizando o programa BLAST (Altschul *et al*, 1997). O processo consiste na busca de sequências similares, em nucleotídeos ou aminoácidos, com funções conhecidas e já anotadas.

Um total de 1.276.813 ESTs de soja – 393.386 provenientes de sequenciamento SANGER (cerca de 550 bp) e 882.427 provenientes de sequenciamento 454 (cerca de 250 bp) – de diversos cultivares e tecidos foram obtidos no banco de dados *dbEST* do *NCBI*. Estas sequências foram separadas por cultivar e biblioteca e inseridas no banco de dados local do GENOSOJA através de scripts em PERL. As Figuras 5 e 6 abaixo apresentam, respectivamente, o porcentual de cada um dos cultivares e tecidos dentre o conjunto total de sequências. A Tabela 2, por sua vez, apresenta as estatísticas do processo de trimagem dos ESTs e o número total de sequências que foram utilizadas na montagem.

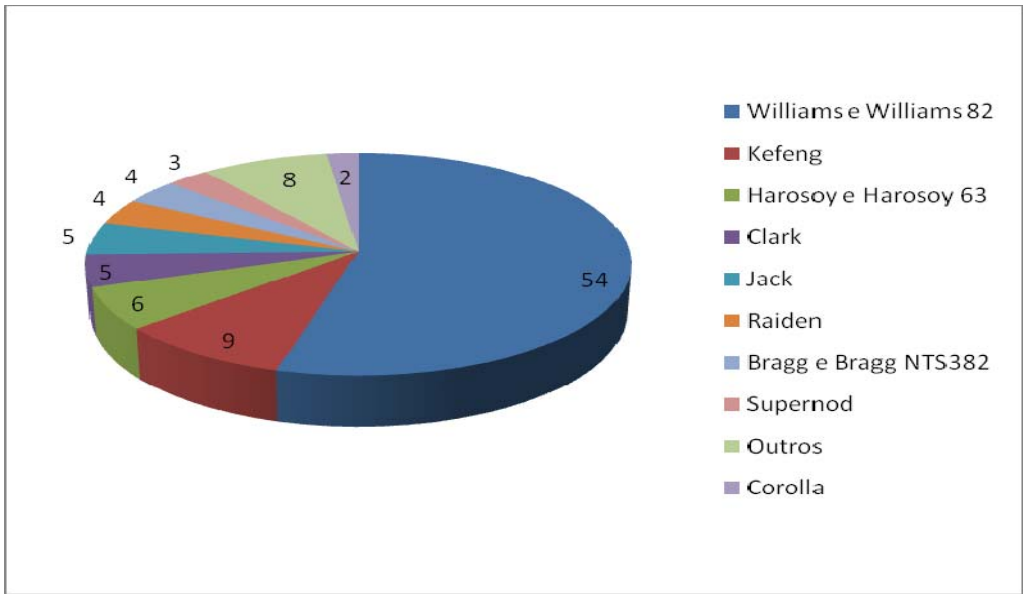


Figura 5: Distribuição (%) dos ESTs de soja entre os diversos cultivares – A maioria das sequências é dos cultivares Williams e Williams 82, que são os mais plantados nos Estados Unidos.

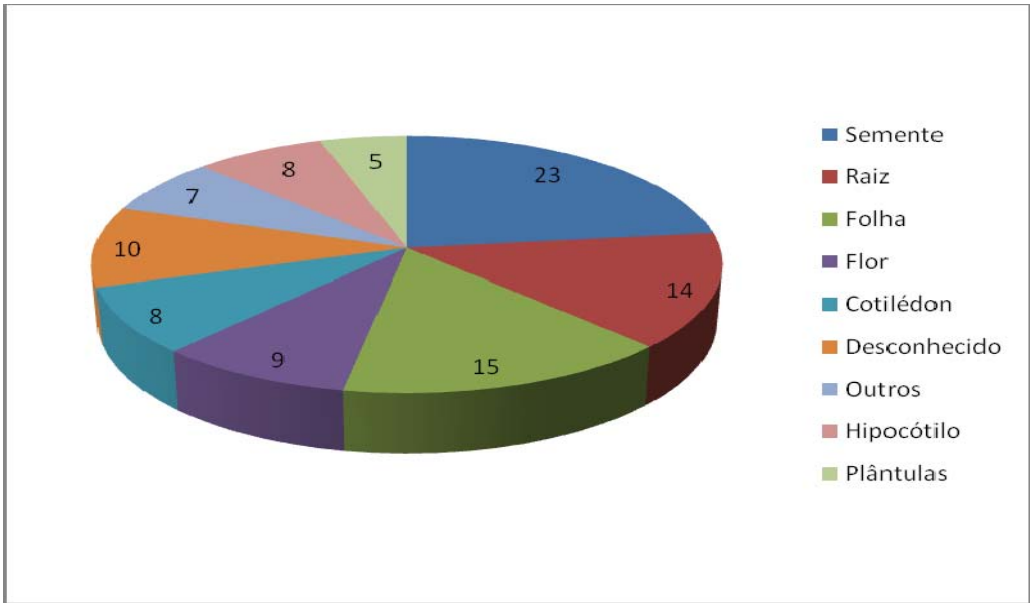


Figura 6: Distribuição (%) dos ESTs de soja entre os diversos tecidos – O gráfico mostra que a maioria dos ESTs é de semente e ainda, que 10% das sequências não tinham tecido conhecido.

Tabela 2: Estatística do processo de trimagem – Número de sequências ribossomais, com vetor, cauda poly A/T encontradas com o bdtrimer. Após este processo, cerca de 170.000 sequências foram descartadas, sendo que a montagem foi realizada, com cerca de 1.100.000 sequências.

Tipo	Número de ocorrências
Total	1.276.813
Ribossomais	86.742
Vetor	83.716
Poly A/T	90.193
Curtas	88.085
Utilizadas	1.101.986

De acordo com Wang *et al* (2004) podem existir dois tipos de erro resultantes do processo de montagem de ESTs. O erro do tipo um ocorre quando sequências relativas ao mesmo gene não são agrupadas no mesmo *contig*. Já o erro do tipo dois ocorre quando ESTs pertencentes a dois genes diferentes são agrupados de maneira conjunta em um só *contig*.

Em busca da clusterização que minimizava ambos os tipos de erro, foram realizadas diversas montagens com diferenciação entre os parâmetros do CAP3. Para todas elas foi utilizado um servidor Dual Xeon 2,66 GHZ com 96 GB de memória RAM. Os parâmetros ajustados foram o tamanho da região de sobreposição entre duas sequências (*o*) e o porcentual de similaridade entre a região de sobreposição (*p*). Além disso, foram utilizados dois tipos de montagem:

i) *Ab-initio:* Montagem típica de ESTs, onde todos os fragmentos são clusterizados em conjunto.

ii) *Reference assembly:* Foi utilizado o programa BLAST (e-value de corte 1e-10) para a busca de similaridade em nucleotídeos entre os ESTs e as sequências do genoma da soja descrito previamente. A partir da saída do BLAST foram realizadas múltiplas montagens onde cada grupo de fragmentos similares com a mesma região do genoma eram clusterizadas entre si (Figura 7).

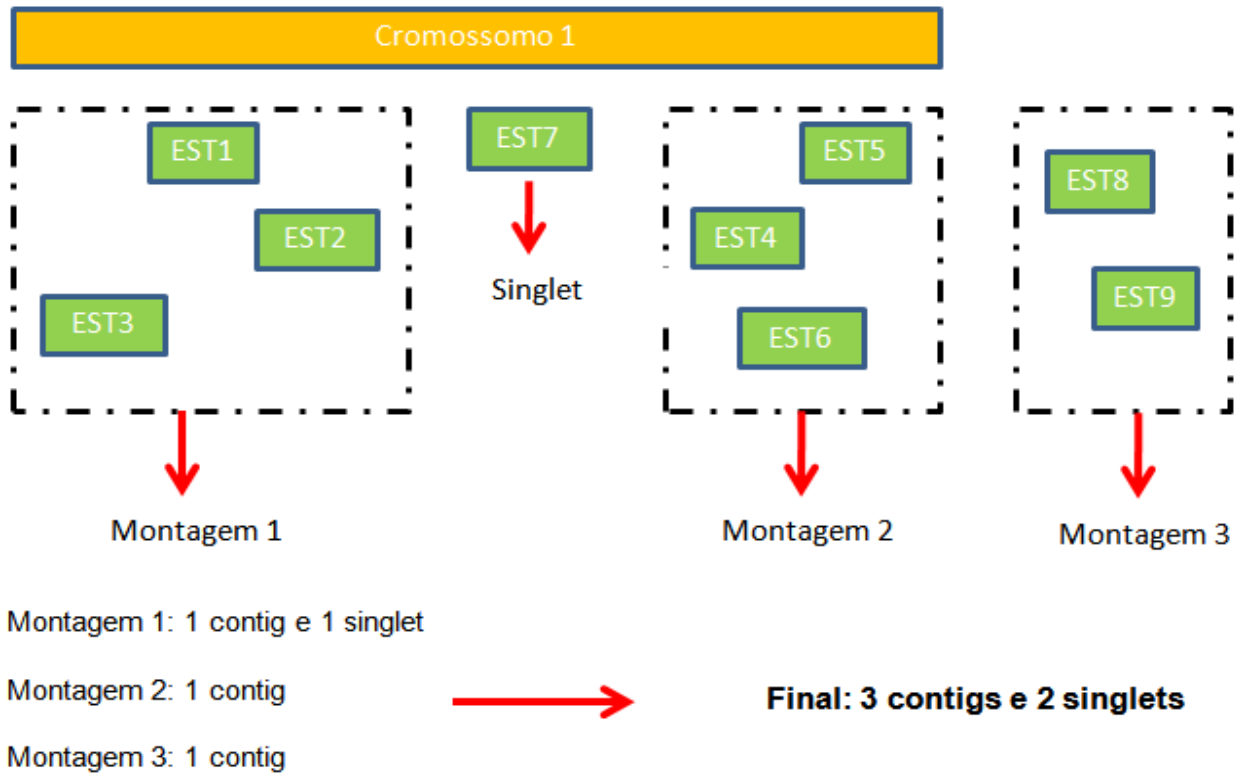


Figura 7: *Reference assembly* - Metodologia empregada nas montagens do tipo *reference assembly*. Todos os ESTs que alinham com a mesma região genômica (e que tem sobreposição entre os seus alinhamentos) são clusterizados em conjunto. Na figura são mostradas duas montagens relativas ao cromossomo 1 (montagem 1 e montagem 2) e mais uma relativa a ESTs que não alinham em nenhum local do genoma (montagem 3). Os fragmentos que alinham sozinhos em uma região genômica são tratados como *singlets*. O resultado final consiste da soma dos *contigs* e *singlets* de todas as montagens.

A Tabela 3 apresenta os parâmetros de cada uma das montagens, definidas como 1, 2, 3, 4, 5 e 6. Já as Figuras 8 a 10 mostram comparações entre as montagens, como número de *contigs* e *singlets*, número de ESTs por *contig* e número de bases por *contig*.

Tabela 3: Parâmetros do CAP3 para as montagens realizadas.

	Tam. da região de comparação (*o*)	Porcentual de smilaridade (*p*)	Tipo
1	40	80	*Ab-initio*
2	100	85	*Ab-initio*
3	100	90	*Ab-initio*
4	100	95	*Ab-initio*
5	40	80	*Reference*
6	100	90	*Reference*

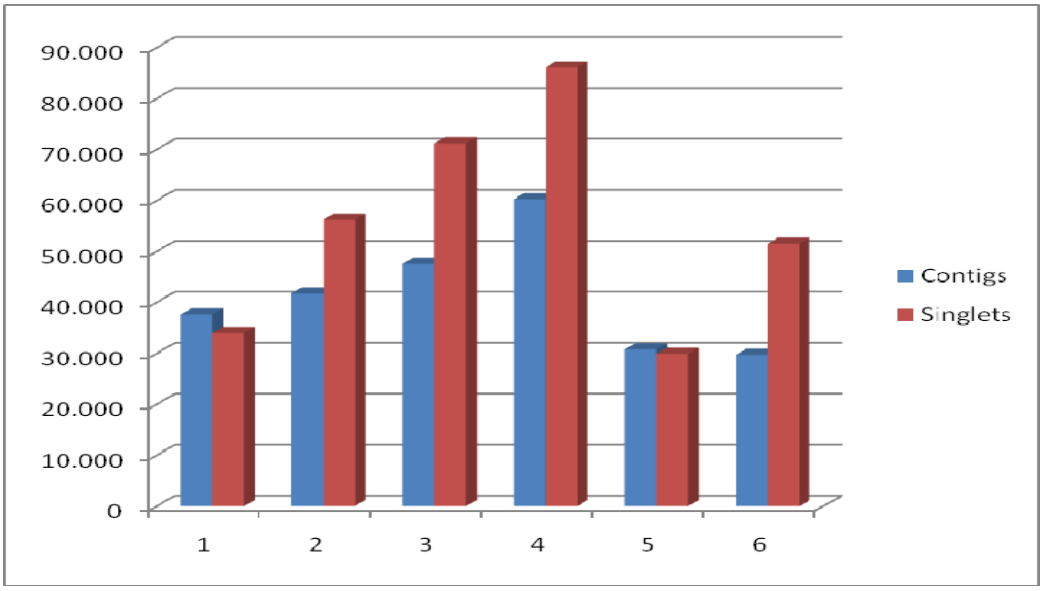


Figura 8: Número de *contigs* e *singlets* de cada montagem – O gráfico mostra que as montagens do tipo *reference assembly* (5 e 6) geram uma quantidade menor de *contigs* e *singlets* que as montagens *ab-initio* com os mesmos parâmetros (1 e 3).

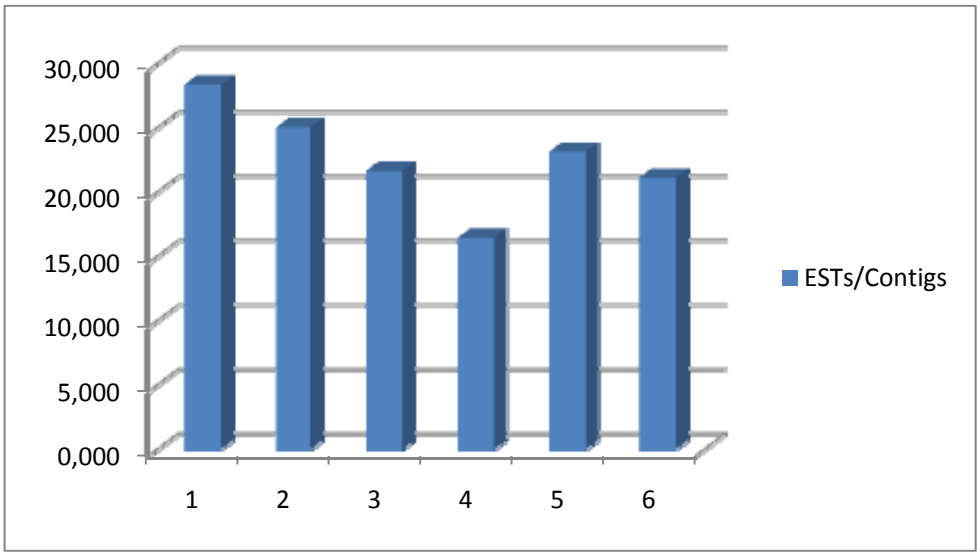


Figura 9: Número de ESTs por contig em cada montagem – Quanto mais restritivos são os parâmetros da montagem, menor o número de ESTs agrupados em um *contig*. Apesar disso, as montagens do tipo *reference assembly* (5 e 6) agrupam menos ESTs nos *contigs* que as montagens *ab-initio* com os mesmos parâmetros (1 e 3).

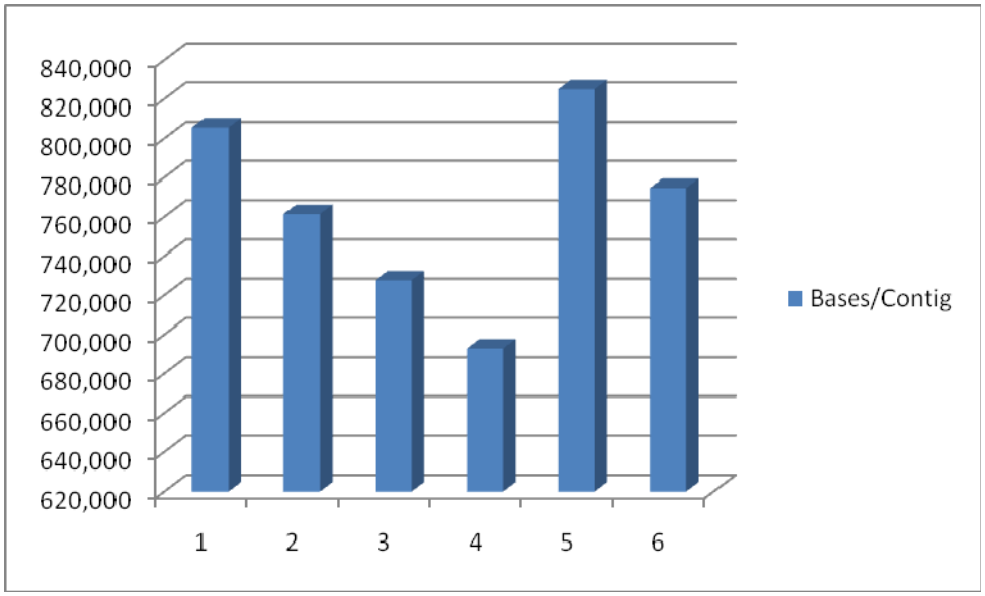


Figura 10: Número de bases por contig em cada montagem – As montagens com parâmetros mais restritivos geram *contigs* menores. As montagens do tipo *reference assembly* são as que geram maiores *contigs*.

Para avaliação da melhor montagem foram utilizadas, complementando as comparações apresentadas acima, as sequências das regiões codantes (genes preditos) do genoma. Foi realizado um BLASTN com e-value de corte 1e-10 para alinhar todos os ESTs utilizados na montagem com um único gene predito (foi considerado somente o primeiro *hit* do BLAST). Após esse processo, um script em PERL analisou cada *contig* das 6 montagens. O objetivo era encontrar qual delas gerou menos *contigs* com ESTs que alinhavam com genes diferentes (erro do tipo 2), como demonstra a Figura 11, onde o "Contig2" é formado por fragmentos que alinharam com dois genes diferentes e por isso é considerado incorreto. Na análise, foi permitido um porcentual de erro de 10% para cada *contig*, isto é, supondo que um *contig* fosse formado por 10 ESTs, era exigido que 9 deles tivessem similaridade com o mesmo gene para ser considerado correto. A Figura 12 apresenta os resultados desta análise.

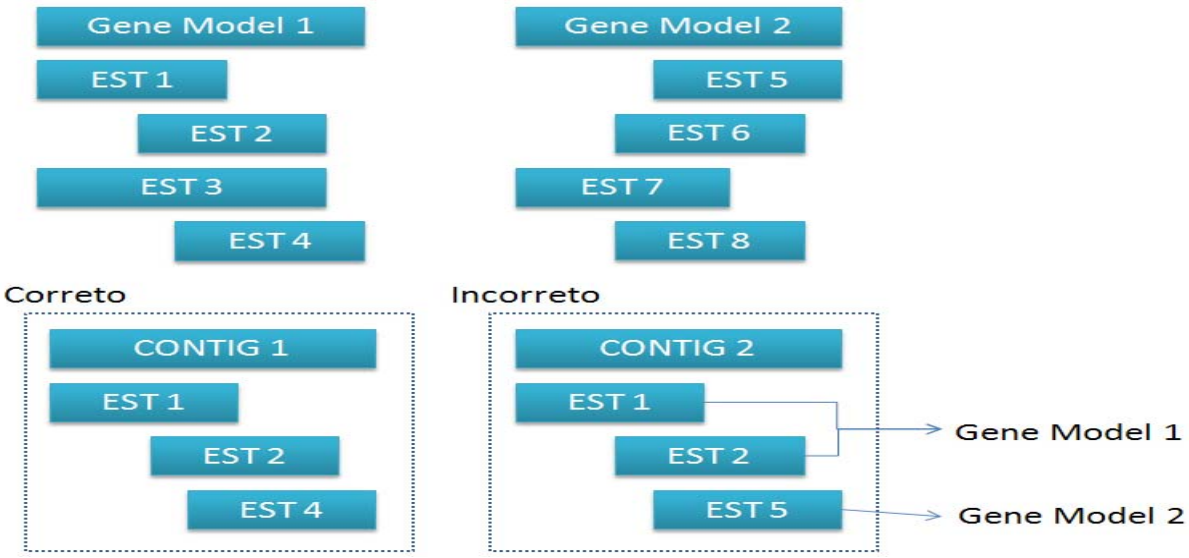


Figura 11: Identificando erros nas montagens - Alinhamento de um conjunto de 8 ESTs com 2 genes preditos do genoma. No caso, o Contig1 foi considerado correto, pois diferentemente do Contig2, é formado por ESTs que alinham somente com um único gene.

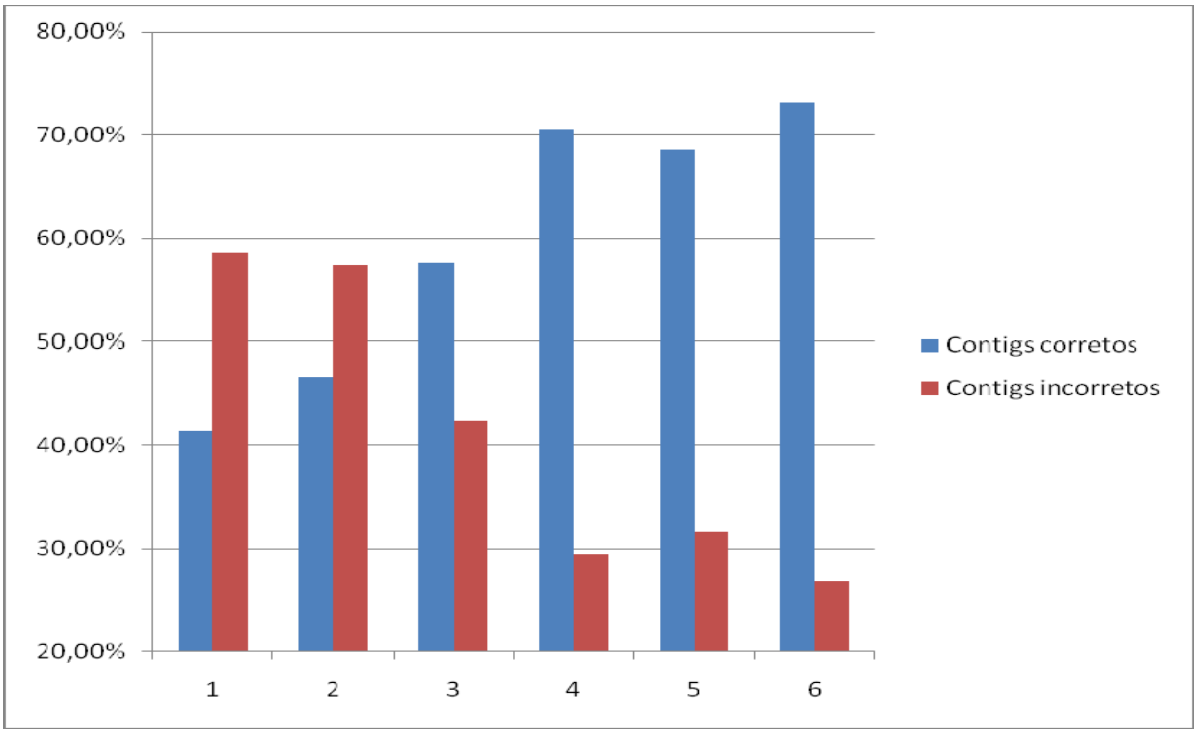


Figura 12: Montagens x Genes preditos do genoma - Resultados das comparações entre os *contigs* de cada uma das 6 montagens com os genes preditos do genoma. No caso das montagens *ab-initio*, os parâmetros menos restritivos (1 e 2) geram mais *contigs* errados do que corretos.

O gráfico apresentado na Figura 12 mostra que as montagens que apresentaram melhores resultados na comparação dos *contigs* com os genes preditos do genoma foram as de número 4, 5 e 6, onde cerca de 70% dos clusters foram considerados corretos. Considerando as outras comparações (Figuras 8, 9 e 10) a montagem considerada como correta foi a de número 5, pois dentre as três, é a que possui o maior número de ESTs e bases por *contig,* além de ter gerado um número de unigenes (60.747 – 30.809 *contigs* e 29.938 *singlets*) parecido com o número de genes preditos (66.153).

Todas as sequências resultantes da montagem e os genes preditos do genoma passaram por um processo de anotação automática utilizando o programa AutoFACT (Koski *et al,* 2005). A função do AutoFACT é, de maneira automatizada, escolher a melhor anotação para uma determinada sequência dentre diversos resultados de alinhamentos contra bancos de dados através do BLAST. Foram utilizados, para esta etapa, o NR - banco de proteínas do *NCBI* contendo informações curadas e não curadas, uniref90 e uniref100 – banco que contém somente proteínas curadas com dados de proteômica (Suzek *et al*, 2007), pfam – banco de famílias de proteínas (Bateman *et al*, 2002) e kegg – banco de vias metabólicas (Kanehisa & Goto, 2000). A Figura 13 apresenta um exemplo da anotação do AutoFACT para um dos unigenes da montagem de ESTs.

Source (Query=Contig9912)	Accession	Description
AutoFACT	Contig9912	Pkinase multi-domain protein
uniref90	UniRef90_A7QTQ6	Chromosome undetermined scaffold_171, whole genome shotgun sequence n=1 Tax=Vitis vinifera RepID=A7QTQ6_VITVI
uniref100	UniRef100_A7QTQ6	Chromosome undetermined scaffold_171, whole genome shotgun sequence n=1 Tax=Vitis vinifera RepID=A7QTQ6_VITVI
nr	CAO69321	unnamed protein product [Vitis vinifera]
kegg	osa:4326106	Os01g0114300; hypothetical protein
smart		
pfam	PF00069	pfam00069, Pkinase, Protein kinase domain

Figura 13: Anotação utilizando o AutoFACT - Resultado de anotação utilizando o AutoFACT para o Contig9912. Apesar de a anotação do NR (a base de dados mais utilizada para

anotação de sequências na literatura) retornar um resultado do tipo "*unnamed",* o Autofact usou a anotação do banco de dados Pfam para definir o contig como um domínio protéico do tipo kinase.

Os *contigs* da montagem foram utilizados para construção de uma interface web (chamada de Eletronic Northern ou Northern digital) que disponibiliza ao usuário a expressão gênica diferencial entre as diversas bibliotecas de ESTs. A interface permite buscas por determinados genes ou por bibliotecas específicas. Além disso, também é possível comparar a expressão de uma biblioteca contra um conjunto de outras. Por último, o usuário pode filtrar os resultados de uma pesquisa através da inserção de uma ou mais palavras-chave.

O porcentual de expressão gênica para cada *contig* é calculado através da razão normalizada do número de ESTs no gene e do número total de sequências utilizadas na montagem para cada uma das bibliotecas. Tal processo visa eliminar discrepâncias nos dados devido a existência de mais sequências de uma ou outra biblioteca. A Figura 14 apresenta uma das telas de resultados da interface.

Eletronic Northern - Contigs analysis

Contigs	UK1	Expression UK1	L08	Expression L08	Gene	Autofact
Contig9909	1	44.276	1	55.724	gi|255647555|gb|ACU24241.1| unknown [Glycine max]	Zinc finger (C3HC4-type R protein-like (Os01g09262(n=3 Tax=Oryza sa RepID=Q5JK23_OR

Figura 14: Expressão gênica diferencial entre bibliotecas de ESTs - Resultados da interface Eletronic Northern para o Contig9909. Através desta interface é possível visualizar o

número de ESTs utilizados de cada biblioteca para formar o cluster (colunas 2 e 4), bem como a expressão gênica normalizada de cada biblioteca no *contig* (colunas 3 e 5). As duas últimas colunas são relativas a anotação do gene, mostrando o "first hit" no banco de dados NR do *NCBI* (coluna 6) e o resultado do AutoFACT (coluna 7).

cDNAs Full-Length

Apesar da maior facilidade e do menor custo envolvido no sequenciamento de ESTs, estes não representam, na maioria dos casos, genes completos da espécie em estudo. Desta forma, as bibliotecas de cDNAs full-length são recursos extremamente úteis para a correta anotação de sequências genômicas e para a análise funcional de seus genes e produtos (Seki *et al*, 2002; Amano *et al*, 2010). A maior vantagem desta metodologia é que a grande maioria das sequências contém a totalidade dos mRNAs, incluindo a região codante (CDS) e as regiões não traduzidas (UTRs) (Umezawa *et al*, 2008).

Um total de 4.712 sequências de uma biblioteca de cDNA full-length do cultivar Nourin nº 2 (desenvolvido a partir do cruzamento de diversos outros cultivares japoneses) foram obtidas no "Soybean full-length cDNA database" (Umezava *et al*, 2008) através do link http://rsoy.psc.riken.jp/ e armazenadas no banco de dados local do GENOSOJA.

Através da comparação da biblioteca com os genes preditos do genoma e com os unigenes da montagem de ESTs utilizando o programa BLASTN, foi verificado que 100

sequências não possuíam *hits* em ambos os bancos de dados e, por isso, podem ser considerados novos genes.

Resumo dos dados públicos de soja

Os dados descritos no decorrer deste capítulo representaram um recurso importante para identificação dos genes relativos aos dados gerados pelo próprio projeto GENOSOJA, que serão descritos no próximo capítulo. A Tabela 4 apresenta um resumo de todas as sequências de soja que foram obtidas nos diversos bancos de dados públicos.

Tabela 4: Dados públicos de soja – Incluindo a montagem de ESTs, gerada a partir de dados públicos.

	Tipo de dado	Número de seqs.	Fonte
1	Genoma	20 cromossomos	JGI
2	Genes preditos	66.153 genes	JGI
3	ESTs	1.276.813 sequências	NCBI
4	cDNA full length	4.712 sequências	Soybean full-length cDNA database
5	Montagem de ESTs	60.747 unigenes	Dados gerados a partir dos ESTs (3)

CAPÍTULO 3: DADOS GERADOS PELO GENOSOJA

Este capítulo descreve todos os dados que foram gerados pelo projeto GENOSOJA, os métodos de análise e a forma como foram inseridos no banco de dados do projeto. Todos os dados são baseados na tecnologia de sequenciamento Illumina/Solexa (http://www.illumina.com).

Tags de SuperSAGE

A técnica de SAGE (*Serial Analysis of Gene Expression*), descrita por Velculescu *et al* (1995), é um experimento de quantificação da expressão dos genes baseado na contagem de sequências (*tags)* obtidas de cada transcrito de uma população de células. Tal metodologia tem sido amplamente usada para análises de expressão de genes de diversas espécies como *Arabdopsis thaliana* (Jung *et al*, 2003) e *Oryza sativa* (Song *et al*, 2007), além de doenças como a leucemia (Lee *et al*, 2006).

Apesar do largo uso de ESTs para estudos de expressão gênica, o SAGE é capaz de representar com maior precisão o nível de expressão de um transcrito na amostra (Leyritz *et al*, 2008), além de ser 10 vezes mais eficiente para a detecção de transcritos raros (Sun *et al*, 2004). O experimento consiste, basicamente, em utilizar uma enzima de restrição para cortar as regiões próximas a cauda poly-A de RNAs mensageiros isolados das células (Matsumura *et al*, 2005). Outra enzima cliva tal região em 15 pares de bases, sendo o fragmento resultante sequenciado e denominado tag, conforme mostra a Figura 15. Após o sequenciamento, as tags são agrupadas pela sua

sequência de nucleotídeos, gerando um conjunto chamado de tags únicas. O número de repetições de cada tag única é a representação do transcrito correspondente a ela na amostra. Para identificação de tags diferencialmente expressas entre duas bibliotecas, geralmente é usado o valor *p* (Audic & Claverie, 1997).

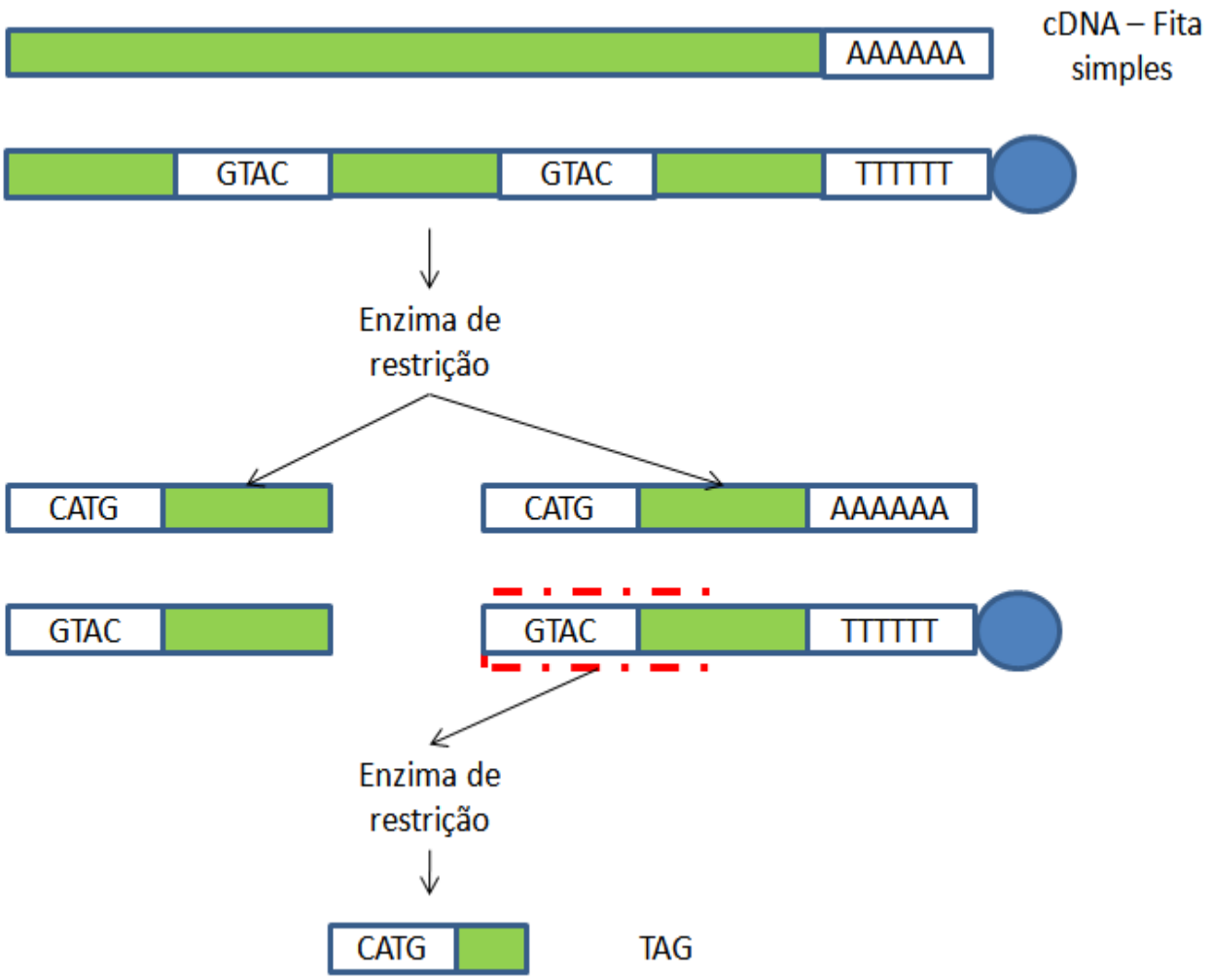


Figura 15: Experimento de SAGE – Após uma enzima de restrição "cortar" os sítios CATG de RNAs mensageiros das células, as sequências resultantes são filtradas pela presença da esfera magnética (região em azul). Uma nova enzima corta 15 bases a partir do CATG da sequência que contém a esfera magnética (próxima a cauda poly-A).

A última etapa do processo consiste na identificação dos genes relativos as tags únicas através de comparações das mesmas com bancos de dados de organismos de interesse, geralmente de ESTs. Apesar de o BLAST não retornar alinhamentos satisfatórios quando são utilizadas sequências curtas, existem diversos programas próprios para esta função, como o MAQ (Li *et al,* 2008), o Bowtie (Langmead *et al*, 2009) e o SOAP (Li *et al*, 2009a). Todos eles retornam resultados satisfatórios, porém neste trabalho foi utilizado o SOAP devido a maior facilidade de trabalho com os seus arquivos de saída.

Na última década foram criadas diversas otimizações do experimento de SAGE, como o LongSAGE (Saha *et al*, 2002) e o SuperSAGE (Matsumura *et al*, 2003). Ambos utilizam metodologia semelhante ao SAGE, porém conseguem gerar tags maiores, com 20 e 26 pares de bases respectivamente, facilitando assim a identificação de genes.

O projeto GENOSOJA gerou três amostras de SuperSAGE com tratamentos de interesse, sendo uma para análise de expressão de genes em plantas sob ação da doença da Ferrugem Asiática - causada pelo fungo *Phakopsora pachyrhizi* - (cultivar PI561356) e outras duas para plantas submetidas a estresse de seca (cultivares BR 16 – suscetível – e Embrapa 48 – resistente). A Tabela 5 apresenta o número total de tags de cada biblioteca e o número de tags únicas obtidas para cada amostra.

As tags únicas de cada uma das amostras foram alinhadas com três dos bancos de dados descritos no capítulo 2: unigenes da montagem de ESTs (30.809 *contigs* e 29.938 *singlets*), genoma (20 cromossomos) e genes preditos a partir do genoma (66.153 genes) utilizando o programa SOAP. O programa foi configurado para permitir

até 2 mismatches nos alinhamentos (alinhamentos não exatos podem se tratar de SNPs, já que as sequências são de cultivares diferentes) e para retornar todos os possíveis alinhamentos ótimos. Em casos que existiam mais de um alinhamento ótimo, foi dada preferência aos alinhamentos com os unigenes da montagem por estes conterem as regiões UTR (grande parte das tags de SuperSAGE estão nas regiões 3' UTR), seguido por alinhamentos com os genes preditos e, só em último caso, foram considerados resultados contra o genoma. As tags que não alinharam com nenhum destes bancos de dados podem ser provenientes de erros de sequenciamento, genes desconhecidos ou, no caso da amostra sob ação da doença da ferrugem asiática, ser genes do fungo *Phakopsora pachyrhizi*.

Tabela 5: Amostras de SuperSAGE do projeto GENOSOJA – Número de tags de cada biblioteca e de tags únicas das amostras.

	Ferrugem Asiática	Seca - BR 16	Seca - Embrapa 48
Biblioteca sadia	813.205	1.092.374	653.352
Biblioteca infectada	885.439	509.465	419.218
Tags únicas	104.725	89.205	74.833

Para a amostra submetida ao estresse de seca do cultivar BR 16 foram alinhadas 75.233 tags únicas (84,34%) contra os bancos de dados de soja. Destas, somente 40.330 (53,61%) foram alinhadas com bancos de dados de genes (unigenes da montagem ou genes preditos). Para a amostra do cultivar Embrapa 48 foram obtidos dados similares, com 33.322 (52,82% do total de tags alinhadas) correspondentes a

bancos de dados de genes. Esses resultados podem indicar a existência de um grande número de genes ainda desconhecidos da planta. A Tabela 6 apresenta um resumo dos alinhamentos para cada uma das amostras. Já a Figura 16 apresenta a porcentagem de tags que alinhou com cada um dos bancos de dados de soja.

Tabela 6: Alinhamentos das tags com bancos de dados de soja – A amostra de ferrugem é que tem mais tags não alinhadas (cerca de 20%), podendo indicar a existência de genes do fungo *Phakopsora pachyrhizi* na amostra.

	Ferrugem	Seca BR 16	Seca Embrapa 48
Tags alinhadas	83.337 (79,58%)	75.233 (84.34%)	63.083 (84,30%)
Tags alinhadas com genes	42.823 (51,38%)	40.330 (53,61%)	33.322 (52,82%)
Tags alinhadas com 0 mismatches	43.107 (51,73%)	47.354 (62,94%)	39.227 (62,18%)
Tags alinhadas com 1 mismatch	27.373 (32,85%)	20.962 (27,86%)	18.082 (28,67%)
Tags alinhadas com 2 mismatches	12.857 (15,42%)	6.917 (9,2%)	5.774 (9,15%)

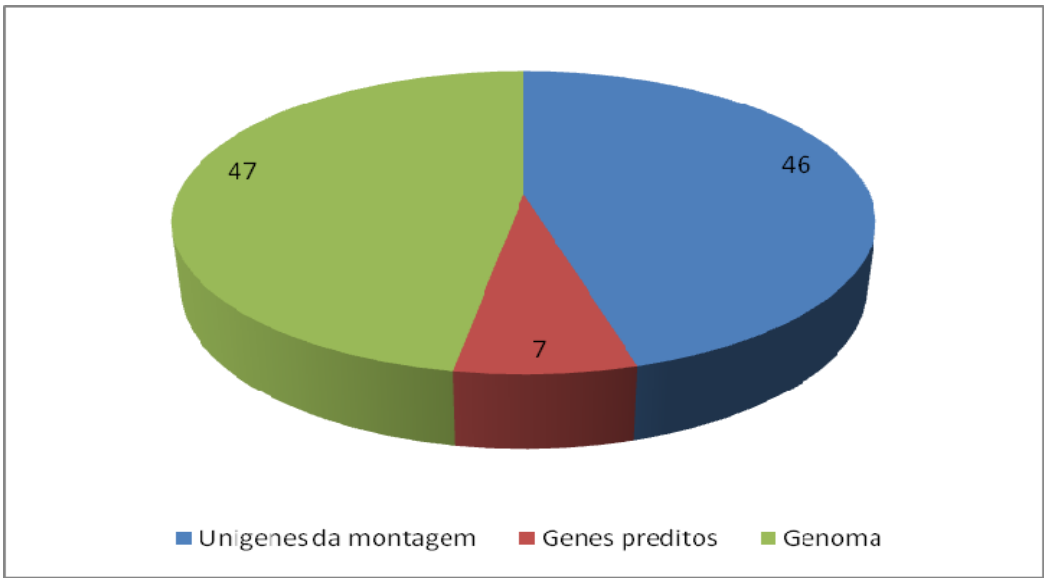


Figura 16: Bancos de dados onde as tags foram alinhadas (%) – O gráfico mostra que cerca de 50% das tags alinham somente com dados do genoma da soja. A porcentagem apresentada foi obtida com base na média dos alinhamentos das três amostras.

Para a amostra de ferrugem asiática, as 21.388 tags (20.42%) únicas que não alinharam com soja foram alinhadas com dois bancos de dados do fungo *Phakopsora pachyrhizi*: genoma (185 *scaffolds*) e 11.101 unigenes (5.950 *contigs* e *singlets*) de uma montagem de 48.567 ESTs disponíveis no *NCBI*. Esta etapa não gerou resultados satisfatórios, podendo ser em decorrência da pouca quantidade de dados disponíveis da espécie. O genoma do fungo, por exemplo, tem tamanho estimado em 600 mb e apenas 50 mb sequenciado.

Para se verificar o potencial e a especificidade da técnica de SuperSAGE foi analisado, além do número de tags alinhadas, o número de alinhamentos para cada uma das tags (Figura 17), o número total de genes obtidos e o número de tags alinhando em cada um desses genes identificados (Tabela 7).

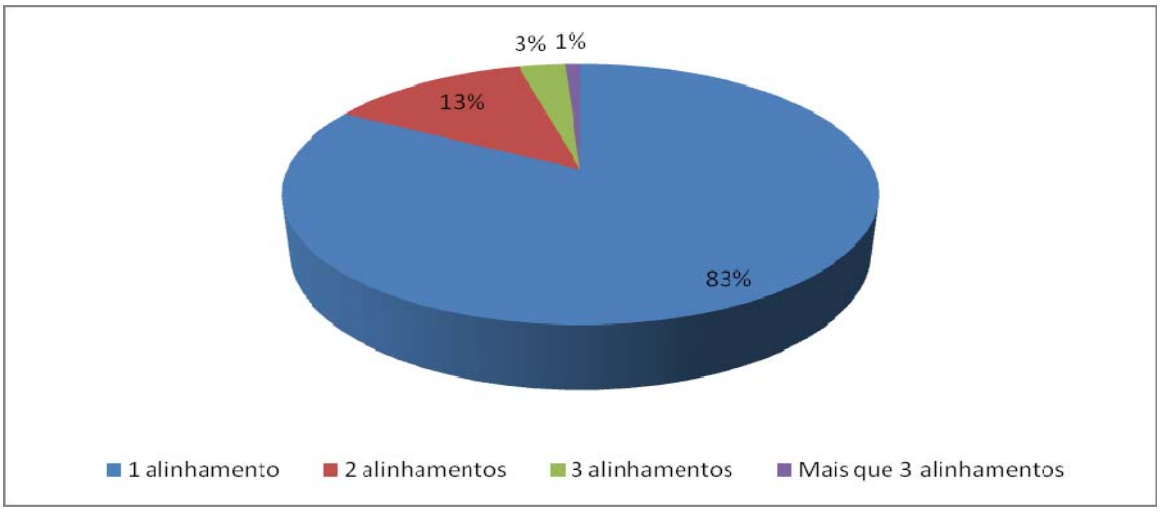


Figura 17: Número de alinhamentos por tag – As porcentagens do gráfico foram obtidas com base na média das três amostras. O grande número de tags com um único alinhamento demonstra a especificidade do experimento.

Tabela 7: Número de genes identificados nas amostras - Em todas as amostras cerca de 60% dos genes identificados são correspondentes a uma única tag.

	Ferrugem	Seca BR 16	Seca Embrapa 48
Número de genes identificados	19.572	21.763	17.821
Genes correspondentes a 1 tag	11.209 (57,27%)	13.300 (61,11%)	11.101 (62,29%)
Genes correspondentes a 2 tags	3.523 (18%)	3.129 (14,37%)	3.266 (18,32%)
Genes correspondentes a 3 tags	1.602 (8,19%)	2.208 (10,14%)	765 (4,29%)
Genes correspondentes a mais de 3 tags	3.238 (16,54%)	3.126 (14,36%)	2.689 (15,08%)

Tendo os genes correspondentes a cada uma das tags pode-se identificar quais deles estão diferencialmente expressos nas condições de cada um dos tratamentos. Utiliza-se o número de repetições da tag relativa a um determinado gene em ambas as bibliotecas (sadia ou controle e infectada ou tratada) para se calcular o valor *p*. Caso este valor seja menor ou igual a 0,05 o gene pode ser considerado *down-regulated* (mais expresso na biblioteca sadia do que na infectada) ou *up-regulated* (mais expresso na biblioteca infectada do que na sadia). Estes genes são interessantes para estudos de potencial biotecnológico, pois podem estar relacionados a resposta da planta ao estresse estudado. O número de tags diferenciamente expressas em cada amostra é apresentado na Tabela 8.

As tags diferenciamente expressas de cada uma das amostras podem ser visualizadas pelos usuários do projeto através de uma interface *web* que resume todos os resultados das análises de SuperSAGE. A interface apresenta o número de repetições das tags em cada uma das bibliotecas e os genes reletivos a cada uma delas, permitindo a inserção de filtros por mismatches e valor *p*, além de buscas por palavras-chave de interesse na anotação dos genes correspondentes as tags. A Figura

18 apresenta um resumo da tela de resultados desta interface para a amostra de ferrugem asiática.

Tabela 8: Tags diferencialmente expressas nas amostras – Para a amostra do cultivar Embrapa 48 (suscetível) foram identificados cerca de 18.000 tags *up-regulated,* mostrando diferenças entre o mecanismo de resposta da planta ao estresse em relação ao cultivar BR16 (resistente).

	Ferrugem	Seca BR16	Seca Embrapa 48
Tags *down-regulated*	7.430	4.901	6.976
Tags *up-regulated*	8.331	9.549	18.388

	Tag	Count Lib. 1	Count Lib. 2	Fold-change	P-value	Mapped on	Position of the alignment	Direction	Mismatches
	tag20014	3807	626	-6.62166416113201	0	EST - Contig5799 See others	635	-	0
	tag29846	17310	11402	-1.65300679164799	0	EST - SJ07-E1-S10-174-H10-UC.F See others	386	+	0
	tag11488	6107	2143	-3.10287553873029	0	Gene Models - Glyma13g07610.1	369	+	0
	tag62374	9822	4306	-2.48361635480317	0	Genoma - Gm11	18750733	+	0
	tag22606	3267	384	-9.26353009657774	0	Genoma - Gm17	40197888	-	0
	tag29503	3555	10172	2.6278957933267	0	Genoma - Gm08	13906273	+	0

Figura 18: Interface de visualização das amostras de SuperSAGE - A primeira coluna representa se a tag é *down-regulated* (verde) ou *up-regulated* (vermelho). A tela traz outras

informações sobre a tag, como o número de repetições em cada uma das bibliotecas (colunas 3 e 4), o valor *p* (p-value) calculado a partir do número de repetições da tag nas bibliotecas (coluna 6), o gene correspondente a tag (coluna 7), a posição onde a tag foi encontrada no gene (coluna 8) e o número de mismatches do alinhamento (coluna 10).

Bibliotecas subtrativas de cDNA

Entre as várias estratégias utilizadas para o estudo de dados de transcriptoma, o sequenciamento de bibliotecas subtrativas de cDNA é um recurso extremamente útil para identificação de genes mais expressos em uma determinada condição do que em outra. O experimento consiste em misturar na mesma reação sequências relativas a biblioteca de controle (*driver*) e sequências relativas a biblioteca de estudo ou tratada (*tester*), além de *primers* específicos para a segunda condição. Com isso, através da reação de PCR, os transcritos relativos a biblioteca tratada terão amplificação exponencial. Alguns transcritos relativos a condição de controle serão amplificados de maneira linear por terem similaridade com transcritos da biblioteca tratada (Wieland *et al*, 1990). Dessa forma, genes identificados no experimento estão muito mais expressos na biblioteca tratada e por isso, podem estar relacionados a resposta do organismo ao estresse estudado.

Para o projeto GENOSOJA, foram construídas 22 bibliotecas subtrativas de cDNA de diferentes cultivares, utilizando vários tratamentos com diferentes tempos de submissão da planta ao estresse. As sequências são da tecnologia Illumina/Solexa com 36 ou 76 bp, dependendo da biblioteca. A Tabela 9 apresenta a descrição e o número de fragmentos de cada uma delas.

Tabela 9: Bibliotecas subtrativas do projeto GENOSOJA – Descrição das bibliotecas, incluindo: cultivar/genótipo, estresse, tempo de estresse e número de sequências geradas.

	Biblioteca	Tempo	Cultivar/Genótipo	Sequências	Tamanho
1	Seca - Folha	25-50 min após o estresse	BR16 - resist.	1.854.641	36 bp
2	Seca - Folha	75-100 min após o estresse	BR16 - resist.	519.031	36 bp
3	Seca - Folha	125-150 min após o estresse	BR16 - resist.	2.035.320	36 bp
4	Seca - Raiz	25-50 min após o estresse	BR16 - resist.	2.486.569	36 bp
5	Seca - Raiz	75-100 min após o estresse	BR16 - resist.	2.458.847	36 bp
6	Seca - Raiz	125-150 min após o estresse	BR16 - resist.	2.428.923	36 bp
7	Ferrugem	12, 24 e 48 horas após a infecção	PI61356 - resist.	5.185.015	76 bp
8	Ferrugem	72 e 96 horas após a infecção	PI61356 - resist.	5.000.616	76 bp
9	Ferrugem	192 horas após a infecção	PI61356 - resist.	4.700.869	76 bp
10	Vírus	5 e 13 dias após a infecção	CD206 - resist.	5.963.145	76 bp
11	Vírus	5 e 13 dias após a infecção	BRSGO - susc.	5.345.985	76 bp
12	Nitrogênio	-	MG/BR46	4.621.072	76 bp
13	Nitrogênio	-	MG/BR46	5.343.969	76 bp
14	Seca - Folha	25-50 min após o estresse	Embrapa48 - susc.	5.144.645	76 bp
15	Seca - Folha	75-100 min após o estresse	Embrapa48 - susc.	5.644.473	76 bp
16	Seca - Folha	125-150 min após o estresse	Embrapa48 - susc.	5.359.395	76 bp
17	Seca - Raiz	25-50 min após o estresse	Embrapa48 - susc.	3.095.694	76 bp
18	Seca - Raiz	75-100 min após o estresse	Embrapa48 - susc.	5.731.156	76 bp
19	Seca - Raiz	125-150 min após o estresse	Embrapa48 - susc.	5.545.375	76 bp
20	Ferrugem	1 e 6 horas após a infecção	PI230970 - resist.	4.679.963	76 bp
21	Ferrugem	12 e 24 horas após a infecção	PI230970 - resist.	4.878.530	76 bp
22	Ferrugem	48 e 72 horas após a infecção	PI230970 - resist.	4.355.862	76 bp

Existem formas diferentes de se trabalhar com sequências curtas de transcritos (RNA-seq) descritas na literatura. Uma delas é o alinhamento das sequências contra uma referência (Yassour *et al*, 2009; Hashimoto *et al*, 2009), que pode ser o genoma ou o conjunto de genes da espécie. Usando esta metodologia, é possível obter a posição dos fragmentos no genoma e a expressão dos possíveis transcritos através de programas apropriados, como o Cufflinks (Trapnell *et al*, 2010). Por outro lado, pode-se utilizar o método *ab-initio*, isto é, montar as sequências sem utilizar uma referência (Birol *et al*, 2009). Existem diversos programas específicos para montagem de

sequências curtas, como o Velvet (Zerbino & Birney, 2008), o Edena (Hernandez *et al,* 2008) e o ABySS (Simpson *et al*, 2009). Apesar disso, todos eles foram desenvolvidos para montagem de sequências genômicas e são baseados no parâmetro *k* (*k-mer*), cuja otimização varia de acordo com a cobertura do genoma estudado. No caso de cDNA - diferentemente de genomas onde a cobertura é uniforme, variando somente em regiões repetitivas - a cobertura de um gene varia de acordo com o nível de expressão do mesmo. Neste caso, altos *k-mers* são apropriados para montagem de genes altamente expressos e baixos *k-mers* para a montagem de genes pouco expressos, de modo que o melhor resultado é obtido através do uso de vários *k-mers* (Nascimento *et al*, 2009). Neste trabalho, foram utilizadas ambos os métodos, visando identificar tanto os genes já conhecidos quanto genes novos nas bibliotecas. Para as montagens *ab-initio*, o Edena produziu melhores resultados em testes com dados de transcritos e, por isso, foi escolhido para ser utilizado neste trabalho.

As sequências de cada uma das 22 bibliotecas foram alinhadas, inicialmente, com os unigenes da montagem de ESTs utilizando o programa SOAP configurado para permitir até 2 mismatches (assim como no caso de SuperSAGE as sequências são de cultivares diferentes e alinhamentos não exatos podem ser SNPs), descartar fragmentos que continham bases não identificadas durante o sequenciamento (“N”s) e para retornar todos os alinhamentos ótimos. Os fragmentos que não alinharam com a montagem de ESTs foram alinhados com os genes preditos do genoma utilizando os mesmos parâmetros. As sequências relativas a cada gene (unigene ou gene predito) foram montadas utilizando o Edena com o *k-mer* ótimo e descartando *contigs* menores que 100 bp. Os *contigs* resultantes são específicos do cultivar de cada uma das

bibliotecas e podem ser úteis para desenhos de primers, por exemplo. A Tabela 10 apresenta um resumo dos resultados dos alinhamentos de cada biblioteca, mostrando o número de genes e de *contigs* encontrados em cada uma delas.

Tabela 10 – Número de genes encontrados nas bibliotecas subtrativas – Inclui também o número de contigs obtidos com o Edena para cada uma das bibliotecas.

	Biblioteca	Cultivar/Genótipo	Genes	Contigs
1	Seca - Folha	BR16 - resist.	1.560	2187
2	Seca - Folha	BR16 - resist.	2.009	3769
3	Seca - Folha	BR16 - resist.	3.124	4922
4	Seca - Raiz	BR16 - resist.	258	358
5	Seca - Raiz	BR16 - resist.	600	810
6	Seca - Raiz	BR16 - resist.	657	1288
7	Ferrugem	PI61356 - resist.	1.994	3.857
8	Ferrugem	PI61356 - resist.	802	1.328
9	Ferrugem	PI61356 - resist.	754	990
10	Vírus	CD206 - resist.	1.109	1.907
11	Vírus	BRSGO - susc.	862	1.796
12	Nitrogênio	MG/BR46	4.775	8.998
13	Nitrogênio	MG/BR46	5.989	11.020
14	Seca - Folha	Embrapa48 - susc.	3.643	6.899
15	Seca - Folha	Embrapa48 - susc.	4.603	9.786
16	Seca - Folha	Embrapa48 - susc.	3.109	7.048
17	Seca - Raiz	Embrapa48 - susc.	1.313	1.799
18	Seca - Raiz	Embrapa48 - susc.	1.364	3.104
19	Seca - Raiz	Embrapa48 - susc.	1.775	3.767
20	Ferrugem	PI230970 - resist.	490	1.242
21	Ferrugem	PI230970 - resist.	447	1.606
22	Ferrugem	PI230970 - resist.	2.097	4.616

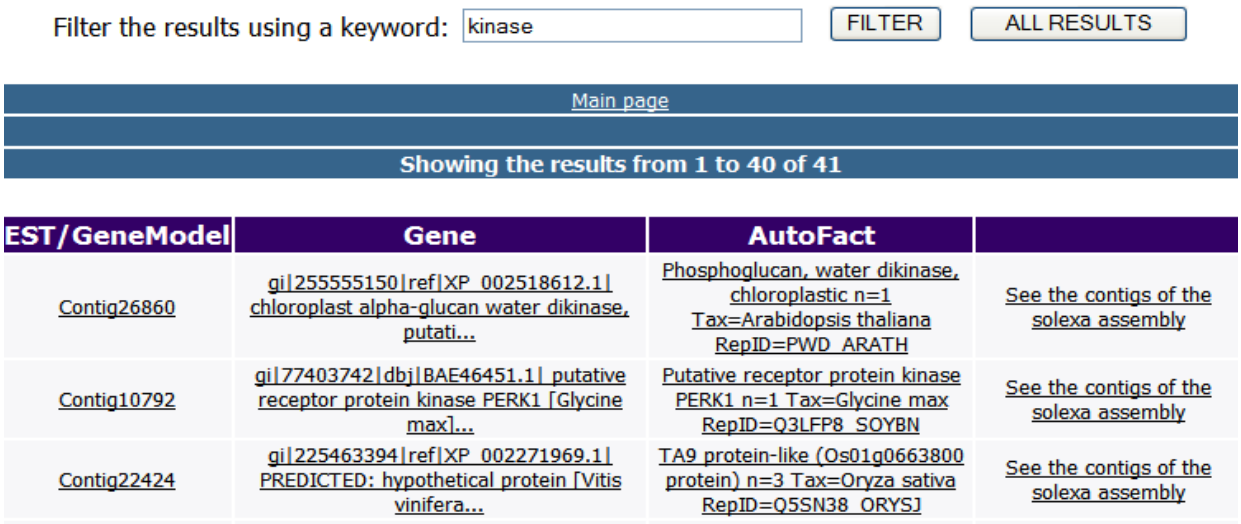
O cálculo de expressão dos genes nas bibliotecas foi feito utilizando o valor *RPKM* (Reads Por Kilobase por Milhão de reads) (Mortazavi *et al*, 2008) do programa Cufflinks. O valor *RPKM* é baseado no número de fragmentos que alinharam com um gene. Esse número é normalizado pelo tamanho do gene (utiliza o valor proporcional a

1Kb) e pelo número total de fragmentos na biblioteca (utiliza o valor proporcional a 1 milhão de fragmentos). Além disso, sequências que alinham com mais de uma referência tem menor peso no cálculo final. Para utilizar o resultado do SOAP como *input* do Cufflinks foi preciso convertê-lo para o formato SAM (Sequence Alignment/Map Format) (Li *et al*, 2009b) utilizando scripts disponíveis no pacote SAMtools (http://samtools.sourceforge.net/) (Li *et al,* 2009b). Este cálculo facilita a escolha de genes com potencial biotecnológico, já que, altos *RPKMs* significam alta expressão do transcrito na condição de estudo (*tester*).

Foi construída uma interface web que apresenta, para todas as bibliotecas, todos os genes que foram identificados na mesma, bem como a anotação (banco de dados NR e do AutoFACT) relativa a cada um dos genes. Através de um link, também é possível obter os *contigs* de solexa relativos a cada um deles. Por último, existe a possibilidade de o usuário filtrar a busca pelo nome do gene ou através de inserção de uma ou mais palavras-chave para consulta na anotação dos genes. A Figura 19 apresenta uma tela de resultados desta interface para uma das bibliotecas subtrativas submetidas ao estresse de seca do cultivar BR16 (biblioteca 5 da Tabela 9) após ser utilizada como filtro a palavra-chave “kinase”.

As montagens *ab-initio* foram realizadas utilizando todas as bibliotecas referentes ao mesmo cultivar, independente do tratamento. Desta forma, foram obtidos 5 conjuntos de dados relacionados aos seguintes cultivares: BR16 (bibliotecas 1 a 6), PI61356 (bibliotecas 7 a 9), MG/BR46 (bibliotecas 12 e 13) Embrapa48 (bibliotecas 14 a 19) e PI230970 (bibliotecas 20 a 22). Para cada cultivar foi utilizado o montador

Edena para gerar 3 montagens com *k-mers* diferentes (baixo, médio e alto) eliminando *contigs* menores que 100 bp. As 3 montagens do Edena foram fundidas em uma única montagem final através do programa Minimus (Sommer *et al*, 2007). Tal abordagem teve dois objetivos: (*i*) encontrar genes específicos dos cultivares brasileiros que, possivelmente, não estariam mapeados no genoma do cultivar *Williams 82;* (*ii*) encontrar genes que estão presentes no genoma, mas que ainda não haviam sido identificados, tanto nos unigenes da montagem de ESTs, quanto nos genes preditos. Para isto, os *contigs* de cada montagem foram comparados com os genes preditos e os unigenes da montagem de ESTs utilizando o BLASTN com e-value de corte 1e-10. As sequências "No hits" foram comparadas com o genoma com o mesmo programa e parâmetros. A Figura 20 apresenta um esquema resumido da metodologia empregada nesta etapa.


Filter the results using a keyword: kinase FILTER ALL RESULTS

Main page

Showing the results from 1 to 40 of 41

EST/GeneModel	Gene	AutoFact	
Contig26860	gi|255555150|ref|XP_002518612.1| chloroplast alpha-glucan water dikinase, putati...	Phosphoglucan, water dikinase, chloroplastic n=1 Tax=Arabidopsis thaliana RepID=PWD_ARATH	See the contigs of the solexa assembly
Contig10792	gi|77403742|dbj|BAE46451.1| putative receptor protein kinase PERK1 [Glycine max]...	Putative receptor protein kinase PERK1 n=1 Tax=Glycine max RepID=Q3LFP8_SOYBN	See the contigs of the solexa assembly
Contig22424	gi|225463394|ref|XP_002271969.1| PREDICTED: hypothetical protein [Vitis vinifera...	TA9 protein-like (Os01g0663800 protein) n=3 Tax=Oryza sativa RepID=Q5SN38_ORYSJ	See the contigs of the solexa assembly

Figura 19: Interface para visualização dos dados de bibliotecas subtrativas - A primeira coluna mostra o nome do gene identificado na biblioteca, enquanto a segunda e a terceira apresentam a anotação do NR e do AutoFACT para o mesmo.

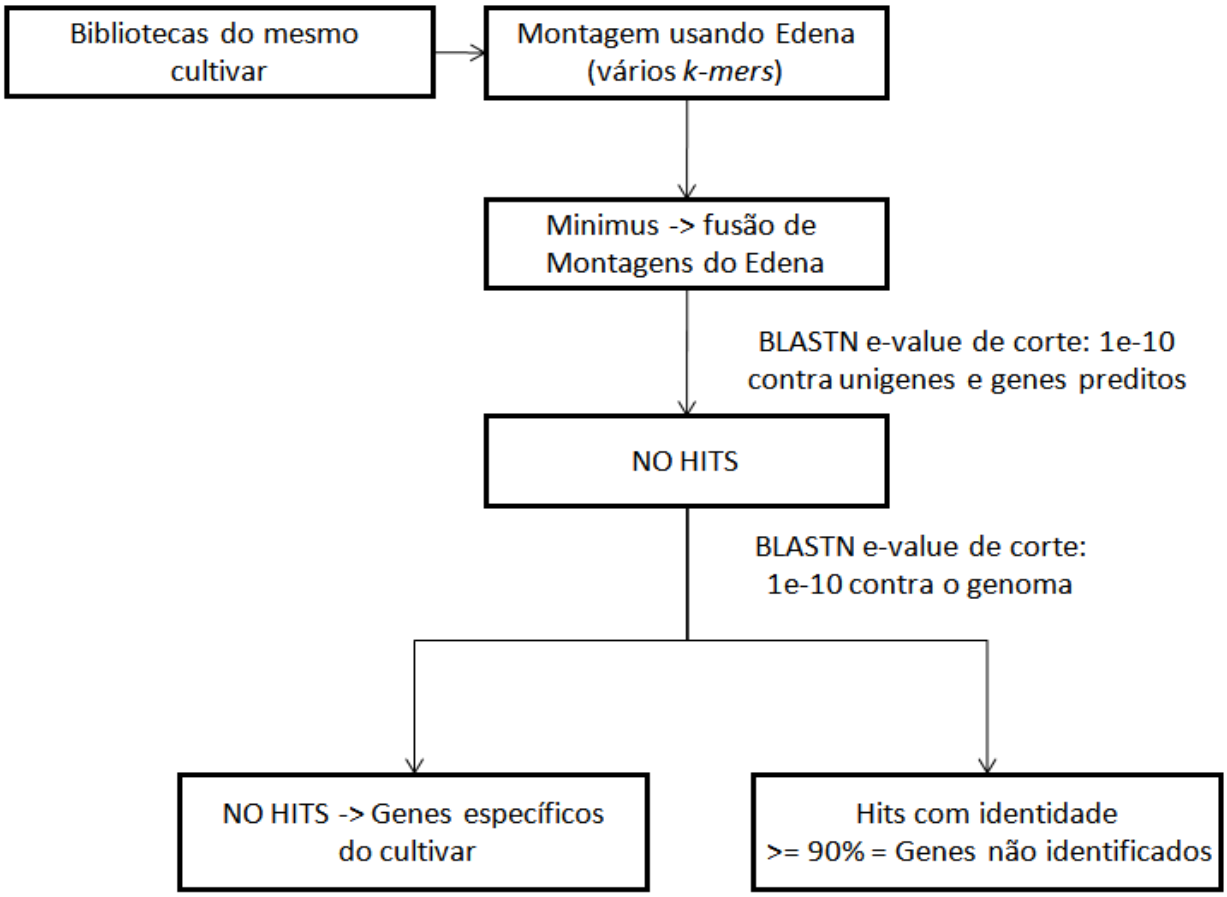


Figura 20: Pipeline da análise *ab-initio* das bibliotecas subtrativas – A abordagem *ab-initio* visava identificar genes específicos dos cultivares brasileiros (sequências sem alinhamento com o genoma do cultivar Williams 82) e genes ainda não identificados, ou seja, que alinhavam com o genoma em regiões que não houvessem outros genes correspondentes.

No caso do cutlivar BR 16, por exemplo, foram encontrados *contigs* com alinhamento no genoma, mas sem genes correspondentes nos bancos de dados utilizados. Além disso, os genes mais próximos estão a mais de 3 Kb de distância, eliminando a possibiblidade de se tratarem de UTRs. Tais *contigs* precisam passar por uma análise experimental para confirmação *in vivo* dos dados obtidos *in silico*. A Figura

21 apresenta uma visualização de 3 kb do cromossomo 11, onde um desses contigs está alinhado. Além de não existir nenhum gene identificado nesta região, é possível observar que somente existem sequências de solexa relativas ao *contig* para a biblioteca com o estágio inicial do estresse de seca (25-50 minutos).

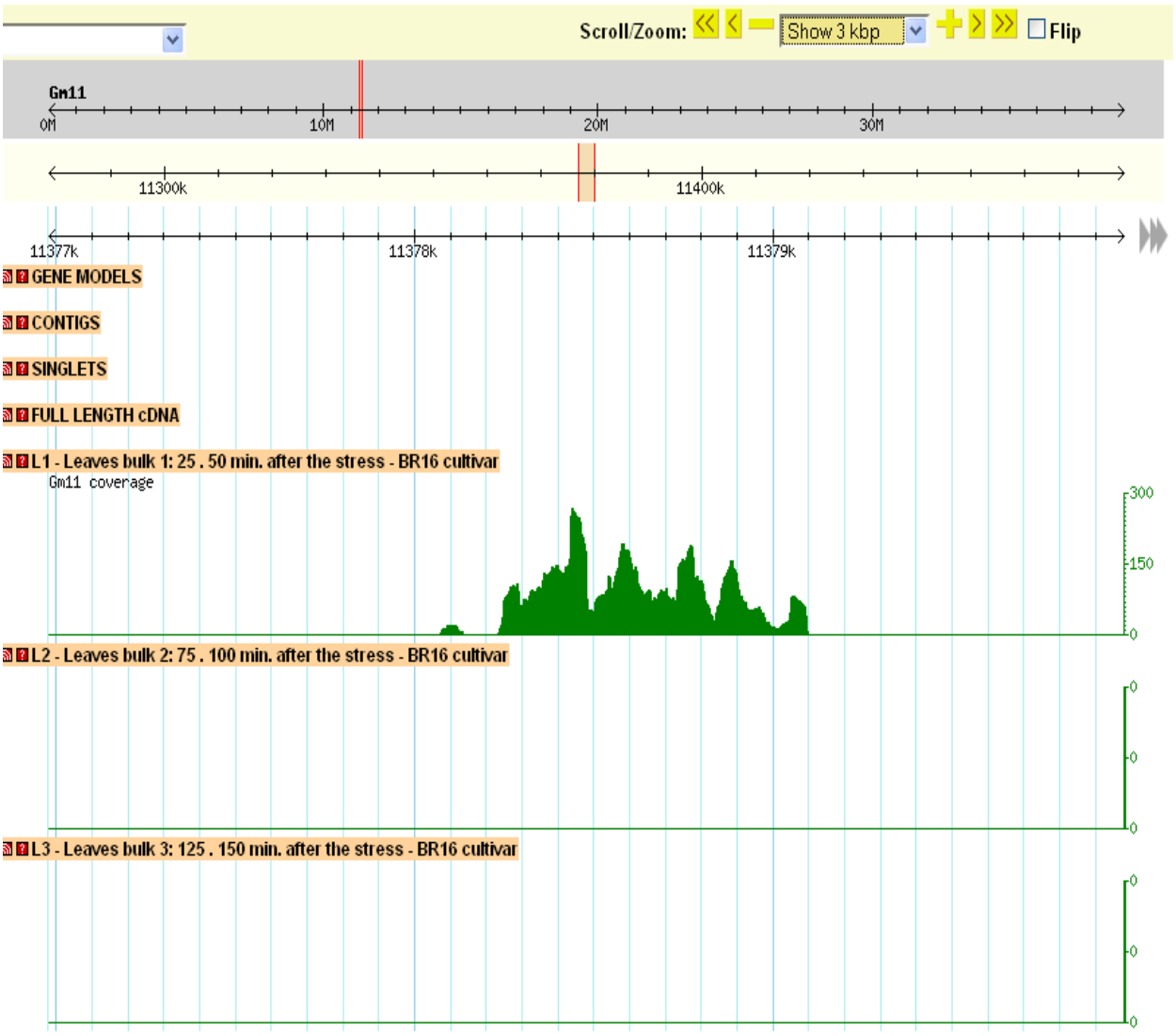


Figura 21: Visualização da região de um possível gene – A área em verde representa a existência de sequências que alinham com aquela região do genoma. O número apresentado a direita é exatamente o número de reads de solexa que possuem similaridade com a região.

MicroRNAs

MicroRNAs são pequenas moléculas (entre 19 e 25 bp) de RNA de fita simples, não-codificantes, que agem como reguladores da expressão ou como inibidores da tradução de genes em diversos organismos, incluindo plantas e animais. Os miRNAs agem através da ligação com a região 3' não traduzida (UTR) do RNA mensageiro alvo formando uma estrutura abaulada (conhecida como hairpin) (Lagos-Quintana *et al*, 2003). O alinhamento do miRNA com o gene alvo pode ser exato, inibindo a tradução do mesmo, ou inexato (com gaps e mismatches), interferindo na expressão do RNA mensageiro. Neste último caso, exatamente pelo pareamento não precisar ser exato, um único miRNA pode regular a expressão de múltiplos alvos, bem como a expressão de um gene pode ser regulada por múltiplos microRNAs. Por outro lado, a expressão de um determinado miRNA pode variar entre diferentes tecidos ou condições biológicas (Mazière and Enright, 2007), ocasionando a repressão ou expressão do gene entre diferentes estágios de estresse.

Apesar de os miRNAs não terem suas funções totalmente esclarecidas, muitas delas tem sido identificadas recentemente. Dentre elas, pode-se citar: o controle do desenvolvimento floral em plantas (Aukerman and Sakai, 2003; Palatnik *et al*, 2003), a progressão de cânceres em humanos (Volinia *et al*, 2006), a diferenciação da morfogênese dentária (Michon *et al*, 2010), entre outras. Com esses estudos, a quantidade de miRNAs conhecidos tem aumentado consideravelmente. O banco de dados miRBase (Griffith-Jones, 2006) foi criado com o objetivo de reunir esses

microRNAs conhecidos de maneira organizada e de fácil consulta. Em julho de 2010, haviam 216 miRNAs de soja identificados no miRBase.

A identificação de microRNAs é baseada, inicialmente, na utilização de programas de alinhamento configurados com parâmetros específicos, dependendo da espécie estudada. No caso de plantas, diferentemente de animais, os alinhamentos tendem a ser mais conservados, ou seja, com menos gaps e mismatches (Lagos-Quintana *et al*, 2003). Após a identificação exata da região do genoma onde o miRNA se encontra, é necessária uma análise manual das condições de entalpia e de estrutura secundária (hairpin) desta região para validar a sequência. Por fim, a identificação dos alvos geralmente é feita computacionalmente – *in silico* - sendo baseada na comparação do microRNA validado com o conjunto de genes da espécie estudada.

O projeto GENOSOJA utilizou a tecnologia Illumina/Solexa para gerar oito bibliotecas de sequências de RNAs pequenos de soja. Foram utilizadas plantas submetidas ao estresse de seca (cultivares suscetível e resistente) e plantas infectadas com a doença da ferrugem asiática (cultivares suscetível e resistente). Para cada biblioteca, foram obtidas sequências de diferentes tamanhos, variando entre 19 e 24 bp. O número de sequências (reads) obtidas para cada biblioteca é apresentado na Tabela 11.

Para eliminar a redundância e facilitar a análise dos dados, os reads foram agrupados em sequências únicas, sendo a contagem do número de sequências únicas realizada separadamente para cada uma das bibliotecas. As sequências com contagem menor ou igual a dois foram excluídas da lista final. Para análise de

expressão diferencial entre as bibliotecas, foi aplicado um pipeline estatístico baseado no software DEGseq (Wang *et al*, 2010), considerando uma taxa de confiança de 95% (corte em 0,05). A tabela 12 apresenta o número de sequências únicas e diferenciais para cada uma das bibliotecas.

Tabela 11: Número de sequências das bibliotecas de RNAs pequenos

			Tamanho das sequências					
			19 bp	20 bp	21 bp	22 bp	23 bp	24 bp
Soja submetida a seca	Tolerante	Controle	327.448	271.772	531.595	357.980	203.722	208.377
		Tratado	71.011	72.628	154.808	87.326	77.045	177.626
	Suscetível	Controle	89.040	91.816	215.419	128.524	142.446	200.087
		Tratado	266.165	220.714	353.003	244.641	138.051	250.213
Soja infectada com ferrugem asisática	Tolerante	Controle	91.908	205.404	1.177.303	394.378	175.063	285.064
		Tratado	100.045	155.849	779.788	426.383	187.926	859.624
	Suscetível	Controle	115.824	236.750	921.964	340.129	167.306	540.949
		Tratado	123.423	190.799	962.676	363.983	86.230	176.753

Tabela 12: Número de sequências únicas e diferenciais das bibliotecas

			Tamanho das sequências					
			19 bp	20 bp	21 bp	22 bp	23 bp	24 bp
Soja submetida a seca	Tolerante	Únicas	725	665	448	522	448	231
		Diferenciais	79,3%	77,3%	78,1%	76,6%	76,8%	89,2%
	Suscetível	Únicas	719	652	448	516	442	170
		Diferenciais	75,5%	75,6%	88,0%	81,8%	82,1%	95,3%
Soja infectada com ferrugem asisática	Tolerante	Únicas	588	524	427	456	386	208
		Diferenciais	54,3%	60,9%	72,1%	55,5%	46,4%	57,7%
	Suscetível	Únicas	590	537	435	461	373	220
		Diferenciais	63,4%	57,9%	76,3%	69,0%	67,8%	73,2%

As sequências únicas de cada uma das bibliotecas foram alinhadas com o genoma da soja utilizando o programa SOAP, configurado para retornar todos os

alinhamentos ótimos e não permitir mismatches. A partir desses alinhamentos, foram utilizados scripts em PERL para extrair 300 bp flanqueando a região genômica a partir da posição do alinhamento do read. As sequências resultantes desta etapa foram alinhadas com o complementar reverso do read original usando o algoritmo de Smith-Waterman (Smith and Waterman, 1981), permitindo 2 gaps e 4 mismatches. As sequências resultantes foram consideradas microRNAs candidatos, aos quais foram curadas manualmente pelo grupo da UFRGS (conveniado ao GENOSOJA e gerenciado pelo professor Rogério Margis), gerando uma lista final de 269 miRNAs, sendo que 42 destes são novos, ou seja, ainda não estavam catalogados no miRBase.

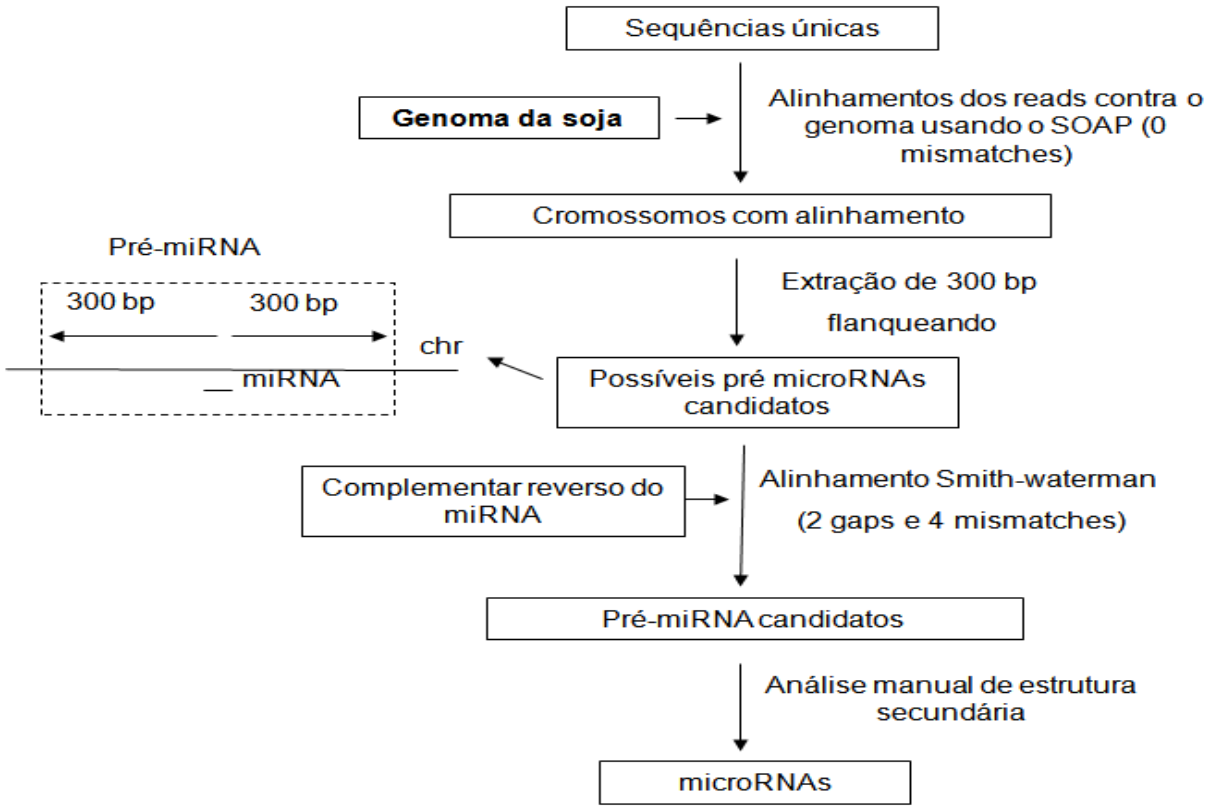


Figura 22: Pipeline utilizado na identificação de microRNAs – Incluindo a parte de análise de estrutura secundária, feita manualmente pelo grupo da UFRGS.

Para identificação dos genes alvo a lista final de miRNAs foi alinhada com os unigenes da montagem de ESTs com o algoritmo de Smith-Waterman configurado para permitir 3 mismatches, sem gaps e retornar todos os alinhamentos. Somente foram considerados alinhamentos na fita anti-sense (3'-5'). A direção 3'-5' dos clusters de ESTs foi identificada através da anotação dos mesmos com bancos de dados de proteínas. Para 169 dos 269 miRNAs identificados foi encontrado um ou mais genes alvo. A figura 23 apresenta o número de genes alvo por microRNA.

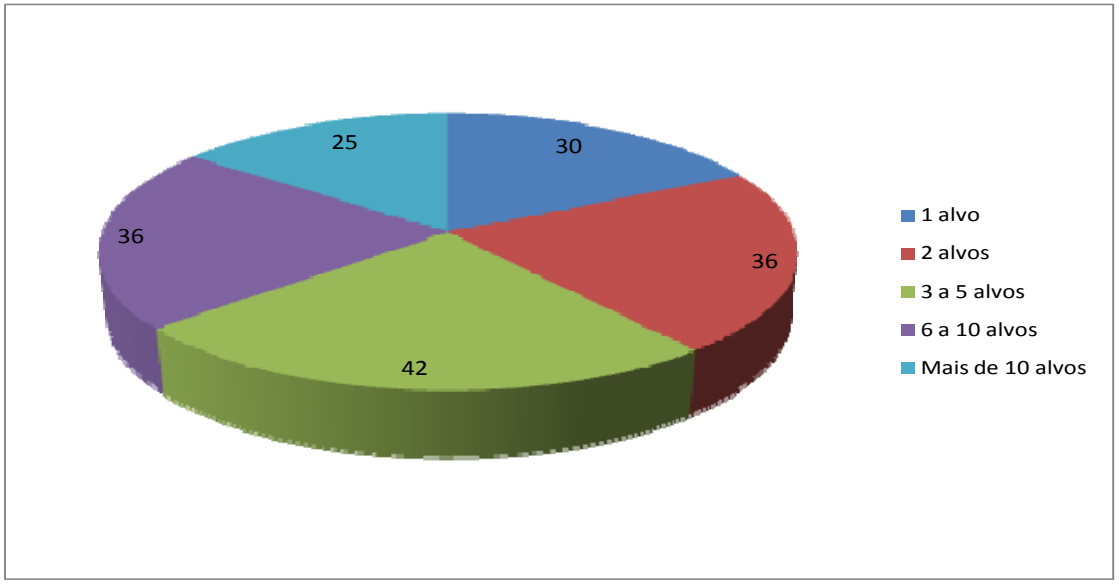


Figura 23: Número de genes alvo por MicroRNA.

Resumo dos dados gerados pelo projeto GENOSOJA

As diversas fontes de dados utilizadas no projeto GENOSOJA – apresentadas no decorrer deste capítulo - tiveram como objetivo a obtenção de um perfil de expressão de genes da soja sobre diferentes tipos de estresse, todos eles relacionados ao cultivo

da soja no Brasil. A tabela 13 apresenta um resumo de todo o conjunto de dados gerados pelo projeto.

Neste capítulo, também foram apresentadas as metodologias utilizadas na análise de cada um desses dados. Para todos os casos, os dados públicos de soja – descritos no capítulo 2 deste trabalho – foram utilizados como referencial, facilitando assim a integração dos dados das diversas fontes, etapa esta que será descrita no próximo capítulo.

Tabela 13: Resumo dos dados gerados pelo GENOSOJA

	Tipo de dado	Número de seqs.	Referencial utilizado na análise
1	SuperSAGE	3 amostras com cerca de 90.000 tags cada	Unigenes da montagem, genoma e genes preditos
2	Bibliotecas sibtrativas de cDNA	22 bibliotecas com cerca de 4 milhões de sequências cada	Unigenes da montagem e genes preditos
3	MicroRNAs	269	Unigenes da montagem

CAPÍTULO 4: INTEGRAÇÃO DE DADOS

Durante os primeiros capítulos deste trabalho foram descritos os diversos tipos de dados utilizados durante o projeto GENOSOJA, sendo que cada um dos experimentos foi tratado de maneira individual, envolvendo uma análise de bioinformática específica. Para cada análise foram criadas diferentes interfaces web, também já apresentadas. Tais interfaces disponibilizam diversas maneiras de recuperar informações do banco de dados, incluindo funcionalidades como: buscas por palavras-chave, análises estatísticas e anotação dos genes utilizando vários bancos conhecidos de proteínas. Além disso, cada interface possui funcionalidades específicas que tornam possível a obtenção de dados sobre a expressão de um gene em cada um dos múltiplos experimentos realizados.

Apesar da aparente distinção entre os experimentos de transcriptoma realizados pelo projeto, as bibliotecas utilizadas nas diferentes metodologias são, não só relativas a estresses em comum, como também a tempos semelhantes de estresse. Como exemplo, observa-se que os dados da biblioteca de SuperSAGE infectada com a doença da ferrugem asiática são relativos ao genótipo PI561356 com 12, 24 e 48 horas após a ação do fungo. No caso das bibliotecas subtrativas e dos microRNAs, existem bibliotecas relativas ao mesmo cultivar, estresse e tempos de infecção. A mesma semelhança pode ser notada para os experimentos de seca, tanto do cultivar BR 16 quanto do Embrapa 48. Dessa forma, espera-se que, genes relacionados a resposta da planta a estes estresses apareçam como diferencialmente expressos em qualquer um dos experimentos analisados. A Figura 24 demonstra tal hipótese, onde se verifica que o aumento do *RPKM* nas bibliotecas subtrativas resulta no aumento da mediana do

fold-change dos genes *up-regulated* em SuperSAGE, revelando assim genes altamente expressos em ambas as análises.

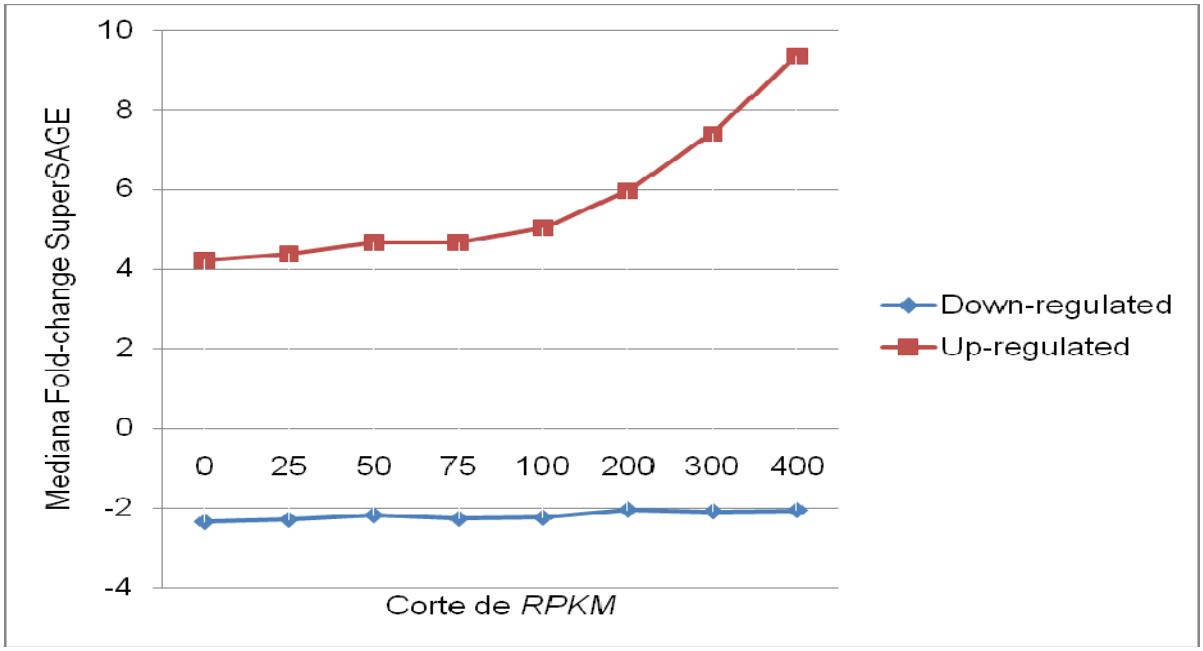


Figura 24: SuperSAGE x Bibliotecas subtrativas - Comparação entre os resultados do experimento de SuperSAGE e das bibliotecas subtrativas para o estresse de seca do cultivar BR 16 (resistente). Observa-se que, o aumento do *RPKM* das bibliotecas subtrativas ocasiona o aumento do fold-change dos genes *up-regulated* no experimento de SuperSAGE.

Tendo esse cruzamento entre as metodologias, a integração dos dados dos diversos experimentos pode facilitar a mineração de dados do projeto e simplificar a busca por genes com possível potencial biotecnológico. Além disso, tendo o genoma da soja com referência, é possível obter informações sobre a posição de cada um dos genes no mesmo através de programas de alinhamento, como o exonerate (Slater and Virney, 2005). A vantagem do exonerate em relação a outros programas como o

BLAST, por exemplo, é que seu alinhamento traz a posição exata dos exons e introns na referência.

Para facilitar a integração das informações, o banco de dados foi organizado de maneira específica. Todos os dados estão ligados a três tabelas principais, contendo: (*i*) genoma da soja (cromossomos e scaffolds), (*ii*) unigenes (*contigs* e *singlets*) da montagem de ESTs, e (*iii*) genes preditos do genoma. Todos os dados de transcriptoma do GENOSOJA (SuperSAGE, bibliotecas de cDNA – subtrativas e full-length -, microRNAs) fazem referência as tabelas *(ii*) e (*iii*) através de metodologias específicas que foram descritas no capítulo 3. Essas tabelas, (*ii*) e (*iii*), por sua vez, estão unidas entre si através de um BLASTN com e-value de corte 1e-50, buscando elucidar sobreposições entre os dois conjuntos de genes. Além disso, elas estão "linkadas" com a primeira através das posições dos genes no genoma, obtidas com o exonerate. A Figura 25 apresenta o esquema resumido do banco de dados, mostrando as principais tabelas.

A partir do modelo de banco de dados proposto foi criada uma interface web simples e amigável para disponibilizar todas as informações do projeto. Através dela é possível realizar buscas por genes preditos, unigenes da montagem ou sequências de cDNA full-lenght. Em qualquer uma das buscas, a interface se encarregará de buscar os outros dados correspondentes, isto é, uma busca por um gene predito ou um unigene da montagem correspondente ao primeiro retornará a mesma tela de resultados. Por exemplo, o Contig1 tem 100% similaridade com o gene predito "Glyma09g42180.1". Ao buscarmos por Contig1 na interface ela trará todos os dados

relativos ao gene predito. Dessa forma, a busca sempre será baseada no mesmo referencial.

A tela de resultados da interface está dividida em 6 diferentes blocos. A primeira ("*Gene information*") apresenta todas as informações básicas sobre o gene, incluindo nome, tamanho, posição no genoma, variantes de splicing, além de diversos links para suas sequências em formato FASTA (gene, reverso complementar e proteína). A segunda seção ("*Blast Results*") mostra os resultados de anotação do gene nos diversos bancos de dados, como NR, Uniref, GO, além do resultado do AutoFACT. Todos os resultados deste bloco são links que apontam para o arquivo de BLAST. A terceira ("*NCBI ESTs clusters*") lista todos os unigenes da montagem que possuem similaridade com o gene selecionado. Para cada cluster são exibidos o resultado de anotação do mesmo contra o NR e um link para a interface que aponta sua expressão diferencial (Northern digital), já descrita previamente. O nome do cluster também é um link para uma página onde são exibidos os ESTs que foram agrupados para formar o gene. O quarto bloco apresenta todas as bibliotecas subtrativas do projeto que possuem o gene presente. A expressão do gene em cada biblioteca é apresentada através do valor *RPKM*. Os dados de SuperSAGE, incluindo número de repetições em cada biblioteca, valor *p* e mismatches do alinhamento são apresentados no quinto bloco. As cores verde e vermelho mostram, respectivamente, se a tag foi considerada *down-regulated* ou *up-regulated* em cada uma das amostras. Por fim, a última seção mostra todos os microRNAs que tem o gene selecionado como alvo, incluindo sua expressão na biblioteca à qual ele foi sequenciado. A Figura 26 mostra um exemplo desta tela após uma busca pelo unigene "Contig1058".

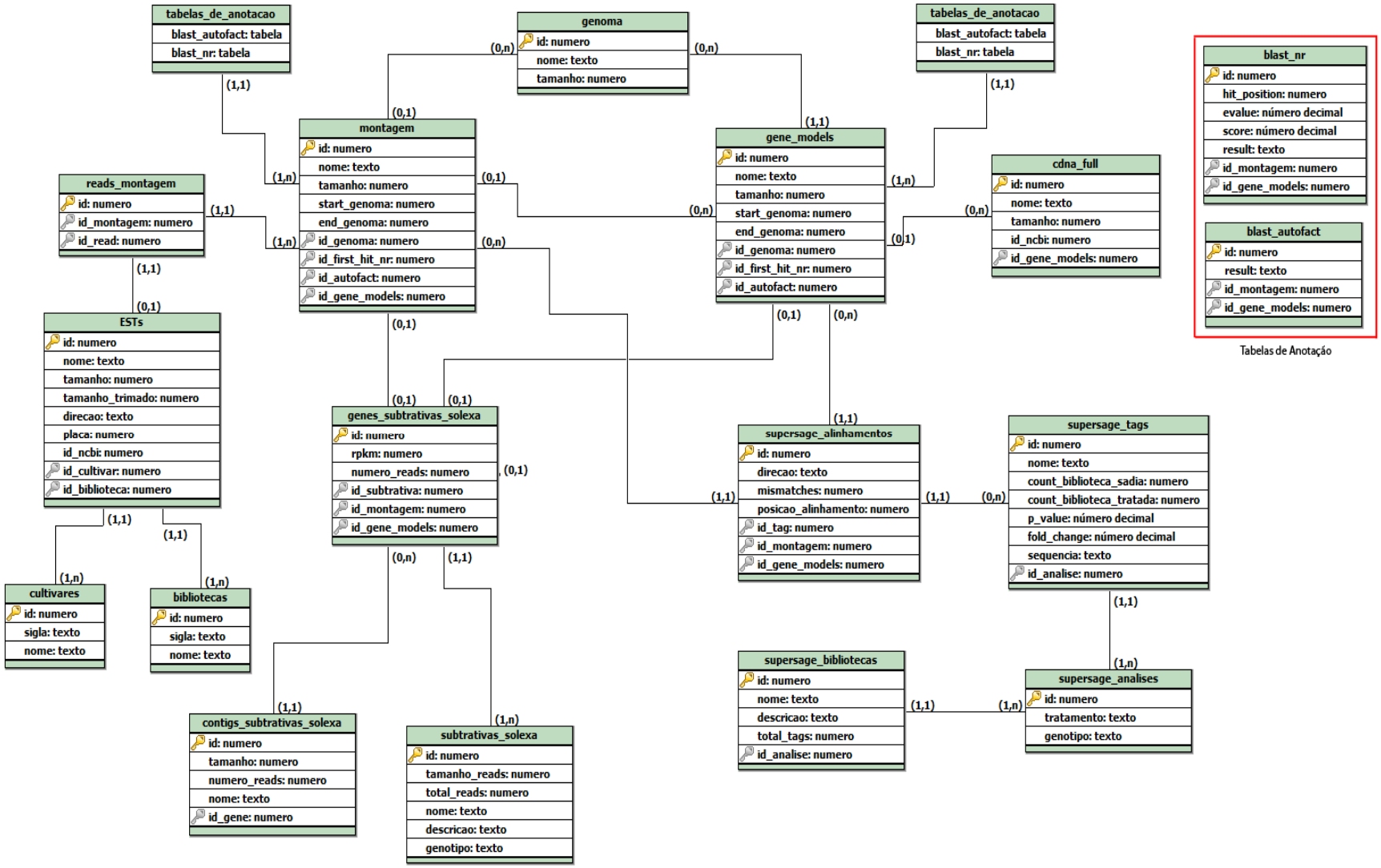
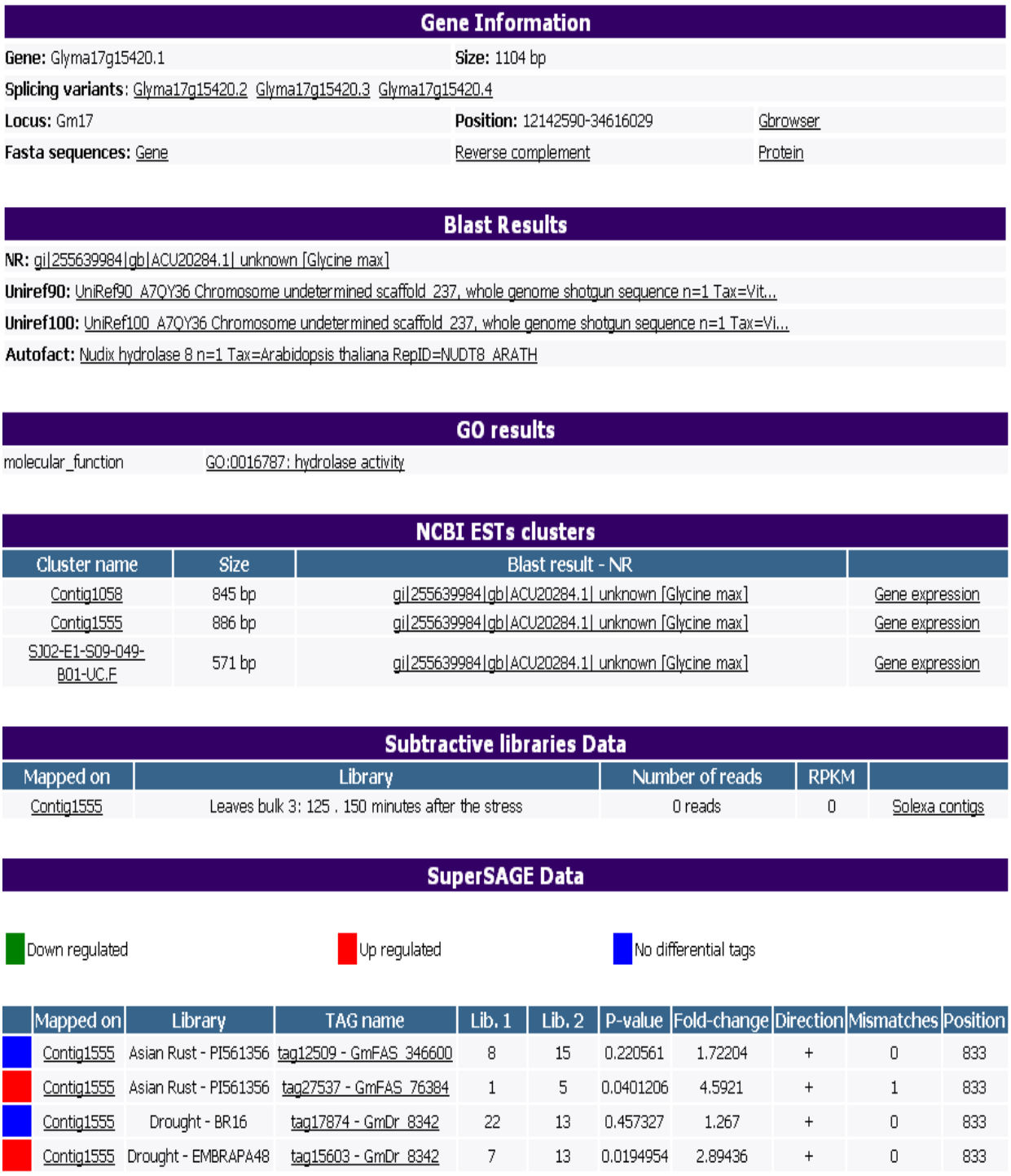


Figura 25: Modelo relacional resumido do banco de dados – Contém somente as principais tabelas. Observa-se que todos os dados estão ligados as tabelas montagens e gene_models. Estas, por suas vez, estão ligadas a tabela genoma.



Gene Information

Gene: Glyma17g15420.1	**Size:** 1104 bp	
Splicing variants: Glyma17g15420.2 Glyma17g15420.3 Glyma17g15420.4		
Locus: Gm17	**Position:** 12142590-34616029	Gbrowser
Fasta sequences: Gene	Reverse complement	Protein

Blast Results

NR:	gi|255639984|gb|ACU20284.1| unknown [Glycine max]
Uniref90:	UniRef90_A7QY36 Chromosome undetermined scaffold_237, whole genome shotgun sequence n=1 Tax=Vit...
Uniref100:	UniRef100_A7QY36 Chromosome undetermined scaffold_237, whole genome shotgun sequence n=1 Tax=Vi...
Autofact:	Nudix hydrolase 8 n=1 Tax=Arabidopsis thaliana RepID=NUDT8_ARATH

GO results

molecular_function	GO:0016787: hydrolase activity

NCBI ESTs clusters

Cluster name	Size	Blast result - NR	
Contig1058	845 bp	gi|255639984|gb|ACU20284.1| unknown [Glycine max]	Gene expression
Contig1555	886 bp	gi|255639984|gb|ACU20284.1| unknown [Glycine max]	Gene expression
SJ02-E1-S09-049-B01-UC.F	571 bp	gi|255639984|gb|ACU20284.1| unknown [Glycine max]	Gene expression

Subtractive libraries Data

Mapped on	Library	Number of reads	RPKM	
Contig1555	Leaves bulk 3: 125 . 150 minutes after the stress	0 reads	0	Solexa contigs

SuperSAGE Data

Down regulated Up regulated No differential tags

	Mapped on	Library	TAG name	Lib. 1	Lib. 2	P-value	Fold-change	Direction	Mismatches	Position
	Contig1555	Asian Rust - PI561356	tag12509 - GmFAS_346600	8	15	0.220561	1.72204	+	0	833
	Contig1555	Asian Rust - PI561356	tag27537 - GmFAS_76384	1	5	0.0401206	4.5921	+	1	833
	Contig1555	Drought - BR16	tag17874 - GmDr_8342	22	13	0.457327	1.267	+	0	833
	Contig1555	Drought - EMBRAPA48	tag15603 - GmDr_8342	7	13	0.0194954	2.89436	+	0	833

Figura 26: Interface web para integração dos dados - Tela de resultados da interface. A figura apresenta cada um dos blocos da interface, incluindo informações do gene, resultados de BLAST, unigenes relativos ao gene, além de dados de bibliotecas subtrativas e SuperSAGE.

Além da interface apresentada, outra forma de integrar as diversas fontes de dados é através da utilização do programa Gbrowse (Generic Genome Browser). O Gbrowse é um dos programas mais utilizados para visualização de dados quando se tem uma sequência de referência (Podicheti *et al*, 2009), tendo se tornado bastante popular por ter seu código aberto e ser altamente configurável. Ele integra as diversas informações disponíveis através de trilhas clicáveis. A trilha superior é a de referência, comumente a sequência genômica da espécie de estudo. As demais trilhas correspondem aos diversos tipos de dados mapeados na trilha de referência. O usuário pode escolher a região do genoma e as trilhas que deseja visualizar.

Uma das grandes vantagens do Gbrowse é a adaptação ao formato GFF (Generic Feature Format), um formato texto, com campos separados por tabulações, muito utilizado por diversos programas de alinhamento. Além disso, as versões mais recentes do programa também trabalham com arquivos no formato SAM. Através de resultados de alinhamento de transcritos curtos com este formato é possível criar trilhas que representam picos de expressão no genoma. Com isso, genes diferencialmente expressos em condições ou tempos de estresse são identificados com maior facilidade.

Para o GENOSOJA o genoma do cultivar Williams82 foi utilizado como referência para o Gbrowse. Utilizando o exonerate foram criadas trilhas específicas para os genes preditos do genoma, para os unigenes da montagem de ESTs e para a biblioteca de cDNA full-lenght. Além disso foram criadas trilhas para as bibliotecas subtrativas do projeto, uma para cada. Para isso foi utilizado o programa SOAP para alinhar os reads com o genoma, permitindo 2 mismatches e retornando todos os alinhamentos ótimos.

O arquivo de saída do SOAP foi convertido para o formato SAM utilizando scripts do pacote SAMtools. Os arquivos SAM foram utilizados como entrada para o Gbrowse. A Figura 27 apresenta um exemplo de visualização de dados com o Gbrowse. Nela, é possível observar o gene “Glyma13g10330.1” somente é expresso no cultivar BR16 após um grande tempo de submissão ao estresse de seca (125-150 min.).

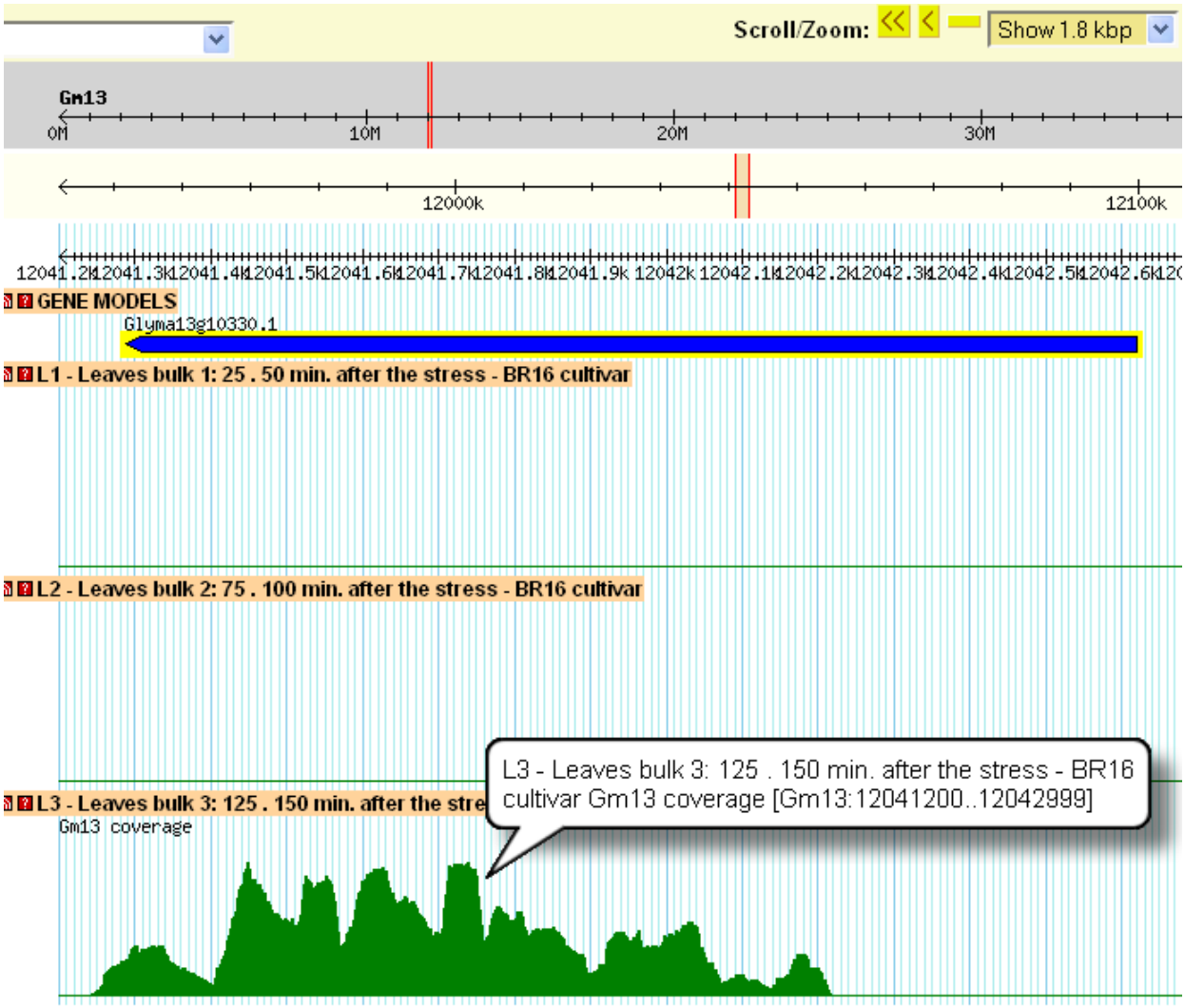


Figura 27: Visualização de dados com o Gbrowse - Com o Gbrowse, é possível comparar a expressão de um gene durante a evolução do tempo de submissão a um determinado estresse.

CAPÍTULO 5 - CONCLUSÕES E PERSPECTIVAS

O ferramental de bioinformática apresentado neste trabalho revela-se como um importante avanço para identificação de genes com potencial biotecnológico para incrementar a produção brasileira de soja. Através de consultas simples em interfaces web amigáveis é possível obter características de expressão dos diversos genes nas diversas condições estudadas pelo projeto GENOSOJA, ou seja, nas condições que se revelam problemáticas para a cultura no Brasil. Por outro lado, as metodologias de bioinformática, as interfaces web e o modelo de banco de dados propostos neste trabalho podem ser utilizados como referenciais para outros projetos de estudo expressão de genes, independentemente da espécie em estudo.

A utilização de referenciais comuns (o genoma, os unigenes da montagem de ESTs e os genes preditos) para a análise dos diversos experimentos garante facilidade para integração de novos dados, sejam de outros tipos de experimentos ou de outras condições de estudo. A inclusão de possíveis novos genes (ainda passíveis de validação experimental) também é simples, não necessitando de nenhuma alteração no modelo de banco de dados proposto e utilizado. Além disso, a integração dos dados aqui apresentados com os gerados pelo Consórcio Internacional de soja (*ISGC*), proposta pelo GENOSOJA, está facilitada, podendo se basear nos genes preditos do genoma e nas posições dos genes no mesmo.

Atualmente, a utilização dos dados apresentados neste trabalho está restrita aos pesquisadores do projeto GENOSOJA, incluindo importantes instituições brasileiras, como: Embrapa Soja, Universidade Federal de Pernambuco, Universidade Federal de

Viçosa, Universidade Federal do Rio Grande do Sul. Esses pesquisadores foram treinados sobre as melhores formas de acessar as informações em dois cursos ministrados por membros da equipe de bioinformática (incluindo o autor do trabalho) do Laboratório de Genômica e Expressão, realizados em Londrina – Paraná (2009) e em Recife – Pernambuco (2010).

Por fim, visando uma maior facilitação na identificação de genes com potencial biotecnológico propõe-se como perspectivas para continuação deste trabalho:

1 – Análise de SNPs dentre os cultivares brasileiros. Existem SNPs específicos que ao causar alterações em proteínas levam a resistência a determinados estresses?

2 – Análise de expressão de genes dentre os diversos cultivares. Genes relativos a cultivares específicos podem ser a “chave” da resistência de um cultivar a determinado estresse.

3 – Análise de expressão de genes entre os diversos tecidos. Assim como no caso de genes de cultivares específicos, genes identificados em um só tecido também são importantes para estudos biotecnológicos.

REFERÊNCIAS BIBLIOGRÁFICAS

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF *et al* (1991). **Complementary DNA sequencing: expressed sequence tags and human genome project**. Science. 252:1651–1656.

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

Amano N, Tanaka T, Numa H, Sakai H and Itoh T (2010). **Efficient Plant Gene Identification Based on Interspecies Mapping of Full-Length cDNAs.** DNA Res 17 (5): 271-279.

Audic S and Claverie JM (1997). **The significance of Digital Gene Expression Profiles.** Genome Res., 7:986-995.

Aukerman, MJ and Sakai, H (2003). **Regulation of flowering time and floral organ identify by a MicroRNA and its APETALA2-like target genes.** Plant Cell Vol. 15, 2730-2741.

Arumuganathan K and Earle ED (1991). **Nuclear DNA content of some important plant species.** Plant. Mol. Biol. 9(3):208-218.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer ELL (2002). **The Pfam Protein Families Database.** Nucl. Acids Res. 30(1): 276-280.

Baudet C and Dias Z. **New EST Trimming Strategy** in: Brazilian Symposium on Bioinformatics, 2005. Lecture Notes in Bioinformatics – Berlin – Alemanha: Springer – Verlag, 2005. v. 3594 p. 206-209.

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE *et al* (2009). ***De novo* transcriptome assembly with ABySS.** Bioinformatics 25:2872-2877.

Brandão ASP, Rezende GC and Marques, RWC. **Crescimento agrícola no período 1999-2004, explosão da área plantada com soja e meio ambiente no Brasil.** IPEA, 2005. Disponível em: http://cdi.mecon.gov.ar/biblio/doc/ipea/td/1062.pdf. Acesso em 01/10/2010.

Coward L, Barnes NC, Setchell KDR and Barnes S (1993). **Genistein, daidzen, and their .beta.-glycoside conjugates: antitumor isoflavones in soybean foods from American and Asian diets.** Journal of Agricutural and Food Chemistry, 41 (11), 1961-1967.

Crick FHC, Barnett L, Brenner S and Watts-Tobin RJ (1961). **General Nature of the Genetic Code for proteins.** Nature nº 4809 Vol. 192 1227-1232.

Forsdyke DR (2009). **Scherrer and Jost's symposium: the gene concept in 2008.** Theory in Biosciences Vol. 128, Number 3, 157-161.

Goldberg RB (1978). **DNA sequence organization in the soybean plant.** Biochem Genet 16, 45-68.

Griffiths-Jones, S (2006). **The MicroRNA sequence database.** Methods in Molecular Biology, Vol. 342, 129-138.

Hashimoto S, Qu W, Ahsan B, Ogoshi K, Sasaki A, Nakatani Y, Lee Y, Ogawa A, Ametani A, Suzuki Y *et al* (2009). **High-resolution analysis of the 5'end transcriptome using a next generation DNA sequencer.** PLoS One, 4:e4108.

Hernandez D, François P, Farinelli L *et al* (2008). **De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer.** Genome Res. 18:802-809 2008.

Huang X and Madan A (1999). **CAP3: A DNA Sequence Assembly Program**. Genome Res., 9:868-877.

Jongeneel CV. (200) **Searching the expressed sequence tag (EST) databases: Panning for genes.** Bioinformatics 1:76-92.

Jung SH, Lee JY and Lee DH (2003). **Use of SAGE technology to reveal changes in**

gene expression in *Arabidopsis* leaves undergoing cold stress. Plant Molecular Biology, Volº 52, Nº 3, 553-567.

Kanehisa M and Goto S (2000). **KEGG: Kyoto Encyclopedia of Genes and Genomes.** Nucl. Acids Res. 28 (1): 27-30.

Koski LB, Gray LW, Lang BF and Burger G (2005). **AutoFACT: An Automatic Functional Annotation and Classification Tool.** BMC Bioinformatics, 6:151.

Langmead B, Trapnel C, Pop M and Salzberg SL (2009). **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** Genome Biology, 10:R25.

Lagos-Quintana M, Rauhut R, Meyer J *et al* (2003). **New microRNAs from mouse and human.** RNA Vol. 9: 175-179.

Lee S, Chen J, Zhou G, Zhou G, Zhang Shi R, Bouffard GG, Kocherginsky M, Ge X, Sun M, Jayathilaka N, Kim YC *et al* (2006). **Gene expression profiles in acute myeloid leukemia with common translocations using SAGE.** PNAS, vol. 103 no. 4 1030-1035

Leyritz J, Schicklin S, Blachon S, Keime C, Robardet C, Boulicaut J-F, Besson J, Pensa R and Gandrillon O (2008). **SQUAT: A web tool to mine human, murine and avian SAGE data.** BMC Bioinformatics 2008, 9**:**378.

Li H, Ruan J and Durbin R (2008). **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** Genome Res. 18: 1851-1858.

Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K and Wang J (2009a). **SOAP2: an improved ultrafast tool for short read alignment.** Bioinformatics 25 (15): 1966-1967.

Li H, Handsaker B, Wysoker A, Fenneli T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R (2009b) **The Sequence Alignment/Map format and SAMtools.** Bioinformatics 25(16): 2078-2079.

Liang F, Holt I, Pertea G, Karamycheva S, Salaberg SL and Quackenbush J (2000). **An optimized protocol for analysis of EST sequences.** Nucleids Acids Research

28(18):3657-3665.

Libault M, Farmer A, Brechenmacher L, Drnevich J, Langley J, Bilgin DD, Radwan O, Neece DJ, Clough SJ, May GD and Stacey G (2010). **Complete Transcriptome of the Soybean Hair Cell, a Single-Cell Model, and Its Alteration in Response to *Bradyhizobium japonicum* Infection.** Plant Physiology 152:541-552.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z *et al* (2005). **Genome sequencing in microfabricated high-density picolitre reactors.** Nature 437, 376-380.

Matsumura H, Reich S, Ito A, Saitoh H, Kamoun S, Winter P, Kahl G, Reuter M, Kruger DH and Terauchi R (2003). **Gene expression analysis of plant host-pathogen interactions by SuperSAGE.** Proc. Natl. Acad. Sci., 100:15718-15723.

Matsumura H, Ito A, Saitoh H, Winter P, Kahl G, Reuter M, Kruger DH and Terauchi R (2005). **SuperSAGE.** Cellular Microbiology, Vol. 7, Issue 1, pages 11-18.

Mazière P and Enright AJ (2007). **Prediction of microRNA targets.** Drug Discovery Today, Vol. 12: 452-458.

Michon F, Tummers M, Kyyronen M, Frilander MJ and Thesleff I (2010). **Tooth morphogenesis and ameloblast differentiation are regulated by micro-RNAs.** Developmental Biology, Vol. 340, 355-368.

Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B (2008). **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** Nature Methods – 5, 621-628.

Nascimento LC, Vidal RO, Costa GGL, Pereira GAG and Carazzolle MF. **Ab-initio and mapping assemblies of transcriptome using short sequences technologies.** In 5ª International Conference of the Brazilian Association for Bioinformatics and Computational Biology – X-meeting, 2009.

Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, Carrington JC and Weigel D (2003). **Control of leaf morphogenesis by microRNAs.** Nature 425, 257-263.

Podicheti R, Gollapudi R and Dong Q (2009). **WebGBrowse – a web server for**

GBrowse. Bioinformatics 25(12): 1550-1551.

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW and Velculescu VE (2002). **Using the transcriptome to annotate the genome.** Nature Biotechnology 20, 508 - 512 .

Santos, AL, Weber LM and Moreira TZT. **A matriz energética brasileira e o aproveitamento das fontes renováveis.** Disponível em: http://www.ipardes.gov.br/pdf/bol_ana_conjuntural/bol_28_1g.pdf. Acesso em 01/10/2010.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Xu D *et al* (2010). **Genome sequence of the palaeopolyploid soybean.** Nature 463, 178-183.

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y *et al* (2002). **Functional Annotation of a Full-Length *Arabidopsis* cDNA Collection.** Science, 296, 141-145.

Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP and Shoemaker RC (2010). **RNA-Seq Atlas of *Glycine max:* A guide to the soybean transcriptome.** BMC Plant Biology, 10:160.

Shoemaker RC, Potzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP *et al* (1996). **Genome Duplication in Soybean (Glycine Subgenus Soja).** Genetics 144, 329-338.

Slater GSC and Birney E (2005). **Automated generation of heuristics for biological sequence comparison.** BMC Bioinformatics 6:31.

Shinozaki K, Yamaguchi-Shinozaki K and Seki M (2003). **Regulatory network of gene expression in drought and cold stress responses.** Current Opinion in Plant Biology, Vol. 6, Issue 5, 410-417.

Shinozaki K and Yamaguchi-Shinozaki K (2007). **Gene networks involved in drought stress and tolerance.** Journal of Experimental Botany, Vol. 58, Issue 2, 221-227.

Simpson, JT, Kim W, Jackman SD, Schein JE, Jones SJM, Birol I (2009). **ABySS: A parallel assembler for short read sequence data.** Genome Res. 19:1117-1123.

Smith TF and Waterman MS (1981). **Identification of Common Molecular Subsequences.** J. Mol. Biol. Vol. 147, 195-197.

Song S, Qu H, Chen C, Hu S and Yu J. **Differential gene expression in an elite hybrid rice cultivar (*Oryza sativa, L*) and its parental lines based on SAGE data.** BMC Plant Biology 2007, 7**:**49

Sommer DD, Delcher AL, Salzberg SL and Pop M (2007). **Minimus: a fast, light weight genome assembler.** BMC Bioinformatics, 8:64.

Sun M, Zhou G, Lee S, Chen S, Shi RZ and Wang SM (2004). **SAGE is far more sensitive than EST for detecting low-abundance transcripts.** BMC Genomics, 5:1

Sutton GG, White O, Adams MD and Kerlavage AR (1995). ***TIGR* Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects.** Genome Science and Technology 1(1):9-19.

Suzek BE, Huang H, McGarvey P, Mazumber R, Wu CH (2007). **Uniref: comprehensive and non-redundant UniProt reference clusters.** Bioinformatics 23 (10): 1282-1288.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJ, Salzberg SL, Wold BJ and Pachter L (2010). **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** Nature Biotechnology Vol. 28, 511-515.

Umezawa T, Sakurai T, Totoki Y , Toyoda A, Seki M, Ishiwata A, Akiyama K, Kurotani A, Yoshida T, Mochida K *et al* (2008). **Sequencing and Analysis of Approximately 40000 Soybean cDNA Clones from a Full-Length-Enriched cDNA Library.** DNA Res., 15 (6): 333-346.

Useche FJ, Gao G, Harafey M and Rafalski A (2001). **High-throughput identification, database storage and analysis of SNPs in EST sequences.** Genome

inform. 12:194-203.

Van de Mortel M, Recknor JC, Graham MA, Netteton D, Dittman JD, Nelson RT, Godoy CV, Abdelnoor RV, Almeida AMR, Baum TJ and Witham SA (2007). **Distinct Biphasic mRNA Changes in Response to Asian Soybean Rust Infection.** Molecular Plant-Microbe Interactions. Vol. 20, Number 8, 887-899.

Velculescu VE, Zhang L, Vogelstein B and Kinzler Kw (1995). **Serial Analysis of Gene Expression.** Science Vol. 270 nº 5235, pages 484-487.

Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M *et al* (2006). **A microRNA expression signature of human solid tumors defines cancer gene targets.** PNAS Vol. 103 nº 7, 2257-2261.

Wang JP, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC and Pamphilis CW (2004). **EST clustering evaluation and correction.** Bioinformatics Vol. 20 nº 27, pages 2973-2984.

Wang L, Feng Z, Wang X and Zhang X (2010). **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.** Bioinformatics 26(1): 136-138.

Watson JD and Berry A (2003). **DNA: The Secret of the Life.** Alfred A. Knopf: New York, New York, USA, 446 pp., ISBN: 0-375-41546-7.

Watson JD and Crick FHC (1953). **A structure for Deoxyribose Nucleid Acid.** Nature, 171: 737-738.

Wieland I, Bolger G, Asouline G and Wigler M (1990). **A method for difference cloning: Gene amplification following subctrative hybridization.** PNAS, 87:2720-2724.

Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A *et al* (2009). **Ab initio construction of eukaryotic transcriptome by massively parallel mRNA sequencing.** PNAS, 106:3254-3269.

Zerbino, DR and Birney, W (2008). **Velvet: Algorithms for de novo short read**

assembly using de Bruijn graphs. Genome Res. 18: 821-829.