UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

Rodrigo Martins Brandão

Abordagem computacional aplicada ao desenvolvimento de um SAGEmap de *Apis mellifera*

Ribeirão Preto – SP
2009

Catalogação na Publicação
Serviço de Documentação
Faculdade de Medicina de Ribeirão Preto

Brandão, Rodrigo Martins
Abordagem computacional aplicada ao desenvolvimento de um SAGEmap de *Apis mellifera* / Rodrigo Martins Brandão; orientador: Wilson Araújo da Silva Júnior. – Ribeirão Preto -SP, 2009.
52 f.:fig.

Dissertação (Mestrado – Programa de Pós-Graduação em Genética. Área de Concentração: Genética) – Faculdade de Medicina de Ribeirão Preto.

1. *Apis mellifera*. 2. SAGE. 3. Anotação.

Rodrigo Martins Brandão

Abordagem computacional aplicada ao desenvolvimento de um SAGEmap de *Apis mellifera*

Dissertação apresentada ao Programa de Pós-Graduação em Genética da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Mestre em Ciências.

Área de Concentração: Genética
Orientador: Prof. Dr. Wilson Araújo da Silva Júnior

Ribeirão Preto – SP
2009

Dedicatória

Dedico esta dissertação a meus pais,
Marco Antônio Cividanes Brandão e Edith de Castro Martins Brandão
cujo exemplo de honestidade e trabalho
tem sido um norteador para a minha vida.
Dedico também as minhas irmãs Renata Martins Brandão e
Fernanda Martins Brandão

Agradecimentos

Dedico meus sinceros agradecimentos para:

– O Prof. Dr. Wilson Araújo da Silva Júnior, pela orientação e incentivo;

– À equipe do Laboratório de Bioinformática, em especial aos colegas Alynne Oya, Daniel Guariz Pinheiro, Gislaine Pereira, Israel Tojal da Silva, Rodrigo Lucena Borges, Olsen Rodrigo e Thiago Yukio Oliveira, pela ajuda e companheirismo em diversos momentos;

– O pessoal do Laboratório de Genética Molecular pela convivência e ajuda durante a elaboração do projeto. Em especial, Carla Martins Kaneto, Cristiane Ayres Ferreira, Greice Molfetta, Dalila, Adriana e Anemari.

– O Departamento de Genética da FMRP;

– À Adriana, Aline Carolina e Liliane pelo empréstimo de material biológico de abelha;

– À Susie, secretária do programa de pós-graduação em genética;

– À Meire Tarla pela paciência, ajuda e dedicação nas tarefas burocráticas e administrativas.

– À Dalvinha pela ajuda nas mais diversas tarefas burocráticas e pelo o seu exemplo de bondade e honestidade.

– À Fundação Hemocentro de Ribeirão Preto que ofereceu uma excelente infra-estrutura para o desenvolvimento do projeto;

– O Centro de Terapia Celular (CEPID/FAPESP) pelo apoio financeiro;

"Prefiro ser um otimista e estar errado
a ser um pessimista e estar certo."
Albert Einstein

Resumo

A *Apis mellifera* é uma espécie que desperta grande interesse dos Biólogos por possuir um mecanismo complexo e organizado relacionado ao comportamento social, capacidade de aprendizado e memória. Muitas dessas características possuem uma base genética determinada pelas regulação e variação estrutural e funcional de um conjunto de genes. Variações do nível de expressão gênica também contribuem para a regulação desses mecanismos. Uma das técnicas mais usadas para avaliar o padrão de expressão global de um tecido ou linhagem celular é o SAGE (*Serial Analysis of Gene Expression*), que permite estudar as variações na expressão gênica causadas por estímulos externos ou pela fisiologia de um tecido em vários organismos. O método de SAGE se baseia na análise do perfil de expressão gênica de um tecido pela geração e contagem de sequências de nucleotídeos curtas (10 bases) denominadas *tags*. Para tornar-se informativo, primeiramente é necessário associar cada *tag* a um gene (processo de anotação). O objetivo deste trabalho foi desenvolver uma metodologia para anotar cada *tag* de uma biblioteca de SAGE produzida a partir de amostras de tecido cerebral de *A. mellifera*. As análises para o desenvolvimento dessa metodologia seguiram as seguintes etapas: 1) gerar uma lista de *tags* confiáveis (LTC) da biblioteca de SAGE; 2) extração das *tags* virtuais em base de dados de transcritos de *A. mellifera*; 3) relacionar as *tags* geradas experimentalmente com as *tags* virtuais; 4) associar *tag* ao gene; 5) reduzir a ambiguidade das *tags*; 6) mapear as *tags* da biblioteca no genoma; 7) validar os resultados. As *tags* virtuais foram extraídas pela da identificação computacional das 10 bases adjacentes ao sítio da enzima de ancoragem *NlaIII* em sequências de transcritos da abelha. Uma base de dados relacional foi modelada para armazenar os dados da biblioteca de SAGE de cérebro e do SAGE virtual. Experimentalmente foram geradas 60.536 *tags* de SAGE, sendo 20.483 *tags* únicas. Foi gerada uma lista de *tags* mais confiáveis, reduzindo o número de *tags* para 45.674, sendo 5.683 *tags* únicas. Para anotarmos a biblioteca de SAGE de *A. mellifera*, as 5.683 *tags* da lista LTC foram mapeadas contra a base de dados do UniGene desse organismo resultando em 27,5% das *tags* anotadas. As *tags* foram classificadas de acordo com alguns critérios estabelecidos para associarmos com maior confiança a melhor *tag* ao gene, diminuindo a ambiguidade das *tags*. Todas as *tags* da lista LTC foram mapeadas no genoma da abelha nas duas orientações e foram encontradas 85% das *tags*. Encontramos 63% de *tags* no genoma que também foram anotadas no UniGene e 94% das *tags* LTC anotadas no UniGene foram encontradas no genoma. Selecionamos seis genes para avaliar suas expressões através da técnica de RT-PCR, e os resultados corresponderam ao observado na biblioteca de SAGE. O presente estudo apresenta a primeira iniciativa de anotação de uma biblioteca de SAGE de *A. mellifera*. O resultado mostrou ser uma metodologia eficaz e que será de grande utilidade nos estudos envolvendo a análise do perfil de expressão gênica desse organismo.

Palavras-Chave: *Apis mellifera*, SAGE e anotação.

Abstract

Computational approach applied to development of the SAGEmap of Apis mellifera.

The *Apis mellifera* is a specie that arouses great interest of biologists by its organized and complex social behavior, learning capacity and memory. Many of these characteristics have a genetic basis which can be determined by the structural changes and a set of functional genes. Through the technique of quantification of gene expression in large scale SAGE (Serial Analysis of Gene Expression) it is possible to study the genetic changes caused by external stimuli or the physiology of a given tissue in various organisms. The SAGE method is based on the analysis of gene expression profile of a tissue by identifying and counting small nucleotide sequences (tags), that represent a unique transcript. In order to become functional, it is first necessary to associate each tag to a gene (annotation process). Our objective was to develop a methodology to map a library of SAGE built from brain tissue samples of *A. mellifera*. The analyses for the construction of this methodology followed these steps: generate the reliable SAGE tag list (LTC) of the SAGE library, extract virtual tags in a database of transcripts of *A. mellifera*, link the tags experimentally generated to virtual tags, associate the tag to the gene, reduce the ambiguity of the tags, map the tags of the library in the genome and validate the results. The virtual tags were extracted through computational identification of 10 adjacent bases to the site of the anchor enzyme in the transcribed sequences of bee. A relational database was modeled to store data from the SAGE library of the brain and the virtual SAGE. Around 60,536 SAGE tags were experimentally generated and 20,483 unique tags. A more reliable list of tags was generated, reducing the number of tags to 45,674, with 5,683 unique tags. To note the SAGE library of the *A. mellifera*, the list of 5,683 LTC tags were mapped against the UniGene database of the organism resulting in 27.5% of annotated tags. The tags were classified according an established criteria in order to associate with higher confidence the best tag to the gene, decreasing the ambiguity of the tags. All the list LTC tags were mapped in the bee genome in both orientations and we found 85% of tags. We found 63% of tags in the genome that were annotated in the UniGene and 94% of LTC tags annotated in UniGene were found in the genome. We selected six genes to assess their expression through the technique of RT-PCR, and the results corresponded to that observed in the SAGE library. We present the first initiative for the annotation of the SAGE library of *A. mellifera*, showing an effective methodology that will be useful in studies involving the analysis of gene expression profile of this organism.

Keyword: *Apis mellifera*, SAGE and annotation.

Sumário

Lista de Abreviaturas e Siglas

AE *Anchoring Enzyme*

API *Application Program Interface*

BLAST *Basic Local Alignment Search Tool*

CGAP *Cancer Genome Anatomy Project*

CGI *Common Gateway Interface*

DNA *Deoxyribonucleic Acid*

EST *Expressed Sequence Tag*

FFCLRP Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

FMRP Faculdade de Medicina de Ribeirão Preto

GDM *Genome Data Mining*

GEO Gene Expression Omnibus

HBGSP *Honey Bee Genome Sequencing Project*

LGMB Laboratório de Genética Molecular e Bioinformática

LTC Lista de *Tags* Confiáveis

NCBI *National Center for Biotechnology Information*

NHGRI *National Human Genome Research Institute*

PERL *Practical Extraction and Report Language*

pb Pares de bases

PCR *Polymerase Chain Reaction*

Poli-A Poliadenilação

RT-PCR *Reverse Transcriptase PCR*

SAGE *Serial Analysis of Gene Expression*

SGBD Sistema de Gerenciamento de Banco de Dados

TE *Tagging Enzyme*

TIGR *The Institute for Genomic Research*

UTR *Untranslated Region*

Lista de Figuras

Lista de Tabelas

1 *Introdução*

O avanço da Bioinformática e a evolução das técnicas de sequenciamento genético aumentaram demasiadamente o número de sequências gênicas e genomas nos bancos de dados públicos. Isso reforça a necessidade do desenvolvimento de novas e eficientes metodologias para a análise funcional de alguns aspectos da biologia de um organismo com base na identificação e caracterização de padrões de expressão gênica e seu relacionamento com determinado fenótipo de interesse, bem como o estado ou uma condição biológica de interesse (DONSON *et al.*, 2002; MEYERS *et al.*, 2004).

1.1 Biologia de *Apis mellifera*

A espécie de abelha *A. mellifera* pertence ao reino Animalia, do filo Arthropoda, da classe Insecta, da ordem Hymenoptera e da família Apidade. É também conhecida popularmente como abelha do mel.

A biologia de *A. mellifera* ainda possui obstáculos para aplicação de várias técnicas experimentais da genética com organismos modelo, já que a abelha não é um animal comum de laboratório. Os intervalos de procriação são longos, a manutenção da reprodução é trabalhosa, o melhoramento de populações é pequeno e os efeitos da endogamia são marcantes, que impedem o desenvolvimento de linhagens isogênicas de abelha. Apesar disso, não há dúvidas em qualifica-la como um organismo modelo pelo seu rico repertório genético e comportamental, motivando pesquisadores a continuarem seus esforços no sentido de desvendar melhor suas características genotípicas e fenotípicas (PAGE; GADAU; BEYE, 2002).

1.1.1 Organização social

As abelhas *A. mellifera* são insetos sociais que despertam grande interesse dos biólogos, por possuírem um complexo e organizado comportamento social, capacidade de apren-

dizado e memória. Essa complexidade etológica é sustentada por um sistema nervoso competente e simples do ponto de vista anatômico, mas com alto nível de organização (OLESKEVICH; CLEMENTS; SRINIVASAN, 1997).

As abelhas vivem em colônias organizadas em que os indivíduos se dividem em três diferentes castas: rainha, operária e zangão (Figura 1). Possuem funções bem definidas que são executadas visando sempre à sobrevivência e manutenção da colmeia. Em uma colônia, em condições normais, existe uma rainha, cerca de dez a trinta mil operárias e de zero a quatrocentos zangões (PAGE; PENG, 2001).

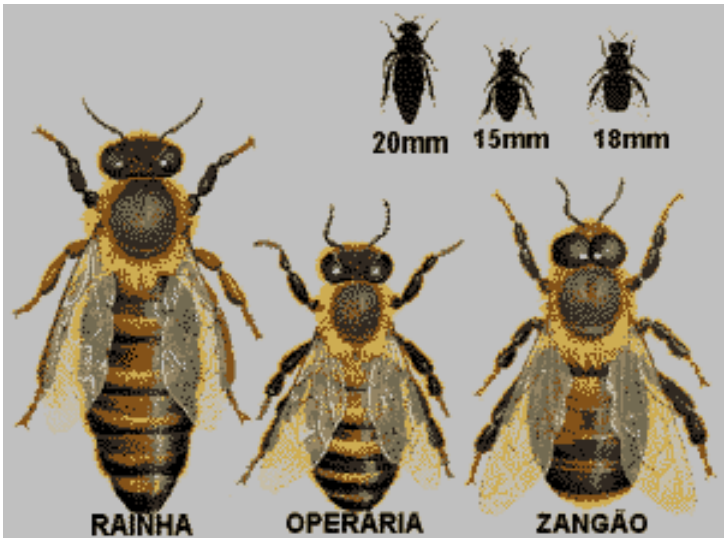


Figura 1: Relação de tamanho entre a Rainha, operária e zangão respectivamente (Figura obtida na internet pelo sítio: http://www.expoanimais.com.br/apicultura/rainha.htm).

A variabilidade genética numa colônia é alta considerando que apenas a rainha se reproduz, sendo fecundada por até dezessete zangões. As diferenças que se estabeleceram entre rainha e as operárias são de várias naturezas e resultam em um dimorfismo completo entre as duas castas (Figura 1). A organização social das abelhas é caracterizada pela divisão de trabalho e reprodução entre as fêmeas da colônia.

A rainha tem por função a postura de ovos e a manutenção da ordem social na colmeia. A larva da rainha é criada num alvéolo modificado bem maior que o das larvas de operárias e zangões, de formato cilíndrico, denominado realeira, sendo alimentada pelas operárias com a geleia real (produto rico em proteínas e vitaminas). A rainha adulta possui quase o dobro do tamanho de uma operária e é a única fêmea fértil da colmeia, apresentando o aparelho reprodutor bem desenvolvido.

As operárias (fêmeas estéreis ou semi-estéreis) desempenham uma função na colônia de acordo com a idade (polietismo etário), o desenvolvimento glandular e com a necessidade

da colônia (HUANG; ROBINSON, 1992) descritos na Tabela 1.

Tabela 1: Polietismo etário das operárias.

Idade	**Função**
1^{o} ao 5^{o} dia	Limpeza dos alvéolos e de abelhas recém-nascidas.
5^{o} ao 10^{o}	Cuidam da alimentação das larvas em desenvolvimento, apresentando grande desenvolvimento das glândulas hipofaringeanas e mandibulares (produtoras de geleia real).
11^{o} ao 20^{o} dia	Produzem cera para construção de favos (quando há necessidade) e recebem e desidratam o néctar trazido pelas campeiras, elaborando o mel.
18^{o} ao 21^{o} dia	Defesa da colmeia. As operárias apresentam os órgãos de defesa bem desenvolvidos, com grande acúmulo de veneno. Podem também participar do controle da temperatura na colmeia.
22^{o} dia até a morte	São denominadas campeiras. Coleta de néctar, pólen, resinas e água.

Caso necessário, as operárias podem reativar algumas das glândulas atrofiadas para realizar outras atividades, ou seja, em algumas situações uma abelha mais nova pode sair para a coleta no campo e uma abelha mais velha pode encarregar-se de alimentar a cria. As operárias possuem os órgãos reprodutores atrofiados, não sendo capazes de se reproduzirem. Isso acontece porque, na fase de larva, elas recebem alimento pouco nutritivo e em menor quantidade que a rainha. Além disso, a rainha produz feromônios que inibem o desenvolvimento do sistema reprodutor das operárias na fase adulta. Em compensação, elas possuem órgãos de defesa e trabalho perfeitamente desenvolvidos, muitos dos quais não são observados na rainha e no zangão, como por exemplo a corbícula (onde é feito o transporte de materiais sólidos) e as glândulas de cera.

Os zangões, machos haplóides, são resultado de óvulos não fecundados e sua única função é fecundar a abelha rainha durante o voo nupcial, morrendo logo após a cópula (cruzamento entre zangão e a rainha) (PAGE; PENG, 2001). As larvas de zangões são criadas em alvéolos maiores que os alvéolos das larvas de operárias, e levam vinte e quatro dias para completarem seu desenvolvimento de embrião (ovo) a adulto. Eles são maiores e mais fortes do que as operárias, entretanto, não possuem órgãos para trabalho nem ferrão e, em determinados períodos, são alimentados pelas operárias. Em contrapartida, os zangões apresentam os olhos compostos mais desenvolvidos e antenas com maior capacidade olfativa. Além disso, possuem asas maiores e musculatura de voo mais desenvolvida. Essas características lhes permitem maior orientação, percepção e rapidez para a localização de

rainhas virgens durante o voo nupcial.

1.1.2 Desenvolvimento

Durante o ciclo de vida, as abelhas passam por cinco diferentes fases: embrião (ovo), larva (dividida em cinco estágios), pré-pupa, pupa e adulto (Figura 2).

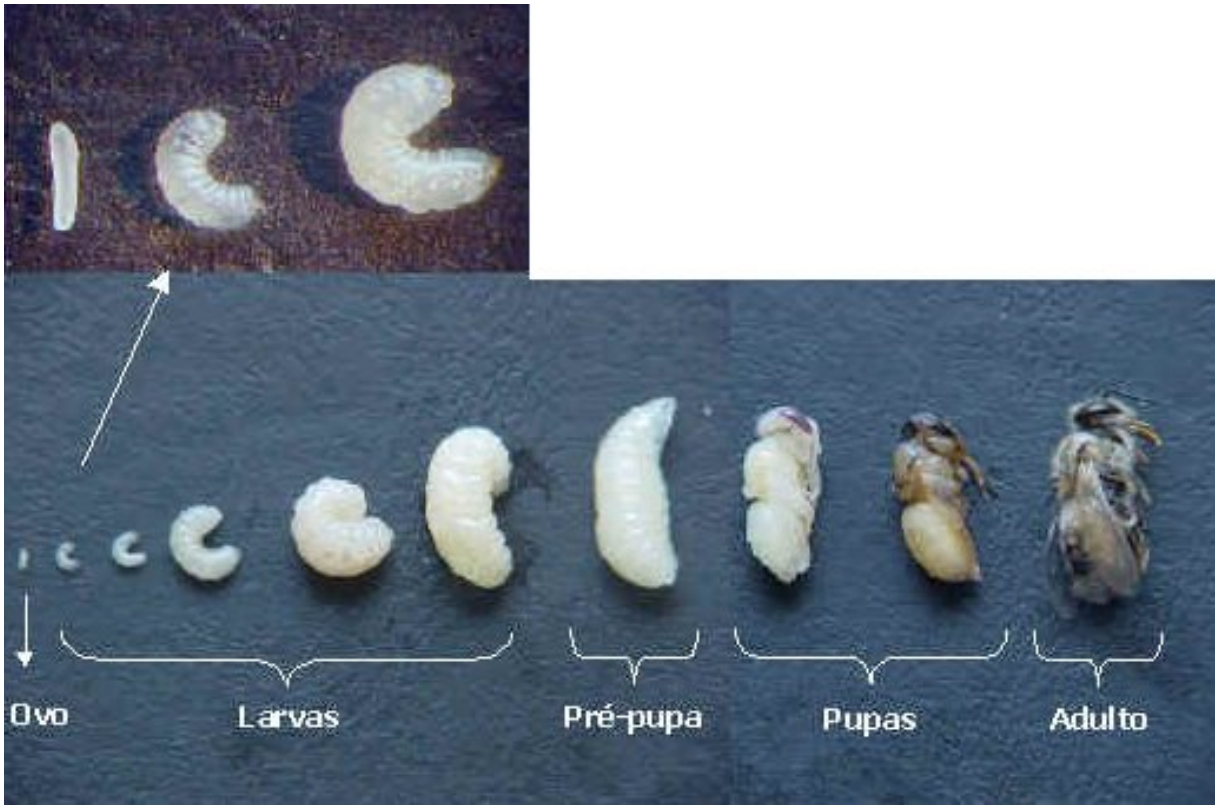


Figura 2: Fases do ciclo de desenvolvimento da *Apis mellifera* (Figura obtida na internet pelo sítio: `http://sistemasdeproducao.cnptia.embrapa.br/FontesHTML/Mel/SPMel/organizacao.htm`).

A rainha inicia a postura de ovos geralmente após o terceiro dia de sua fecundação, depositando um ovo em cada alvéolo. O ovo tem formato cilíndrico, de cor branca e, quando recém colocado, fica em posição vertical no fundo do alvéolo. Três dias após a postura de ovos, ocorre o nascimento da larva, que tem cor branca, formato vermiforme e fica posicionada no fundo do alvéolo, com corpo recurvado (Figura 2). Durante essa fase, a larva passa por cinco estágios de crescimento, trocando sua cutícula (pele) após cada estágio.

No final da fase larval, cinco a seis dias após a eclosão, a célula é operculada e a larva muda de posição, ficando reta e imóvel. Nessa fase, ela não se alimenta mais, tece seu casulo, sendo comumente chamada de pré-pupa. Na fase de pupa já podem ser distinguidos a cabeça, o tórax e o abdome, visualizando-se olhos, pernas, asas, antenas e

partes bucais. Os olhos e o corpo passam por mudanças de coloração até a emersão da abelha adulta (Figura 2). Toda a transformação pela qual a abelha passa até chegar ao estágio adulto denomina-se metamorfose. A duração de cada uma das fases é diferenciada para rainhas, operárias e zangões (Tabela 2).

Tabela 2: Período em dias do desenvolvimento das abelhas.

Casta	**Ovo**	**Larva**	**Pupa**	**Total**
Rainha	3	5	7	15
Operária	3	5	12	20
Zangão	3	6,5	14,5	24

A longevidade dos adultos das três castas também é diferente: a rainha pode viver até dois anos ou mais apesar de que, em clima tropical, sua vida reprodutiva dura, em média, um ano; as operárias, em condições normais, vivem de vinte a quarenta dias. Os zangões que não acasalam podem viver até oitenta dias, se houver alimento na colmeia. Durante o período de escassez de alimento, as operárias costumam expulsar ou matar os zangões.

1.1.3 Genoma

Em maio de 2002 o *National Human Genome Research Institute* (NHGRI) incluiu a *A. mellifera* em sua lista de prioridade para o sequenciamento completo do genoma devido a sua importância para estudos de neurobiologia, no que diz respeito aos instintos sociais, e características comportamentais únicas, bem como a sua relevância para a agricultura, estudo biológico e saúde humana. Destacaram também que a *A. mellifera* é um modelo para estudos de resistência a antibióticos, imunidade, reações alérgicas, desenvolvimento, saúde mental, doenças ligadas ao cromossomo X e longevidade (`http://www.genome.gov/page.cfm?pageID=10002851`).

A *A. mellifera* foi o terceiro inseto a ter seu genoma sequenciado por completo. O primeiro foi o da *Drosophila melanogaster* (ADAMS *et al.*, 2000) e em segundo o do mosquito da malária da espécie *Anopheles gambiae* (AULTMAN *et al.*, 2002), insetos considerados menos complexos que o da abelha.

O projeto do sequenciamento da abelha, *Honey Bee Genome Sequencing Project* (HBGSP), foi iniciado em 2003 pelo Centro de Sequenciamento do Genoma Humano da *Baylor College of Medicine* no Texas, EUA (BCM-HGSC). O sequenciamento com-

pleto do genoma e a anotação de famílias gênicas da abelha foram anunciados em outubro de 2006 pelo *The Honey Bee Genome Sequencing Consortium*, e encontra-se disponível no sítio `http://www.hgsc.bcm.tmc.edu/projects/honeybee/`, na versão 4.0 (CONSORTIUM, 2006). Esse projeto recebeu ajuda financeira do NHGRI e do Departamento da Agricultura dos Estados Unidos (USDA).

Foram identificados aproximadamente dez mil genes, um número inferior ao que os estudos anteriores esperavam (de aproximadamente quinze mil genes). O genoma da abelha tem o dobro do genoma da drosófila apesar de possuírem quase o mesmo número de genes e quando comparado com o genoma humano, a abelha possui metade do número de genes (CONSORTIUM, 2006).

O genoma da abelha está dividido em dezessete grupos cromossômicos, sendo dezesseis grupos referentes a cada cromossomo do tipo LG (*Linkage Group*) e um cromossômo mitocôndrial (MT). A Figura 3 mostra o tamanho dos dezessete grupos. Existe ainda, sequências com localização não definidas, e são denominadas de LGUn.

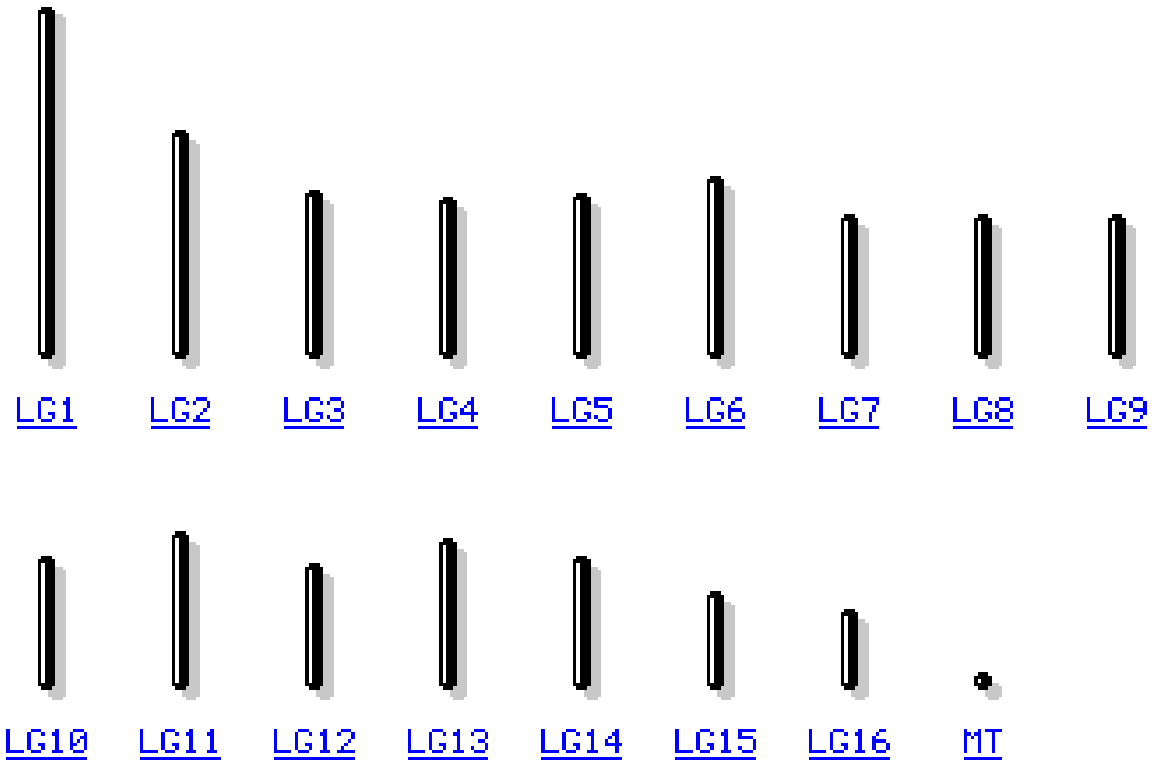


Figura 3: Representação esquemática de tamanho dos 16 cromossomos e do genoma mitocondrial de *A. mellifera*. LG: *Linkage Group*; MT: mitocondrial;

No genoma da abelha existem mais genes relacionados à utilização de pólen e néctar e para o olfato, fundamentais para a comunicação por meio de feromônios na colônia, do

que nos genomas de outros insetos. Os genes indicam que o relógio biológico das abelhas e outros processos celulares são mais parecidos com os dos vertebrados do que com os de outros insetos. Além disso, elas possui menos genes ligados à imunidade inata, à formação da cutícula (ou exoesqueleto) e ao paladar. Essas diferenças podem ser atribuídas à vida em sociedade desse grupo de insetos, que pode proporcionar alguma proteção contra doenças, ameaças físicas e alimentos venenosos (CONSORTIUM, 2006).

O BCM-HGSC em colaboração com o *The Honey Bee Genome Sequencing Consortium* desenvolveu um banco de dados, chamado de BeeBase (`http://racerx00.tamu.edu/bee_resources.html`), que disponibiliza diversos dados relacionados à *A. mellifera* e ferramentas baseadas no GBrowse (visualização genômica) e no CMAP (mapeamento comparativo).

1.1.4 Transcriptoma

O termo transcriptoma refere-se ao conjunto de genes que são expressos em um dado tecido ou tipos celulares. O perfil do transcriptoma pode variar segundo a fase do desenvolvimento de um organismo, estado fisiológico, estímulos físicos, químicos e biológicos ou em caso de doenças.

O sequenciamento do transcriptoma da abelha foi iniciado pelo grupo da Universidade de Illinois (EUA), com o sequenciamento de 20.256 clones de cDNA de bibliotecas de cérebro de operárias adultas. Após processamento em programas que excluem vetores, leituras de baixa qualidade e sequências menores que duzentos pares de bases, restaram 15.311 ESTs de alta qualidade que foram agrupados em 3.136 *contigs* e 5.830 *singlets*, e depositados no GenBank (WHITFIELD *et al.*, 2002). Em seguida, o nosso laboratório em colaboração com vários laboratórios da Faculdade de Medicina de Ribeirão Preto (FMRP) e da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP), sequenciaram 5.021 ESTs provenientes de bibliotecas produzidas das diferentes fases do desenvolvimento da *A. mellifera*. Na ocasião, o estudo representou a segunda maior contribuição de ESTs de *A. mellifera* depositada no dbEST, além de fornecer dados originais sobre a diversidade de expressão gênica ao longo do desenvolvimento da *A. mellifera* (NUNES *et al.*, 2004).

Atualmente o repositório de sequências *Expressed Sequence Tag* (EST) pertencente ao *National Center for Biotechnology Information* (NCBI), chamado dbEST (BOGUSKI; LOWE; TOLSTOSHEV, 1993), que atualmente possui 78.191 sequências depositadas de *A. mellifera* com dados de diversos centros de estudos.

As sequências do dbEST fazem parte do UniGene, uma base de dados que fornece uma estimativa do número de genes e do seu "nível de expressão" em uma dada espécie. Cada registro contém transcritos vindo do mesmo gene (JU WAGNER L, 2003). O UniGene conta com genes conhecidos depositados no GenBank e ESTs depositados no dbEST e em sua versão *Build* 6 (setembro 2008) conta com 64.502 sequências de *A. mellifera* agrupados em 9.747 *cluster*.

1.2 Análise da expressão gênica

O acúmulo exponencial de sequências gênicas e genomas depositados em bancos de dados públicos tem aumentado consideravelmente a demanda por metodologias que permitam sua caracterização funcional ou confirmação de homologia, além da elucidação dos padrões de expressão (CALSA JR; BENEDITO; FIGUEIRA, 2004).

As análises convencionais para os estudos de expressão gênica como *Northern Blotting* ou transcrição reversa e reação em cadeia da polimerase (RT-PCR), mesmo sendo confiáveis e precisas, são aplicadas a um único gene ou a um conjunto pequeno de genes. Abordagens mais eficientes de análise da expressão gênica em larga escala, tornaram-se um desafio na identificação e no estudo simultâneo de um grande número de genes envolvidos em diversos processos biológicos, desde o desenvolvimento dos organismos até suas interações com fatores ambientais (DONSON *et al.*, 2002). Nesse contexto, as técnicas atuais de genética molecular como cDNA *microarray* (SCHENA *et al.*, 1995), SAGE (VELCULESCU *et al.*, 1995), MPSS (BRENNER *et al.*, 2000) e novas tecnologias de sequenciamento (METZKER, 2005) são adequadas, pois possibilitam uma análise eficiente do transcriptoma de diferentes organismos.

1.2.1 O método de SAGE

O método baseia-se na ideia de que um pequeno fragmento de sequência de nucleotídeos de cerca de dez pares de bases, denominado *tag* (também conhecida como etiqueta, marcador ou assinatura do gene), possui informação suficiente para a identificação unívoca de um transcrito. Essas *tags* são, em grande maioria, localizadas na região 3' *Untranslated Region* (UTR) de cada gene codificador. Uma vez obtidos esses pequenos fragmentos, eles podem ser concatenados e sequenciados, permitindo a análise de múltiplos transcritos contidos em um único clone. Com isso, é possível ter uma ideia da expressão dos transcritos baseando-se apenas na frequência das *tags* (VELCULESCU *et*

al., 1995).

O SAGE é um método eficiente e poderoso na análise de padrões de expressão gênica e possui certas vantagens em relação à técnica de *microarrays*, como o menor custo de sequenciamento por transcrito amostrado, a possibilidade de detecção de transcritos ainda não caracterizados e de transcritos com baixa abundância, dentre outros fatores.

Uma das desvantagens do SAGE é a necessidade de sequenciamento de um grande número de clones para gerar um número de *tags* suficientes para a análise de expressão gênica.

Uma sequência de transcrito com a orientação 5'-3', tem seu *tag* identificador bem definido nas dez bases variáveis características do transcrito adjacentes à enzima de ancoragem (*NlaIII*), posicionada na região mais próxima da extremidade 3' UTR.

Existe uma convenção para definir a orientação de uma molécula de mRNA, que é lê-la a partir da extremidade 5' UTR em direção à extremidade 3' UTR. Em geral os transcritos de seres eucariotos possuem também uma característica conhecida como cauda de Poli-A, isto é, uma sequência repetida em mais de oito bases de adenina (A), que por sua vez permite determinar a orientação do transcrito, pois a extremidade 3' UTR é a extremidade onde se encontra a cauda de Poli-A. Além disso, existe uma sequência sinal, chamada de sinal Poli-A, localizada também na extremidade 3', geralmente nas últimas cinquenta bases, normalmente identificadas pelas sequências de nucleotídeos ATTAAA e AATAAA.

Resumidamente o método de SAGE inicia-se com a produção de cDNAs ligados à partículas magnéticas pela cauda Poli-A. O cDNA total é dividido em dois grupos e, em cada grupo, é ligado um adaptador diferente (na extremidade oposta da cauda poli-A), possibilitando a clivagem com a enzima de etiquetagem (*Bsmf1*), que reconhece o sítio no adaptador e cliva a quatorze pares de bases adjacente ao adaptador. Faz-se a ligação das *tags*, formando as *ditags* e amplificando-as por *Polymerase Chain Reaction* (PCR) através dos *primers* localizados nos adaptadores. Posteriormente, retiram-se os adaptadores e as múltiplas *tags* podem ser concatenadas e sequenciadas, revelando a sequência de milhares de *tags* simultaneamente. Esse resultado é uma estimativa quantitativa e qualitativa da expressão gênica dada pela determinação da abundância de *tags* individuais e a identificação do gene correspondente a cada respectiva *tag* (Figura 4).

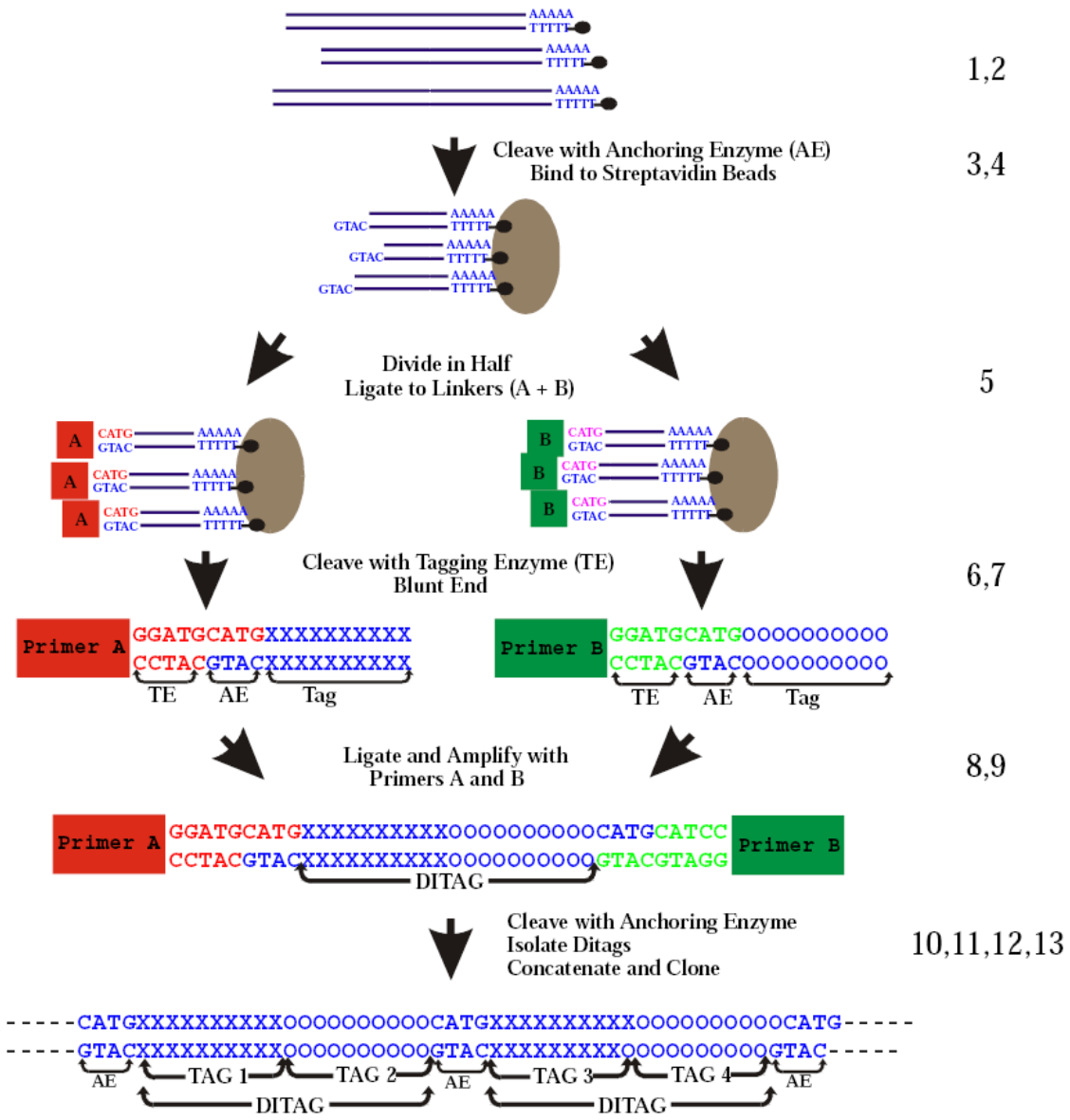


Figura 4: Esquema da abordagem experimental adotada pela técnica de SAGE (Figura obtida do protocolo original disponível em `http://www.sagenet.org`).

As sequências geradas pelo sequenciador estão no formato de cromatograma e devem ser analisadas quanto a qualidade das bases estimando a confiabilidade (EWING *et al.*, 1998). Em seguida, é feita a extração e contagem das *tags*. Existem diversos *softwares* para esse tipo de análise e extração das *tags*, os principais são: SAGE300 (ZHANG *et al.*, 1997), SAGE2000 (I-SAGE *kit*, Invitrogen), eSAGE (MARGULIES; INNIS, 2000) e USAGE (KAMPEN *et al.*, 2000). Outros programas têm sido desenvolvidos para detectar a presença de erros potenciais nos conjuntos de *tags* e suas contagens, analisar comparativamente bibliotecas e facilitar a anotação das *tags* originárias de espécies modelo. São exemplos o POWER_SAGE (MAN; WANG; WANG, 2000), o ExProView (LARSSON *et al.*, 2000) e o

SAGEScreen (AKMAEV; WANG, 2004).

A tecnologia de SAGE fornece a contagem de um dado transcrito, apresentada como a fração relativa ao total de transcritos observados, e não um resultado relativo a outro experimento ou a um gene *housekeeping* particular, como acontece nas técnicas baseadas em hibridação. Isso é uma vantagem, pois evita processos sujeitos a erros como o de normalização entre experimentos. Outra vantagem é que o SAGE determina os níveis de expressão diretamente das amostras de mRNA, não sendo necessário dispor de fragmentos de genes específicos de DNA imobilizados para medir o nível de expressão de cada gene, também necessários nas técnicas baseadas em hibridação. Por outro lado, uma *tag* de dez bases não representa perfeitamente um gene transcrito, com isso, dois ou mais genes podem compartilhar uma mesma *tag* (chamada de *tag* ambígua) e um gene pode ter mais de uma *tag* (através de terminação alternativa do transcrito ou polimorfismo em uma população).

Na técnica original (VELCULESCU *et al.*, 1995) o tamanho da *tag* era de nove bases e a enzima de ancoragem era a *NlaIII*. O protocolo atual de SAGE (ZHANG *et al.*, 1997) gera *tags* de dez a onze bases, e embora a enzima de ancoragem mais utilizada seja a *NlaIII*, é possível utilizar outras enzimas, como a *Sau3A*.

O uso de *tags* maiores, com a técnica *Long* SAGE (*tags* de vinte e um pares de bases), pode ser utilizada para resolver o problema de ambiguidade das *tags* anotadas (WAHL; HEINZMANN; IMAI, 2005; SAHA *et al.*, 2002). Porém existem outras barreiras para o uso de *long* SAGE como o custo elevado, menor eficiência no sequenciamento da *tag* e um aumento significativo da taxa de erro de sequenciamento (AKMAEV; WANG, 2004).

Um simples erro no sequenciamento acarreta um papel importante na produção de *tags*, podendo levar a geração de *tags* com contagem baixa ou poderá aumentar a contagem de outra *tag* existente. Porém, o efeito causado não é de grande preocupação para *tags* com alta contagem, pois quanto maior a contagem da *tag* menor a chance dela ser um erro, portanto maior a chance dela representar verdadeiramente um gene expresso (BEISSBARTH *et al.*, 2004).

1.3 Anotação da biblioteca de SAGE

A anotação das *tags* de um experimento de SAGE é uma importante etapa no processo de análise que nos permite dar sentido biológico aos resultados ao relacionar a *tag* ao gene correspondente.

A anotação das *tags* de SAGE difere da anotação gênica do sequenciamento de ESTs em dois aspectos principais. Em primeiro lugar, o menor tamanho da *tag* de SAGE exige que sua identificação fundamente-se na identidade completa, isto é, para haver associação entre elas é necessário que as dez ou quatorze bases (quatro bases do sítio da enzima de ancoragem e dez bases da *tag*) sejam exatamente iguais à região esperada na sequência completa do transcrito. Em segundo lugar, a confiabilidade da anotação é maior quando esta é conduzida comparativamente a um banco de sequências expressas (cDNAs) ou mesmo à sequência do genoma. Todavia, dados genômicos e ou transcricionais de organismos-modelo filogeneticamente próximos à espécie de interesse podem ser eventualmente utilizados para a obtenção de uma anotação preliminar. Neste contexto existe a base pública do *Gene Index*, com listas de coleções de sequências de cDNA anotadas e sistematicamente depositadas no *The Institute for Genomic Research* (TIGR) `http://www.tigr.org` (CALSA JR; BENEDITO; FIGUEIRA, 2004).

Uma alternativa útil nos casos de organismos modelos já bastante estudados via SAGE é a sua identificação em bibliotecas SAGE já construídas, depositadas e anotadas contra os acessos do UniGene. Essa identificação pode ser realizada no banco de dados público de bibliotecas de SAGE denominado SAGEmap (LASH *et al.*, 2000) disponível no sítio `http://www.ncbi.nlm.nih.gov/sage`, específico para depósito e análise de dados de SAGE no NCBI, integrado à coleção de dados de expressão gênica do GenBank que inclui também informações oriundas de microarrays, Gene Expression Omnibus (GEO) `http://www.ncbi.nlm.nih.gov/geo` (EDGAR; DOMRACHEV; LASH, 2002).

Entretanto, a maioria dos bancos de dados de SAGE disponíveis é de tecido humano e camundongo, assim como as ferramentas *on-line* para gerenciamento, análise e anotação das *tags* de SAGE que também são destinadas a esses organismos. São exemplos o SAGE-Genie (BOON *et al.*, 2002), SAGEnet (`http://www.sagenet.org`) e o The Mouse SAGE Site (DIVINA; FOREJT, 2004). Outros organismos incluindo a *A. mellifera*, ainda não foram anotados e disponibilizados em bancos de dados públicos.

O SAGEmap não dispõe de recursos *on-line* que permitam, por exemplo, realizar a anotação de uma coleção de *tags* em alta escala de maneira similar ao programa MegaBLAST. É possível obter os dados depositados no SAGEmap e compará-los aos dados locais para sua anotação, pois muitas *tags* depositadas no SAGEmap já se encontram associadas a um *cluster* do UniGene já anotado, com uma função presumível atribuída.

A abordagem utilizada para anotar uma biblioteca de SAGE consiste em extrair *tags* computacionalmente do transcriptoma ou genoma, as chamadas *tags* virtuais. A *tag*

virtual é uma sequência de dez pares de bases adjacente ao sítio da enzima *NlaIII* na região 3' UTR, geralmente a mais próxima a esta região. Eventualmente podem ocorrer *tags* internas em um experimento de SAGE.

As *tags* internas podem ser geradas quando ocorrer sinal de poliadenilação alternativa usado para produzir um transcrito curto. Encadeamento alternativo próximo a região 3' e polimorfismos na enzima *NlaIII* também podem produzir transcritos mais curtos. No entanto, as *tags* internas poderiam ser artefatos experimentais. Por exemplo, se uma síntese de cDNA for anelada em algum lugar que não seja a cauda de Poli-A, ou se a digestão for incompleta pela enzima de ancoragem, será produzido um *tag* que não é mais próxima a região 3'.

Após a obtenção das *tags* virtuais deve-se relacionar a *tag* virtual a *tag* de SAGE, obtendo assim a associação entre *tag* e gene.

2 Objetivos

O objetivo deste trabalho foi construir uma metodologia eficiente para a anotação das *tags* de uma biblioteca de SAGE de *A. mellifera*, avaliar a precisão da anotação, modelar uma base de dados relacional para armazenar os dados, selecionar alguns genes para validação experimental e desenvolver uma interface *web* que disponibilizará os resultados gerados.

3 Material e métodos

Esse trabalho foi desenvolvido com base na metodologia descrita no trabalho de (LASH *et al.*, 2000) e implementado na ferramenta chamada SAGEmap, um repositório público de dados de expressão gênica incluindo dados de mapeamento que permite o acesso e análise dos dados *on-line*.

O procedimento inclue gerar a lista de *tags* mais confiáveis (LTC) da biblioteca de SAGE, extrair as *tags* virtuais a partir de bases de dados de transcritos, construir um *ranking* das *tags* encontradas, relacionar *tags* de SAGE aos genes, mapear no genoma da abelha as *tags* de SAGE e validar experimentalmente os resultados. Esse procedimento está representado na Figura 5.

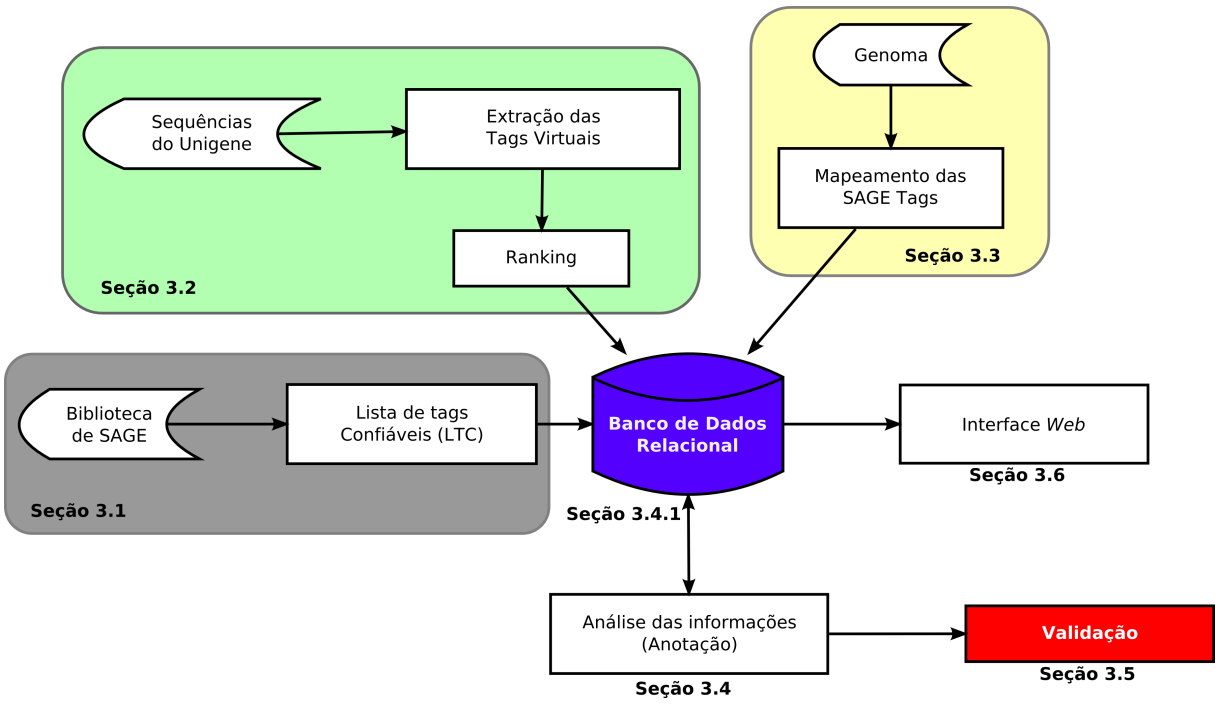


Figura 5: Fluxograma representando as principais etapas do projeto.

Os algoritmos para extrações de informações, inserções e consulta ao banco de dados, foram implementados utilizando a linguagem de programação *Practical Extraction*

and Report Language (PERL) na versão 5.8.8, em associação com alguns módulos dessa linguagem desenvolvidos especificamente para se trabalhar com a análise de sequências gênicas, os módulos do pacote BioPerl (`http://www.bioperl.org`) na versão 1.4. A linguagem de programação PERL oferece uma *Application Program Interface* (API) que facilita a comunicação com a base de dados e proporciona vantagens (quando comparada a outras linguagens de programação) para análise e extração de informações em arquivos no formato texto.

3.1 A biblioteca de SAGE

Foi utilizada uma biblioteca de SAGE de tecido de cérebro de operária de *A. mellifera* adulta gerada pelo Laboratório de Genética Molecular e Bioinformática (LGMB) a partir de amostras cedidas pelo Prof. Dr. Klaus Hartfelder, docente do Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos da FMRP, da Universidade de São Paulo.

A biblioteca foi gerada a partir do RNA total isolado pelo método de TRIzol Reagent® (*Life Technologies*, *Molecular Research Center, Inc.*). Foram obtidos 25 μg de RNA, e a enzima de restrição utilizada foi a *BsmFI* e a enzima de ancoragem foi a *NlaIII*.

Os cromatogramas resultantes do sequenciador foram submetidos ao *software* SAGE2000 na versão 4.5 (`http://sagenet.org`) que realizou a leitura das bases (*base calling*), a extração das *ditags* e contagem das *tags*. Foram utilizados os parâmetros padrão: Enzima de ancoragem *NlaIII*, tamanho da *tag* dez e tamanho da *ditag* vinte e quatro.

A partir desta biblioteca de SAGE, foi gerada uma lista das *tags* mais confiáveis, LTC, obedecendo aos seguintes critérios:

- *Tags* com frequência igual a 1 foram eliminadas;
- *Tags* provenientes de adaptadores de sequências e suas variações de 1 par de bases foram eliminadas;

A listagem de *tags* provenientes de adaptadores de sequências foi obtida pelo sítio do SAGEGenie pertencente ao *Cancer Genome Anatomy Project* (CGAP). Essa listagem contém *tags* com variação de um par de bases (substituição, inserção, ou deleção) das seguintes sequências: TCCCTATTAA e TCCCCGTACA.

3.2 Extração das *tags* virtuais

As *tags* virtuais foram extraídas a partir da base de dados do UniGene de *A. mellifera* na versão *Build* 6 de 10 de setembro de 2008, obtida através do sítio do NCBI `http://www.ncbi.nlm.nih.gov/unigene`.

O algoritmo para extração das *tags* virtuais executou os seguintes procedimentos:

- Verificou a existência de cauda e sinal de Poli-A;
- Buscou o sítio de restrição da enzima *NlaIII* (CATG) mais próximas a região 3' UTR.
- Extraiu as dez bases adjacentes de cada enzima de restrição encontrada (as quatro mais próximas a região 3') de cada transcrito.

A existência do sinal Poli-A aumenta a confiança de que a *tag* extraída é de uma sequência com a região 3' completa. O primeiro indicativo da existência da cauda Poli-A é o sinal Poli-A, que foi identificado através de pelo menos um dos dois sinais mais frequentes nas últimas 50 bases de cada RNA mensageiro (região 3' UTR): AATAAA e ATAAAA. Para este estudo a cauda Poli-A foi identificada por meio de linguagem formal de expressão regular, obedecendo aos seguintes critérios:

- Quando for EST: existência de pelo menos oito bases adenina (A) no final da sequência, podendo conter em seguida e até no máximo três vezes o seguinte padrão: uma ou mais bases As seguidas de uma a cinco bases diferentes de A (C, T, G ou N) seguidas ou não de pelo menos uma base A.
- Quando não for EST: existência de pelo menos quatro bases adenina (A) no final da sequência, podendo conter em seguida e até no máximo três vezes o seguinte padrão: uma ou mais bases As seguidas de uma a cinco bases diferentes de A (C, T, G ou N) seguidas ou não de pelo menos uma base A.

As quatro *tags* mais próximas à região 3' UTR de cada transcrito, foram numeradas de um (a *tag* mais próxima da região 3' UTR) a quatro (a *tag* mais distante da região 3' UTR), como mostra a Figura 6.

Figura 6: Tags extraídas de cada sequência de transcrito e numeradas de 1 (mais próxima à região 3') até 4 (mais distante à região 3').

3.2.1 *Ranking*

Devido à existência de *tags* ambíguas (uma *tag* estar relacionada a mais de um gene ou um gene possuir duas ou mais *tags*), as *tags* virtuais foram classificadas com valores de um a cinco (*ranking*) e priorizadas as *tags* com maior frequência e que pertençam à lista LTC:

1. a *tag* mais próxima à região 3' de mRNAs com cauda poli-A;
2. a *tag* mais próxima à região 3' de ESTs com cauda de poli-A (ou cabeça de poli-T);
3. a *tag* mais próxima à região 3' de mRNAs com sinal de poli-A;
4. a *tag* mais próxima à região 3' de mRNAs sem sinal e sem cauda poli-A;
5. as *tags* internas (as *tags* de número 4, 3 e 2) de mRNAs;

3.3 Mapeamento no genoma

O genoma da *A. mellifera* utilizado nesse estudo está disponível no NCBI (Build 4, versão 1) com data da última atualização (*release*) de 11 de agosto de 2006 (Amel_4.0).

Dessa versão foram extraídas todas as informações relevantes do genoma tais como: coordenadas de regiões codificadoras (CDS), gênicas e de mRNA, juntamente com a orientação, gene id, símbolo, etc.

Todas as *tags* produzidas pela biblioteca de SAGE foram mapeadas no genoma e sua localização (coordenadas) foram armazenadas no banco de dados.

3.4 Relacionamento entre *tag* e gene

3.4.1 Modelagem da base de dados relacional

Todos os dados gerados neste projeto foram armazenados em uma base de dados relacional que foi modelada utilizando a ferramenta DBDesigner `http://fabforce.net/dbdesigner4/` na versão 4.0.5.4 Beta. O modelo do banco está representado na Figura 7.

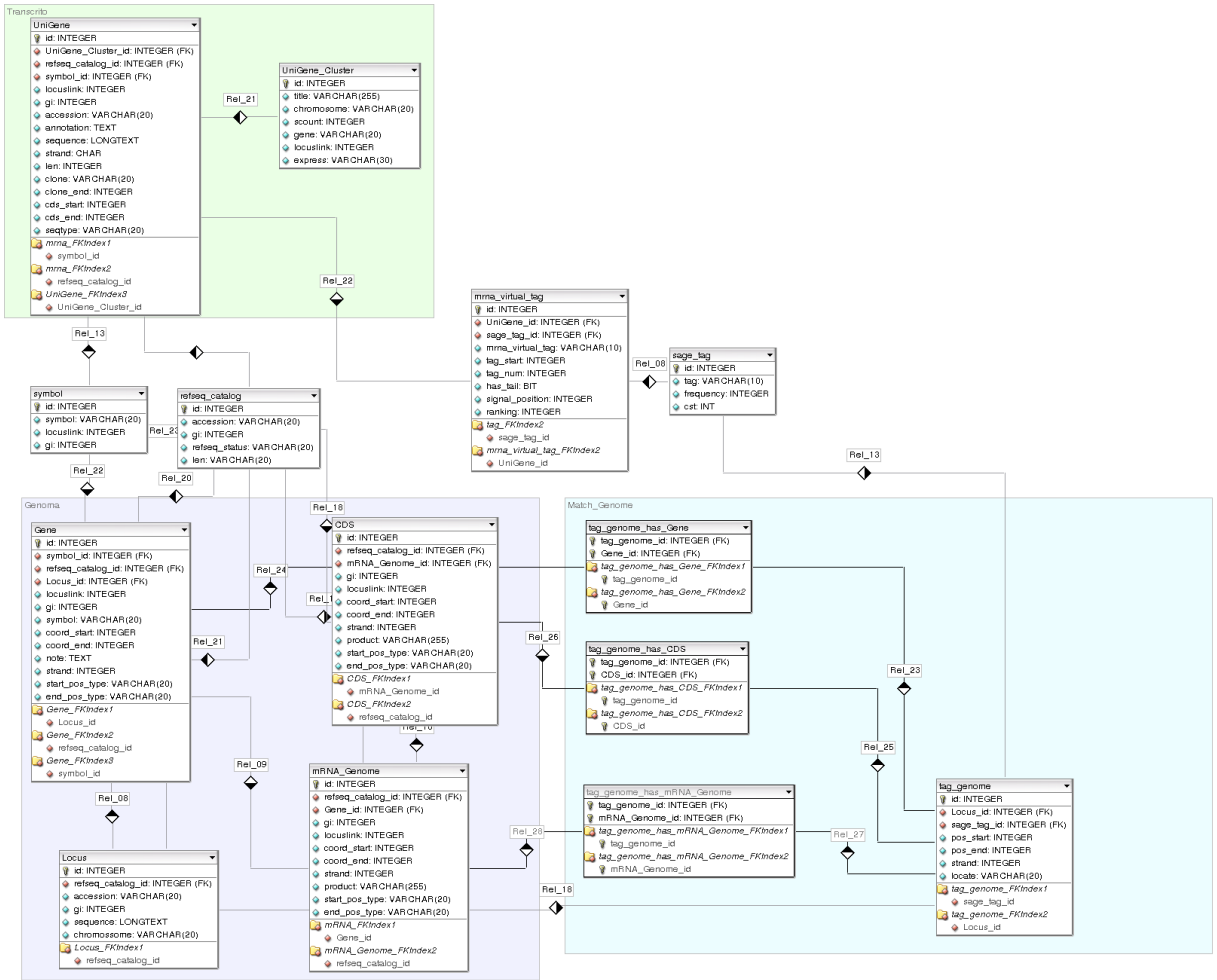


Figura 7: Modelo do banco de dados relacional.

O Sistema de Gerenciamento de Banco de Dados (SGBD) utilizado foi o MySQL (`http://www.mysql.org`) na versão 5.0.32. O MySQL é um sistema bastante empregado na área de bioinformática por ser um sistema gratuito, estável, extremamente eficiente e utiliza uma API de conexão entre o banco de dados e a linguagem de programação utilizada neste projeto (PERL).

3.4.2 Catálogo do RefSeq

O RefSeq é um banco de sequências completas de RNA mensageiro de todos os genes já identificados, além das informações estruturais de cada gene. O RefSeq contem todas as informações de função, mapeamento, associação com doenças, mutações, etc. Todas essas informações foram incluídas no banco de dados. A versão utilizada foi a 31 obtida pelo NCBI.

3.5 Validação dos resultados

3.5.1 *RT-PCR*

Seis genes foram selecionados para a validação experimental: *DEF* (*defensin*), *LOC409906* (*similar to TBP-associated factor 2 CG6711-PA*), *LOC552773* (*similar to CG14661-PA*), *VG* (*Vitellogenin*), *TRF* (*Transferrin*) e *RP49* (*Ribosomal protein 49*, utilizado como controle endógeno).

Para a validação o RNA total de embrião, estágios um e três de larva, pupa, adulto e cérebro de abelha foram isolados pelo método de TRIzol Reagent® (*Life Technologies, Molecular Research Center, Inc.*), segundo instruções do fabricante. A reação de transcrição reversa foi feita a partir de 1 μg de RNA total utilizando-se o *kit High Capacity cDNA Reverse Transcription (Applied Biosystems)*, segundo instruções do fabricante. Os *primers* usados para a validação de cada gene foram desenhados de modo a amplificar os genes com alta especificidade.

As reações de PCR foram realizadas no mesmo termociclador (Perkin Elmer 7700) utilizando 30 ciclos. Os produtos de cada amplificação foram visualizados em gel de agarose 1,2%, corado com brometo de etídeo.

3.6 Interface *web*

Foi criada uma interface *web* para facilitar a visualização dos dados da anotação. Essa ferramenta *web* foi nomeada de STAMP.

A STAMP está disponível na internet através do servidor *web* Apache (`http://www.apache.org`) na versão 2.2 e foi desenvolvida com a linguagem de programação PERL juntamente com o módulo *Common Gateway Interface* (CGI) `http://hoohoo.`

`ncsa.uiuc.edu/cgi/`.

4 Resultados

4.1 Biblioteca de SAGE

A Figura 8 mostra a quantidade de *tags* da biblioteca de SAGE por em intervalos de frequências, como esperado é uma biblioteca em que a maioria das *tags* é pouco frequente.

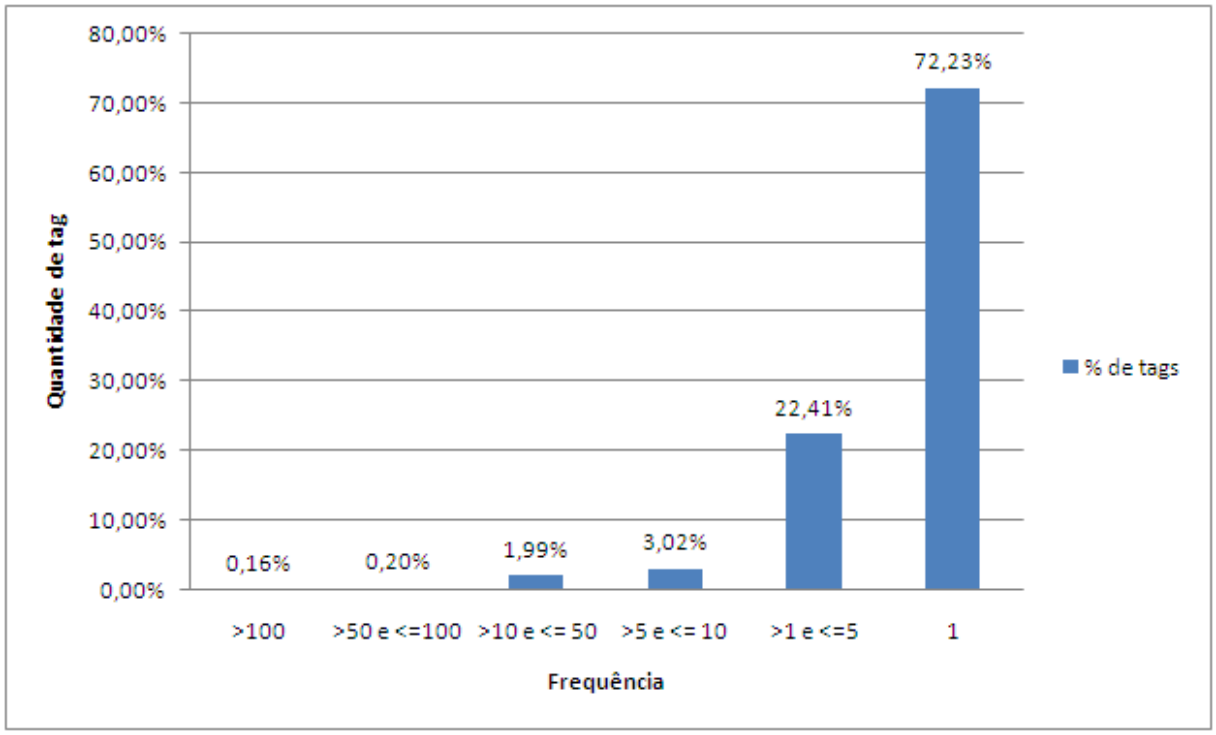


Figura 8: Quantidade de *tags* da biblioteca de SAGE existentes em intervalos de frequências.

A taxa de contaminação por adaptadores de sequências na biblioteca de SAGE foi de 0,02%, esse valor é equivalente a de outras bibliotecas publicadas (VELCULESCU *et al.*, 1997).

Foram geradas 60.536 *tags*, sendo que após a análise de redundância, obtivemos 20.483 *tags* únicas. Com a lista LTC reduzimos o número de *tags* para 45.674, sendo 5.683 foram *tags* únicas.

4.2 Extração das *tags* virtuais

Foram extraídas 93.200 *tags* virtuais do UniGene incluindo as quatro *tags* possíveis para cada transcrito. Comparando as *tags* virtuais extraídas com as *tags* da biblioteca de SAGE, identificamos correspondência com 3.859 *tags* únicas, sendo 1.563 *tags* classificadas como LTC. A Figura 9 informa a quantidade de *tags* virtuais encontradas nas quatro posições de cada transcrito e o total de *tags* correspondentes geradas pela biblioteca de SAGE.

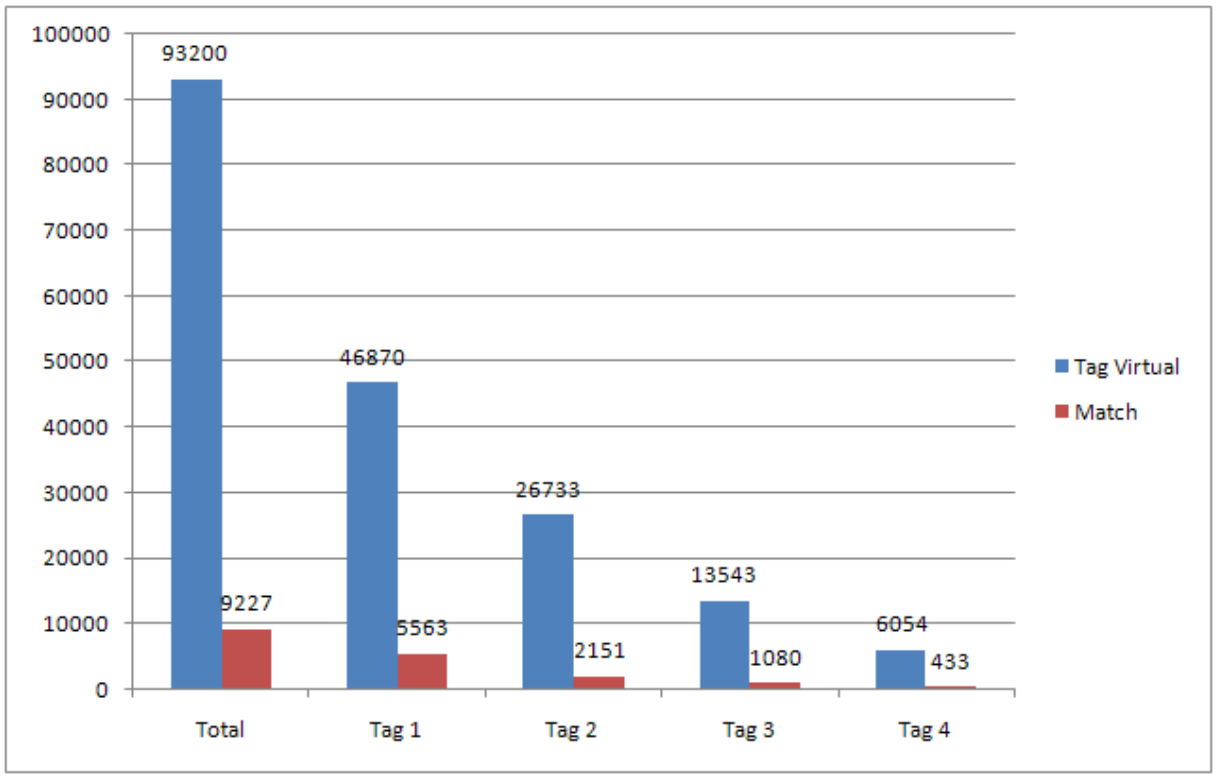


Figura 9: Quantidade de *tags* virtuais encontradas e a quantidade de *tags* virtuais anotadas nas quatro *tags* de cada transcrito.

Foram anotadas 27,5% das *tags* da lista LTC, enquanto que 19% da biblioteca total de SAGE foi anotada. A quantidade de *tags* de SAGE que não encontramos uma *tag* virtual relacionada foi de 81% e considerando apenas as *tags* da lista LTC foi de 72,5% (Figura 10).

As *tags* da lista LTC foram mapeadas contra a base de dados de Unigene desse organismo e verificada a existência de cauda e sinal de Poli-A (Tabela 3). Podemos observar que ocorreu uma maior quantidade de anotação quando existiu o sinal e a cauda de Poli-A e em menor quantidade quando não existia a cauda e o sinal de Poli-A.

Na Figura 11, A e B representam a relação de todos os clusters do UniGene rela-

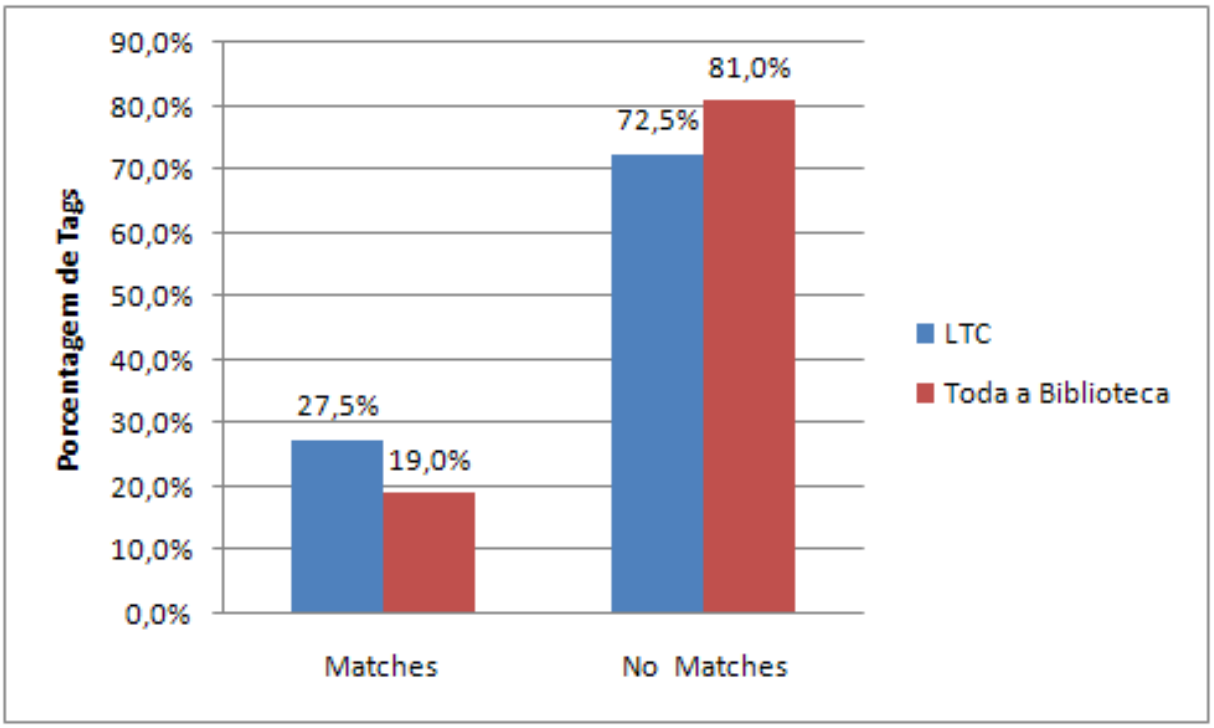


Figura 10: Quantidade de *tags* de SAGE anotadas (*matches*) e as não anotadas (*no matches*).

Tabela 3: Representação das *tags* anotadas de acordo com a existência de cauda e sinal de Poli-A.

Total	**Sinal Poli(A)**	**Cauda Poli(A)**
14,71%	Sim	Sim
6,67%	Sim	Não
8,33%	Não	Sim
4,68%	Não	Não

cionados com as *tags* da lista LTC, mostrando uma grande quantidade de ambiguidade provenientes de erros de sequenciamento de EST, e C e D representam a relação somente dos *clusters* do UniGene de mRNA bem caracterizados e relacionados com as *tags* da lista LTC, mostrando um aumento significativo na relação de um para um e uma diminuição no relacionamento maior que quatro.

As dez *tags* mais expressas na biblioteca de SAGE anotadas (associada a um gene), são mostradas na Tabela 4.

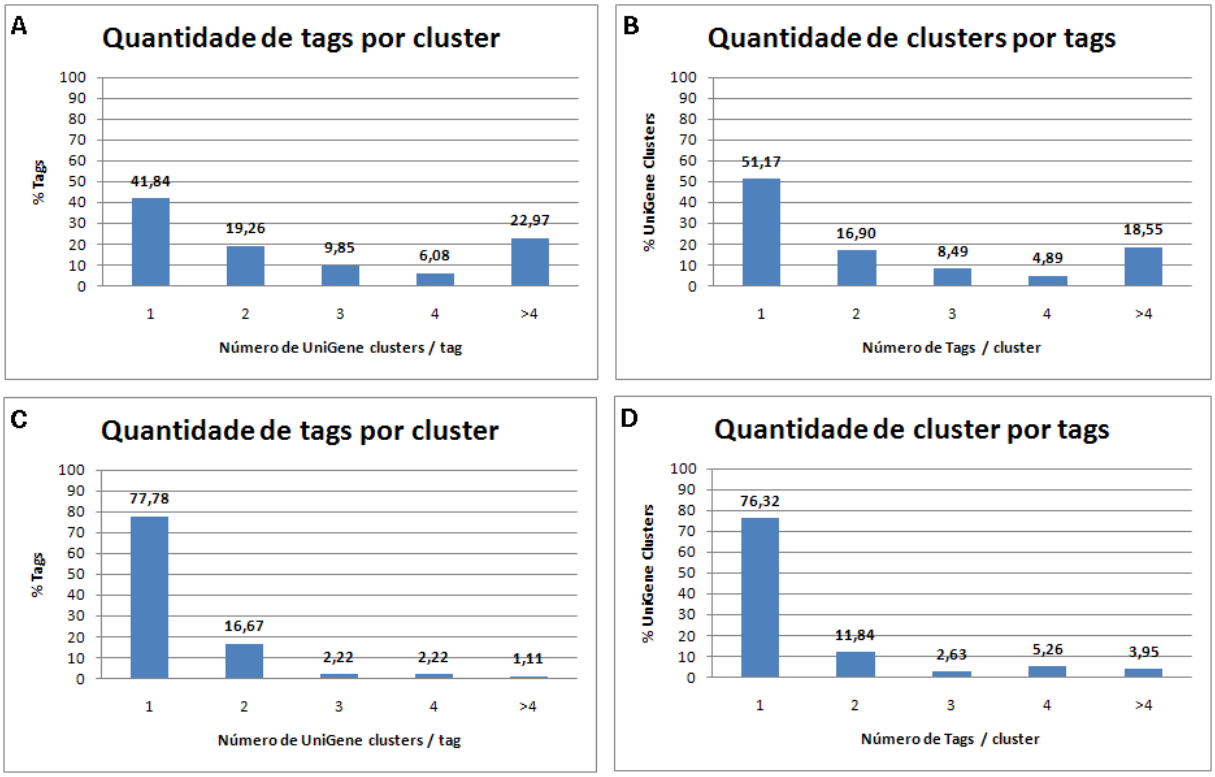


Figura 11: Porcentagem de anotação de *tags* por gene e genes por *tag* respectivamente. (A,B) todos os UniGene *cluster* contra a lista LTC e (C,D) somente mRNA bem caracterizados do UniGene contra a lista LTC.

Tabela 4: Lista de anotação das dez *tags* mais frequentes da biblioteca.

Tag	Frequência	UniGene Cluster	Símbolo	RefSeq Accession	Cromossomo
AAGATTAAGG	1733	Ame.208	*Mrjp1*	NM_001011579	LG11
TTGGTCAGCG	1657	Ame.4275	*LOC406093*	DB756321	LG6
GCAGACCATA	1451	Ame.209	*Mrjp2*	NM_001011580	LG11
TTAGGATGCG	528	Ame.718		DB779024	LG7
TTTTTGATAA	476	Ame.11104		DB758412	MT
TAAGATTTTA	384	Ame.11104		DB757091	MT
CCGAATGTAA	329	Ame.676		BI515316	LG14
TAACACGTTC	265	Ame.10949		DB735692	LG2
AGGGATCTGG	214	Ame.808	*LOC406081*	NM_001011574	LG5
CTAGCGATCA	185	Ame.208	*Mrjp1*	BI515543	LG11

4.3 *Ranking*

As *tags* anotadas foram classificadas por um *ranking* de 1 até 5 e a quantidade de *tags* em cada posicão no *ranking* são mostradas na Tabela 5. A maioria das *tags* foi classificada com valor 3 pelo *ranking*, que são aquelas provenientes de mRNA com sinal Poli-A mas sem a cauda Poli-A. As *tags* com classifcadas com valor 5 (*tags internas*), que

são as menos confiáveis, tiveram quantidade significante.

Tabela 5: Quantidade de *tags* da lista LTC anotadas no UniGene em cada posição no *ranking*.

Ranking	**Lista LTC**
1	1%
2	1,4%
3	60.4%
4	0,2%
5	37%

As dez *tags* mais frequentes com melhor pontuação no *ranking* foram anotadas como mostra a Tabela 6.

Tabela 6: Lista das 10 *tags* mais frequentes com melhor pontuação no *ranking*.

Tag	**Frequência**	**UniGene Cluster**	**Símbolo**	**RefSeq Status**	**RefSeq Accession**	**Cromossomo**
AAAAAAAAAA	109	Ame.1999	*LOC409906*	MODEL	XM_393397	LG3
AATTAAAAAT	99	Ame.3	*LOC406145*	VALIDATED	NM_001011618	LG5
GACCACTTGA	90	Ame.4717	*LOC409589*	MODEL	XM_393092	LGUn
GTAGACCAGG	69	Ame.7534	*LOC551644*	MODEL	XM_624035	LG7
CGATAATCGA	43	Ame.4921	*LOC408284*	MODEL	XM_391836	LG10
AAAAATAAAA	29	Ame.3184	*LOC411112*	MODEL	XR_015005	LG5
CACGTGTTAA	23	Ame.1172	*LOC408289*	MODEL	XM_391841	LG10
TATTAATAAA	17	Ame.8746	*LOC724210*	MODEL	XM_001119980	LG2
TTGGACGTTT	15	Ame.7143	*LOC411515*	MODEL	XM_394987	LG4
AACGAAGAAA	14	Ame.3891	*LOC724428*	MODEL	XM_001120284	LG4

Foi gerada uma lista de melhor *tag* para o gene (*Best Tag*) no qual os genes (UniGene *cluster*) são únicos, podendo haver uma mesma *tag* relacionada a um ou mais genes diferentes, e resultou em 3.640 UniGene *clusters* com uma *tag* relacionada (Tabela 7).

A lista de melhor gene para a *tag* (*Best Gene*) também foi gerada. Nessa lista as *tags* são únicas podendo haver um mesmo gene relacionado a uma ou mais *tags* diferentes, e resultou em 3.859 *tags* com um UniGene *cluster* relacionado (Tabela 8).

4.4 Mapeamento no genoma

Cada *tag* da lista LTC foram mapeadas nas duas orientações do genoma da abelha, resultando em um total de 21.538 *tags* encontradas (*matches*) sendo 4.863 *tags* únicas. O total de 820 *tags* não foram encontradas no genoma com a localização cromossômica definida (Figura 12).

Dentre as *tags* da lista LTC que foram anotadas no UniGene, 94% delas também foram encontradas no genoma, e apenas 347 dessas *tags* apresentam RefSeq *status* definidos e são mostradas na Tabela 9.

Tabela 7: Os primeiros 30 registros da lista de melhor *tag* para o gene (*Best tag*).

UniGene cluster	Tag	Símbolo	Nome do Cluster	Cromossomo
Ame.1	AAGGTCTCAA	*Asp3c*	Antennal-specific protein 3c	LG5
Ame.3	AATTAAAAAT	*LOC406145*	Secapin	LG5
Ame.5	AAAGAATTAA	*LOC552632*	Similar to Ribosomal protein L22 CG7434-PA	LG14
Ame.9	TAATTTCAAT	*LOC410837*	Similar to Microsomal glutathione S-transferase-like CG1742-PA, isoform A	LG2
Ame.10	TAATTTCAAT	*LOC410791*	Similar to uncoupling protein 2	LG1
Ame.12	TTCAATAAAT	*LOC408695*	Hypothetical LOC408695	LG2
Ame.14	AAAAAAGGGA	*LOC409090*	Similar to apontic CG5393-PB, isoform B	LG15
Ame.15	TTCCTTTAAT	*LOC550645*	Similar to Phosphorylase kinase CG1830-PB, isoform B	LG1
Ame.19	TTAACAGAAG	*LOC550975*	Hypothetical LOC550975	LGUn
Ame.21	TAATGACAAA	*LOC725382*	Chemosensory protein 1	LG8
Ame.23	GAAAATATGT	*LOC409410*	Similar to vasa intronic gene CG4170-PA, isoform A	LGUn
Ame.31	CAAAATACTT		Transcribed locus	LG1
Ame.37	TAAATTTAAT	*LOC408980*	Similar to AlkB CG33250-PA	LG9
Ame.43	TATCCTTGTT	*LOC409263*	Similar to CG9520-PB, isoform B	LG6
Ame.50	CTTTAAATTG	*LOC725249*	Similar to T06E6.10	LG15
Ame.51	CTTTAAATTG	*LOC726262*	Similar to Adenylyl cyclase 78C CG10564-PA	LGUn
Ame.52	TAATATATAT		Transcribed locus	LG1
Ame.60	GCGTATCGTG		Transcribed locus	LG15
Ame.62	AAACAGACAC	*LOC409733*	Similar to Myb protein	LG1
Ame.68	CTATGGTATG		Transcribed locus	LG1
Ame.74	TGCCGAAAGA	*LOC726321*	Similar to CG7781-PA	LG10
Ame.84	GAATGTCGGG		Transcribed locus	LG1
Ame.86	TGACAAGAAC		Transcribed locus	LG7
Ame.87	GACCCAGGCA		Transcribed locus	LG11
Ame.88	GATCGTCTCT		Transcribed locus	LG1
Ame.90	TTATGTAAAA		Transcribed locus	LG1
Ame.95	AGGAATGTGG		Transcribed locus	LG16
Ame.96	ACTTAATATT		Transcribed locus	LG4
Ame.97	ATATGAATTT	*LOC411724*	Similar to CG5508-PA, isoform A	LG15
Ame.99	TATACAATCT	*LOC413108*	Similar to CG3542-PA, isoform A	LG11

Tabela 8: Os primeiros 30 registros da lista de melhor gene para a *tag* (*Best Gene*).

UniGene cluster	Tag	Símbolo	Nome do Cluster	Cromossomo
Ame.1	AAGGTCTCAA	Asp3c	Antennal-specific protein 3c	LG5
Ame.3	AATTAAAAAT	LOC406145	Secapin	LG5
Ame.5	GGTGGTCGGT	LOC552632	Similar to Ribosomal protein L22 CG7434-PA	LG14
Ame.5	TAAATATTTT	LOC552632	Similar to Ribosomal protein L22 CG7434-PA	LG14
Ame.5	AAAGAATTAA	LOC552632	Similar to Ribosomal protein L22 CG7434-PA	LG14
Ame.9	TAATTTCAAT	LOC410837	Similar to Microsomal glutathione S-transferase-like CG1742-PA, isoform A	LG2
Ame.9	TAAAATAATA	LOC410837	Similar to Microsomal glutathione S-transferase-like CG1742-PA, isoform A	LG2
Ame.10	TAGAATATTT	LOC410791	Similar to uncoupling protein 2	LG1
Ame.10	AAAAAAAAAA	LOC410791	Similar to uncoupling protein 2	LG1
Ame.10	CACTATCTAC	LOC410791	Similar to uncoupling protein 2	LG1
Ame.12	ATGGTGGTTA	LOC408695	Hypothetical LOC408695	LG2
Ame.12	GCCAACTCGG	LOC408695	Hypothetical LOC408695	LG2
Ame.12	AACCTGTTCT	LOC408695	Hypothetical LOC408695	LG2
Ame.12	TTCAATAAAT	LOC408695	Hypothetical LOC408695	LG2
Ame.14	TATTTACTCT	LOC409090	Similar to apontic CG5393-PB, isoform B	LG15
Ame.14	AAAAAAGGGA	LOC409090	Similar to apontic CG5393-PB, isoform B	LG15
Ame.14	AGCGGCTTCC	LOC409090	Similar to apontic CG5393-PB, isoform B	LG15
Ame.15	AAGGAAAGGA	LOC550645	Similar to Phosphorylase kinase CG1830-PB, isoform B	LG1
Ame.15	TTCCTTTAAT	LOC550645	Similar to Phosphorylase kinase CG1830-PB, isoform B	LG1
Ame.19	TTAACAGAAG	LOC550975	Hypothetical LOC550975	LGUn
Ame.21	GTGTTTGAAA	LOC725382	Chemosensory protein 1	LG8
Ame.21	TAATGACAAA	LOC725382	Chemosensory protein 1	LG8
Ame.21	ATATAATTAA	LOC725382	Chemosensory protein 1	LG8
Ame.23	GAAAATATGT	LOC409410	Similar to vasa intronic gene CG4170-PA, isoform A	LGUn
Ame.31	CAAAATACTT		Transcribed locus	LG1
Ame.37	TAAATTTAAT	LOC408980	Similar to AlkB CG33250-PA	LG9
Ame.43	TATCCTTGTT	LOC409263	Similar to CG9520-PB, isoform B	LG6
Ame.50	CTTTAAATTG	LOC725249	Similar to T06E6.10	LG15
Ame.52	TAATATATAT		Transcribed locus	LG1
Ame.60	GCGTATCGTG		Transcribed locus	LG15

Das 89 *tags* (6%) anotadas no UniGene que não foram encontradas no genoma, a maioria delas (66 *tags*) foram classificadas no *ranking* com posição 3, 17 *tags* com posição 5, 2 *tags* com posição 2 e 4 *tags* com posição 1, como mostra a Figura 13. Somente uma

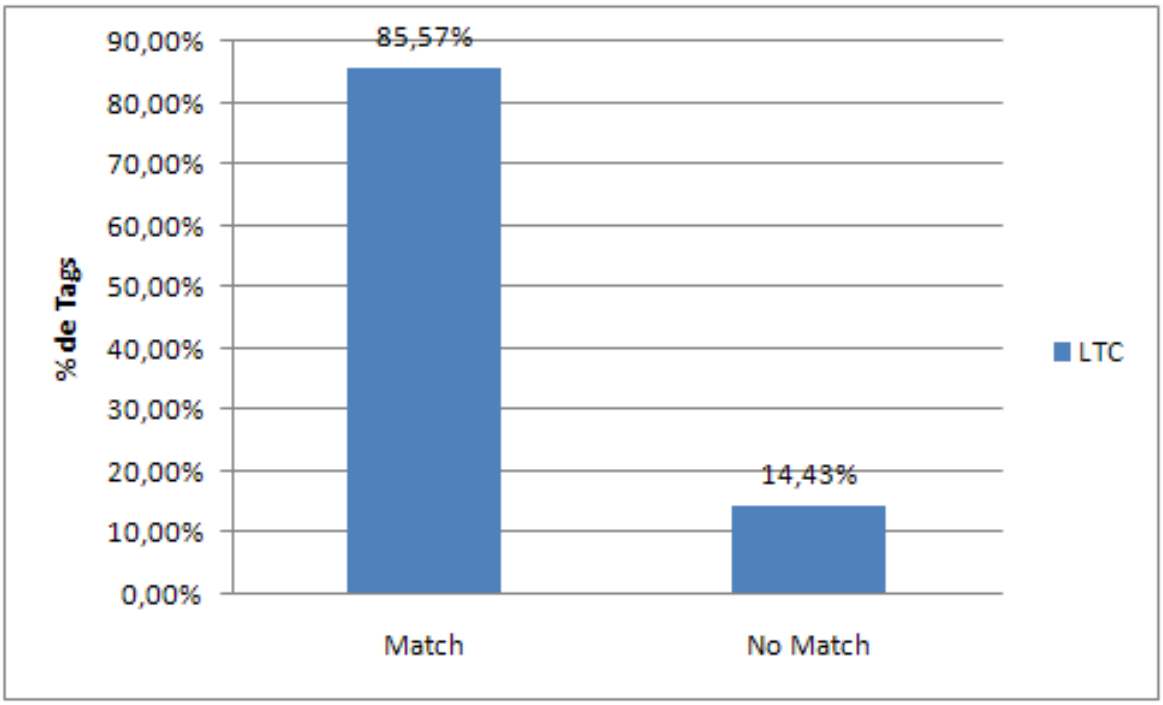


Figura 12: Lista LTC mapeadas no Genoma.

Tabela 9: Relação das *tags* encontradas no UniGene e no genoma com o RefSeq *status* definido.

Quantidade de Tags	**RefSeq Status**
7	Validada
1	Predita
35	Provisória
0	Inferida
304	Modelo
0	na

tag foi classificada na posiçao 1 do *ranking* e com expressao alta, e essa *tag* está no grupo de contigs não localizados (LGUn).

Todas as *tags* LTC que foram encontradas no genoma da abelha, 63% delas também foram encontradas no UniGene.

Das *tags* da lista LTC encontradas no genoma, 70% não foram encontradas no UniGene.45% dessas *tags* foram encontradas em região gênica predita computacional pelo método GNOMON no genoma, essas *tags* são evidências de que esses genes preditos são verdadeiros. A Tabela 10 mostra a lista das 10 *tags* mais frequentes anotadas somente no genoma em região gênica.

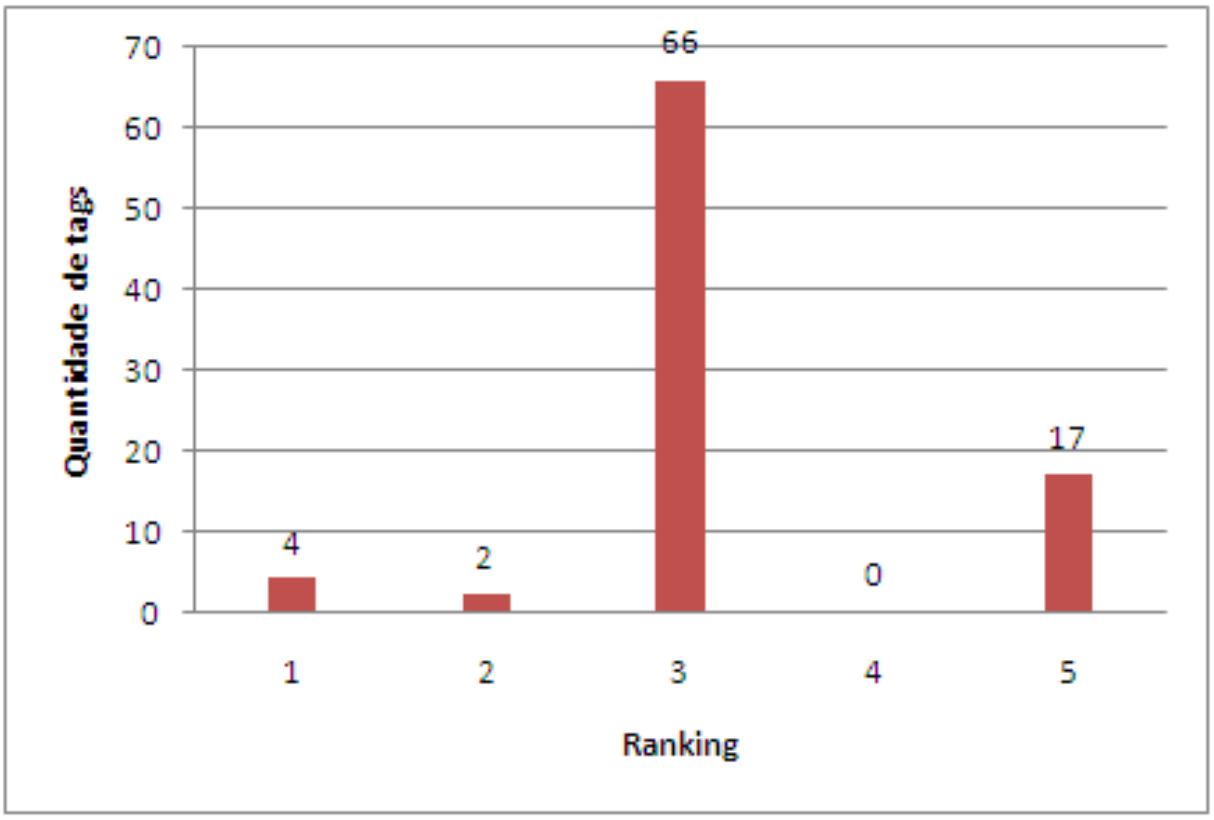


Figura 13: Relação das *tags* anotadas no UniGene não encontradas no genoma de acordo com suas respectivas posições no *ranking*.

Tabela 10: Relação das 10 *tags* mais frequentes anotadas somente no genoma em região gênica.

Tag	**Frequência**	RefSeq *Accession*	Símbolo	Cromossomo
ATTTGAGTTC	5505	NW_001253392	LOC725799	LG6
AACAGTTCCA	931	NW_001253008	LOC412893	LG11
GAAAATCTTC	642	NW_001252994	Mrjp4	LG11
TCCTGTTGAT	464	NW_001253056	LOC413224	LG12
AAATCCTGTT	193	NW_001253123	LOC410633	LG15
TTAATAAAAT	123	NW_001253484	LOC408922	LG7
ATACATACTA	121	NW_001252986	LOC726895	LG11
TAGTCCGATA	93	NW_001253479	LOC408928	LG7
ATAAAAAAAA	91	NW_001253142	LOC724849	LG15
GTAAATAAAT	91	NW_001253180	LOC725852	LG1

4.5 Validação dos resultados

Para validarmos experimentalmente nossos resultados utilizamos a técnica de *Reverse Transcriptase PCR* (RT-PCR). Utilizamos dois grupos de genes para validação: genes cujos primes já existiam em nosso laboratório (Figura 14 A), e genes cujos *primers* foram desenhados baseados na classificação do *ranking* (classificadas com valor 1) e pela sua

alta taxa de expressão que apresentaram no mapeamento (Figura 14 B). Como controle endógeno utilizamos o gene *RP49*, o qual foi mapeado em nossa biblioteca com uma *tag* de frequência 66 e, como esperado, mostrou-se expresso em todas as fases de desenvolvimento da *A. mellifera*. O gene da transferrina (*TRF*) também se mostrou expresso em todas as fases do desenvolvimento. Os genes vitelogenina (*VG*) e *LOC552773* apresentaram expressão variada dependendo da fase em que se encontravam. O gene *LOC552773* não teve expressão em nosso experimento em cérebro e também não foi encontrado em nossa biblioteca de SAGE. Os genes da defensina (*DEF*) e *LOC409906*, foram selecionados por apresentarem maior confiabilidade em nosso mapeamento, e os resultados confirmaram a elevada expressão no tecido cérebral.

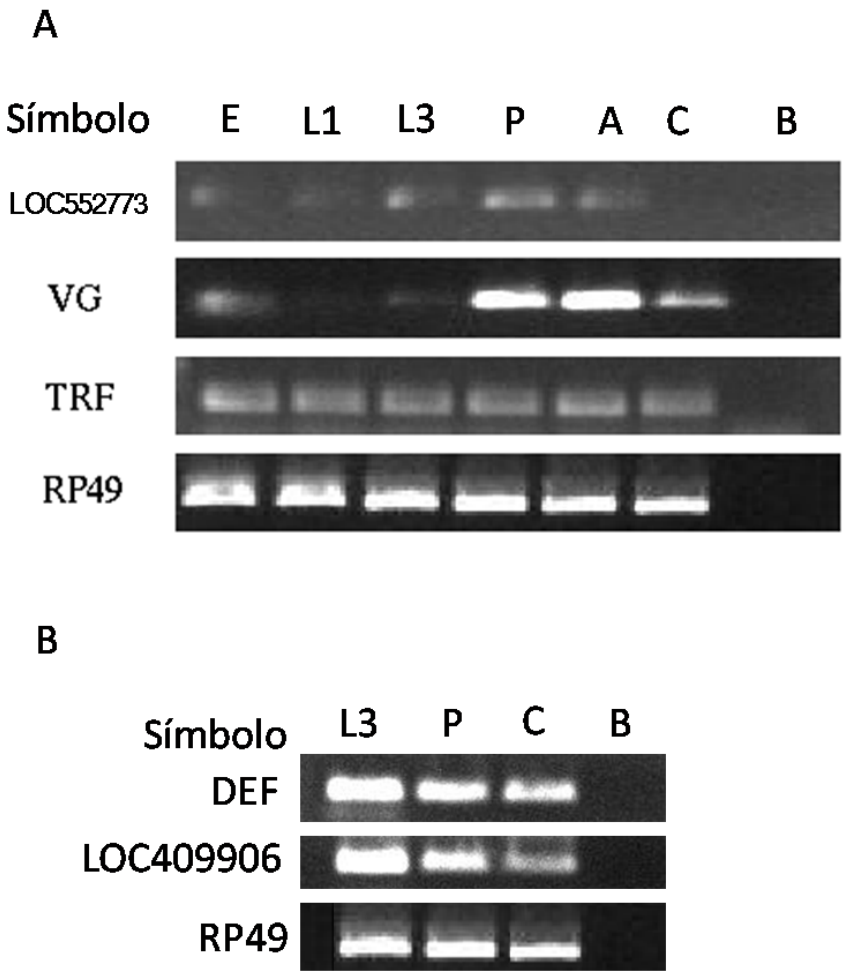


Figura 14: Foto do gel agarose 1,2%. E - embrião; L1 - larva 1; L3 - larva 3; P - pupa; A - adulto; C - cérebro; B - Branco; (A) Genes cujos *primers* já se encontravam no laboratório. (B) Genes selecionados cujos *primers* foram desenhados baseados na classificação no *ranking* e valor de expressão.

A Tabela 11 mostra os genes utilizados na validação experimental e suas frequências correspondentes na biblioteca de SAGE.

Tabela 11: Genes utilizados na validação experimental e suas frequências correspondentes na biblioteca de SAGE.

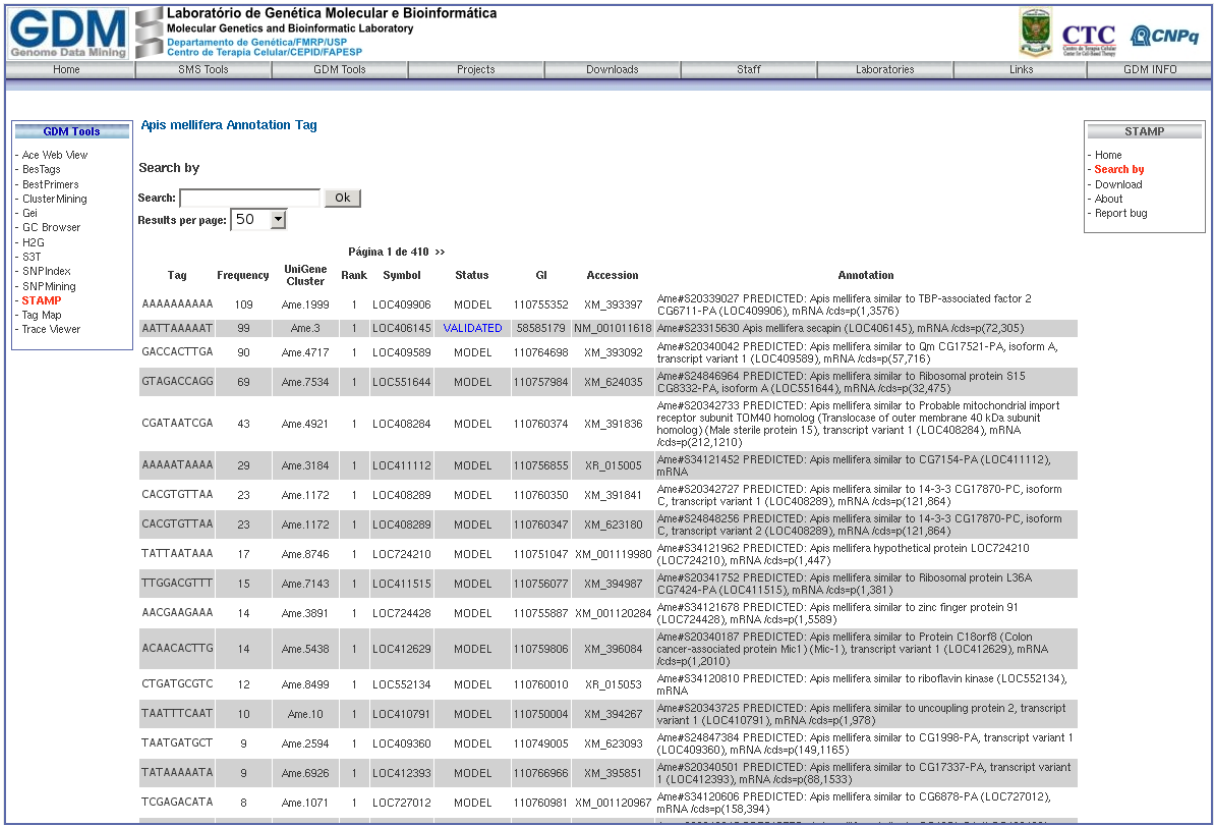
Gene	**Frequência na biblioteca**
LOC552773	0
VG	2
TRF	20
RP49	66
DEF	130
LOC409906	109

4.6 Interface *web*

A interface *web*, STAMP, está integrada ao site de ferramentas do LGMB, o *Genome Data Mining* (GDM), disponível no endereço `http://gdm.fmrp.usp.br`.

O STAMP disponibiliza a lista de todas as *tags* de SAGE anotadas e ordenadas pela a classificação das *tags* no *ranking* (seção *Search by*), permitindo a consulta ao mapeamento do símbolo do gene ou pela *tag* (Figura 15). O resultado da lista ou consulta de *tags* mostra a associação de *tag* com o gene, a frequência, o UniGene *Cluster*, a posição no *ranking*, o RefSeq *Status*, o *locuslink*, o RefSeq *Accession*, o cromossomo e a descrição.

Os arquivos das listas de melhor *tag* para o gene (*best tag*) e melhor gene para a *tag* (*best gene*) estão disponíveis na seção de *Download* protegido por senha em arquivo texto no formato tabular (Figura 16).



Tag	Frequency	UniGene Cluster	Rank	Symbol	Status	GI	Accession	Annotation
AAAAAAAAAA	109	Ame.1999	1	LOC409906	MODEL	110755352	XM_393397	Ame#S20339027 PREDICTED: Apis mellifera similar to TBP-associated factor 2 CG6711-PA (LOC409906), mRNA /cds=p(1,3576)
AATTAAAAAT	99	Ame.3	1	LOC406145	VALIDATED	58585179	NM_001011618	Ame#S23315630 Apis mellifera secapin (LOC406145), mRNA /cds=p(72,305)
GACCACTTGA	90	Ame.4717	1	LOC409589	MODEL	110764698	XM_393092	Ame#S20340042 PREDICTED: Apis mellifera similar to Qm CG17521-PA, isoform A, transcript variant 1 (LOC409589), mRNA /cds=p(57,716)
GTAGACCAGG	69	Ame.7534	1	LOC551644	MODEL	110757984	XM_624035	Ame#S24846964 PREDICTED: Apis mellifera similar to Ribosomal protein S15 CG8332-PA, isoform A (LOC551644), mRNA /cds=p(32,475)
CGATAATCGA	43	Ame.4921	1	LOC408284	MODEL	110760374	XM_391836	Ame#S20342733 PREDICTED: Apis mellifera similar to Probable mitochondrial import receptor subunit TOM40 homolog (Translocase of outer membrane 40 kDa subunit homolog) (Male sterile protein 15), transcript variant 1 (LOC408284), mRNA /cds=p(212,1210)
AAAAATAAAA	29	Ame.3184	1	LOC411112	MODEL	110756855	XR_015005	Ame#S34121452 PREDICTED: Apis mellifera similar to CG7154-PA (LOC411112), mRNA
CACGTGTTAA	23	Ame.1172	1	LOC408289	MODEL	110760350	XM_391841	Ame#S20342727 PREDICTED: Apis mellifera similar to 14-3-3 CG17870-PC, isoform C, transcript variant 1 (LOC408289), mRNA /cds=p(121,864)
CACGTGTTAA	23	Ame.1172	1	LOC408289	MODEL	110760347	XM_623180	Ame#S24848256 PREDICTED: Apis mellifera similar to 14-3-3 CG17870-PC, isoform C, transcript variant 2 (LOC408289), mRNA /cds=p(121,864)
TATTAATAAA	17	Ame.8746	1	LOC724210	MODEL	110751047	XM_001119980	Ame#S34121962 PREDICTED: Apis mellifera hypothetical protein LOC724210 (LOC724210), mRNA /cds=p(1,447)
TTGGACGTTT	15	Ame.7143	1	LOC411515	MODEL	110756077	XM_394987	Ame#S20341752 PREDICTED: Apis mellifera similar to Ribosomal protein L36A CG7424-PA (LOC411515), mRNA /cds=p(1,381)
AACGAAGAAA	14	Ame.3891	1	LOC724428	MODEL	110755887	XM_001120284	Ame#S34121678 PREDICTED: Apis mellifera similar to zinc finger protein 91 (LOC724428), mRNA /cds=p(1,5589)
ACAACACTTG	14	Ame.5438	1	LOC412629	MODEL	110759806	XM_396084	Ame#S20340187 PREDICTED: Apis mellifera similar to Protein C18orf8 (Colon cancer-associated protein Mic1) (Mic-1), transcript variant 1 (LOC412629), mRNA /cds=p(1,2010)
CTGATGCGTC	12	Ame.8499	1	LOC552134	MODEL	110760010	XR_015053	Ame#S34120810 PREDICTED: Apis mellifera similar to riboflavin kinase (LOC552134), mRNA
TAATTTCAAT	10	Ame.10	1	LOC410791	MODEL	110750004	XM_394267	Ame#S20343725 PREDICTED: Apis mellifera similar to uncoupling protein 2, transcript variant 1 (LOC410791), mRNA /cds=p(1,978)
TAATGATGCT	9	Ame.2594	1	LOC409360	MODEL	110749005	XM_623093	Ame#S24847384 PREDICTED: Apis mellifera similar to CG1998-PA, transcript variant 1 (LOC409360), mRNA /cds=p(149,1165)
TATAAAAATA	9	Ame.6926	1	LOC412393	MODEL	110766966	XM_395851	Ame#S20340501 PREDICTED: Apis mellifera similar to CG17337-PA, transcript variant 1 (LOC412393), mRNA /cds=p(88,1533)
TCGAGACATA	8	Ame.1071	1	LOC727012	MODEL	110760981	XM_001120967	Ame#S34120606 PREDICTED: Apis mellifera similar to CG6878-PA (LOC727012), mRNA /cds=p(158,394)

Figura 15: Seção de consulta (por símbolo ou *tag*) à base de dados de anotação (*Search by*).

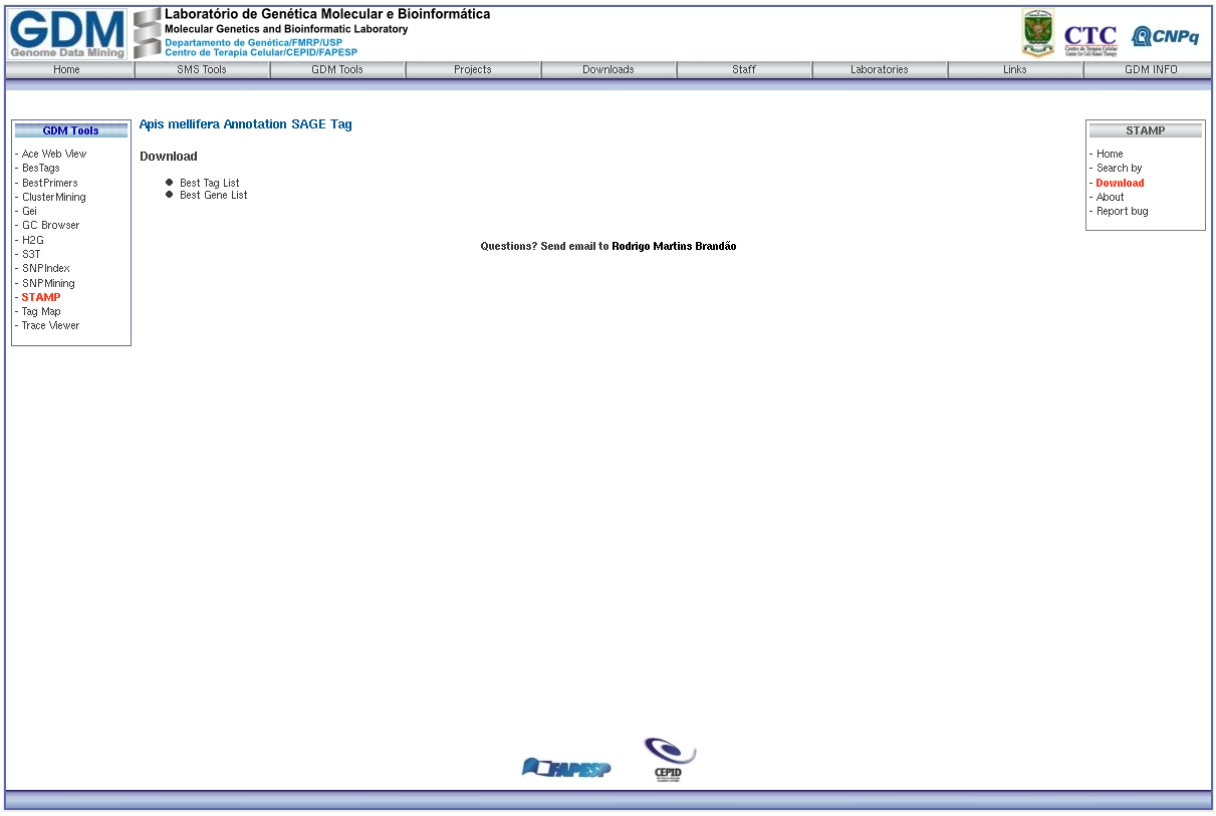


Figura 16: Seção de *download* da interface *web*.

5 Discussão

A abordagem utilizada para anotação de biblioteca de SAGE é similar a utilizada no SAGEmap. Trabalhos semelhantes já foram publicados para outros organismos (humano, plantas e etc.), entretanto este é o primeiro trabalho a anotar uma biblioteca de SAGE de *A. mellifera*. Apesar de ser baseada na metodologia do SAGEmap, a bordagem empregada no presente estudo difere nos seguintes aspectos:

- Ao contrário de outros organismos que possuem diferentes bases de dados públicas de transcritos para o mapeamento e anotação das *tags*, foi utilizado apenas a base de dados do UniGene.
- O SAGEmap considera como cauda de poli-A apenas se as últimas oito bases do transcrito na região 3' UTR forem bases adenina (A). Nesse trabalho foram consideradas possíveis variações que possam ocorrer nessas oito bases de adenina.
- Foi realizado o mapeamento das *tags* de SAGE no genoma inteiro de *A. mellifera*.

5.1 Biblioteca de SAGE

Esse trabalho utilizou uma biblioteca de SAGE de cérebro de *A. mellifera* produzida em nosso laboratório, que consiste da primeira biblioteca de SAGE de *A. mellifera* descrita até a presente data. Para obter resultados mais precisos da anotação das *tags*, seria importante aumentar o número de bibliotecas diminuindo assim os erros associados a *tags* de frequência baixa. Não foi necessário normalizar os dados, pois esse procedimento somente é aplicado com duas ou mais bibliotecas.

Tags de baixa contagem (menor que cinco) possuem aumento significativo na probabilidade de erro (BEISSBARTH *et al.*, 2004), entretanto as *tags* menos frequentes não foram excluídas do estudo, pois representam a maior parte do total de *tags* e em algumas

situações podem caracterizar genes com baixa frequência (Figura 8). Foram removidas apenas *tags* com contagem igual a um (LORENZ; DEAN, 2002).

5.2 Anotação das *tags*

Dos 64.511 transcritos usados nesse projeto, apenas 793 (1,2%) são sequências de mRNA bem caracterizadas, enquanto que a maioria é sequência EST. O problema de usar EST é que existe uma taxa de erro de 1% em base das dez bases das *tags* (HILLIER *et al.*, 1996). Assim existe uma chance de 10% de uma *tag* de dez bases gerada tenha uma ou mais bases erradas. Sequências de mRNA bem caracterizados tem maior confiança na anotação, como mostra a Figura 11, e por isso atribuímos uma classificação maior no *ranking* quando comparada a sequências provenientes de EST.

Independente de como as *tags* internas são geradas, identificamos a posição em que as *tags* de SAGE são localizadas e a maioria delas está localizada na região mais próxima à região 3' UTR, como mostra a Figura 9. Apesar da baixa expressão das *tags* internas, elas devem ser consideradas para estabelecermos todas as possibilidades de associação de um dado gene.

Na tentativa de anotar todas as *tags* da *A. mellifera*, as *tags* não mapeadas com o UniGene foram mapeadas contra o genoma de abelha para identificar genes não expresso no cérebro da abelha.

Dentre as *tags* da lista LTC mapeadas no genoma, 6% não foram encontradas no UniGene. A Figura 13 mostrou que essas *tags* possuem frequencias baixas e também tiveram baixa classificaçao no ranking, indicando *tags* não confiáveis, possivelmente provenientes de erros no sequenciamento. Apenas uma dessas *tags* possui alta expressão (90), e foi classificada no ranking com valor de 1. Essa por sua vez deve ser de um gene não expresso em cérebro.

A porcentagem de 63% de *tags* mapeadas no genoma que também foram anotadas no UniGene, é um dado satisfatório que aprova a abordagem usada no estudo.

Com as novas atualizações no UniGene (última em janeiro 2009) já foram descritos a maioria dos genes dessa lista (Tabela 12). Isso mostra que o mapeamento das *tags* de SAGE no genoma, pode ser usado para encontrar genes ainda não validados e descritos no UniGene.

Tabela 12: Lista das dez *tags* mais frequentes que foram mapeadas no genoma e que não existem na versão do UniGene utilizado nesse projeto, mas que já foram descritas na nova versão do UniGene.

Tag	Frequência	Símbolo	UniGene Cluster	Cromossomo
ATTTGAGTTC	5505	*LOC725799*	Ame.8605	LG6
AACAGTTCCA	931	*LOC412893*		LG11
GAAAATCTTC	642	*Mrjp3*	Ame.731	LG11
TCCTGTTGAT	464	*LOC413224*	Ame.5807	LG12
AAATCCTGTT	193	*LOC410633*		LG15
TTAATAAAAT	123	*LOC408922*	Ame.3839	LG7
ATACATACTA	121	*LOC726895*	Ame.8466	LG11
TAGTCCGATA	93	*LOC408928*	Ame.242	LG7
GTAAATAAAT	91	*LOC409722*	Ame.4201	LG8
GCAAGACGGG	81	*LOC551330*	Ame.2348	LG2

5.3 Validação

A validação experimental utilizada nesse projeto pela técnica de RT-PCR visou verificar se a expressão dos genes indicados na anotação eram concordantes com a análise laboratorial. O experimento não teve como objetivo quantificar a expressão dos genes, pois a técnica RT-PCR não permite essa análise como. O mais adequado seria usar a técnica de *Real-time* PCR. Com a validação experimental, aumentamos a confiança dos resultados de anotação e valorizamos a abordagem computacional.

O UniGene de *A. mellifera* ainda não está bem definido e vem sofrendo constantes atualizações. A versão do UniGene utilizada nesse projeto, foi a de setembro de 2008. Verificamos manualmente a anotação das *tags* nos genes utilizados na validação experimental e constatamos que um deles, o *LOC409906*, foi anotado com outra *tag* devido a atualização do UniGene. A *tag* anotada para o gene *LOC409906* nesse projeto foi a AAAAAAAAAA com frequência 109 na biblioteca de SAGE, e com a nova atualização de novembro de 2008, a *tag* anota foi a GTAAAGATAA com frequência 5 na biblioteca de SAGE.

6 Conclusão

Descrevemos nesse estudo a primeira iniciativa de anotar uma biblioteca de SAGE de *A. mellifera*, que será de grande utilidade nos estudos envolvendo a análise do perfil de expressão gênica desse organismo. Os resultados do estudo permite-nos concluir que:

- A lista LTC mostrou-se importante para eliminarmos possíveis erros de sequênciamento e com isso aumentamos a confiança das *tags* da biblioteca de SAGE.
- a extração de *tags* virtuais em transcritos mostrou-se ser uma abordagem eficiente, porém existe uma quantidade considerável de *tags* ambíguas que devem ser ainda tratadas.
- a classificação (*ranking*) para a associação entre as *tags* e os genes, reduziu a ambiguidade das *tags* permitindo indicar com maior eficiência a melhor *tag* para o gene e o melhor gene para a *tag*.
- O mapeamento das *tags* de SAGE no genoma demonstrou ser importante para indicar novos genes para validação, em casos de uma tag não anotada no UniGene.
- A validação experimental através da técnica RT-PCR correspondeu com o observado na biblioteca de SAGE, aumentando a confiança da metodologia de anotação desenvolvida.
- A construção de uma interface *web* facilitou a visualização dos dados, permitindo compartilhar os resultados com a comunidade cietífica.

Referências

ADAMS, M. *et al.* The genome sequence of Drosophila melanogaster. *Science*, v. 287, n. 5461, p. 2185–2195, 2000.

AKMAEV, V.; WANG, C. Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics*, v. 20, p. 1254–1263, May 2004.

AULTMAN, K. *et al.* Anopheles gambiae genome: completing the malaria triad. *Science*, v. 298, p. 13, Oct 2002.

BEISSBARTH, T. *et al.* Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics*, v. 20 Suppl 1, p. i31–39, Aug 2004.

BOGUSKI, M.; LOWE, T.; TOLSTOSHEV, C. dbEST–database for "expressed sequence tags". *Nat. Genet.*, v. 4, p. 332–333, Aug 1993.

BOON, K. *et al.* An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci U S A*, v. 99, n. 17, p. 11287–11292, Aug 2002.

BRENNER, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, v. 18, p. 630–634, Jun 2000.

CALSA JR, T.; BENEDITO, V. A.; FIGUEIRA, A. Análise Serial da Expressão Gênica - Análise Serial da Expressão Gênica (SAGE) em plantas. *Revista Biotecnologia Ciência e Desenvolvimento*, v. 33, p. 86–98, 2004.

CONSORTIUM, H. G. S. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, v. 443, n. 7114, p. 931–949, Oct 2006.

DIVINA, P.; FOREJT, J. The Mouse SAGE Site: database of public mouse SAGE libraries. *Nucleic Acids Res*, v. 32, n. Database issue, p. 482–483, Jan 2004.

DONSON, J. *et al.* Comprehensive gene expression analysis by transcript profiling. *Plant Mol Biol*, v. 48, n. 1-2, p. 75–97, Jan 2002.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, v. 30, n. 1, p. 207–210, Jan 2002.

EWING, B. *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, v. 8, p. 175–185, Mar 1998.

HILLIER, L. *et al.* Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.*, v. 6, p. 807–828, Sep 1996.

HUANG, Z.; ROBINSON, G. Honeybee colony integration: worker-worker interactions mediate hormonally regulated plasticity in division of labor. *Proc. Natl. Acad. Sci. U.S.A.*, v. 89, p. 11726–11729, Dec 1992.

JU WAGNER L, S. G. P. UniGene: a unified view of the transcriptome. *The NCBI Handbook*, Aug 2003.

KAMPEN, A. van *et al.* USAGE: a web-based approach towards the analysis of SAGE data. Serial Analysis of Gene Expression. *Bioinformatics*, v. 16, p. 899–905, Oct 2000.

LARSSON, M. *et al.* Expression profile viewer (ExProView): a software tool for transcriptome analysis. *Genomics*, v. 63, p. 341–353, Feb 2000.

LASH, A. E. *et al.* SAGEmap: a public gene expression resource. *Genome Res*, v. 10, n. 7, p. 1051–1060, Jul 2000.

LORENZ, W.; DEAN, J. SAGE profiling and demonstration of differential gene expression along the axial developmental gradient of lignifying xylem in loblolly pine (Pinus taeda). *Tree Physiol.*, v. 22, p. 301–310, Apr 2002.

MAN, M.; WANG, X.; WANG, Y. POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, v. 16, p. 953–959, Nov 2000.

MARGULIES, E. H.; INNIS, J. W. eSAGE: managing and analysing data generated with Serial Analysis of Gene Expression (SAGE). *Bioinformatics*, v. 16, n. 7, p. 650–651, 2000. Disponível em: <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/16/7/650>.

METZKER, M. L. Emerging technologies in DNA sequencing. *Genome Res.*, v. 15, p. 1767–1776, Dec 2005.

MEYERS, B. C. *et al.* Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. *Nat Biotechnol*, v. 22, n. 8, p. 1006–1011, Aug 2004. Comparative Study.

NUNES, F. M. *et al.* The use of Open Reading frame ESTs (ORESTES) for analysis of the honey bee transcriptome. *BMC Genomics*, v. 5, p. 84, Nov 2004.

OLESKEVICH, S.; CLEMENTS, J. D.; SRINIVASAN, M. V. Long-term synaptic plasticity in the honeybee. *J Neurophysiol*, v. 78, n. 1, p. 528–532, Jul 1997.

PAGE, R.; PENG, C. Aging and development in social insects with emphasis on the honey bee, Apis mellifera L. *Exp. Gerontol.*, v. 36, p. 695–711, Apr 2001.

PAGE, R. E. J.; GADAU, J.; BEYE, M. The emergence of hymenopteran genetics. *Genetics*, v. 160, n. 2, p. 375–379, Feb 2002. Historical Article.

SAHA, S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.*, v. 20, p. 508–512, May 2002.

SCHENA, M. *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, v. 270, p. 467–470, Oct 1995.

VELCULESCU, V. *et al.* Characterization of the yeast transcriptome. *Cell*, v. 88, p. 243–251, Jan 1997.

VELCULESCU, V. E. *et al.* Serial analysis of gene expression. *Science*, v. 270, n. 5235, p. 484–487, Oct 1995.

WAHL, M.; HEINZMANN, U.; IMAI, K. LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse. *Bioinformatics*, v. 21, p. 1393–1400, Apr 2005.

WHITFIELD, C. *et al.* Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Res.*, v. 12, p. 555–566, Apr 2002.

ZHANG, L. *et al.* Gene expression profiles in normal and cancer cells. *Science*, v. 276, p. 1268–1272, May 1997.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo