UNIVERSIDADE ESTADUAL DE CAMPINAS

EDUARDO FERNANDES FORMIGHIERI

"GENOMA DE *Moniliophthora perniciosa*: MONTAGEM E ANOTAÇÃO DA MITOCÔNDRIA E DESENVOLVIMENTO DE SISTEMA DE ANOTAÇÃO SEMI-AUTOMÁTICO DE GENES"

Tese apresentada ao Instituto de Biologia para obtenção do Título de Doutor em Biologia funcional e Molecular, na área de Bioquímica.

Orientador: Prof. Dr. Gonçalo Amarante Guimarães Pereira.

Campinas, 2006.

**FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP**

Formighieri, Eduardo Fernandes
F767g Genoma de *Moniliophthora perniciosa*: montagem e anotação da mitocôndria e desenvolvimento de sistema de anotação semi-automática de genes / Eduardo Fernandes Formighieri. – Campinas, SP: [s.n.], 2006.

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

1.Genoma. 2.Mitocôndria. 3.Bioinformática. 4.Vassoura-de-bruxa. I. Pereira, Gonçalo Amarante Guimarães. II.Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Amr/ib

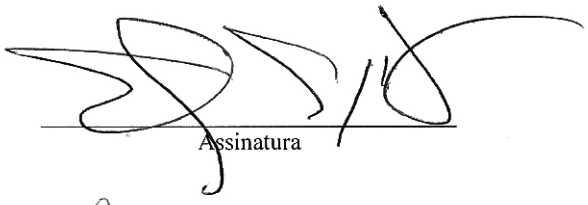
(scs/ib)

Título em inglês: *Moniliophthora perniciosa* genome: assembly and annotation of mitochondrion and development of a semi-automatic system of genes annotation.
Palavras-chave em inglês: Mitochondria; Genome; Bioinformatics; Whitches´ broom.
Área de concentração: Bioquímica.
Titulação: Doutor em Biologia Funcional e Molecular .
Banca examinadora: Gonçalo Amarante Guimarães Pereira, Carlos Augusto Colombo, José Camillo Novello, Francisco Xavier Medrano Martin, Sérgio Furtado dos Reis.
Data da defesa: 21/11/2006.
Programa de Pós-Graduação: Biologia Funcional e Molecular**.**

Campinas, 21 de novembro de 2006.

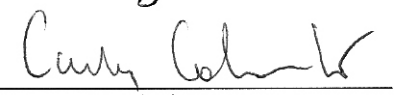
BANCA EXAMINADORA

Prof. Dr. Gonçalo Amarante Guimarães Pereira (Orientador)



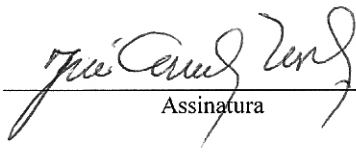
Assinatura

Prof. Dr. Carlos Augusto Colombo



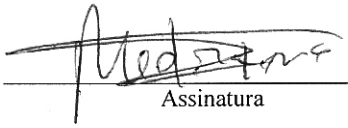
Assinatura

Prof. Dr. José Camilo Novello



Assinatura

Prof. Dr. Francisco Javier Medrano Martin



Assinatura

Prof. Dr. Sérgio Furtado dos Reis

Assinatura

Prof. Dr. Michel Georges Albert Vincentz

Assinatura

Prof. Dr. Marcelo Menossi Teixeira

Assinatura

Prof. Dr. Marcelo Brocchi

Assinatura

Dedico este trabalho a meus pais Gentil e Márcia,
irmãos Paulo e Érica, avós Manoel, Palmira e Luiza,
à namorada Carol
e a todos os meus demais amigos,
pelo amor.
AMO VOCÊS!

AGRADECIMENTOS

Ao Prof. Dr. Gonçalo Amarante Guimarães Pereira pelas oportunidades, broncas, incentivo, conhecimento e paciência.

A todos os amigos com quem convivi na UNICAMP, especialmente a todos do LGE (Ana, Andréia, Bruno, Carla, Cláudio, Dani, Deyse, Javier, Gabriel, Isa, Joan, Johana, Jorge, Jorge (hehe), Carol S., Michelle, Odalys, Silvia, Sula, e Welbe. Também para quem já saiu de lá: Alê, Ana Paula, Anderson, Babi, Camila, Cacá, Cínthia, Diana, Diana (sim, duas), Buda, Hugo, João, José Pedro, Leandra, Márcio, Naiara, Raquel, Rodrigo, Victor e Vitor). Aprendi muito e guardo comigo muitos amigos.

Ok ☺, um agradecimento especial à galera com quem convivi na Bioinfo (ordem alfabética também: Danilo, Diego, Gustavo, Lucas, Luciano, Marcelo, Marcos Renato, Taís e Tiba), responsáveis diretos pelo que me tornei (seja isto bom ou mau :) e para Eliane L. Dias.

À Siu Mui Tsai e ao pessoal do CENA/USP, onde realmente iniciou minha vida de pesquisador e onde deixei muitos amigos (se estão lá ou não são outros 500).

Aos que tentaram me ajudar ao contrário, por me ajudarem a melhorar sempre.

Ao curso de Pós-Graduação em Biologia Funcional e Molecular, especialmente à Andréia Aparecida Vigilato e às coordenadoras Dra. Eneida de Paula, Dra. Helena Coutinho Franco de Oliveira e Dra. Alba Regina Monteiro Souza Brito.

À CAPES pela bolsa concedida.

Aos professores Carlos Augusto Colombo, Eduardo Galembeck e Marcelo Brocchi pelas contribuições na qualificação. E aos membros da banca da defesa, outro obrigado.

À minha família, para quem não preciso dizer nada...

Mas digo: muito obrigado, mesmo!

À Carol, cada dia mais presente em minha vida, à sogrinha Vânia e à nova vó Lídia.

Enfim, a todos os que esqueci de citar, e com certeza foram muitos, que me influenciaram direta ou indiretamente.

MUITO OBRIGADO!

Esta tese tem um pouco de cada um de vocês.

(não, não estou chamando vocês de fungos...).

“Embora ninguém possa voltar atrás para fazer um novo começo,
qualquer um pode começar agora a fazer um novo fim.”
(Chico Xavier)

“Quem busca a verdade, quem obedece à lei do amor,
não pode estar preocupado com o amanhã.”
(Mahatma Gandhi)

RESUMO

O genoma mitocondrial (mtDNA) do fungo *Moniliophthora perniciosa* foi completamente seqüenciado e contém 109103 pb, com 31% de bases GC, porcentagem menor que a encontrada nas seqüências do genoma nuclear (47%). É o maior genoma mitocondrial de fungos descrito até o momento, e seu tamanho é conseqüência de grande espaço intergênico, que contém diversas ORFs com possibilidade de serem confirmadas como novos genes. Análises computacionais indicam a presença de variação no número de mtDNAs/célula nas diferentes bibliotecas, com tendência significativa de menor número de mtDNAs/célula no grupo de bibliotecas proveniente de culturas submetidas a repetidas repicagens. A maioria dos genes típicos (*atp6*, *atp9*, *nad1-6*, *nad4L*, *cox1-3*, *cob*, sendo a exceção o *atp8*), todos os rRNAS, tRNAS (foi encontrado pelo menos um para cada aminoácido) e genes das ORFs intrônicas estão orientados no sentido horário. Foram identificados também um gene *rps3* e um grupo de ORFs com características semelhantes às dos genes típicos. Surpreendentemente o mtDNA apresenta uma região ocupada por uma estrutura de *invertron* característica de plasmídeos *kalilo-like*, integrado de maneira estável ao genoma em todas as variedades do biótipo C, e presente nos demais biótipos testados. Esta seqüência está disponível no GenBank através do número de acesso: AY376688. A outra linha de trabalho foi desenvolvida juntamente com outros bioinformatas do Laboratório de Genômica e Expressão. Foram desenvolvidas ferramentas de mineração e anotação de genes para projetos genoma, sendo os maiores destaques o Gene Projects, que permite mineração e anotação de genes durante o processo de seqüenciamento, e a nova interface de anotação, desenvolvida para otimizar a qualidade e a eficiência da anotação de genes.

ABSTRACT

The mitochondrial genome (mtDNA) of the fungus *Moniliophthora perniciosa* was completely sequenced and it contains 109103 bases pair, with 31% of bases GC, smaller percentage than found in the sequences of the nuclear genome (47%). It is the largest mitochondrial genome of fungus described to the moment, and its size is consequence of great intergenic space, with several ORFs who can be confirmed as new genes. Computational analyses show the presence of variation in the number of mtDNAs / cell in different libraries, with significant tendency of smaller mtDNAs / cell number in group of libraries originating from cultures undergoes to repeatedly reply. Most of the typical genes (*atp6*, *atp9*, *nad1-6*, *nad4L*, *cox1-3*, *cob*, being the exception the *atp8*), all of the rRNAS, tRNAS (it was found at least one for each amino acid) and genes of the intronic ORFs are guided in the hourly sense. Surprisingly the mtDNA presents one region occupied for a structure of invertron, characteristic of plasmids kalilo-like, integrated in stable way to the genome in all of the varieties of the biotype C, and present in other tested biotypes. This sequence is available in the GenBank through the accession number: AY376688. The other work line was developed together with other bioinformatics of the Genomic and Expression Laboratory. Data mining and annotation of genes tools were developed for projects genome, being the largest prominences the Gene Projects, that allows mining and annotation of genes during the sequencing process, and the new annotation interface, developed to optimize the quality and the efficiency of the annotation of genes.

SUMÁRIO

ABREVIATURAS

AOX – Oxidase alternativa (*Alternative OXidase*)

BLAST – Ferramenta de busca de alinhamentos locais (*Basic Local Aligmnent Search Tool*)

cDNA – DNA complementar

CEPLAC – Comissão Executiva do Plano da Lavoura Cacaueira

DGE – Departamento de Genética e Evolução

DNA – Ácido desoxirribunucléico (*Deoxyribonucleic Acid*)

EMBnet – Rede de Biologia Molecular Européia (*European Molecular Biology network*)

ESTs – Etiquetas de seqüências expressas (*Expressed Sequence Tags*)

LGE – Laboratório de Genômica e Expressão

mtDNA – Genoma mitocondrial

NCBI – Centro Nacional para informação biotecnológica (*National Center for Biotechnology Information*)

NR – Banco de proteínas não redundante do NCBI (*Non Redundant*)

Número EC – Número na comissão de enzimas (*Enzyme Comission*)

Número TC – Número na classificação de proteínas de transporte (*Transport Classification*)

ORF – Quadro aberto de leitura (*Open Reading Frame*)

RNA – Ácido ribonucléico (*Ribonucleic Acid*)

rRNA – RNA ribossomal

SHAM – Ácido salicilhidroxâmico (*Salicylhydroxamic Acid*)

SNP – Polimorfismo de um único nucleotídeo (*Single Nucleotide Polymorphism*)

tRNA – RNA transportador

UNICAMP – Universidade Estadual de Campinas

1. INTRODUÇÃO

1.1. A vassoura-de-bruxa

A doença vassoura-de-bruxa em *Theobroma cacao* L. (cacaueiro), causada pelo fungo basidiomiceto *Moniliophthora perniciosa* (Stahel) Aime & Phillips-Mora é o maior problema fitopatológico do hemisfério sul das recentes décadas (Griffith, G.W. et al. 2003). A doença é originária da bacia amazônica e só foi detectada no sul da Bahia em 1989 (Pereira, J.L. et al. 1996). De 1991 para 2000 o Brasil teve sua produção anual reduzida de 320,5 mil toneladas para 191,1 mil toneladas, caindo a sua participação no mercado internacional de 14,8% para 4%, passando de exportador a importador. Esse quadro, associado aos baixos preços do produto praticados após a introdução da doença, tem fragilizado consideravelmente a situação sócio-econômica e o equilíbrio ecológico das regiões produtoras do cacau no país, onde cerca de 2,5 milhões de pessoas dependem dessa atividade (http://www.agricultura.gov.br/). Este problema atinge o Brasil como um todo ao afetar toda a cadeia produtiva de cacau.

1.2. O fungo *Moniliophthora perniciosa*

O fungo causador da doença foi descrito em 1915 por Stahel como *Marasmius perniciosus*. Em 1942 Singer o reclassificou como *Crinipellis perniciosa* (Stahel) (Singer, R. 1942) e em 2005 ele recebeu o nome atual: *Moniliophthora perniciosa* (Stahel) Aime & Phillips-Mora (Aime, M.C. et al. 2005).

Taxonomicamente, *M. perniciosa* é classificado na divisão Eumycota, subdivisão Basidimycotina, classe Basidiomycetes, subclasse Homobasidiomycetidae, ordem Agaricales e família Tricholomataceae (Purdy, L.H. et al. 1996). Baseado em caracteres morfológicos, (Pegler, D.N. 1978) descreveu três variedades para esta espécie: *Crinipellis perniciosa* var. *perniciosa*, *C. perniciosa* var. *ecuatoriensis* e *C. perniciosa* var. *citriniceps*, todas fitopatógenos de cacaueiro (*Theobroma cacao* L.) causando a doença vassoura-de-bruxa.

Embora se acreditasse que *M. perniciosa* estava restrito a hospedeiros pertencentes aos gêneros *Theobroma* e *Herrania*, da família Sterculiaceae (Hedger *et al*. 1987), posteriormente o fungo foi encontrado associado a outros hospedeiros. (Griffith & Hedger 1994a) consideram que o fungo pode ser classificado em quatro biótipos dependendo do hospedeiro: i) o Biótipo C, que

infecta cacau e outros membros da família Sterculiaceae (Bastos *et al*. 1988); ii) o Biótipo S, que infecta várias espécies da família Solanaceae (Bastos & Evans 1985); iii) o Biótipo B, que foi encontrado numa plantação de urucuzeiro (*Bixa orellana*, Bixaceae) (Bastos & Anderbrhan 1986); e iv) o Biótipo L, que foi encontrado infectando lianas das espécies *Arrabideae verrucosa* (Bignomiaceae) e *Entada gigas* (Leguminosae) assim como galhos mortos de diversas espécies de árvores suspensos na copa de florestas primárias na bacia amazônica (Evans 1978; Hedger *et al*. 1987). Os Biótipos C, S e B causam os sintomas característicos da vassoura-de-bruxa (hiperplasia e hipertrofia do tecido afetado) (Griffith, G.W. et al. 1994a) e o Biótipo L não causa este tipo de sintoma, parecendo ser um saprófito não seletivo (Hedger, J.N. et al. 1987; Griffith, G.W. et al. 1994b). Mais recentemente foi descrito o Biótipo H, encontrado associado aos sintomas de vassoura-de-bruxa na espécie *Heteropterys acutifolia* (família Malpighiaceae) no Estado de Minas Gerais, Brasil (Resende, M.L.V. et al. 2000).

M. perniciosa é um fungo hemibiotrófico com dois tipos de micélio: o biotrófico (parasítico) e o necrotrófico (saprofítico). O ciclo de vida do fungo inicia com a germinação dos basidiósporos sobre cutículas e base dos tricomas, com posterior penetração por estômatos, tecidos lesados ou de forma direta (Sreenivasan, T.N. et al. 1989). Estes tubos germinativos penetram unicamente em tecidos meristemáticos formando um micélio uninuclear e haplóide que invade os espaços intercelulares do tecido com hifas irregulares, monocarióticas e com ausência de grampos de conexão (Silva, S.D.V.M. et al. 1999). A planta atacada sofre perda de dominância apical, resultando em superbrotação chamada de vassoura verde, além de anomalias em frutos e almofadas florais (Evans, H.C. et al. 1979). O crescimento de *M. perniciosa* dicariotizado dá origem a um micélio de fase secundária (saprotrófico), no qual as hifas são mais finas e apresentam grampos de conexão (Silva, S.D.V.M. et al. 1999). Nesta fase, *M. perniciosa* causa necrose, apodrecimento e morte dos tecidos afetados da planta, formando as vassouras secas. Unicamente nesta fase da vida do fungo e após um período de seca aparecem os basidiomas, os quais produzem numerosos esporos que disseminam cada vez mais a doença (Andebrhan, T. et al. 1994; Orchard, J. et al. 1994). As condições climáticas do Estado da Bahia favorecem a produção de esporos durante o ano todo.

1.3. Estratégias de controle

Quando a doença é estabelecida em uma plantação a produção normalmente sofre queda de mais de 90% (Griffith, G.W. et al. 2003). Devido à grande importância econômica da vassoura-de-bruxa, numerosos esforços têm sido realizados na tentativa de estabelecer um plano de controle efetivo e economicamente viável para esta doença (Purdy, L.H. et al. 1996). Um exemplo é o estudo de controle biológico, como o caso de *Trichoderma stromaticum* (Hypocreales), um parasita de micélio e basidiocarpos de *M. perniciosa* (Sanogo, S. et al. 2002). A estratégia de controle considerada mais promissora foi o uso de variedades de cacaueiro resistentes a *M. perniciosa* e alguns clones de cacaueiros resistentes já estão sendo distribuídos para os produtores da Bahia através da CEPLAC. Porém, no Equador demonstrou-se que variedades resistentes podem tornar-se suscetíveis ao longo de várias gerações (Bartley, B.G.D. 1986) e um estudo recente demonstrou que a variabilidade genética de *M. perniciosa* na Bahia é muito baixa (foram encontrados apenas dois genótipos), ao contrário da variabilidade encontrada na região amazônica. Estes resultados indicam que estes clones resistentes podem ser muito sensíveis a novas introduções do fungo provenientes da Amazônia (Rincones, J. et al. 2006).

As pesquisas realizadas até o momento não têm mostrado opções de controle efetivo, devido principalmente à falta de conhecimentos sobre a biologia básica do fungo e da sua interação com o hospedeiro. Para suprir esta demanda foi lançado o Programa de Genoma da Vassoura-de-Bruxa (http://www.lge.ibi.unicamp.br/vassoura), que visa coordenar um conjunto de pesquisas de diferentes áreas, como biologia celular, morfologia, bioquímica, fisiologia vegetal e genética molecular, tendo os diversos pesquisadores envolvidos o apoio de um banco de dados de seqüências genômicas e de cDNA do fungo. O objetivo deste projeto é conseguir as bases para compreensão da doença com vistas à intervenção tecnológica para o seu combate.

1.4. **Mitocôndria**

Há um interesse particular na compreensão do metabolismo mitocondrial, o qual é freqüentemente utilizado no controle de doenças causadas por fungos (Gisi, U. et al. 2002). Distúrbios neste metabolismo estão normalmente associados a mutações supressivas ou rearranjo de seqüências do mtDNA, causando prejuízos na cadeia respiratória (Tudzynski, P. et al. 1979; Rieck, A. et al. 1982; Griffiths, A.J. et al. 1992; Osiewacz, H.D. 2002a). Portanto, mutações no

mtDNA podem reduzir a eficiência respiratória, taxa de crescimento e virulência, e representam potencialmente uma nova estratégia de controle de fungos fitopatogênicos (Monteiro-Vitorello, C.B. et al. 1995).

Embora a maioria das espécies de fungos filamentosos sejam aeróbias obrigatórias, fungos são capazes de compensar defeitos respiratórios severos através da indução da oxidase alternativa (AOX), a qual se apresenta como uma via alternativa, desviando o fluxo de elétrons dos complexos III e IV e auxiliando na proteção das células contra o excesso de elétrons causado por uma cadeira respiratória defectiva (Osiewacz, H.D. 2002b; Gredilla, R. et al. 2006). Entretanto, esta proteção é incompleta e o fenômeno de indução da AOX tem sido sistematicamente associado com o princípio do processo de degeneração conhecido como senescência fúngica (Bertrand, H. 2000).

Plasmídeos, íntrons e seqüências repetitivas são conhecidos como agentes de instabilidade no mtDNA, sendo comumente associados à senescência (Griffiths, A.J. 1992). Em *Neurospora* spp. alguns isolados que tendem à senescência contém um plasmídeo linear kalilo-like, o qual é ausente em variedades de vida longa (Bertrand, H. et al. 1985; Bertrand, H. et al. 1986). Este plasmídeo tem uma típica estrutura *invertron* com longas repetições terminais invertidas, dois genes que codificam para a DNA polimerase e RNA polimerase e proteínas terminais vinculadas a extremidade 5´ (Chan, B.S. et al. 1991; Court, D.A. et al. 1992). Uma vez inserido no interior de qualquer gene do mtDNA, este plasmídeo induz disrupção mitocondrial e estas mitocôndrias com funcionalidade alterada apresentam desvantagens em relação a mitocôndrias normais; logo, devido a uma respiração defectiva tem-se a ocorrência de senescência e morte (Rieck, A. et al. 1982; Griffiths, A.J. 1992; Griffiths, A.J.F. 1998; Bertrand, H. 2000). Contraditoriamente, em linhagens AL2 de *Podospora anserina* e no fungo *Physarum polycephalum*, espécies que normalmente senescem, a inserção de plasmídeos mitocondriais no mtDNA foi correlacionada com uma aumento da duração da vida (Hermanns, J. et al. 1994; Nakagawa, C.C. et al. 1998). Em alguns casos, a habilidade dos plasmídeos mitocondrais em causar senescência depende de fatores ambientais. Por exemplo, em *P. anserina* a restrição calórica estende a duração da vida em variedades normais, mas esta condição leva a efeitos opostos nas células que contém o plasmídeo linear pAL2-1 (Maas, M.F. et al. 2004). Embora nem todos os mecanismos estejam

bem descritos, estes dados indicam que existe uma interação entre os plasmídeos mitocondriais, a longevidade fúngica e o metabolismo mitocondrial. Dentre os Basidiomycota, plasmídeos de DNA com estrutura invertron foram completamente seqüenciados em *Agaricus bitorquis* (Robison, M.M. et al. 1999), *Flammulina velutipes* (Nakai, R. et al. 2000) e *Pleurotus ostreatus* (Kim, E.K. et al. 2000), entretanto, nenhuma função específica foi associada com a presença dos mesmos nos hospedeiros.

Genomas mitocondriais de eucariotos apresentam diversidade em tamanho, conteúdo de genes e organização do genoma. As subunidades que codificam o complexo um da cadeira respiratória (*nad1-6* e *nad4L*) estão presentes na maior parte dos casos, mas ausentes em algumas espécies, como em *Saccharomyces cerevisiae* e *Schizosaccharomyces pombe*. O código genético dos mtDNAs de fungos é bastante conservado, ocorrendo apenas uma exceção em relação ao código genético clássico – UGA codifica triptofano ao invés de códon de terminação (Kerscher, S. et al. 2001).

A seqüência completa do genoma mitocondrial de *M. perniciosa* foi obtida (GenBank, número de acesso AY376688) e foi encontrado um plasmídeo linear integrado de forma estável ao mesmo. O conjunto de trabalhos desenvolvidos no Laboratório de Genômica e Expressão (LGE) permitiu a identificação de senescência no fungo, processo que foi relacionado com alterações na morfologia da colônia, redução do número de mtDNA por célula e com ativação da oxidase alternativa. Este é a primeira vez que a senescência é descrita em um basidiomiceto e foram discutidas possíveis implicações da senescência no ciclo de vida de *M. perniciosa*.

1.5. **Sistema de mineração e anotação**

Para facilitar a compreensão do leitor, a seguir será dada uma rápida revisão sobre os principais conceitos envolvidos. Detalhes sobre análise de qualidade de cromatogramas, montagem, Blast (Altschul, S.F. et al. 1990), mineração e anotação podem ser encontrados nos anexos B, C e D.

A tecnologia mais utilizada atualmente para seqüenciamento, utilizando o método Sanger (Sanger, F. et al. 1977), gera cromatogramas com cerca de 1000 pb, dos quais apenas parte possui boa qualidade, e esta qualidade depende de diversos fatores do processo de seqüenciamento. Em

projetos genoma estes cromatogramas são gerados em laboratórios de seqüenciamento, normalmente distribuídos no estado ou país. A primeira ferramenta de bioinformática necessária é o sistema de submissão, que permite que todos os cromatogramas sejam enviados para centros de processamento onde serão analisados.

Diversas análises são realizadas, e normalmente a primeira é a verificação da nomenclatura, importantíssima para que se possa traçar a origem exata da seqüência gerada, por exemplo: a descrição de organismo, variedade/cepa, estratégia de seqüenciamento, detalhes da construção da biblioteca genômica ou de ESTs, placa, primer e ainda que laboratório realizou o seqüenciamento. A seguir é realizada a análise de qualidade de cada seqüência, num processo chamado de *base calling*, normalmente utilizando-se o programa Phred (Ewing, B. et al. 1998a; Ewing, B. et al. 1998b). Este processo calcula e atribui valores de qualidade para cada base de cada seqüência, e esta informação será utilizada posteriormente para otimizar processos como trimagem e montagem. A trimagem é o próximo passo, e se trata da busca de regiões que devem ser marcadas para que não atrapalhem a montagem. Trechos de vetores, contaminações, caudas poli-A e DNA ribossomal estão entre os principais trechos que devem ser mascarados para possibilitar a montagem o mais correta possível.

A montagem de que se fala é o processo de clusterização (formação de clusters ou grupos) por similaridade de seqüências e formação de contigs, e se faz necessária em projetos genoma devido à quebra do DNA necessária ao seqüenciamento. Os cromatogramas agrupados em contigs são equivalentes à reconstrução do fragmento original. O programa normalmente utilizado para este fim é o Phrap (Gordon, D. et al. 2001). Nos casos de projetos de ESTs, além de agrupar seqüências relativas ao mesmo mRNA, busca-se encontrar parálogos e ainda verificar padrões de expressão virtuais, que podem direcionar estudos posteriores em bancada. Para montagem de ESTs, utilizasse mais o programa CAP3 (Huang, X. et al. 1999). Ambos os programas geram clusters, que são contigs ou singlets. Singlets são seqüências que não foram agrupadas em contigs, e o resultado das montagens pode ser analisado e editado com os programas Consed (visualização da estrutura de cada contig) e Phrapview (relações entre contigs) (Gordon, D. et al. 1998). A montagem e sua análise também podem ser utilizadas para outros

fins, como a localização de microsatélites e de polimorfismos. Exemplos destas análises são alguns dos trabalhos citados no anexo F.

Os resultados das montagens são utilizados como fonte de informação para diversas análises, e as principais normalmente são mineração e anotação. Minerar (o equivalente em português a *data mining*) é buscar informações específicas num banco de dados biológico, no nosso caso, principalmente genes. Normalmente é necessário que se finalize o processo de anotação para que possa ser iniciada a mineração das informações de um processo genoma. A anotação é a definição de identidade e classificação a determinado trecho de DNA, inferindo a função da proteína codificada pelo mesmo. O processo de anotação é um dos mais demorados de um projeto genoma, e exige o tempo precioso de especialistas para anotar e verificar a anotação, o que motiva a construção de análises prévias e interface de anotação que facilitem e otimizem ao máximo a capacidade de trabalho destes pesquisadores.

Outra ferramenta muito importante para a mineração de genomas é o *Microarray* (Schena, M. et al. 1995), que gera grande quantidade de informações (milhares de genes num mesmo ensaio de hibridização) e exige estrutura de bioinformática para que seja bem aproveitada. O ideal é que se possa trabalhar com o máximo de ferramentas de forma integrada, para otimizar a capacidade de localização de informação relevante.

É importante lembrar que projetos genoma podem ser de seqüenciamento de genomas completos ou parciais, de ESTs ou ainda (e muito comumente) de DNA genômico e de ESTs. Projetos de seqüenciamento em larga escala geram uma grande quantidade de informação biológica, que exige uma estrutura de bioinformática para serem analisadas de forma eficaz e eficiente. Embora muitos sistemas tenham sido desenvolvidos, dificilmente existem estruturas de bioinformática disponíveis que cumpram todas as necessidades específicas de cada projeto. Em alguns casos é o custo da utilização do programa que inviabiliza a aquisição, em outros a estrutura exigida, necessidades específicas de análise ou ainda de estrutura.

Diversos programas foram estudados, e um resumo desta análise é encontrado na tabela 2 do capítulo 3, onde o sistema desenvolvido no Laboratório de Bioinformática do Laboratório de Genômica e Expressão (LGE) é descrito detalhadamente. Cada programa é desenvolvido em seu próprio contexto, visando suprir a demanda escolhida. Dos programas estudados, apenas o

GENOTRACE (Berezikov, E. et al. 2002) não possui interface Web e não trabalha com ESTs. O ESTAnnotator (Hotz-Wagenblatt, A. et al. 2003) é o único que não demonstrou possuir ferramentas de busca nas informações anotadas. Outras características são menos comuns, como a utilização de projetos temáticos e o controle de acesso aos dados por login e senha pessoais, só descritos no ESTIMA (Kumar, C.G. et al. 2004). O programa GENOTRACE trabalha apenas com seqüenciamento de genomas, e nenhum dos programas estudados permitia a utilização tanto em projetos de genoma quanto em de ESTs. Processamento de cromatogramas, trimagem e filtragem de seqüências e anotação de reads são características comuns dos programas ESTWeb (Paquola, A.C.M. et al. 2003), ESTAP (Mao, C.H. et al. 2003), ESTAnnotator, Parti-Gene (Parkinson, J. et al. 2004) e EST-PAGE (Matukumalli, L.K. et al. 2004).

Alguns dos programas estudados realizam clusterizações, e alguns trabalham com genomas durante o seqüenciamento, porém nenhum dos programas realiza todas as operações necessárias aos projetos suportados pelo LGE. Para suprir esta demanda foi iniciado no Projeto Vassoura de Bruxa o Laboratório de Bioinformática do LGE, visando o desenvolvimento de ferramentas que suprissem esta demanda de forma integrada e segura, com baixo orçamento (utilização de computadores PC e programas de utilização acadêmica livre), e que pudessem ser adaptados rapidamente a diferentes projetos suprindo as demandas específicas de cada um. Outro objetivo é a disseminação do conhecimento e da estrutura, através de colaborações, palestras e cursos (anexo F mostra colaborações do autor, e anexo G os principais cursos e palestras ministrados). Parte dos resultados deste trabalho conjunto está descrita nesta tese.

2. CAPÍTULO 1

The mitochondrial genome of the phytopathogenic basidiomycete *Moniliophthora perniciosa*

Formighieri, E. F.; Tiburcio, R. A.; Armas, E. D.; Shimo, H. M. [1]; Carels, N.; Góes-Neto, A.; Araujo, M. R. R.; Cotomacci, C.; Carazzolle, M. F.; Sardinha-Pinto, N.; Thomazella, D. P. T.; Rincones, J.; Digiampietri, L. A.; Carraro, D. M.; Azeredo-Espin, A. M.; Reis, S.F.; Deckmann, A. C.; Gramacho, K.; Gonçalves, M. S.; Moura Neto, J. P.; Barbosa, L. V.; Meinhardt, L. W.; Cascardo, J. C. M. & Pereira, G. A. G.

The mitochondrial genome of the phytopathogenic basidiomycete *Moniliophthora perniciosa*

Formighieri, E. F.[1]; Tiburcio, R. A.[1]; Armas, E. D.[2]; Shimo, H. M. [1]; Carels, N.[3]; Góes-Neto, A.[4]; Araujo, M. R. R.[1]; Cotomacci, C.[1]; Carazzolle, M. F.[1]; Sardinha-Pinto, N.[1]; Thomazella, D. P. T.[1]; Rincones, J.[1]; Digiampietri, L. A.[5]; Carraro, D.M. [6]; Azeredo-Espin, A. M. [7]; Reis, S.F. [8]; Deckmann, A. C[1].; Gramacho, K.[9]; Gonçalves, M. S.[10]; Moura Neto, J. P. [10]; Barbosa, L. V.[11]; Meinhardt, L. W. [13]; Cascardo, J. C. M.[12] & Pereira, G. A. G.[1]*

[1] Laboratório de Genômica e Expressão – Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, 13083-970, Campinas – SP, Brazil.

[2] Laboratório de Ecotoxicologia, Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, 13400-970, Piracicaba – SP, Brazil.

[3] Laboratório de Bioinformática da Universidade Estadual de Santa Cruz, 45650-000, Ilhéus – BA, Brazil.

[4] Laboratório de Pesquisa em Microbiologia (LAPEM), Departamento de Ciências Biológicas, Universidade Estadual de Feira de Santana (UEFS), 44031-460, Feira de Santana – BA, Brazil.

[5] Instituto de Computação, Universidade Estadual de Campinas, 13084-971, Campinas – SP, Brazil.

[6] Ludwig Institute For Cancer Research, 01509-010, São Paulo – SP, Brazil.

[7] Departamento de Genétiva e Evolução e Laboratório de Genética Animal, Centro de Biologia Molecular e Engenharia Genética – CBMEG, Universidade Estadual de Campinas, 13035-875, Campinas – SP, Brazil.

[8] Departamento de Parasitologia, Instituto de Biologia, Universidade Estadual de Campinas, 13083-970, Campinas – SP, Brazil.

[9] CEPLAC/CEPEC/SEFIT, 45600-970, Itabuna – BA, Brazil.

[10] Laboratório de Biologia Molecular – Faculdade de Farmácia, Universidade Federal da Bahia, 40170-290, Salvador – BA, Brazil.

[11] Laboratório de Biologia Molecular – Departamento de Biologia Geral, Instituto de Biologia, Universidade Federal da Bahia, 40170-290, Salvador – BA, Brazil.

[12] Departamento de Ciências Biológicas, Universidade Estadual de Santa Cruz, 45650-000, Ilhéus – BA, Brazil.

[13.] Sustainable Perennial Crops Laboratory, USDA-ARS, BARC-W, Beltsville MD, USA 20740

* Corresponding author: Gonçalo Amarante Guimarães Pereira, phone: +55 19 37886237, fax: +55 19 37886235, e-mail: goncalo@unicamp.br.

Abstract

We show that the necrotrophic mycelia of the hemibiotrophic basidiomycete *Moniliophthora perniciosa*, the causal agent of the witches' broom disease in *Theobroma cacao*. The mtDNA presents 109,103 bp and is the largest mitochondrial genome sequenced so far. It contains 14 typical genes, two rRNA genes and 26 tRNA genes, with anti-codons for all amino acids. Except for *atp8*, all typical genes are transcribed from the same DNA strand. The large genome size is related to a number of hypothetical genes and to the probably recent integration of a complete kalilo-like mitochondrial plasmid. Analysis of GC content and codon usage indicates that several hypothetical genes could be active.

Index Descriptors and Abbreviations: *Moniliophthora perniciosa;* mitochondrial genome; kalilo; codon usage; mtDNA: mitochondrial DNA; *hyp:* hypothetical ORF; *hypP*: hypothetical ORF associated to a plasmid-like region; *oiXgene:* ORF found in intron X of a given gene; CWO: clockwise orientation; CCWO: counter clockwise orientation.

1. Introduction

The witches' broom disease (WBD) of cacao (*Theobroma cacao* L.) is one of the most important phytopathological problems to afflict the Southern Hemisphere in recent decades (Griffith, Nicholson et al. 2003). In Brazil, the disease is endemic to the Amazon region and in 1989 was introduced into southern Bahia, the largest area of cocoa production in the country (Pereira, deAlmeida et al. 1996). This resulted in a severe drop in the Brazilian production of this commodity and, within a decade, Brazil shifted from the second largest cocoa exporter to a cocoa importer (Luz 1997; Andebrhan, Figueira et al. 1999).

Moniliophthora perniciosa (Stahel) Aime & Phillips-Mora (*Agaricales*, *Marasmiaceae*), previously classified as *Crinipellis perniciosa* (Stahel) Singer, the causal agent of WBD, was recently determined to be closely related to *Moniliophthora roreri* (HC, JA et al. 1978; Evans, Holmes et al. 2002; Aime and Phillips-Mora 2005), the causal agent of frosty pod rot of cacao (FPR). Prior to this reclassification, the fungal phytopathogen had been one of 112 known species of *Crinipellis* (*Tricholomataceae*, Agaricales, Basidiomycota, Fungi) (www.indexfungorum.org), however, its current association with the other members of the *Crinipellis* genus has not yet be reexamined. The infection is accompanied by a concerted series of physiological and biochemical events (Scarpari, Meinhardt et al. 2005). Initially, the fungus is monokaryotic and lives in the apoplastic fluid. Though present at low density in this phase of the disease (Penman, Britton et al. 2000), the biotrophic mycelia cause very pronounced symptoms, with hypertrophy of the infected tissue. After a few weeks, the fungus converts to the dikaryotic necrotrophic/saphrotrophic phase, rapidly invades the cells of the infected tissues and is involved in the necrosis/death of the infected tissue (Purdy and Schmidt 1996; Smith and Read 1997; Scarpari, Meinhardt et al. 2005; Meinhardt, Bellato Cde et al. 2006).

In view of its importance, *M. perniciosa* is currently the object of a genome project (www.lge.ibi.unicamp.br/vassoura) and its biology is presently under intensive investigation (Rincones, Meinhardt et al. 2003; Scarpari, Meinhardt et al. 2005; Meinhardt, Bellato Cde et al. 2006). In this context, an investigation of the *M. perniciosa* mitochondrial metabolism is essential, since this organelle is frequently a target for fungal disease control (Gisi, Sierotzki et al. 2002). Moreover, the sequence of its genome can reveal important evolutionary relationships between isolates and biotypes (de Arruda, Ferreira et al. 2003).

Mitochondria are the cytoplasmic energy-transducing organelles of eukaryotes and their genomes, which are vestigials of endosymbiotic proteobacterial ancestors (Gray, Burger et al. 1999; Lang, Gray et al. 1999), can evolve faster and independently of their nuclear counterpart (Burger, Gray et al. 2003; Burger and Lang 2003). In view of this evolutionary independence, the number of mitochondrial genes can fluctuate enormously between species, from 3 to 67 for protein genes and from 0 to 27 for tRNAs (Adams and Palmer 2003). Genes encoding proteins that account for essential processes have been conserved throughout evolution and are widely employed in the study of phylogenetic relationships between species. In contrast, there is an outstanding diversity in gene organization and gene expression (Bullerwell, Forget et al. 2003; Burger, Gray et al. 2003) that can reveal details of recent events of speciation. In view of this, an increasing number of mtDNA genomes have been sequenced or are near completion, which include *Hyaloraphidium curvatum*, a probable taxon of the fungal clade (Ustinova, Krienitz et al. 2000; Bullerwell, Forget et al. 2003), and at least 41 complete fungal mtDNAs encompassing species within the Chytridiomycota, Ascomycota, and Basidiomycota (www.ncbi.nlm.nih.gov/genomes/ORGANELLES/fu.html).

The present work is a comprehensive study of the *M. perniciosa* mtDNA, which includes the description of its organization, gene content, gene order, and comparative and phylogenetic analyses.

2. Material and methods

2.1. Fungal isolate, culture and mtDNA sequencing

Total DNA was extracted as described previously (Talbot 2001). The total DNA was sheared by sonication or nebulization (Surzcki 1990; Surzcki 2000), and fragments ranging in size from 1 to 2 and 2 to 4 kbp were cloned into the *Sma*I site of pUC18 or pCR4Blunt (TOPO Shotgun Subcloning kit, Invitrogen – Life technologies). The inserts were sequenced and analyzed following the Whole Genome Shotgun (WGS) approach (Venter, Adams et al. 1998). The reads were assembled using the Phred/Phrap/Consed software package (Ewing and Green 1998; Ewing, Hillier et al. 1998) and assemblage accuracy was confirmed with the CAP3 software (Huang and Madan 1999). The density of reads along the mitochondrial genome (mtDNA) was defined as the number of reads that form the consensus of each single nucleotide. This estimation was made considering 500 bp windows. In order to confirm the consensus sequence of the mtDNA

generated by contig assemblage, we designed a set of primers (Table 1) that covered three mtDNA regions.

2.2 Gene annotation

Putative mitochondrial genes and ORFs longer than 66 amino acids residues were detected with ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/) followed by similarity searches for known genes present in other fungal mitochondrial genomes using TBLASTN and BLASTX (Altschul, Gish et al. 1990). Their exon/intron boundaries were defined by BLAST and CLUSTAL (Higgins 1994) alignments. The annotation of putative gene function was obtained by similarity searches using BLAST, CD-SEARCH (Marchler-Bauer and Bryant 2004) and PFAM (Bateman, Coin et al. 2004). tRNAscan-SE (Lowe and Eddy 1997) and FASTtRNA (el-Mabrouk and Lisacek 1996) was used in order to identify tRNAs. rRNAs sequences were identified by comparison with homologous Basidiomycota counterparts. Repeats were analyzed using Reputer (Kurtz, Choudhuri et al. 2001) and Tandem Repeats Finder (Benson 1999). GC level in the first (GC1), second (GC2), third (GC3) codon positions and over the whole coding sequence (GCm) was obtained using CODONW (Peden 1999).

2.3 Statistical analyses

Principal Component Analysis was used to search for patterns of codon usage, using the codon frequency table generated from the analysis of 91 ORFs with the program CODONW. These data were submitted to analysis by the method of singular value decomposition of non-centered and non-scaled data matrix by means of the R software. The number of principal components to be retained was determined by scree-plot and the percentage of accumulated variation explained by the components.

2.4 Comparative and phylogenetic analyses

The taxonomic data features reported in the GenBank (release 144) under the heading "ORGANISM" (www.infobiogen.fr) was used to classify the species used in this study. The validity of this classification was tested by a phylogenetic analysis using the protein sequences of three cytochrome oxidase genes (*cox1*, *cox2* and *cox3)* from 30 fungal species with complete mtDNA sequence. Individual protein sequences were aligned using ClustalW (Thompson, Higgins et al. 1994) with default options, except for the PAM matrix (Dayhoff, Schwartz et al. 1978). Regions of uncertain alignment at the terminus of the sequences were removed before

phylogenetic analyses. Misalignments in internal regions were manually edited, resulting in 923 aligned positions from the concatenation of the three *cox* genes. Phylogenetic analyses using Minimum Evolution (ME) were performed in MEGA 3 (Kumar, Tamura et al. 2004) with distances obtained from PAM matrix (Dayhoff and Orcutt 1979). The robustness of tree topology was tested using 5000 bootstrap resamplings (Felsenstein 1985). A Maximum Likelihood-based phylogeny (ML) was constructed using PROML and SEQBOOT, available in the PHYLYP package (Felsenstein 1997), using the same PAM matrix. The tree branch support was obtained using 100 replicated datasets. Evolutionary rates were presumed to be homogeneous among sites in both phylogenetic analyses. The gene order for synteny analysis was established for each mtDNA using GenBank annotations.

3. Results

3.1 Mitochondrial genome sequencing

The genome of *Moniliophthora perniciosa* is being sequenced using the whole genome shotgun approach. For this, we constructed approximately 50 genomic libraries using total DNA extracted from cells of the strain CP02. Each library corresponded to independent cloning events using DNA obtained from individually growing cultures of CP02.

After assembling 124,565 reads derived from plasmids of different libraries, a large circular contig enclosing 5,448 reads was generated, this corresponded to the mitochondrial genome (mtDNA) of the fungus. This sequence was confirmed by further shotgun sequencing, with the final assemblage containing 6,920 reads, and by the extension of three unrelated regions. The sequence has been submitted to GenBank (accession number **AY376688**).

3.2 Protein coding genes

As expected, we identified in the *M. perniciosa* mtDNA the 14 protein-coding genes typically found in fungal mtDNA. They are directly involved in oxidative phosphorylation/energy metabolism, i.e. genes encoding for NADH dehydrogenases (7 subunits), cytochrome c oxydase (3 subunits), ATP synthases F0 (3 subunits) and apocytochrome b. We also found the *rps3*, involved in ribosome assembly (Bullerwell, Burger et al. 2000), and 2 polymerases: a DNA-dependent RNA polymerase (*rpo*) and a DNA-directed DNA polymerase (*dpoB*). It is noteworthy that the polymerases are in opposite orientation and are flanked by two sets of small inverted repeats (Fig. 1), a structure very similar to that described for kalilo-like mitochondrial plasmids

(Griffiths 1995). Furthermore, 12 conserved (9 intronic) and 59 non-conserved hypothetical ORFs were found. These results are summarized in Fig. 1.

3.3 Introns, intronic ORFs, codon usage and ORF orientation

Several of the protein-coding genes were interrupted by introns; 12 introns were found in the genes *cox1, cox2, cob, nad1, nad4* and *nad5* and nine of these introns present one conserved intronic ORF (Fig. 1), bearing domain that is characteristic of endonuclease involvement in the intron processing. Seven introns contained two LAGLIDADG domains each, one included one LAGLIDADG domain (Heath, Stephens et al. 1997) and one contained the GIY-YIG_C terminal domain (Van Roey, Waddling et al. 2001).

Preferential codon utilization was revealed in the *M. perniciosa* mitochondrial genome, by the analysis of ORFs from typical genes or conserved intronic ORFs (Table 2). To evaluate whether the unknown ORFs have the potential to be true genes, the codon usage of all 91 ORFs, including those from typical genes, were analyzed by CODONW (Peden 1999). Subsequently, the output table presenting the relative utilization of each codon for each ORF was evaluated by Principal Component Analysis (PCA) employing a scree-plot, which yielded a two-component solution accounting for 78.1% of the total variance (65.44% and 12.68% from components one and two, respectively). This revealed two separated groups of sequences (Fig. 2A), and the second component was mainly responsible for this separation, which was due to the importance of the codons UUA, AGA, AGG, AGC, CUU, UGC and UCC in its composition (Fig. 2B). The majority of the ORFs from hypothetical genes, including the non-conserved intronic ORFs, were clustered into group A. Group B included the ORFs from all of the typical protein-coding genes plus *rps3,* along with ORFs from 3 unknown conserved genes, which have been found in other mitochondrial genomes, and 17 other hypothetical gene ORFs.

There is an evident correlation between ORF orientation and codon usage. Except ATP8, all typical genes and most hypothetical ORFs were in the clockwise orientation (Fig 1 and 4). In counter clockwise orientation, we found the ORF encoding the DNA polymerase *dpoB,* typically found in mitochondrial plasmids, and the ORFs *HypP3, 58* and *57,* located immediately upstream to *dpoB* (see Fig. 1).

3.4 Ribosomal and transfer RNAs

The ribosomal RNAs were identified by similarity searches (paired-BLASTN; (Tatusova and Madden 1999) with *Suillus sinuspaulianus* (rnl) and *Lentinula edodes* (rns), and were located at positions 463 to 4,748 and 40,091 to 41,988, respectively (Fig. 1).

The software tRNAscan-SE (Lowe and Eddy 1997) identified 26 tRNAs, grouped into three regions of the mtDNA corresponding to GC% peaks (data not show), with at least one tRNA for each amino acid. Additionally, one tRNA was found for an undetermined anti-codon and three tRNAs were present for methionine, with one of them probably encoding for the initiation codon. Apart from the three Met tRNAs, we found two tRNAs with different anti-codons for Arginine, Serine and Leucine. The tRNAs positions are indicated in Table 2 indicate the presumed anticodon for each tRNA.

3.5 GC content

The mitochondrial genome was analyzed for GC content. While the nuclear genome presents a GC content of 47.7% (calculated from the non-mitochondrial clusters; data not shown), in the mtDNA it was only 31.9%. In the plasmid-like region there was a drop in the GC%. Comparing GC1, GC2, GC3 and GCm, we found that GC3 was significantly biased toward lower values, which has been observed for the mitochondrial ORFs of other fungi.

3.6 Comparative and phylogenetic analyses

A comparison of the mtDNA of several fungi according to the presence or absence of known genes (Fig. 3) showed that *M. perniciosa* mtDNA has the same set of genes as *Schizophyllum commune*, another member of the Agaricales family, except for the plasmid genes. Recently, Kouvelis et al. (2004) found synteny between genes in mitochondrial genomes of Sordariomycetes species. Therefore, we searched for any possible synteny between the mitochondrial genes of *M. perniciosa* and additional related fungal species (Fig. 4), but the results showed no evident of a conservation of gene order.

A second interesting difference was found when comparing the mitochondrial genomes of *M. perniciosa* and *Podospora anserina*, both over 100 kb (Fig. 4). We observed that while in *P. anserin*a this increased size was due to the presence of several introns, in *M. perniciosa* it was the consequence of large intergenic regions, which are occupied by 62 hypothetical ORFs, as mentioned above.

Phylogenetic analyses were performed using the three cytochrome oxidase genes (*cox*), which produced, individually, almost the same tree for the 30 fungal species analyzed (not shown). Another analysis was then performed with concatenated *cox* genes, producing a higher, although not significant, bootstrap support, especially in the basal branches, which separated the four major fungal phyla. Phylogenetic trees obtained from concatenate genes by maximum likelihood (ML) and minimum evolution (ME) largely agreed with each other and with the ordinary fungal taxonomy (Fig. 3). Figure 5 shows the ME phylogeny, rooted using *Allomyces macrogynus* as an out-group (Lang, Burger et al. 1997; Paquin, Laforest et al. 1997; Kouvelis, Ghikas et al. 2004).

4. Discussion

MtDNAs are characterized by sequences rich in AT (Campbell, Mrazek et al. 1999), and show a systematic AT mutational bias in unicellular parasites and endosymbionts (Musto, Rodriguez-Maseda et al. 1995; Andersson, Zomorodipour et al. 1998; McInerney 1998; Ghosh, Gupta et al. 2000; Romero, Zavala et al. 2000; Rispe, Delmotte et al. 2004). Accordingly, the mitochondrial genome of *M. perniciosa* has a 31.9% GC-content that contrasts with the 47.7% content observed for the nuclear DNA. We also found that the ORFs of the typical genes generally employed A or T at the third codon position, a situation that has been verified for most fungal mtDNAs. We have found tRNAs for each amino acid (Table 2), but not for all codons found in the ORFs of the typical genes. Even assuming that an unmodified U in the first anti-codon position (wobble position) can pair with all four bases, while G can pair with U or C (Kerscher, Durstewitz et al. 2001), tRNAs for the codons UGA and AUA were not found. This fact suggests that these tRNAs may be imported from the cytoplasm, a phenomenon that has been considered for similar situations observed in other fungi (Specht, Novotny et al. 1992). For example, the mtDNA of *Schizophyllum commune* does not code for the tRNA of tyrosine, which is an amino acid frequently found in proteins encoded in the mitochondrial genome (Paquin, Laforest et al. 1997).

The mitochondrial genome of *M. perniciosa* contains 14 typical protein-coding genes, common to fungal mtDNA (Fig. 3), and the *rps3*, which is an ancient gene encoding a ribosomal protein that persists in the mtDNA of some Ascomycota, Basidiomycota, Zygomycota and a few Chytridiomycota (Bullerwell, Burger et al. 2000). Using the amino acid sequences encoded by the *cytochrome oxidase (cox)* genes, we were able to reconstruct the phylogeny of *M. perniciosa* (Fig. 5). Although the bootstrap value supporting the separation between Zygomycetes, Basidiomycetes and Ascomycetes was not very high, the topology of this tree was identical to the phylogeny obtained elsewhere (Kouvelis, Ghikas et al. 2004), which used 14 protein-coding genes. The analysis using other typical genes showed the same results (data not shown), indicating that the evolution of the typical genes found in the *M. perniciosa* mitochondrial genome occurred with a vertical inheritance pattern.

We also tested the occurrence of synteny between the mitochondrial genes of *M. perniciosa* in comparison to other related species of the Basidiomycota, such as *Schyzophylum commune* (Fig. 4) or *Cryptococcus neoformans* (not shown), but no conservation in gene order was detected.

Synteny has been recently described for the mtDNA of *Lecanicillium muscarium* (Fig. 4) in comparison with other species of Sordariomycetes (Kouvelis, Ghikas et al. 2004). A characteristic of these genomes is the relatively small size, with a high density of genes (58% of coding DNA), which may limit the possibility of recombination or rearrangement due to the risk of gene disruption. On the contrary, *M. perniciosa* and *Schizophyllum commune* present genomes with lower gene density (around 45% of coding DNA for *S. commune*), with intergenic spaces that may provide a greater structural flexibility. Therefore, in view of these structural features, the absence of synteny is not surprising.

Interestingly, while most fungal mtDNA present a size range of 20-80 kb (Bullerwell and Lang 2005), the *M. perniciosa* mitochondrial genome is 109,103 kb, which is the largest fungal mtDNA sequenced so far (http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/fu.html). Moreover, the senescence-prone *Podospora anserina* is also a species with a sequenced mitochondrial genome over 100 kb. The increased size has been correlated with a high frequency of introns (Cummings, McNally et al. 1990), some of which were considered to trigger senescence. For example, aging cells of *P. anserine* are known to accumulate small mtDNA sequences that correspond to intron I of the *cox1* gene, which constitutes a mobile intron of group II (Dai, Toor et al. 2003). This class of molecules are able to retrotranspose in the mitochondrial chromosome and cause mtDNA rearrangements, which was initially considered to be the cause of senescence (Sellem, Lecellier et al. 1993). Later, it has been found that the accumulation of small mtDNA-derived fragments, designated as senDNA, were a consequence of rearrangement and possible accelerators of the senescence process, but not the primary cause (Begel, Boulay et al. 1999; Dufour, Boulay et al. 2000). SenDNA have been found in several senescent-prone fungi, such as several species of the genera *Neurospora* (Bertrand, Collins et al. 1980) and *Aspergillus* (Lazarus, Earl et al. 1980).

We searched for parallels in the *M. perniciosa* mitochondrial genome. In this case, all intronic ORFs present the typical endonuclease LAGLIDADG domain, thus suggesting that they are type I introns (Toor and Zimmerly 2002), unable to transpose in the genome. Moreover, we were unable to detect over-representation of reads for any defined mtDNA region, which would be expected in the case of an accumulation of senDNA-like elements in old mycelia. Therefore, we

concluded that these phenomena, although described for several filamentous fungi, do not seem to be associated with the senescence observed in *M. perniciosa*.

In contrast to the mitochondrial genome of *P. anserina*, the large size of the *M. perniciosa* mtDNA is a consequence of long intergenic sequences (Fig. 4). We investigated the existence of potential unknown genes in these regions from all of the 62 hypothetical ORFs found in this genome. Based on PCA analysis of the frequency of codon usage for all ORFs, conserved and hypothetical, it was possible to separate them into two groups. Consistently, most of the hypothetical ORFs clustered separately from the typical and conserved genes, with the exception of 17 ORFs (Fig. 2, groups A and B, respectively). Also, a tendency in the orientation of genes and ORFs of the second group was identified, since apart from *atp8* and *dpoB*, all typical and conserved genes were found in the clockwise orientation. The same was observed for the hypothetical ORFs, except for a cluster of 3 ORFs found immediately upstream of *dpoB*. Finally, the coding potential of the ORFs located in the intergenic regions was inspected by searching for tRNAs genes in the vicinity, with the rationale being that tRNAs are not expected to be superimposed with true coding regions. We found that tRNAs form 5 main clusters along the mtDNA (Fig. 1) and that most of the hypothetical ORFs of the second group were located in regions not populated by tRNAs. These analyses suggest that the *M. perniciosa* mtDNA may present several uncharacterized genes, which should be carefully investigated due to their potential to encode new proteins involved in mitochondrial metabolisms.

5. Acknowledgements

This research was supported by the Brazilian agencies CNPq (research fellowship to N. Carels), Capes, CNPq Regional Genoma Program, SEAGRI, and FAPESP (No. 02/09280-1)

6. References

Adams, K. L., Palmer, J. D., 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol. 29, 380-95.

Aime, M. C., Phillips-Mora, W., 2005. The causal agents of witches' broom and frosty pod rot of cacao (chocolate, *Theobroma cacao*) form a new lineage of Marasmiaceae. Mycologia. 97, 1012-22.

Altschul, S. F., et al., 1990. Basic local alignment search tool. J Mol Biol. 215, 403-10.

Andebrhan, T., et al., 1999. Molecular fingerprinting suggests two primary outbreaks of witches' broom disease (*Crinipellis perniciosa*) of *Theobroma cacao* in Bahia, Brazil. European Journal of Plant Pathology. 105, 167-175.

Andersson, S. G., et al., 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature. 396, 133-40.

Bateman, A., et al., 2004. The Pfam protein families database. Nucleic Acids Res. 32, D138-41.

Begel, O., et al., 1999. Mitochondrial group II introns, cytochrome c oxidase, and senescence in *Podospora anserina*. Mol Cell Biol. 19, 4093-100.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573-80.

Bertrand, H., 2000. Role of mitochondrial DNA in the senescence and hypovirulence of fungi and potential for plant disease control. Annu Rev Phytopathol. 38, 397-422.

Bertrand, H., et al., 1980. Deletion mutants of *Neurospora crassa* mitochondrial DNA and their relationship to the "stop-start" growth phenotype. Proc Natl Acad Sci U S A. 77, 6032-6.

Bullerwell, C. E., et al., 2000. A novel motif for identifying rps3 homologs in fungal mitochondrial genomes. Trends Biochem Sci. 25, 363-5.

Bullerwell, C. E., et al., 2003. Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic Acids Res. 31, 1614-23.

Bullerwell, C. E., Lang, B. F., 2005. Fungal evolution: the case of the vanishing mitochondrion. Curr Opin Microbiol. 8, 362-9.

Burger, G., et al., 2003. Mitochondrial genomes: anything goes. Trends in Genetics. 19, 709-716.

Burger, G., Lang, B. F., 2003. Parallels in genome evolution in mitochondria and bacterial symbionts. IUBMB Life. 55, 205-12.

Campbell, A., et al., 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci U S A. 96, 9184-9.

Cummings, D. J., et al., 1990. The complete DNA sequence of the mitochondrial genome of *Podospora anserina*. Curr Genet. 17, 375-402.

Dai, L., et al., 2003. Database for mobile group II introns. Nucleic Acids Res. 31, 424-6.

Dayhoff, M. O., Orcutt, B. C., 1979. Methods for identifying proteins by using partial sequences. Proc Natl Acad Sci U S A. 76, 2170-4.

Dayhoff, M. O., et al., A model of evolutionary change in proteins. In: M. O. Dayhoff, (Ed.), Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, D.C., 1978, pp. 345–352.

de Arruda, M. C., et al., 2003. Nuclear and mitochondrial rDNA variability in *Crinipellis perniciosa* from different geographic origins and hosts. Mycol Res. 107, 25-37.

Dufour, E., et al., 2000. A causal link between respiration and senescence in *Podospora anserina*. Proc Natl Acad Sci U S A. 97, 4138-43.

el-Mabrouk, N., Lisacek, F., 1996. Very fast identification of RNA motifs in genomic DNA. Application to tRNA search in the yeast genome. J Mol Biol. 264, 46-55.

Evans, H. C., et al., 2002. What's in a name? *Crinipellis*, the final resting place for the frosty pod rot pathogen of cocoa? Mycologist. 16, 148-152.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186-94.

Ewing, B., et al., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175-85.

Felsenstein, J., 1985. Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. Evolution. 39, 783-791.

Felsenstein, J., 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol. 46, 101-11.

Ghosh, T. C., et al., 2000. Studies on codon usage in *Entamoeba histolytica*. Int J Parasitol. 30, 715-22.

Gisi, U., et al., 2002. Mechanisms influencing the evolution of resistance to Qo inhibitor fungicides. Pest Manag Sci. 58, 859-67.

Gray, M. W., et al., 1999. Mitochondrial evolution. Science. 283, 1476-81.

Griffith, G. W., et al., 2003. Witches' brooms and frosty pods: two major pathogens of cacao. New Zealand Journal of Botany. 41, 423-435.

Griffiths, A. J., 1995. Natural plasmids of filamentous fungi. Microbiol Rev. 59, 673-85.

Grigoriev, A., 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26, 2286-90.

HC, E., et al., 1978. Taxonomy of *Monilia roreri*, an important pathogen of *Theobroma cacao* in South America. Canadian Journal of Botany. 56, 2528-2532.

Heath, P. J., et al., 1997. The structure of I-Crel, a group I intron-encoded homing endonuclease. Nat Struct Biol. 4, 468-76.

Higgins, D. G., 1994. CLUSTAL V: multiple alignment of DNA and protein sequences. Methods Mol Biol. 25, 307-18.

Huang, X., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Res. 9, 868-77.

Kerscher, S., et al., 2001. The complete mitochondrial genome of *Yarrowia lipolytica*. Comparative and Functional Genomics. 2, 80-90.

Kouvelis, V. N., et al., 2004. The analysis of the complete mitochondrial genome of *Lecanicillium muscarium* (synonym *Verticillium lecanii*) suggests a minimum common gene organization in mtDNAs of Sordariomycetes: phylogenetic implications. Fungal Genet Biol. 41, 930-40.

Kumar, S., et al., 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5, 150-63.

Kurtz, S., et al., 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29, 4633-42.

Lang, B. F., et al., 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature. 387, 493-7.

Lang, B. F., et al., 1999. Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet. 33, 351-97.

Lazarus, C. M., et al., 1980. Amplification of a mitochondrial DNA sequence in the cytoplasmically inherited 'ragged' mutant of *Aspergillus amstelodami*. Eur J Biochem. 106, 633-41.

Lowe, T. M., Eddy, S. R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955-64.

Luz, E. D. M. N. e. a., Doenças do Cacaueiro. In: F. X. R. Vale, (Ed.), Controle de doenças das principais culturas do Brasil. Imprensa Universitária, Viçosa, 1997.

Marchler-Bauer, A., Bryant, S. H., 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 32, W327-31.

McInerney, J. O., 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc Natl Acad Sci U S A. 95, 10698-703.

Meinhardt, L. W., et al., 2006. In Vitro Production of Biotrophic-Like Cultures of *Crinipellis perniciosa*, the Causal Agent of Witches' Broom Disease of *Theobroma cacao*. Curr Microbiol. 52, 191-6.

Meinhardt, L. W., et al., 2001. SYBR green I used to evaluate the nuclei number of fungal mycelia. Biotechniques. 31, 42-4, 46.

Musto, H., et al., 1995. Compositional properties of nuclear genes from *Plasmodium falciparum*. Gene. 152, 127-32.

Paquin, B., et al., 1997. The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression. Curr Genet. 31, 380-95.

Peden, J. F., Analysis of codon usage., Genetics. University of Nottingham, Nottingham, 1999, pp. 226.

Penman, D., et al., 2000. Chitin as a measure of biomass of *Crinipellis perniciosa*, causal agent of witches' broom disease of *Theobroma cacao*. Mycological Research. 104, 671-675.

Pereira, J. L., et al., 1996. Witches' broom disease of cocoa in Bahia: Attempts at eradication and containment. Crop Protection. 15, 743-752.

Purdy, L. H., Schmidt, R. A., 1996. Status of cacao witches' broom: Biology, epidemiology, and management. Annual Review of Phytopathology. 34, 573-594.

Rincones, J., et al., 2003. Electrophoretic karyotype analysis of *Crinipellis perniciosa*, the causal agent of witches' broom disease of *Theobroma cacao*. Mycol Res. 107, 452-8.

Rispe, C., et al., 2004. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Res. 14, 44-53.

Romero, H., et al., 2000. Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoeba histolytica*. Gene. 242, 307-11.

Scarpari, L. M., et al., 2005. Biochemical changes during the development of witches' broom: the most important disease of cocoa in Brazil caused by *Crinipellis perniciosa*. J Exp Bot. 56, 865-77.

Sellem, C. H., et al., 1993. Transposition of a group II intron. Nature. 366, 176-8.

Smith, S. E., Read, D. J., 1997. Mycorrhizal symbiosis. Academic Press, Cambridge.

Specht, C. A., et al., 1992. Mitochondrial DNA of *Schizophyllum commune*: restriction map, genetic map, and mode of inheritance. Curr Genet. 22, 129-34.

Surzcki, S., A fast method to prepare random fragment sequencing libraries using a new procedure of DNA shearing by nebulization and electroporation. The International

conference on the status and future of research on the human genome, San Diego, 1990, pp. CA: 51.

Surzcki, S., 2000. Basic methods in molecular biology. Springer-Verlag, New York.

Talbot, N., 2001. Molecular and Cellular Biology of Filamentous Fungi. Oxford, New York.

Tatusova, T. A., Madden, T. L., 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett. 174, 247-50.

Thompson, J. D., et al., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-80.

Toor, N., Zimmerly, S., 2002. Identification of a family of group II introns encoding LAGLIDADG ORFs typical of group I introns. Rna. 8, 1373-7.

Ustinova, I., et al., 2000. *Hyaloraphidium curvatum* is not a green alga, but a lower fungus; *Amoebidium parasiticum* is not a fungus, but a member of the DRIPs. Protist. 151, 253-62.

Van Roey, P., et al., 2001. Intertwined structure of the DNA-binding domain of intron endonuclease I-TevI with its substrate. Embo J. 20, 3631-7.

Venter, J. C., et al., 1998. Shotgun sequencing of the human genome. Science. 280, 1540-2.

Tables

Table 1 – List of PCR primers used in this work. *Or*.: Orientation of the extension by that primer;. >> and << indicate the primers on the 5' and 3' sides, respectively. *Init. pos.* and *Final pos.* are for the mtDNA location (bp) where sequence is homologous to the 5' and 3' primers, respectively.

Primer	Sequence	Or.	Init. pos	Final pos
A	GACCCTATGCAGCTTCTACTG	>>	3,062	3,082
B	TTATCCCTAGCGTAACTTTTATT	<<	4,033	4,013
C	CTGGTTTAATAGAAGGTGATGG	>>	19,938	19,959
D	GGGAATGAAAGTAGCCGGAGGC	<<	20,906	20,885
E	GCTGATTCCTGCTACCCAC	>>	43,979	43,997
F	CATAAACAGGTCAGGATATAGG	<<	44,399	44,378

Table 2 – Relative frequency of codon (Cod) and corresponding tRNA anticodons (Anticod) in exons (E) from conserved mitochondrial genes and intronic ORFs (I) in *M. perniciosa*. The tRNAs that can potentially match codons by the wooble rule are indicated by small caps between parentheses. '-' indicates the missing tRNAs, and '*' the stop codons. *aa* is for amino acid.

aa	Cod	E	I	Anti-cod	aa	Cod	E	I	Anti-cod	aa	Cod	E	I	Anti-cod	aa	Cod	E	I	Anti-cod
Phe	UUU	334	141	(g)aa	**Ser**	UCU	167	81	(u)ga	**Tyr**	UAU	217	107	(g)ua	**Cys**	UGU	32	21	(g)ca
	UUC	96	15	**GAA**		UCC	16	9	(u)ga		UAC	39	20	**GUA**		UGC	4	6	**GCA**
Leu	UUA	545	172	**UAA**		UCA	140	47	**UGA**	**TER**	UAA	25	7	*	**Trp**	*UGA*	*41*	*30*	-
	UUG	52	19	(u)aa		UCG	9	7	(u)ga		UAG	14	1	*		UGG	21	9	**CCA**
	CUU	79	43	(u)ag	**Pro**	CCU	109	47	(u)gg	**His**	CAU	64	44	(g)ug	**Arg**	CGU	1	9	(u)cg
	CUC	5	3	(u)ag		CCC	3	2	(u)gg		CAC	36	12	**GUG**		CGC	1	1	(u)cg
	CUA	72	33	**UAG**		CCA	47	20	**UGG**	**Gln**	CAA	111	50	**UUG**		CGA	10	14	**UCG**
	CUG	21	10	(u)ag		CCG	3	1	(u)gg		CAG	14	11	(u)ug		CGG	1	1	(u)cg
Ile	AUU	270	120	(g)au	**Thr**	ACU	135	63	(u)gu	**Asn**	AAU	235	176	(g)uu	**Ser**	AGU	114	60	(g)cu
	AUC	44	20	**GAU**		ACC	5	10	(u)gu		AAC	34	30	**GUU**		AGC	19	7	**GCU**
	AUA	*286*	*108*	-		ACA	118	39	**UGU**	**Lys**	AAA	184	209	**UUU**	**Arg**	AGA	78	49	**UCU**
Met	AUG	102	35	**CAU**		ACG	4	7	(u)gu		AAG	18	25	(u)uu		AGG	1	3	(u)cu
Val	GUU	135	39	(u)ac	**Ala**	GCU	164	46	(u)gc	**Asp**	GAU	119	91	(g)uc	**Gly**	GGU	136	78	(u)cc
	GUC	5	5	(u)ac		GCC	13	3	(u)gc		GAC	20	11	**GUC**		GGC	9	5	(u)cc
	GUA	160	42	**UAC**		GCA	91	29	**UGC**	**Glu**	GAA	111	82	**UUC**		GGA	138	46	**UCC**
	GUG	20	3	(u)ac		GCG	13	8	(u)gc		GAG	16	21	(u)uc		GGG	12	5	(u)cc

Figure legends

Figure 1 – *M. perniciosa* mitochondrial genome. The boxes indicate genes, ORFs or inverted repeats. Sequences in clockwise or counterclockwise directions are indicated by their location outside or inside the circles, respectively. Typical genes, rRNAs and rps3 are depicted in the outermost circle. Smaller boxes inside the conserved genes indicate intronic ORFs. The second circle displays the hypothetical ORFs. Dark box indicates ORF with codon usage pattern from conserved genes. White dot indicates hypothetical conserved ORFs. The innermost circle displays the tRNA position with their respective one-letter amino acid code. X is for the undefined anticodon. The small arrows outside the circles indicate the position of the primers used to confirm the position of a putative pKAL-like plasmid (1 to 10; see Fig. 3) and the mitochondria assembly (A to F). Amplification of primers A to F is shown at the top right, with the indication of the 100bp molecular weight marker. Estimated sizes of the amplified fragments are indicated.

Figure 2 – Analysis of codon usage in the identified ORFs. A: Plot of principal component analysis showing the sequence distribution according to codon usage, and relationship with ORF orientation (CWO: clockwise orientation; CCWO: counterclockwise orientation). B: Variable loadings (eigenvectors) of the first (left) and second (right) principal components.

Figure 3 – Linnean classification of fungal species with sequenced mitochondrial genomes. The taxonomical data reported in the GenBank features under the field “ORGANISM” were used to order the species. “+” and “-“ indicate presence and absence of genes in a given species; “ab” indicates presence of isoforms and “2” indicates existence of copies.

Figure 4 – Relative mtDNA size and gene location on the physical maps of representative fungal mitochondrial genomes – *Lecaniclillium muscarium*, *Schizophyllum commune*, *Podospora anserine* and *Moniliophthora perniciosa*. Individual genes are indicated by specific colors and vertical dashed lines inside the gene box represent introns.

Figure 5 – Phylogenic tree resulted from the concatenation of *cox1*, *cox2* and *cox3* obtained by Minimum Evolution (5000 bootstrap resamplings). The numbers under each internal branch are bootstrap values.

Figures

Figure 1

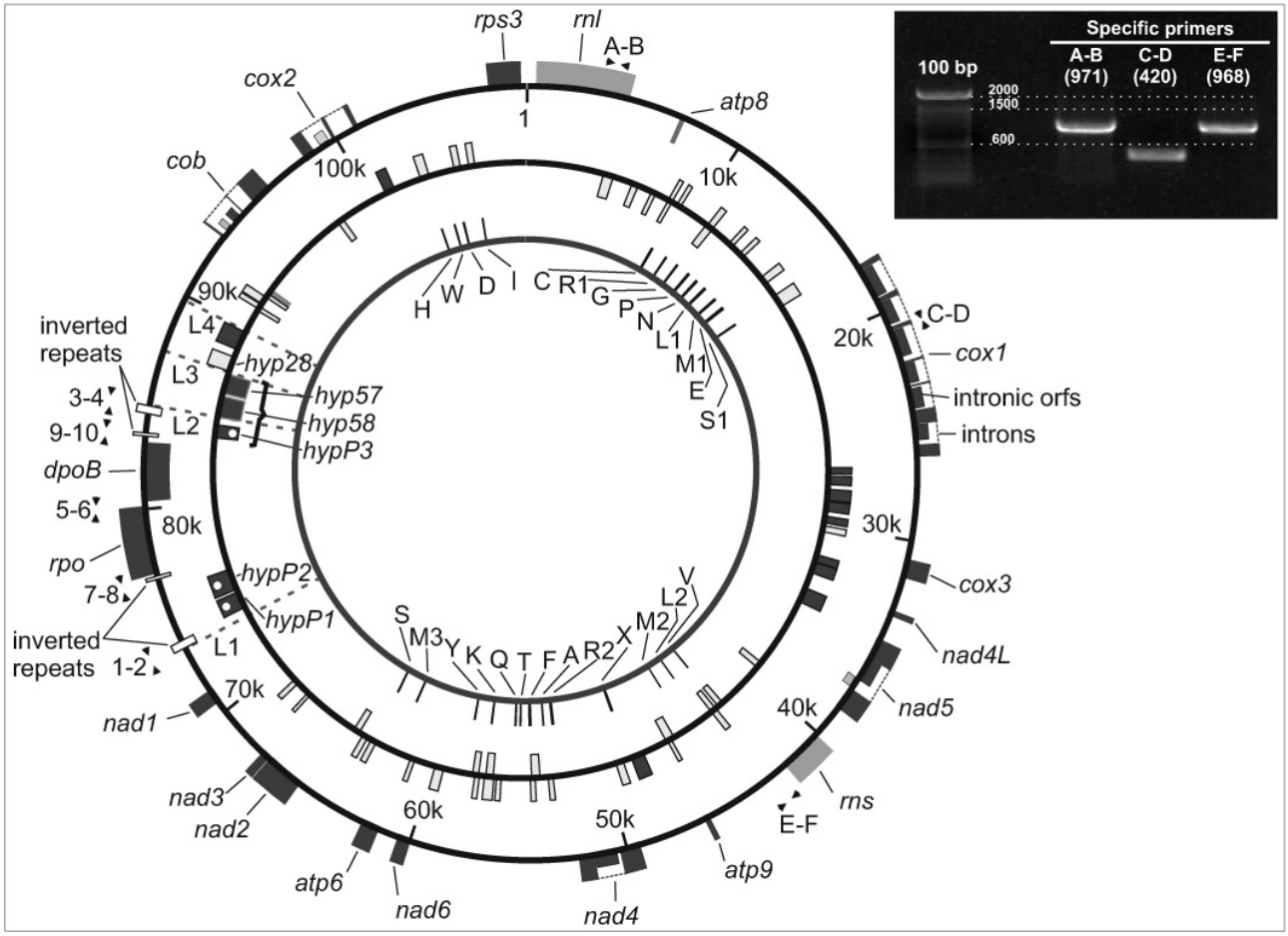
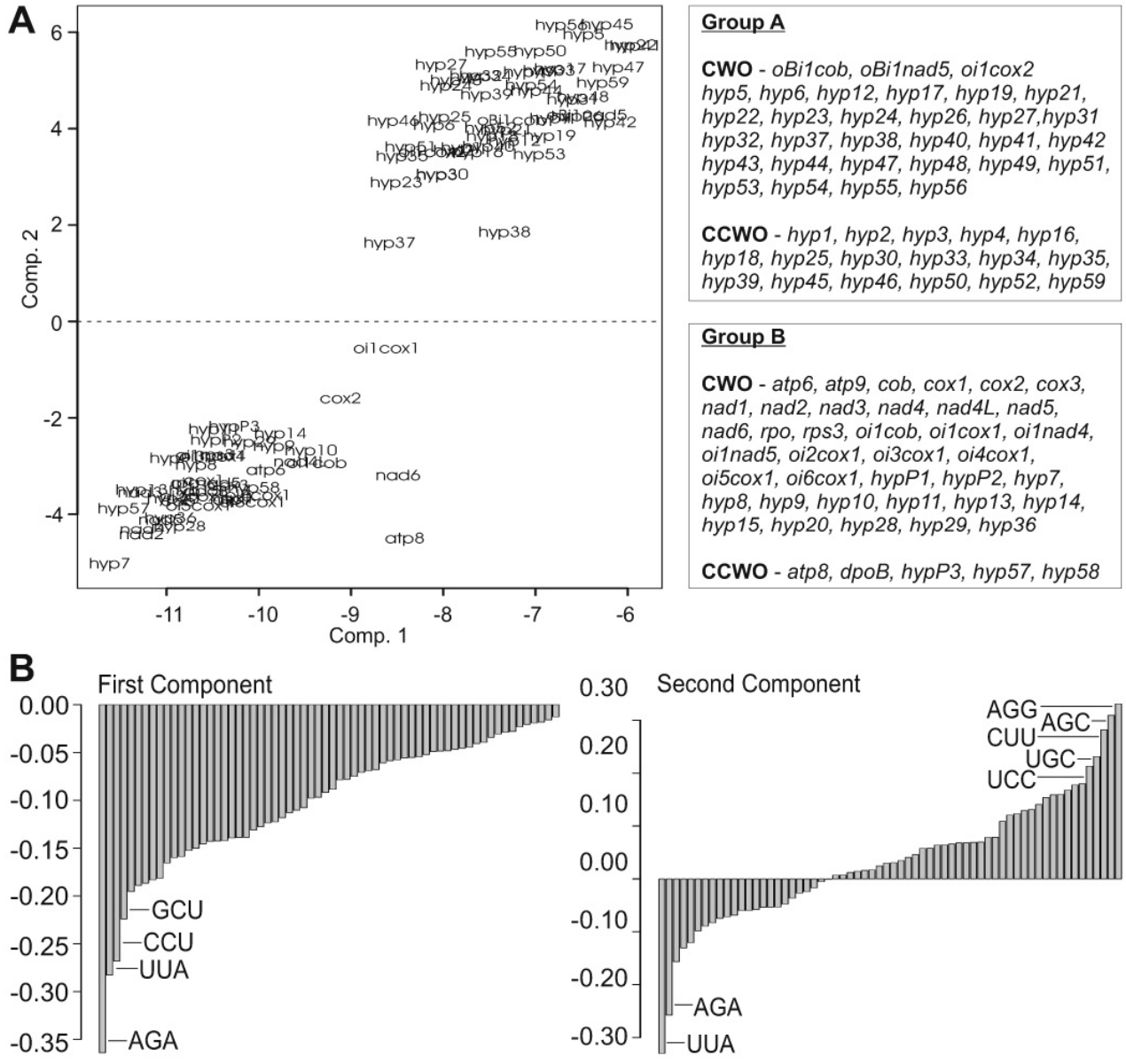


Figure 2



Group A

CWO - *oBi1cob, oBi1nad5, oi1cox2 hyp5, hyp6, hyp12, hyp17, hyp19, hyp21, hyp22, hyp23, hyp24, hyp26, hyp27,hyp31 hyp32, hyp37, hyp38, hyp40, hyp41, hyp42 hyp43, hyp44, hyp47, hyp48, hyp49, hyp51, hyp53, hyp54, hyp55, hyp56*

CCWO - *hyp1, hyp2, hyp3, hyp4, hyp16, hyp18, hyp25, hyp30, hyp33, hyp34, hyp35, hyp39, hyp45, hyp46, hyp50, hyp52, hyp59*

Group B

CWO - *atp6, atp9, cob, cox1, cox2, cox3, nad1, nad2, nad3, nad4, nad4L, nad5, nad6, rpo, rps3, oi1cob, oi1cox1, oi1nad4, oi1nad5, oi2cox1, oi3cox1, oi4cox1, oi5cox1, oi6cox1, hypP1, hypP2, hyp7, hyp8, hyp9, hyp10, hyp11, hyp13, hyp14, hyp15, hyp20, hyp28, hyp29, hyp36*

CCWO - *atp8, dpoB, hypP3, hyp57, hyp58*

Figure 3

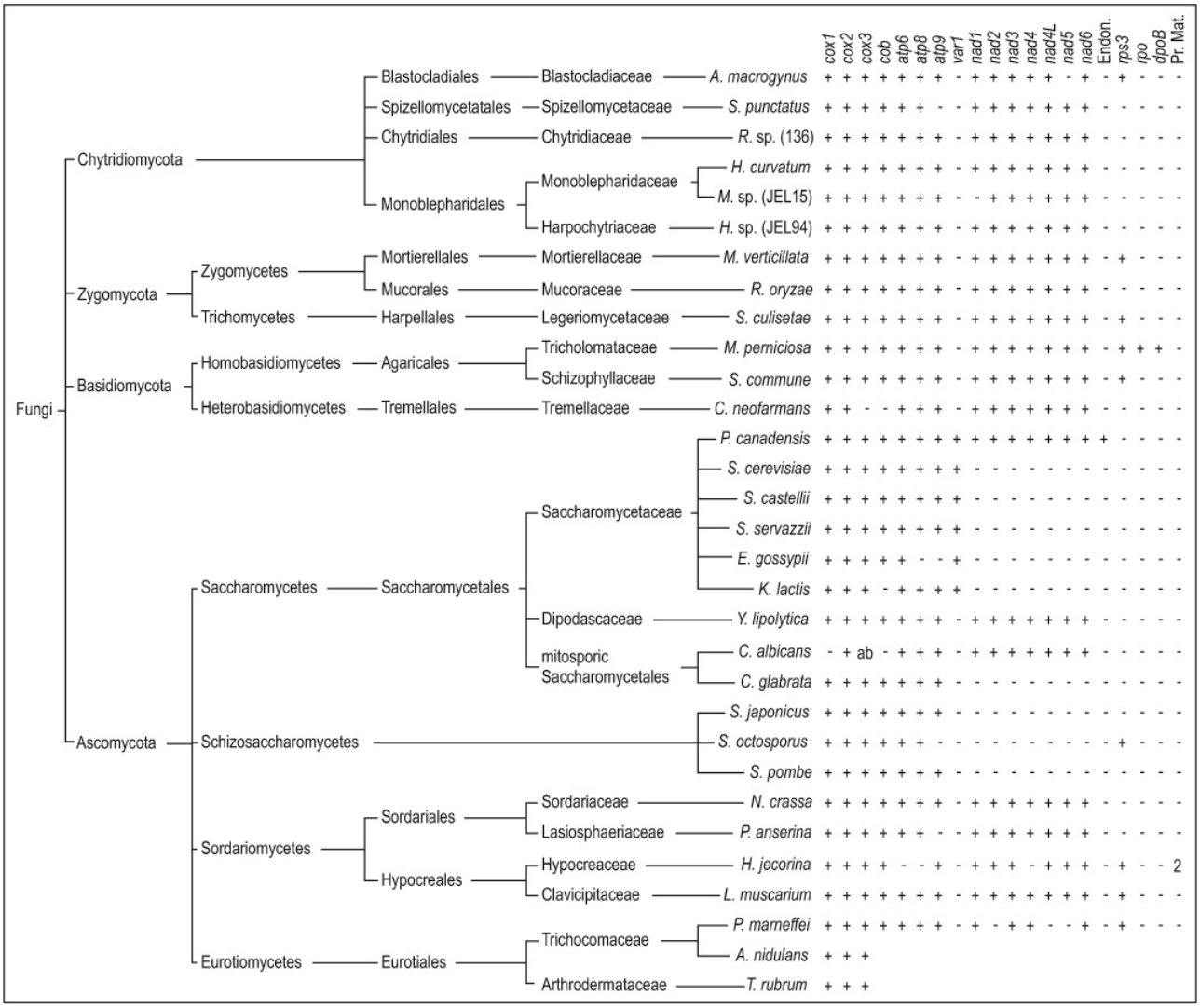

cox1 cox2 cox3 cob atp6 atp8 atp9 var1 nad1 nad2 nad3 nad4 nad4L nad5 nad6 Endon. rps3 rpo dpoB Pr. Mat.
Fungi
Chytridiomycota
Zygomycota
Basidiomycota
Ascomycota
Blastocladiales
Spizellomycetales
Chytridiales
Monoblepharidales
Mortierellales
Mucorales
Harpellales
Homobasidiomycetes
Heterobasidiomycetes
Agaricales
Tremellales
Saccharomycetales
Schizosaccharomycetes
Sordariales
Hypocreales
Eurotiales
Zygomycetes
Trichomycetes
Saccharomycetes
Sordariomycetes
Eurotiomycetes
Blastocladiaceae
Spizellomycetaceae
Chytridiaceae
Monoblepharidaceae
Harpochytriaceae
Mortierellaceae
Mucoraceae
Legeriomycetaceae
Tricholomataceae
Schizophyllaceae
Tremellaceae
Saccharomycetaceae
Dipodascaceae
mitosporic Saccharomycetales
Sordariaceae
Lasiosphaeriaceae
Hypocreaceae
Clavicipitaceae
Trichocomaceae
Arthrodermataceae
A. macrogynus
S. punctatus
R. sp. (136)
H. curvatum
M. sp. (JEL15)
H. sp. (JEL94)
M. verticillata
R. oryzae
S. culisetae
M. perniciosa
S. commune
C. neoformans
P. canadensis
S. cerevisiae
S. castellii
S. servazzii
E. gossypii
K. lactis
Y. lipolytica
C. albicans
C. glabrata
S. japonicus
S. octosporus
S. pombe
N. crassa
P. anserina
H. jecorina
L. muscarium
P. marneffei
A. nidulans
T. rubrum

Figure 4

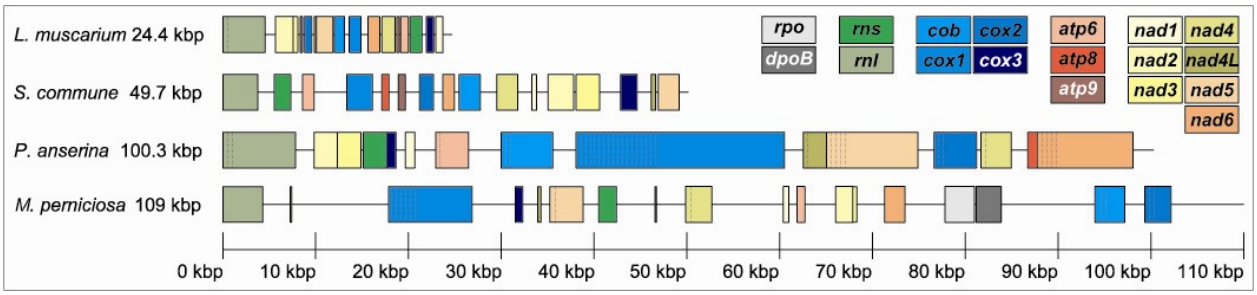
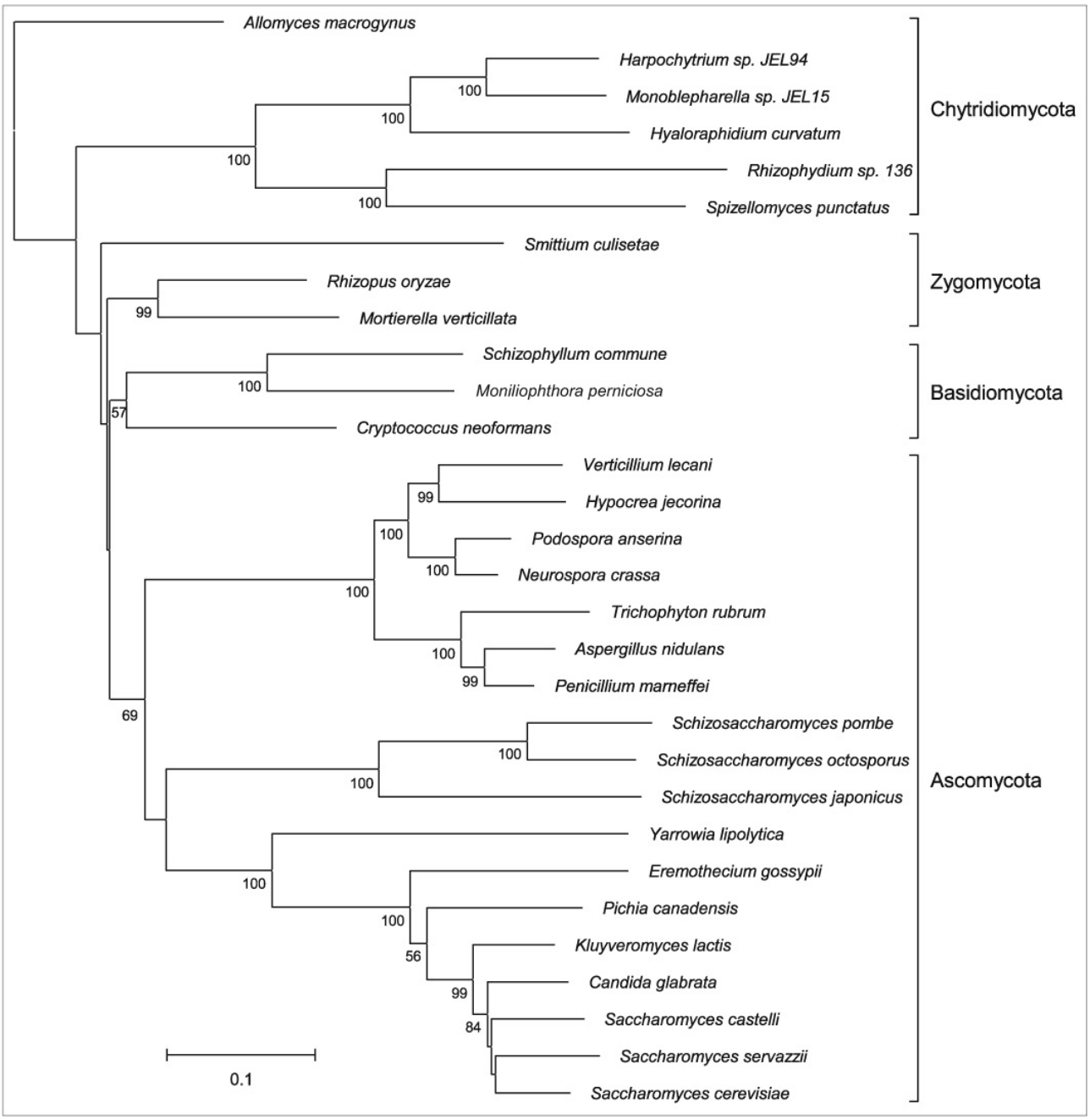

L. muscarium 24.4 kbp
S. commune 49.7 kbp
P. anserina 100.3 kbp
M. perniciosa 109 kbp
rpo
dpoB
rns
rnl
cob
cox1
cox2
cox3
atp6
atp8
atp9
nad1
nad2
nad3
nad4
nad4L
nad5
nad6
0 kbp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
100 kbp
110 kbp

Figure 5


Allomyces macrogynus
Harpochytrium sp. JEL94
Monoblepharella sp. JEL15
Hyaloraphidium curvatum
Rhizophydium sp. 136
Spizellomyces punctatus
Chytridiomycota
Smittium culisetae
Rhizopus oryzae
Mortierella verticillata
Zygomycota
Schizophyllum commune
Moniliophthora perniciosa
Cryptococcus neoformans
Basidiomycota
Verticillium lecani
Hypocrea jecorina
Podospora anserina
Neurospora crassa
Trichophyton rubrum
Aspergillus nidulans
Penicillium marneffei
Schizosaccharomyces pombe
Schizosaccharomyces octosporus
Schizosaccharomyces japonicus
Yarrowia lipolytica
Eremothecium gossypii
Pichia canadensis
Kluyveromyces lactis
Candida glabrata
Saccharomyces castelli
Saccharomyces servazzii
Saccharomyces cerevisiae
Ascomycota
100
100
100
100
99
100
57
99
100
100
100
99
69
100
100
100
100
56
99
84
0.1

3. CAPÍTULO 2

The phytopathogenic basidiomycete *Moniliophthora perniciosa* senesces and contains a linear plasmid stably integrated in its large mitochondrial genome

Formighieri, E. F.; Rincones, J.; Thomazella, D. P. T.; Tiburcio, R. A.; Armas, E. D.; Shimo, H. M.; Araujo, M. R. R.; Cotomacci, C.; Carazzolle, M. F.; Carels, N.; Góes-Neto, A.; Carraro, D.M.; Deckmann, A. C..; Cascardo, J. C. M.; Meinhardt, L. W.; & Pereira, G. A. G.

Aceito na *Fungal Genetics and Biology*. Manuscrito com revisões pedidas já foi submetido.

The phytopathogenic basidiomycete *Moniliophthora perniciosa* senesces and contains a linear plasmid stably integrated in its large mitochondrial genome

Formighieri, E. F.[1]; Rincones, J.[1]; Thomazella, D. P. T.[1]; Tiburcio, R. A.[1]; Armas, E. D.[2]; Shimo, H. M.[1]; Araujo, M. R. R.[1]; Cotomacci, C.[1]; Carazzolle, M. F.[1]; Carels, N.[3]; Góes-Neto, A.[4]; Carraro, D.M.[5]; Deckmann, A. C[1].; Cascardo, J. C. M.[6]; Meinhardt, L. W. [7]; & Pereira, G. A. G.[1]*

[1] Laboratório de Genômica e Expressão – Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, 13083-970, Campinas – SP, Brazil.

[2] Laboratório de Ecotoxicologia, Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, 13400-970, Piracicaba – SP, Brazil.

[3] Laboratório de Bioinformática da Universidade Estadual de Santa Cruz, 45650-000, Ilhéus – BA, Brazil.

[4] Laboratório de Pesquisa em Microbiologia (LAPEM), Departamento de Ciências Biológicas, Universidade Estadual de Feira de Santana (UEFS), 44031-460, Feira de Santana – BA, Brazil.

[5] Ludwig Institute For Cancer Research, 01509-010, São Paulo – SP, Brazil.

[6] Departamento de Ciências Biológicas, Universidade Estadual de Santa Cruz, 45650-000, Ilhéus – BA, Brazil.

[7] Sustainable Perennial Crops Laboratory, USDA-ARS, BARC-W, Beltsville MD, USA 20740

* Corresponding author: Gonçalo Amarante Guimarães Pereira, phone: +55 19 37886237, fax: +55 19 37886235, e-mail: goncalo@unicamp.br.

Abstract

The hemibiotrophic basidiomycete *Moniliophthora perniciosa* causes witches' broom disease in *Theobroma cacao.* Older cells of the saprotrophic mycelia show changes in colony morphology, hyphae mortality, and lower number of mitochondrial DNA (mtDNA) per cell and expression of an alternative oxidase, showing that *M. perniciosa* undergoes senescence, a phenomenon described for only a few ascomycetes. In *Neurospora* spp. senescence has been associated with mtDNA rearrangement caused by the integration of linear plasmids such as Kalilo and Maranhar. Remarkably, a similar plasmid was found integrated in the *M. perniciosa* mtDNA, but no mtDNA rearrangements were detected. The plasmid is an integral part of the mtDNA of the cacao-infecting biotype and GC analysis suggested that its integration is a recent evolutionary acquisition of this biotype. This is the first report of senescence in a basidiomycete and the possible advantage of this process in the life cycle of this fungus is discussed.

Index Descriptors and Abbreviations: *Moniliophthora perniciosa;* witches' broom; senescence mitochondrial genome; linear plasmid; mtDNA: mitochondrial DNA; AOX: Alternative oxidase

1. Introduction

The witches' broom disease (WBD) of cacao (*Theobroma cacao* L.) is one of the most important phytopathological problems to afflict the Southern Hemisphere in recent decades (Griffith et al., 2003). In Brazil, the disease is endemic to the Amazon region and in 1989 it was introduced to southern Bahia, the largest area of cacao production in the country (Pereira et al., 1996). This resulted in a severe drop in the Brazilian production of this commodity and, within a decade, Brazil shifted from the second largest cacao exporter to a cacao importer (Andebrhan et al., 1999).

Moniliophthora perniciosa (Stahel) Aime & Phillips-Mora (*Agaricales, Marasmiaceae*), previously classified as *Crinipellis perniciosa* (Stahel) Singer, is the causal agent of WBD and was recently determined to be closely related to *Moniliophthora roreri* (Aime and Phillips-Mora, 2005; Evans et al., 2002; HC et al., 1978), the causal agent of frosty pod rot of cacao (FPR). Currently, the *Moniliophthora* classification pertains only to the C biotype members of the *Crinipellis perniciosa* complex, which was subdivided into at least 3 biotypes according to their hosts (Griffith et al., 2003). Biotype L is the most diverse in terms of biology. It is heterothallic and infects tropical lianas of the Bignoniaceae and associated plant debris, but this infection produces no apparent symptoms. Biotypes S and C are homothallic, hemibiotrophic and infect Solanaceae and Sterculiaceae, respectively. The infection is accompanied by a concerted series of physiological and biochemical events (Scarpari et al., 2005). Initially, the fungus is monokaryotic and occupies the apoplastic fluid. Though present at low density during this phase of the disease (Penman et al., 2000), the biotrophic mycelia cause very pronounced symptoms, with hypertrophy of the infected tissue. After a few weeks, the fungus converts to the dikaryotic necrotrophic/saphrotrophic phase and rapidly invades the brooms, which become necrotic (Meinhardt et al., 2006; Purdy and Schmidt, 1996; Scarpari et al., 2005; Smith and Read, 1997). The dry brooms can remain attached to the tree for up to several months, until appropriate environmental conditions induce the necrotrophic hyphae to produce basidiomes. Basidiospores are then liberated through an explosive mechanism and germinate on cacao meristems, thus reinitiating the fungal life cycle (Wheeler and Suarez, 1993).

In view of its importance, *M. perniciosa* is currently the object of a genome project (www.lge.ibi.unicamp.br/vassoura) and its biology is under intensive investigation (Meinhardt et

al., 2006; Rincones et al., 2006; Rincones et al., 2003; Scarpari et al., 2005). In particular, we are very interested in understanding its mitochondrial metabolism, since this pathway is frequently a target for fungal disease control (Gisi et al., 2002).

Disorders of the mitochondrial metabolism are normally associated with suppressive mtDNA mutations or sequence rearrangements, which cause impairment in the respiratory chain (Griffiths et al., 1992; Osiewacz, 2002a; Rieck et al., 1982; Tudzynski and Esser, 1979). Although most filamentous fungal species are obligate aerobes, fungi are able to compensate severe respiratory defects by activating a cyanide-insensitive, inducible alternative oxidase (AOX), which bypasses complexes III and IV and thus helps to protect the cells against electron leakage from an impaired respiratory chain (Gredilla et al., 2006; Osiewacz, 2002b). However, this protection is incomplete and the phenomenon of AOX induction has been systematically associated with the onset of the degenerative process known as fungal senescence (Bertrand, 2000).

In *Neurospora* spp., some senescence-prone isolates contain the linear DNA kalilo plasmid (kalDNA), which is absent in long-lived strains (Bertrand et al., 1985; Bertrand et al., 1986). This plasmid has a typical invertron structure, with long terminal inverted repeats, two genes that encode for a DNA polymerase and a RNA polymerase, and 5'-linked terminal proteins (Chan et al., 1991; Court and Bertrand, 1992; Larsen, 1997). Once inserted into any mtDNA gene, kalDNA induces mitochondrial disruption and these functionally altered mitochondria outcompete normal mitochondria, leading to senescence and death due to defective respiration (Bertrand, 2000; Griffiths, 1992; Griffiths, 1998; Rieck et al., 1982). On the contrary, in *P. anserina* strain AL2 and in the slime mold *Physarum polycephalum*, which are species that normally senesce, insertion of mitochondrial plasmids in the mtDNA was correlated with an increased life span (Hermanns et al., 1994; Nakagawa et al., 1998). In some cases, the ability of mitochondrial plasmids to cause senescence depends on environmental factors. For instance, in *P. anserina* caloric restriction extends the life span in normal strains, but this condition leads to the opposite effect in cells that harbor the linear plasmid pAL2-1 (Maas et al., 2004). Though not yet clear, these data indicate that an interaction exists between mitochondrial plasmids, fungal lifespan and mitochondrial metabolism. Among the Basidiomycota, invertron-like DNA plasmids have been completely sequenced in *Agaricus bitorquis* (Robison and Horgen, 1999), *Flammulina*

*velutipe*s (Nakai et al., 2000), and *Pleurotus ostreatus* (Kim et al., 2000), but no specific function has been associated with their presence in these hosts.

Our group obtained the complete sequence of the mitochondrial genome of *M. perniciosa* (GenBank; accession number **AY376688**) and was surprised to find a linear plasmid as an integral part of this sequence. Moreover, we identified that this species undergoes senescence, a process that was connected with alterations in colony morphology, a reduction in the number of mtDNA per cell, and the activation of an alternative oxidase. This is the first time that senescence has been described for a basiodiomycete and we discuss the possible implication of senescence in the life cycle of *M. perniciosa*.

2. Material and methods

2.1. Fungal isolates, culture and evaluation of nuclear condition

The isolates of *Moniliophthora perniciosa* used in this study are listed in Table 1. Isolates were maintained in Malt Yeast Extract Agar and a piece of mycelia from the leading edge was subcultured to a new media plate as soon as it reached the edge of the Petri dish. For DNA extraction, mycelia were grown in malt extract broth at 28^{o}C for over 20 days. In order to confirm the presence of the plasmid and its site of insertion in the mitocondrial genome we designed a set of five primers pairs (Table 2 and Fig. 2). Significant changes in culture morphology were noted after several subculturings (>10), done for a period of over a year (Fig 1D). In order to verify nuclear integrity, we performed a nuclei staining analysis of the cultures by incubating live mycelia with 10 units ml^{-1} SYBR® Green I (Molecular Probes, Eugene, OR) in a solution of 10 mM phosphate buffer (pH 7.6) and 18% glycerol for 2 minutes as described previously (Meinhardt et al., 2001).

2.2. Sequencing, sequence analysis and semi-quantitative PCR analysis

The mitochondrial genome of *M. perniciosa* was sequenced as part of the Witches' Broom Genome Project through the whole shotgun approach (Venter et al., 1998). Total DNA was extracted as described previously (Talbot, 2001) and sheared by sonication or nebulization (Surzcki, 2000). Fragments ranging in size from 1 to 2 and 2 to 4 kb were cloned into the *Sma*I site of pUC18 or pCR4Blunt (TOPO Shotgun Subcloning kit, Invitrogen – Life technologies). The reads were assembled using the Phred/Phrap/Consed software package (Ewing and Green, 1998; Ewing et al., 1998) and assemblage accuracy was confirmed with the CAP3 software (Huang and Madan, 1999). The density of reads along the mitochondrial genome (mtDNA) was defined as the number of reads that form the consensus of each single nucleotide. This estimation was made considering 500 bp windows. The complete mtDNA sequence has been deposited at GenBank (accession number **AY376688**). Analyses of GC level and GC skew (Grigoriev, 1998) were carried out using Analseq (http://ludwig-sun2.unil.ch/~plangend/tp4/analseq.html). Semi-quantitative PCR analysis was performed in order to compare the relative amount of specific mitochondrial, plasmid and genomic sequences in a total DNA sample derived from senescent mycelia (>10 subculturings). Three primer pairs were compared by this approach: P9 and P10

(plasmid DNA), A and B (mtDNA) and URA3F and URA3R (chromosomal DNA) (Table 2). The components of the PCR reactions were: 120 ng total DNA, 200 nM each Forward and Reverse primers, 2.5 mM $MgCl_2$, 200 μM each dNTP, 1X PCR reaction buffer (Invitrogen), and 1 unit Platinum® *Taq* DNA polymerase (Invitrogen). The samples were placed in a Peltier Thermal Cycler PTC-225 (MJ Research) using the following program: 5 min at 94 ºC and 30 cycles of 50 s at 94 ºC, 50 s at the proper annealing temperature (52 ºC for primer pairs 9/10 and A/B and 58 ºC for primer pair URA3F/URA3R), and 80 s at 72 ºC. Aliquots were taken after cycles 15, 21, 24, 27, and 30. Gradual accumulation of amplicon products was verified on a 0.8% agarose gel containing 0.5 μg mL^{-1} ethidium bromide, fragments were separated in 1X TAE buffer at 4 V/cm.

2.3. Analysis of AOX gene expression

The gene coding for an alternative oxidase (AOX) from *M. perniciosa* was recognized in the genome database by sequence similarity with counterparts from other fungi. The partial sequence of the *M. perniciosa* AOX was deposited at GenBank (accession number **AY376688**). Total RNA from young and senescent saprotrophic cultures was extracted using the RNeasy Plant Mini kit according to the manufacturer's protocol (Qiagen, USA, Valencia, CA). 10 μg of total RNA from each extraction were subjected to ordinary northern blotting (Ausubel et al., 1998). The AOX probe corresponds to its complete cDNA, without the signal peptide. This 1.3 kb fragment was extended using the primers AOXF (5'-ATCTCTTCAAGCACCAGTAACAGG-3') and AOXR (5'-ACTACTACCTGAACGCCTGTAATGG-3').The PCR product was confirmed through agarose gel electrophoresis, purified using the Wizard® DNA Clean-up Kit (Promega), and ^{32}P-labelled by random priming following standard procedures (Ausubel et al., 1998) Hybridizations were carried out at 42 ºC in a solution containing 50% formamide, 10% polyethylene glycol (PEG 8000), 7% Sodium Dodecyl Sulfate (SDS), 120 mM Na_2HPO_4 (pH 7.2), 250 mM Sodium Chloride, 1 mM EDTA (pH 8.0), and 0.1 mg mL^{-1} salmon sperm DNA. Washing of the membranes was performed according to manufacturer's instructions. Autoradiography was carried out at –70 ºC, using Kodak X-ODAT film with two intensifying screens.

2.4. Statistical analysis

The differences between the ratios of reads from nuclear DNA to that from mtDNA were evaluated by analysis of variance using R software (Ihaka and Gentleman, 1996). Prior to that analysis, homogeneity of variances was checked by Bartlett test and normality by Shapiro-Wilk test and QQ-Plot, with the data being transformed by asin√ratio and the normality confirmed by means of Shapiro-Wilk test of residues (Zar, 1998).

2.5. Phylogenetic analysis

The evolution of fungal invertron plasmids was analyzed using a phylogeny of the protein sequences for the DNA (*dpo*) and RNA (*rpo*) polymerases of invertron-like plasmids from the Basidiomycota and Ascomycota. Individual protein sequences were aligned using ClustalW (Thompson et al., 1994) with default options, except for the PAM matrix (Dayhoff et al., 1978). Regions of uncertain alignment at the terminus of the sequences were removed before phylogenetic analyses. Misalignments in internal regions were manually edited. Phylogenetic analyses using Minimum Evolution (ME) were performed in MEGA 3 (Kumar et al., 2004) with distances obtained from PAM matrix (Dayhoff and Orcutt, 1979). The robustness of tree topology was tested using 5000 bootstrap resamplings (Felsenstein, 1985). A Maximum Likelihood-based phylogeny (ML) was constructed using PROML and SEQBOOT, available in the PHYLYP package (Felsenstein, 1997), using the same PAM matrix. The tree branch support was obtained using 100 replicated datasets. Evolutionary rates were presumed to be homogeneous among sites in both phylogenetic analyses.

3. Results

3.1 Mitochondrial genome sequencing and estimation of mitochondria copy number

The genome of *Moniliophthora perniciosa* was sequenced using the whole genome shotgun approach. For this, we constructed approximately 50 genomic libraries using total DNA extracted from cells of the isolate CP02 (Table 1). Each library corresponded to independent cloning events using DNA obtained from individually growing cultures of CP02. After assembling 124,565 reads derived from plasmids of different libraries, a large circular contig enclosing 5,448 reads was generated that corresponded to the mitochondrial genome (mtDNA) of the fungus. This sequence was confirmed by further shotgun sequencing, with the final assemblage containing 6,920 reads, and by the extension of three unrelated regions (data not shown). Because mtDNA was assembled from randomly generated-reads, the occurrence of mitochondrial or nuclear reads is directly proportional to the amount of DNA in each of the two genomes. Thus, considering that (i) cells are dikaryotic (two nuclei/cell), (ii) the individual nuclear genome has approximately 30 Mb (Rincones et al., 2003) and (iii) the mitochondrial genome is 109 kb, the occurrence of one read of mtDNA in every 550 reads would indicate the presence of one mtDNA/cell. According to this calculation, we estimated a mean of 25 mtDNA per cell. However, we identified that the proportion between nuclear and mitochondrial reads was not constant, but varied according to age of the cells employed to produce the libraries (Fig. 1A). Libraries produced from young mycelia presented a higher number of mitochondrial reads in their composition, with an estimated frequency of 41 mtDNA/cell, when compared with libraries constructed using DNA from older hypha (16 mtDNA/cell). This difference of composition was statistically significant, as determined by the F-test in the analysis of variance of transformed data ratios at a 1% level of significance (Fig. 1B and 1C). The correlation between number of mitochondria and the age of the culture suggested the existence of different physiological stages in the saprotrophic mycelia.

This fact led us to make a more detailed inspection of young and old cultures growing in solid media. We observed that continuous subculturing of the hyphae lead to a change in colony morphology. In some cases, we were able to observe in the same plate the formation of sectors with different growth patterns (Fig 1D, central picture), which were maintained even when new cultures were started with cells from the different sectors (Fig. 1D, panel left and right). While young cultures show a regular and vigorous growth, with cells containing distinct nuclei (Fig. 1D,

right panels), old cultures present an irregular progression of mycelia, forming concentric borders. Most cells of old cultures contain smeared DNA (Fig. 1D. left panels) and are unable to initiate a new culture. Taken together, these observations indicate the occurrence of the senescence phenomenon, which has been described in a limited number of fungal species (Griffiths, 1992).

3.2 Plasmid sequence

We identified in the *M. perniciosa* mitochondrial genome a structure resembling linear mitochondrial plasmids (Griffiths, 1995): a DNA-dependent RNA polymerase (*rpo*) and a DNA-directed DNA polymerase (*dpoB*), in opposite orientation and flanked by two sets of small inverted repeats (Fig. 2). The mitochondrial kalilo plasmids harbin-3, from *Neurospora intermedia* and maranhar, from *N. crassa*, have been found in dynamic integration processes associated with mtDNA rearrangement and senescence (Bertrand et al., 1986; He et al., 2000). Since *M. perniciosa* undergoes senescence as well, we reasoned that the plasmid we found in the mtDNA could be involved in this phenomenon by leading to the disorganization of mtDNA in old mycelia. If so, the plasmid would be expected to be able of autonomous replication from the mitochondria and that it would cause the formation of rearranged versions of the mtDNA by random integration.

To test this hypothesis, we first investigated the presence of rearranged versions of the mtDNA by inspecting if the assemblage of mtDNA was unique or if non-consensus regions were formed when DNA sequences from old and young hypha were mixed. This analysis clearly showed that the assemblage always resulted in the same consensus, thus refuting the possibility that diverse mtDNA versions were being produced as the mycelia aged. Moreover, the density of reads was homogeneous along the mtDNA assemblage (data not shown), indicating that the plasmid region may not exist in free form. In this case, it would be expected that reads composing its sequence would be overrepresented. In order to further confirm this hypothesis, a semi-quantitative PCR analysis was carried out to compare the relative amount of specific mitochondrial, plasmid and genomic sequences in a total DNA sample derived from senescent mycelia (>10 subculturings) (Fig. 4a). The results confirm a higher proportion of mitochondrial DNA (amplicon can be seen after 21 PCR cycles) in relation to genomic DNA (amplicon can only be seen after 24 cycles) and

also indicate equal proportion of mitochondrial and plasmid sequences. This result indicates that the plasmid is an integral part of the mtDNA.

In order to investigate the existence of this plasmid sequence in other isolates and to survey for mitochondrial genomes without the inserted plasmid, a set of five primer pairs was designed (primers P1 to P10 in Table 2) and tested in specific combinations (Fig. 2, upper panel). Note that, except for P5 and P6, all primer pairs match the sequences immediately adjacent to the inverted repeats.

In Fig. 2, lower panel, we show representative results of the PCR analysis of 37 isolates, including 28 C- and 9 L-biotypes. Using DNA from all C-biotypes isolates, independent of their geographic origin (Table 1), and the PCR extension always resulted in the same profile found in the reference isolate (CP02). This indicates that all of these isolates bear a plasmid-like region integrated at the same site in their mitochondrial genome. On the other hand, PCR analysis of the L-biotype DNA showed significant differences in its profile. Although these isolates may bear a plasmid-like region, as can be concluded from the positive extension of the primer pairs P5-P6 and P7-P8, this region is not integrated at the same site within the mitochondria, due to the fact that the extension of primers P1-P2 and P3-P4 failed to produce discrete fragments. Moreover, extension of P9-P10 produced fragments with sizes differing from those found in the C-biotypes.

Additional fragments were observed at reduced intensities in the gels and were consistently produced in most reactions, such as P5-P6, P7-P8 and P9-P10 in C-biotype isolates (Fig. 2). However, these fragments were produced even when only one primer was independently used in the PCR reaction (data not shown). The probable explanation for these spurious amplifications is the high AT content of the primers, an unavoidable feature due to the low GC content found in the mtDNA, which may allow for a less stringent pairing.

All isolates were also tested with the primer combination P1-P4, which should amplify a small fragment in any mtDNA lacking the plasmid. Irrespective of the biotype tested, we did not succeed in identifying any band of this size.

3.3 GC content

The mitochondrial genome was analyzed for GC content. While the nuclear genome presents a GC content of 47.7% (calculated from the non-mitochondrial clusters; data not shown), in the mtDNA it was only 31.9%. The local content was assessed by sliding windows over 5000/500 bp

and 500/50, which presented compositional variation according to the region (Fig. 3). In the plasmid-like region there was a drop in the GC% (black line: 5000/500), a reduction in the local variation of this composition (blue line: 500/50) and a change in the direction of the cumulative GC skew (limited by the dotted lines 1 and 4), which is an event that might represent a genome rearrangement (Grigoriev, 1998). This region covers not only the core plasmid (between the inverted repeats; Fig. 2), but also the hypothetical ORFs *HypP3, 58, 57, 28* and *29*, upstream to *dpoB* (Fig. 2). Moreover, the cumulative GC skew indicates that the origin of replication may be located around position 1, in which a negative peak was identified (Fig. 3).

3.4. AOX gene expression analysis

AOX gene expression in young and senescent cultures of saprotrophic mycelia of *M. perniciosa* was investigated by northern blot analysis. The results show that AOX expression cannot be detected in the young saprotrophic cultures, while in the senescent cultures its expression is considerably induced (Fig. 4b). This result indicates that the respiratory chain is impaired in older saprotrophic cultures of *M. perniciosa.*

3.5. Phylogenetic analysis of the sequences of the plasmid polymerases

The genes encoding polymerases from all fungal invertrons-like plasmids (including kalilo-like ones) were compared with the putative plasmid polymerases found in the *M. perniciosa* mtDNA. Both ML and ME analysis produced the same topology (Fig. 5 shows ME results), indicating the separation of the sequences into two major groups: kalilo-like and other non-kalilo-like linear plasmids. In this analysis, *M. perniciosa* genes grouped together with the kalilo-like sequences from other fungi.

4. Discussion

In this work we report the phenomenon of senescence in *Moniliophthora perniciosa.* Senescence was detected in fungal cultures during the sequencing of its mitochondrial genome (mtDNA) using the whole shotgun strategy. In order to sequence the mtDNA, several libraries had to be constructed because initial DNA isolations, when only older fungal cultures were available for DNA extraction, proved to be inefficient. We observed that with older cultures, independent of the cell density used for extraction, the final DNA concentration was low and unstable; i.e. it degraded rapidly when incubated at room temperature. On the other hand, stable DNA could be easily obtained from mycelia derived from newly germinated spores or from mycelia that had been subjected to only a few cycles of subculturing. Additionally, the frequency of mitochondrial reads in the genomic libraries could be negatively correlated with the number of subcultures prior to the extraction of DNA; i.e. mycelia subcultured for multiple cycles presented a significantly lower proportion of mtDNA when compared to nuclear DNA. The mycelia from multiple subculturing cycles revealed a lower number of mtDNA/cell, estimated at 16 copies, while newer mycelia showed a higher proportion of mtDNA, estimated at approximately 41-mitochondrial genomes/cell.

In view of this difference in the number of mitochondria, closer inspections of the fungal cultures were made. New cultures originated from germinating spores showed vigorous and homogenous mycelial morphology (Fig. 1D, right panel). However, after ~10 subcultures (approx. 45 cm of linear growth), depending on the isolate, mycelial growth became irregular and slow (Fig. 1D, left panel), showing a concentric and irregular pattern resembling the “stop-start” growth of senescent kalilo and maranhar strains of *Neurospora intermedia and N. crassa*, respectively (Court et al., 1991; Griffiths et al., 1992; Griffiths and Bertrand, 1984; Griffiths et al., 1986).

Continuous subculturing led to the death of the culture. Consistent with this fact, staining of nuclei with SYBR® Green I showed that most cells from older cultures have degraded nuclei in contrast with cells from young cultures (see detail in Fig 1D). Therefore, it appears that saprophytic cultures of *M. perniciosa* undergo senescence, a term that was initially employed for the ascomycete *Podospora anserina* to describe a very similar process (Griffiths et al., 1992; Griffiths and Bertrand, 1984; Marcou, 1961; Osiewacz, 2002b; Rizet, 1953; Rizet, 1957). The explanation for the “stop-start” mycelial morphology may be related to the persistence of latent

living cells in the mycelium that are located several centimeters away from the dead tips and are responsible for reinitiating mycelial growth over the dead tissue (Bertrand and Pittenge, 1969).

To our knowledge, this is the first time that senescence has been described for a basidiomycete. The phenomenon has been intensively studied in the ascomycetes *Neurospora* spp. and *Podospora anserina*, and although there is mitochondrial failure in both cases, the mechanisms involved seem to be diverse.

In *Neurospora* spp. senescence has been associated with mitochondrial plasmids, such as the kalilo and maranhar of *Neurospora intermedia and N. crassa*, respectively (Bertrand et al., 1986; Court and Bertrand, 1992). In these species, there is a strong correlation between the presence and random insertion of these plasmids into the mtDNA and the onset of senescence and death (Griffiths et al., 1992). It has been postulated that in young cultures the plasmids are only found in free form. After a specific time period that depends on the strain, the plasmid begins to proliferate and inserts at random sites in the mtDNA. As senescence progresses a greater number of inserted molecules is found, thus generating new versions of the mtDNA (Griffiths, 1992). For unknown reasons, the rearranged mtDNAs show a higher ability to proliferate and become more frequent in the mycelia, ultimately leading to disruption of mitochondrial function, senescence and death (Bertrand et al., 1986; Court and Bertrand, 1992).

Interestingly, two of the ORFs identified in the mtDNA of *M. perniciosa* were the polymerase encoding genes *dpo* and *rpo*, which are arranged as an invertron structure that is typical of mitochondrial linear plasmids (Griffiths, 1995; He et al., 2000; Larsen, 1997). The plasmid-like sequence found in the mtDNA of *M. perniciosa* seems to be complete, with the two polymerases flanked by inverted repeats.

In order to test whether the linear plasmid found in the mtDNA could account for the senescence seen in *M. perniciosa*, we studied its structure and replication. The phylogenetic analysis of the plasmid polymerases produced a consistent Agaricales group represented by plasmids pFV1 of *Flammulina velutipes* and pMLP2 from *Pleurotus ostreatum* (Fig. 5), thus indicating that the presence of such sequences in *M. perniciosa* follows a natural evolutionary history. Concerning its structure, we found some important differences between this plasmid and true linear plasmids. Firstly, typical inverted repeats of linear plasmids are long (around 1000 bp), but in the plasmid-like sequence of *M. perniciosa* they consist of two short pairs of discontinuous repeats, 347 and

130 bp long (Fig. 2), that are located 11284 and 6743 bp apart, respectively. Three hypothetical ORFs were found in the region between the short and long repeats (Fig. 2). It is possible that such interruptions in the inverted repeats hampered the autonomous replication of the plasmid by preventing recombination between these regions. In fact, we show that plasmid and mtDNA are found in equal proportion in a total DNA sample (Fig. 4a), indicating that this plasmid is indeed unable of autonomous replication, thus being found only integrated in a stable manner into the mtDNA. In addition, no recombinant versions of the mtDNA of *M. perniciosa* was found during the assembly of almost 200,000 reads from different cultures (Fig. 1), indicating that the sequence of the mtDNA of this fungus remains stable and is not subjected to dynamic recombination or any other similar processes.

Stability of the mtDNA was further tested by examining several *M. perniciosa* isolates from different geographic regions and biotypes. Remarkably, the results show that all C-biotype isolates, independent of geographic origin, presented the same PCR extension profile (Fig. 2), further confirming that the plasmid region is an integral part of the mitochondrial genome of this biotype. In contrast, the analysis of L-biotype isolates indicated the existence of a similar sequence, but it was neither identical nor integrated at the same position in the mitochondrial genome of this biotype. Therefore, the stable integration of this plasmid in the mtDNA could be a recent evolutionary event, possibly associated with the development of some biotypes. A detailed analysis of the mitochondrial sequences reinforced this hypothesis. The invertron and the five hypothetical ORFs of the plasmid region (three conserved and two unknown; Fig. 2) showed a very particular GC pattern (Fig. 3), including a preeminent alteration in the direction of the cumulative GC-skew. This fact suggests that the whole plasmid region was integrated into the mitochondrial genome in a single event and that its sequence has not yet adapted to the rest of the genome (Guy and Roten, 2004; Roten et al., 2002). This is a situation similar to the one described for the integration of prophages into bacterial genomes (Lazarevic et al., 1999).

Therefore, the evidence presented here indicates that disruption of mtDNA integrity, caused by plasmid integration, may not be responsible for the senescent phenotype observed in *M. perniciosa*.

In fact, there is no direct relationship between the presence of mitochondrial plasmid and senescence. True linear mitochondrial plasmids have been found in numerous fungal species,

including basidiomycetes, for which no senescence has been detected, or at least described (Griffiths, 1995). However, the fact that this plasmid seems to be complete and it is stably integrated in the mitochondrial genome of the C-biotype isolates is intriguing. In most cases of plasmid integration, only fragments of plasmid ("scars") remain as integral parts of the mtDNA (Cahan and Kennell, 2005).

It has been suggested that mitochondrial plasmids could confer some advantage to their hosts, such as improved resistance to temperature and higher fertility in *Neurospora* spp. (Bok and Griffiths, 2000). Also, in some species, such as *P. anserina* and *Physarum polycephalum*, which undergo senescence through a plasmid-independent process, insertion of mitochondrial plasmids into the mtDNA was correlated with an increased lifespan (Hermanns et al., 1994; Nakagawa et al., 1998). Considering that *M. perniciosa* is a filamentous fungus with coenocytic-like arrangement (Maheshwari, 2005) that would allow competition at the subcellular level (Bertrand, 2000; Taylor et al., 2002), the integration of this plasmid at this precise site in the mtDNA of the C-biotype could have conferred some advantage to this organelle that resulted in its proliferation and ultimate displacement of mitochondria that lacked this integration. Further studies would be needed in order to verify this hypothesis.

Although we failed to detect rearranged versions of the mtDNA of *M. perniciosa*, we did find a correlation between a significantly lower number of mitochondria per cell and the onset of senescence. Moreover, we were able to detect increased AOX expression in senescent saprotrophic cultures when compared to young ones (Fig. 4b). This result is consistent with the observations that alternative oxidase is induced as senescence progresses in senescent fungal species (Bertrand, 2000; Osiewacz, 2002b). Therefore, it is highly probable that the reduction of mitochondrial DNA in older cells of *M. perniciosa* is connected with a process of mitochondrial failure, which results in increased AOX expression.

In conclusion, our findings suggest that the life cycle of *M. perniciosa* described in literature is incomplete, and a further stage should be added to the four phases previously defined by specific cell morphology: (1) spores, (2) biotrophic mycelium (3) necrotrophic/saprotrophic mycelium, and (4) mushroom/fruiting body (Evans, 1980; Purdy and Schmidt, 1996). In fact, the saprotrophic mycelia show at least two distinct age-dependent morphologies that relate to different physiologies with regards to respiration and energy production.

Considering that senescence occurs rarely in fungi, the presence of this phenomenon in *M. perniciosa* leads us to believe that it could play an important role in the biology of Witches' Broom Disease. The degraded nuclei suggest that most mycelial cells die during the late saprotrophic phase. Nevertheless, these dead cells might be valuable to the mycelia as a whole, since, according to (Maheshwari, 2005), nuclei of filamentous fungi may serve as storehouses of nitrogen and phosphorus in the form of DNA. The breakdown products would then be recycled through the "coenocytic-like mycelia", thus providing hyphal tips with the ability to persist and forage into new areas (Maheshwari, 2005). Consequently, the rapid colonization of infected tissues followed by the death of old mycelia could be a strategy used by *M. perniciosa* to allow the quick absorption and storage of large amounts of nutrients that are available promptly after the death of the tissues (brown brooms) and, thus, reduce nutrient availability for competing saprotrophic microbes. Moreover, the "stop-start" growth pattern suggests the existence of cells in stationary phase that could "wait" for the proper conditions to start basidiome formation (Rocha and Wheeler, 1985). Additional studies, such as detailed histology of the nuclear condition of the mycelia *in planta*, a careful survey of the mtDNA from other biotypes, and the analysis of the expression of genes connected with mitochondrial function are the next steps to unravel the fascinating life-cycle of *M. perniciosa*.

Acknowledgements

This research was supported by the Brazilian agencies CNPq (research fellowship to N. Carels), Capes, CNPq Regional Genome Program, Government of the State of Bahia, and FAPESP (No. 05/60432-5; 06/50794-0).

References

Aime, M. C., Phillips-Mora, W., 2005. The causal agents of witches' broom and frosty pod rot of cacao (chocolate, *Theobroma cacao*) form a new lineage of Marasmiaceae. Mycologia. 97**,** 1012-1022.

Andebrhan, T., Figueira, A., Yamada, M. M., Cascardo, J., Furtek, D. B., 1999. Molecular fingerprinting suggests two primary outbreaks of witches' broom disease (*Crinipellis perniciosa*) of *Theobroma cacao* in Bahia, Brazil. Eur. J. Plant Pathol. 105**,** 167-175.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, A. J., Struhl, K., 1998. Current Protocols in Molecular Biology. Wiley & Sons Inc, New York-USA.

Bertrand, H., 2000. Role of mitochondrial DNA in the senescence and hypovirulence of fungi and potential for plant disease control. Annu. Rev. Phytopathol. 38**,** 397-422.

Bertrand, H., Chan, B. S., Griffiths, A. J., 1985. Insertion of a foreign nucleotide sequence into mitochondrial DNA causes senescence in *Neurospora intermedia*. Cell. 41**,** 877-884.

Bertrand, H., Griffiths, A. J., Court, D. A., Cheng, C. K., 1986. An extrachromosomal plasmid is the etiological precursor of kalDNA insertion sequences in the mitochondrial chromosome of senescent *Neurospora*. Cell. 47**,** 829-837.

Bertrand, H., Pittenge, T., 1969. Cytoplasmic mutants selected from continuously growing cultures of *Neurospora crassa*. Genetics. 61**,** 643-659.

Bok, J. W., Griffiths, A. J., 2000. Possible benefits of kalilo plasmids to their *Neurospora* hosts. Plasmid. 43**,** 176-180.

Cahan, P., Kennell, J. C., 2005. Identification and distribution of sequences having similarity to mitochondrial plasmids in mitochondrial genomes of filamentous fungi. Mol. Genet. Genomics. 273**,** 462-473.

Chan, B. S., Court, D. A., Vierula, P. J., Bertrand, H., 1991. The kalilo linear senescence-inducing plasmid of *Neurospora* is an invertron and encodes DNA and RNA polymerases. Curr. Genet. 20**,** 225-237.

Court, D. A., Bertrand, H., 1992. Genetic organization and structural features of maranhar, a senescence-inducing linear mitochondrial plasmid of *Neurospora crassa*. Curr. Genet. 22**,** 385-397.

Court, D. A., Griffiths, A. J. F., Kraus, S. R., Russell, P. J., Bertrand, H., 1991. A new senescence-inducing mitochondrial linear plasmid in field-isolated *Neurospora crassa* strains from India. Curr. Genet. 19**,** 129-137.

Dayhoff, M. O., Orcutt, B. C., 1979. Methods for identifying proteins by using partial sequences. Proc. Natl. Acad. Sci. USA. 76**,** 2170-2174.

Dayhoff, M. O., Schwartz, R. M., Orcutt, B. C., A model of evolutionary change in proteins. In: M. O. Dayhoff, (Ed.), Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, D.C., 1978, pp. 345–352.

Evans, H. C., 1980. Pleomorphism in *Crinipellis perniciosa* causal agent of witches' broom disease of cocoa. Trans. British Mycol. Soc. 74**,** 515-523.

Evans, H. C., Holmes, K. A., Phillips, W., Wilkinson, M. J., 2002. What's in a name? *Crinipellis*, the final resting place for the frosty pod rot pathogen of cocoa? Mycologist. 16**,** 148-152.

Evans, H. C., Stalpers, J. A., Samson, R. A., Benny, G. L., 1978. Taxonomy of *Monilia roreri*, an important pathogen of *Theobroma cacao* in South America. Can. J. Bot. 56**,** 2528-2532.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8**,** 186-194.

Ewing, B., Hillier, L., Wendl, M. C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8**,** 175-185.

Felsenstein, J., 1985. Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. Evolution. 39**,** 783-791.

Felsenstein, J., 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst. Biol. 46**,** 101-111.

Gisi, U., Sierotzki, H., Cook, A., McCaffery, A., 2002. Mechanisms influencing the evolution of resistance to Qo inhibitor fungicides. Pest. Manag. Sci. 58**,** 859-867.

Gredilla, R., Grief, J., Osiewacz, H. D., 2006. Mitochondrial free radical generation and lifespan control in the fungal aging model *Podospora anserina*. Exp. Gerontol. 41, 439-447.

Griffith, G. W., Nicholson, J., Nenninger, A., Birch, R. N., Hedger, J. N., 2003. Witches' brooms and frosty pods: two major pathogens of cacao. New Zeal. J. Bot. 41**,** 423-435.

Griffiths, A. J., 1992. Fungal senescence. Annu. Rev. Genet. 26**,** 351-372.

Griffiths, A. J., 1995. Natural plasmids of filamentous fungi. Microbiol. Rev. 59**,** 673-685.

Griffiths, A. J., 1998. The kalilo family of fungal plasmids. Bot. Bull. Acad. Sinica. 39**,** 147-152.

Griffiths, A. J., Bertrand, H., 1984. Unstable cytoplasms in hawaiian strains of *Neurospora intermedia*. Curr. Genet. 8**,** 387-398.

Griffiths, A. J., Kraus, S., Bertrand, H., 1986. Expression of Senescence in *Neurospora intermedia*. Can. J. Genet. Cytol. 28**,** 459-467.

Griffiths, A. J., Xiao, Y., Barton, R., Myers, C., 1992. Suppression of cytoplasmic senescence in *Neurospora*. Curr. Genet. 21**,** 479-484.

Grigoriev, A., 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26**,** 2286-2290.

Guy, L., Roten, C. A., 2004. Genometric analyses of the organization of circular chromosomes: a universal pressure determines the direction of ribosomal RNA genes transcription relative to chromosome replication. Gene. 340**,** 45-52.

He, C., Nastasja de, G., Bok, J. W., Griffiths, A. J., 2000. Kalilo plasmids are a family of four distinct members with individual global distributions across species. Curr. Genet. 37**,** 39-44.

Hermanns, J., Asseburg, A., Osiewacz, H. D., 1994. Evidence for a life span-prolonging effect of a linear plasmid in a longevity mutant of *Podospora anserina*. Mol Gen Genet. 243**,** 297-307.

Huang, X., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Res. 9**,** 868-877.

Ihaka, R., Gentleman, R., 1996. R: A language for data analysis and graphics. J. Comp. Graph. Stat. 5**,** 299-314.

Kim, E. K., Jeong, J. H., Youn, H. S., Koo, Y. B., Roe, J. H., 2000. The terminal protein of a linear mitochondrial plasmid is encoded in the N-terminus of the DNA polymerase gene in white-rot fungus *Pleurotus ostreatus*. Curr. Genet. 38**,** 283-90.

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5**,** 150-63.

Larsen, F. M. R. S. M., 1997. Microbial linear plasmids. Appl Microbiol Biotechnol. 47**,** 329-333.

Lazarevic, V., Dusterhoft, A., Soldo, B., Hilbert, H., Mauel, C., Karamata, D., 1999. Nucleotide sequence of the *Bacillus subtilis* temperate bacteriophage SPbetac2. Microbiology. 145**,** 1055-1067.

Maas, M. F., de Boer, H. J., Debets, A. J., Hoekstra, R. F., 2004. The mitochondrial plasmid pAL2-1 reduces calorie restriction mediated life span extension in the filamentous fungus *Podospora anserina*. Fungal Genet. Biol. 41**,** 865-871.

Maheshwari, R., 2005. Nuclear behavior in fungal hyphae. FEMS Microbiol. Lett. 249**,** 7-14.

Marcou, D., 1961. Notion de longévité et nature cytoplasmique du déterminat de lá senescence chez quelques champignons. Ann. Sci. Nat. Bot. 12**,** 653-764.

Meinhardt, L. W., Bellato C. M., Rincones, J., Azevedo, R. A., Cascardo, J. C., Pereira, G. A., 2006. In Vitro Production of Biotrophic-Like Cultures of *Crinipellis perniciosa*, the Causal Agent of Witches' Broom Disease of *Theobroma cacao*. Curr. Microbiol. 52**,** 191-196.

Meinhardt, L. W., Bellato, C. M., Tsai, S. M., 2001. SYBR green I used to evaluate the nuclei number of fungal mycelia. Biotechniques. 31**,** 42-46.

Nakagawa, C. C., Jones, E. P., Miller, D. L., 1998. Mitochondrial DNA rearrangements associated with mF plasmid integration and plasmodial longevity in *Physarum polycephalum*. Curr. Genet. 33**,** 178-187.

Nakai, R., Sen, K., Kurosawa, S., Shibai, H., 2000. Basidiomycetous fungus *Flammulina velutipes* harbors two linear mitochondrial plasmids encoding DNA and RNA polymerases. FEMS Microbiol. Lett. 190**,** 99-102.

Osiewacz, H. D., 2002a. Aging in fungi: role of mitochondria in *Podospora anserina*. Mech. Ageing Dev. 123**,** 755-764.

Osiewacz, H. D., 2002b. Genes, mitochondria and aging in filamentous fungi. Ageing Res. Rev. 1**,** 425-442.

Penman, D., Britton, G., Hardwick, K., Collin, H. A., Isaac, S., 2000. Chitin as a measure of biomass of *Crinipellis perniciosa*, causal agent of witches' broom disease of *Theobroma cacao*. Mycol. Res. 104**,** 671-675.

Pereira, J. L., deAlmeida, L. C. C., Santos, S. M., 1996. Witches' broom disease of cocoa in Bahia: Attempts at eradication and containment. Crop Prot. 15**,** 743-752.

Purdy, L. H., Schmidt, R. A., 1996. Status of cacao witches' broom: Biology, epidemiology, and management. Ann. Rev. Phytopathol. 34**,** 573-594.

Rieck, A., Griffiths, A. J., Bertrand, H., 1982. Mitochondrial variants of *Neurospora intermedia* from nature. Can. J. Genet. Cytol. 24**,** 741-759.

Rincones, J., Mazotti, G. D., Griffith, G. W., Pomela, A., Figueira, A., Leal, G. A., Jr., Queiroz, M. V., Pereira, J. F., Azevedo, R. A., Pereira, G. A., Meinhardt, L. W., 2006. Genetic variability and chromosome-length polymorphisms of the witches' broom pathogen *Crinipellis perniciosa* from various plant hosts in South America. Mycol. Res. 110**,** 821-832.

Rincones, J., Meinhardt, L. W., Vidal, B. C., Pereira, G. A., 2003. Electrophoretic karyotype analysis of *Crinipellis perniciosa*, the causal agent of witches' broom disease of *Theobroma cacao*. Mycol. Res. 107**,** 452-458.

Rizet, G., 1953. Sur La Longevite Des Souches De *Podospora anserina*. Cr. Hebd. Acad. Sci. 237**,** 1106-1109.

Rizet, G., 1957. Les modifications qui conduisent `a la senéscence sont-elles de nature cytoplasmatiques? C. R. Acad. Sci. 244**,** 663-665.

Robison, M. M., Horgen, P. A., 1999. Widespread distribution of low-copy-number variants of mitochondrial plasmid pEM in the genus *Agaricus*. Fungal Genet. Biol. 26**,** 62-70.

Rocha, H. M., Wheeler, B. E. J., 1985. Factors influencing the production of basidiocarps and the deposition and germination of basidiospores of *Crinipellis perniciosa*, the causal fungus of witches broom on cocoa (*Theobroma cacao*). Plant Pathol. 34**,** 319-328.

Roten, C. A., Gamba, P., Barblan, J. L., Karamata, D., 2002. Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. Nucleic Acids Res. 30**,** 142-4.

Scarpari, L. M., Meinhardt, L. W., Mazzafera, P., Pomella, A. W., Schiavinato, M. A., Cascardo, J. C., Pereira, G. A., 2005. Biochemical changes during the development of witches' broom: the most important disease of cocoa in Brazil caused by *Crinipellis perniciosa*. J. Exp. Bot. 56**,** 865-877.

Smith, S. E., Read, D. J., 1997. Mycorrhizal symbiosis. Academic Press, Cambridge.

Surzcki, S., 2000. Basic methods in molecular biology. Springer-Verlag, New York.

Talbot, N., 2001. Molecular and Cellular Biology of Filamentous Fungi. Oxford, New York.

Taylor, D. R., Zeyl, C., Cooke, E., 2002. Conflicting levels of selection in the accumulation of mitochondrial defects in *Saccharomyces cerevisiae*. P. Natl. Acad. Sci. USA. 99**,** 3690-3694.

Thompson, J. D., Higgins, D. G., Gibson, T. J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22**,** 4673-4680.

Tudzynski, P., Esser, K., 1979. Chromosomal and extrachromosomal control of senescence in the ascomycete *Podospora anserina*. Mol. Gen. Genet. 173**,** 71-84.

Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O., Hunkapiller, M., 1998. Shotgun sequencing of the human genome. Science. 280**,** 1540-1542.

Wheeler, B. E. J., Suarez, C., 1993. The Pathosystem. In: T. Andebrhan, (Ed.), Disease Management in Cocoa: Comparative Epidemiology of Whitches' Broom. Chapman & Hall, London, pp. 9-19.

Zar, J. H., 1998. Biostatistical Analysis. Prentice Hall, New Jersey.

Figure legends

Figure 1 – Distribution of nuclear and mitochondrial reads from shotgun libraries constructed with total DNA extracted from old (white area) or young (gray area) CP02 mycelia. A – top: number of reads from mitochondrial genome (black bars) in comparison with total reads (gray bars). Number of reads in each library is given in log scale (*y* axis). A – bottom: ratio between mitochondrial and total reads (y axis). B: QQ-Plot analysis showing normal distribution of transformed data (asin√ratio). C: Box-Plot of transformed data (asin√ratio) showing the percent of mitochondrial reads in libraries from young (gray) and old mycelia (white). D. Cultures on plate showing the apparent morphology of the mycelia. Right, cultures of young mycelia; central picture, formation of sectors in a plate recently subcultured from a young mycelia showing the shift to old hyphae; left, growth of old hyphae, forming concentric borders. In the detail, fluorescent light microscopy of cells from both young and old hyphae stained with SYBR® Green I (solid bars = 10 microns). In young mycelia the dikaryotic nuclei are easily identified, while in old cells they appear degraded.

Figure 2 – Plasmid structure and occurrence in different isolates and biotypes. A general scheme of the plasmid is shown at the top; it encompasses two polymerases (*rpo* and *dpoB*) and hypothetical ORFs, which present the same color when they have significant sequence similarity. Identical inverted repeats are connected by dotted lines and their sizes are indicated. The primer positions are indicated along with the expected size of the amplifications for each pair. A representative result of the DNA amplification from 3 different C-biotype isolates, collected in Bahia and Amazon, and two L-biotypes (see Table 1) is shown at the bottom. Note that for each amplification reaction of C-biotype isolates, the most intense band always corresponds to the expected size indicated in the scheme.

Figure 3 – Analysis of GC along the *M. perniciosa* mitochondrial genome. The relative position of conserved genes, intronic ORFs, hypothetical genes in the CWO and CCWO, and tRNAs is shown at the bottom. Local GC percent was obtained by scanning the sequence with sliding windows of 5000/500 bp (black line) and 500/50 bp (blue line). The red line shows the

cumulative GC skew. Traced vertical lines represent areas of the plasmid region: lines 1 and 2 mark the limit of the external inverted repeats; line 3 represents the border of the last ORF in CCWO with conserved codon usage; line 4 indicates a sharp alteration in the amplitude of the GC percentage.

Figure 4 – A – Semi-quantitative PCR analysis of a total DNA sample from senescent mycelia (>10 subculturings) used to compare the relative amounts of specific mitochondrial, plasmid and genomic sequences. PCR conditions were identical for the three reactions, using three different specific primer pairs: P9 and P10 for plasmid DNA; A and B for mtDNA and URA3F and URA3R for nuclear DNA (Table 2). Gradual accumulation of the PCR product was verified by collecting 5 μL aliquots from each reaction after the number of cycles indicated. **M**= size marker λ-*Hind*III (Invitrogen). **B** – Northern blot analysis of the alternative oxidase gene expression in young and senescent cultures of saprotrophic mycelia of *M. perniciosa*. Upper panel shows the 30S and 18S ribosomal bands of the total RNA (10 μg) of each condition. Lower panel shows the northern blot hybridization with a ^{32}P-labelled probe corresponding to the complete cDNA coding for AOX of *M. perniciosa.* **1**= total RNA from young saprotrophic mycelia (<10 suculturings); **2**= total RNA from senescent saprotrophic mycelia (>10 subculturings).

Figure 5 – Unrooted tree of the genes encoding polymerases from mitochondrial plasmids including all kalilo-like plasmids published to date obtained by Minimum Evolution (5000 bootstrap resamplings). Numbers next to each internal branch are bootstraps values. The double lined branch indicates a major separation into two groups. Basidiomycetes are underlined.

Table 1 – *M. perniciosa* isolates used in this study, organized by biotypes (Biot), municipality of collection and institution responsible for this collection (Repository).

Isolate	**Biot**	**Local of collection**	**Repository**
CP02	C	Itabuna, BA, Brazil	UESC[1]
CP09	C	Ilhéus, BA, Brazil	CEPLAC[2]
Belmonte	C	Belmonte, BA, Brazil	UFB[3]
Ilhéus	C	Ilhéus, BA, Brazil	UFB
Santo Amaro	C	Santo Amaro, BA, Brazil	UFB
FA42, FA276, FA277, FA278, FA562, FA563	C	Itabuna BA, Brazil	FAC[4]
FA281	C	Aiquara, BA, Brazil	FAC
FA287	C	Inema, BA, Brazil	FAC
FA293	C	Gandu, BA, Brazil	FAC
FA300	C	Ibirataia, BA, Brazil	FAC
FA311	C	Itagiba, BA, Brazil	FAC
FA317	C	Ilhéus, BA, Brazil	FAC
BP10	C	Itapebi, BA, Brazil	FAC
FA551	C	Tabatinga, AM, Brazil	FAC
ESJOH-1	C	Marituba, PA, Brazil	ESALQ[5]
ESJOH-2	C	Ouro Preto, RO, Brazil	ESALQ
ESJOH-3	C	Belém, PA, Brazil	ESALQ
ESJOH-4	C	Altamira, PA, Brazil	ESALQ
ESJOH-5	C	Medicilandia, PA, Brazil	ESALQ
ESJOH-6	C	Ariquemes, RO, Brazil	ESALQ
ESJOH-7	C	Manaus, AM, Brazil	ESALQ
ESJOH-8	C	Ji-Paraná, RO, Brazil	ESALQ
ESJOH-9	C	Alta Floresta, MT, Brazil	ESALQ
SCFT	L	San Carlos, Napo Prov,	UW[6]

		Equador	
FA322	L	San Carlos,,Napo Prov, Equador	FAC
LEP1, LA10, LA17, LC3, LC11	L	Pichilingue, Equador	UW
SCL4, SCFT48	L	San Carlos, Napo Prov, Equador	UW

(1) UESC (*Universidade Estadual de Santa Cruz*) in Ilhéus, Bahia-Brazil, collected by Júlio Cascardo; (2) CEPLAC (*Comissão Executiva do Plano de Lavoura do Cacau*) in Ilhéus, Bahia-Brazil, collected by Karina Gramacho; (3) UFB (*Universidade Federal de Brasilia*) in Brasilia-Brazil, collected by Maricília Arruda; (4) FAC (*Fazenda Almirante Cacau*) in Ilhéus, Bahia-Brazil, collected by Alan Pomella; (5) ESALQ - CENA (*Centro de Energia Nuclear na Agricultura, in Escola Superior de Agricultura "Luiz de Queiroz"*) in Piracicaba, São Paulo-Brazil, collected by Paulo Albuquerque (*ERJOH – CEPLAC*); (6) UW (University of Wales), in Aberystwyth-U.K., collected by Gareth Griffith. Isolates FA 104, FA 607, FA 608 and FA 609 were kindly provided by Dr. Robert Weigart Barreto (*Universidade Federal de Viçosa*).

Table 2 – List of PCR primers used in this work. *Or.*: Orientation of the extension by that primer;. >> and << indicate the primers on the 5' and 3' sides, respectively. *Init. pos* and *Final pos* are for the mtDNA location (bp) where sequence is homologous to the 5' and 3' primers, respectively. Base count is according to Figure 2. URA3F and R primers are from *M. perniciosa* genome.

Primer	**Sequence**	**Or.**	**Init. pos**	**Final pos**
P1	GCAGGGAAGGGATATATAGG	>>	73,212	73,231
P2	TTTGAGAGAGCATCAAATCC	<<	73,755	73,736
P3	TTTTGAGAGAGCATCAAATCC	>>	84,485	84,505
P4	AAAGAACTGAAATCCGAGG	<<	85,184	85,166
P5	CATTTGTAAAGGAAAAGATGG	>>	79,490	79,510
P6	TTCTTCTTTCTTCGCAGC	<<	80,432	80,415
P7	GAAGAAGCTGCTGGATCAGG	>>	76,645	76,664
P8	GCTCCATTTTGTCAACAAGC	<<	77,446	77,427
P9	AAAAGCGAAGTTAAGCAAGAGG	>>	83,266	83,287
P10	TGGGGAGACTATGGAAATGC	<<	84,098	84,079
A	GACCCTATGCAGCTTCTACTG	>>	3,062	3,082
B	TTATCCCTAGCGTAACTTTTATT	<<	4,033	4,013
URA3F	ACCACAAATGCCTTCACCTC			
URA3R	GGGTCCTTTCCGTATATTCCTC			

FIGURES

Figure 1

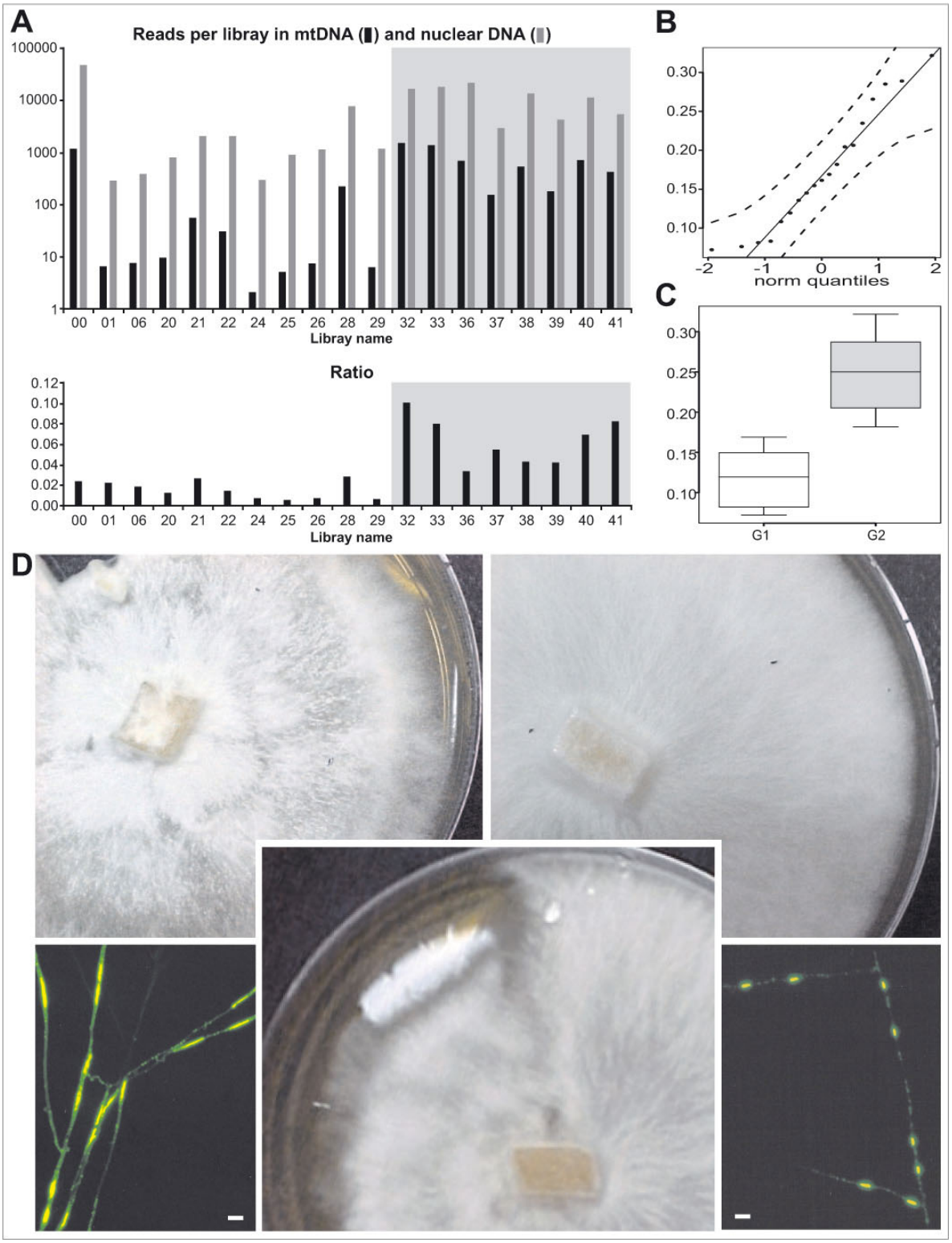


Figure 2

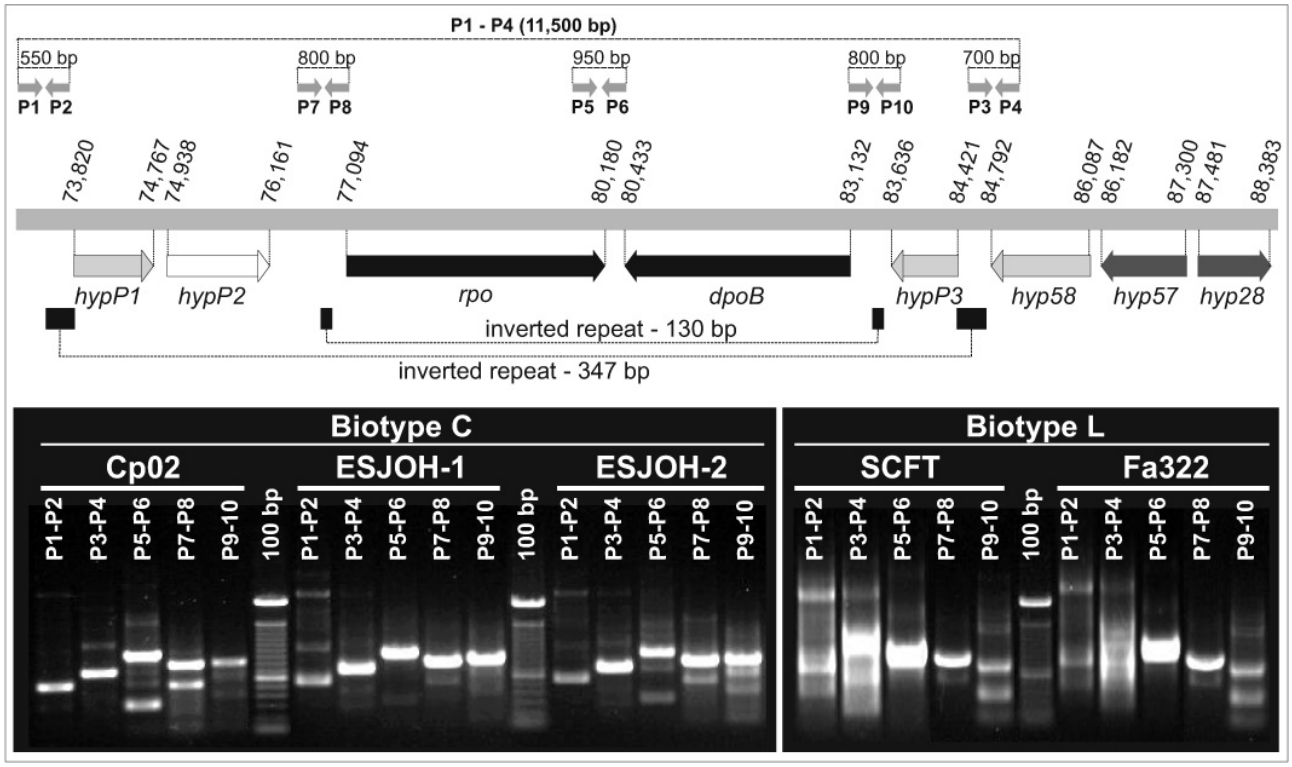

P1 - P4 (11,500 bp)
550 bp
800 bp
950 bp
800 bp
700 bp
P1 P2
P7 P8
P5 P6
P9 P10
P3 P4
73,820
74,767
74,938
76,161
77,094
80,180
80,433
83,132
83,636
84,421
84,792
86,087
86,182
87,300
87,481
88,383
hypP1
hypP2
rpo
dpoB
hypP3
hyp58
hyp57
hyp28
inverted repeat - 130 bp
inverted repeat - 347 bp
Biotype C
Cp02
ESJOH-1
ESJOH-2
Biotype L
SCFT
Fa322
P1-P2
P3-P4
P5-P6
P7-P8
P9-10
100 bp

Figure 3

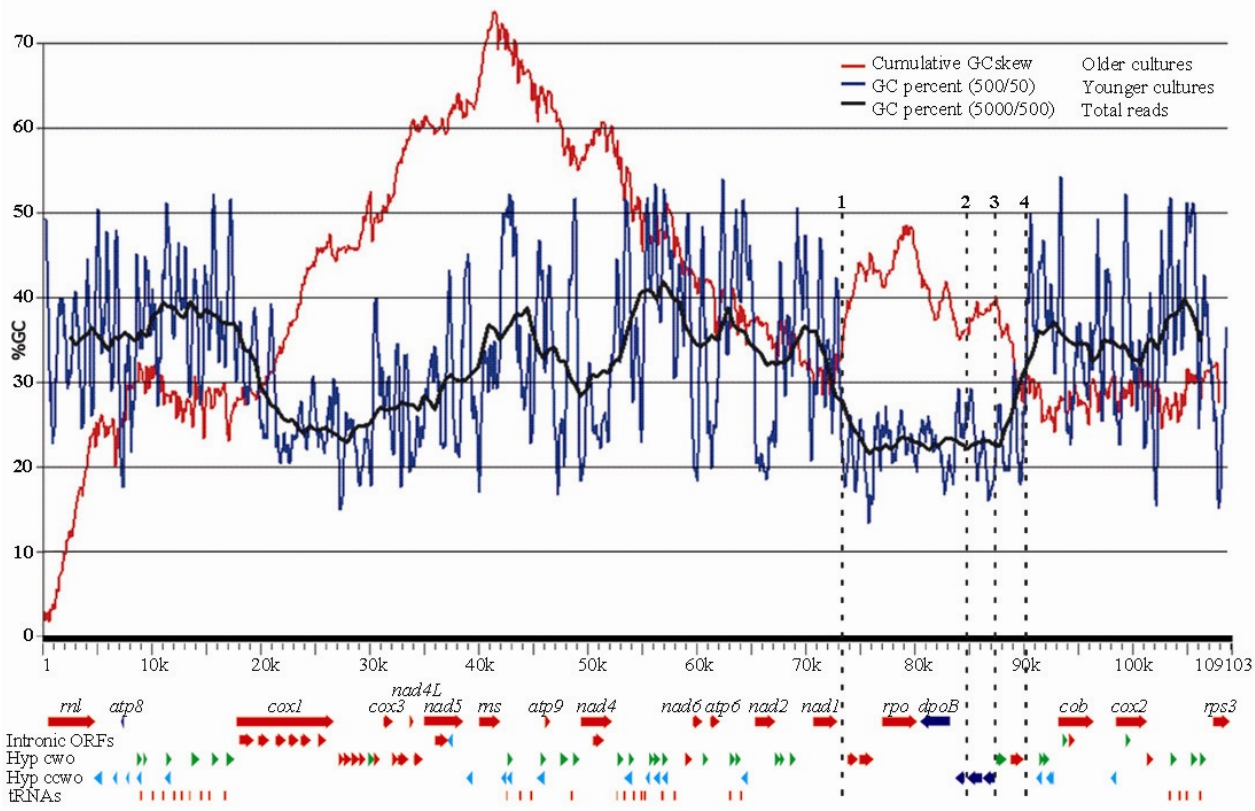

Cumulative GCskew
GC percent (500/50)
GC percent (5000/500)
Older cultures
Younger cultures
Total reads
%GC
70
60
50
40
30
20
10
0
1
2
3
4
1
10k
20k
30k
40k
50k
60k
70k
80k
90k
100k
109103
rnl
atp8
cox1
nad4L
cox3
nad5
ms
atp9
nad4
nad6
atp6
nad2
nad1
rpo
dpoB
cob
cox2
rps3
Intronic ORFs
Hyp cwo
Hyp ccwo
tRNAs

Figure 4

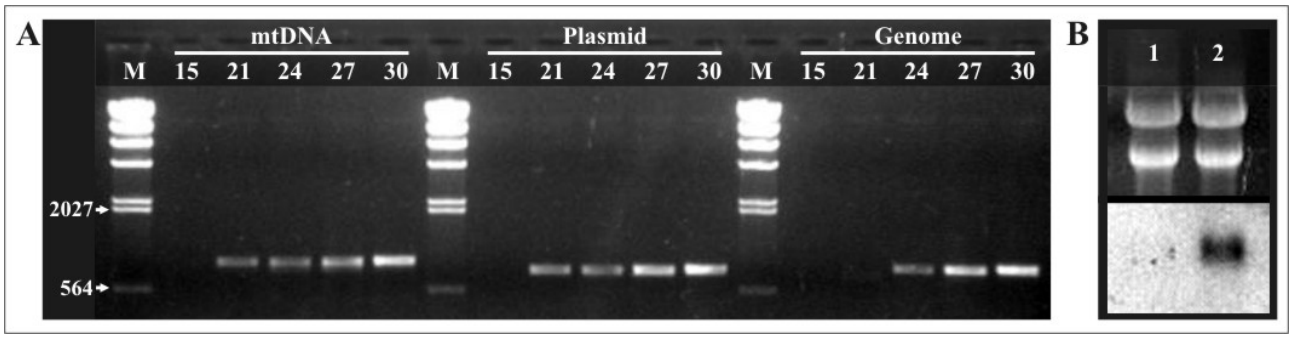
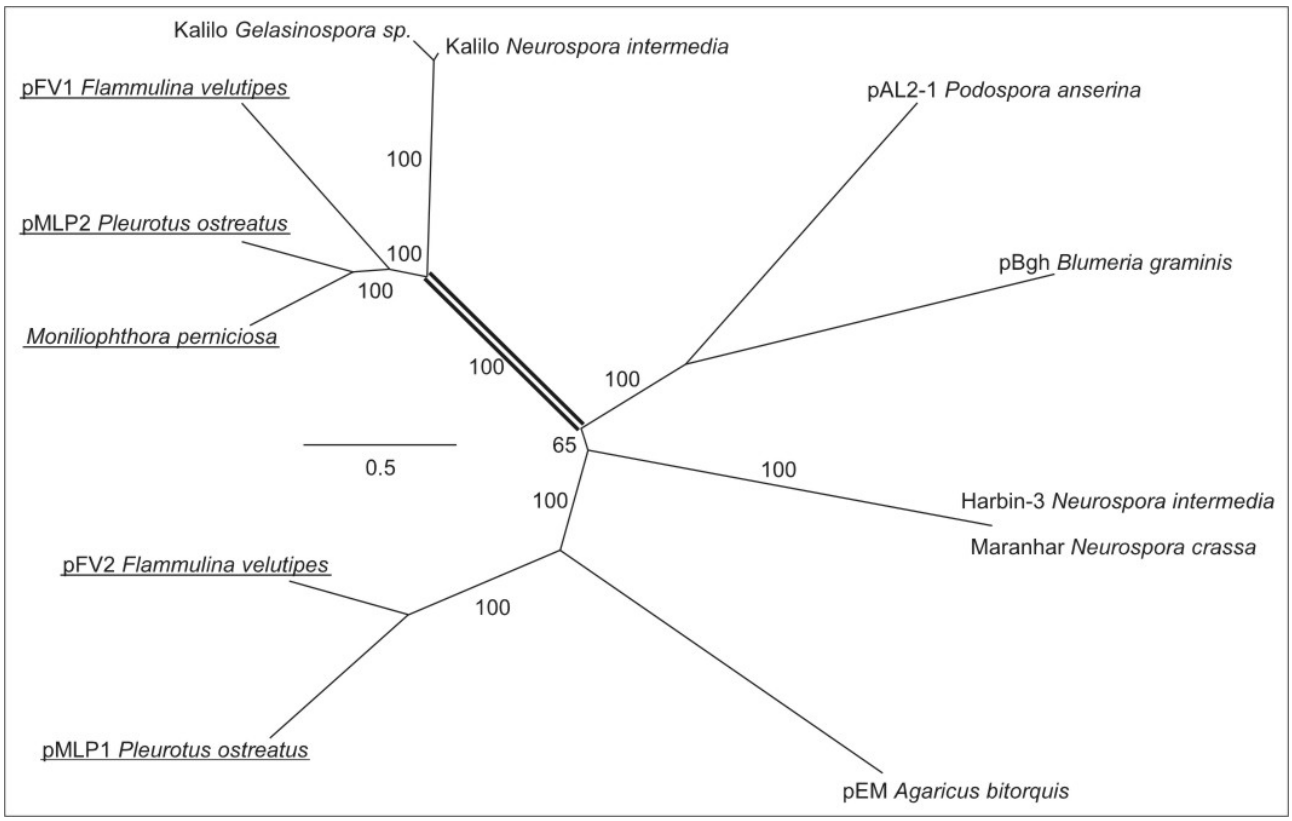


Figure 5



4. CAPÍTULO 3

Gene Projects: a Web application for ongoing annotation in EST and Shotgun genome projects

Marcelo Falsarella Carazzolle, Eduardo Fernandes Formighieri, Luciano Antonio Digiampietri, Marcos Renato Rodrigues Araújo, Gustavo Gilson Lacerda Costa and Gonçalo Amarante Guimarães Pereira

Obs.: Detalhes sobre o programa e o sistema de anotação nos anexos C, D e E.

Gene Projects: a Web application for ongoing annotation in EST and Shotgun genome projects

Marcelo Falsarella Carazzolle[1*], Eduardo Fernandes Formighieri[1], Luciano Antonio Digiampietri[2], Marcos Renato Rodrigues Araujo[1], Gustavo Gilson Lacerda Costa[1] and Gonçalo Amarante Guimarães Pereira[1]

[1] Laboratório de Genômica e Expressão – Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, 13083-970, Campinas – SP, Brazil.
[2] Instituto de Computação, Universidade Estadual de Campinas, 13084-971, Campinas – SP, Brazil.

* Corresponding author: mcarazzo@lge.ibi.unicamp.br, Phone: +55 (19) 3788-6651, FAX: +55 (19) 3788-6235

Abstract

Genome projects, both DNA and ESTs, generate a great amount of information. It demands time and a well-structured bioinformatics laboratory to manage these data. These genome projects use information available in heterogeneous formats from different sources. The amount and heterogeneity of this information, as well as the absence of a world consensus pattern, make the integration of these data a difficult task. At the same time, sub-tasks of these projects are very complex, such as microarray analyses. It creates a demand, easily verified in real projects, for the development of creative solutions for ongoing annotation, thematic projects, and microarrays experiments, among others. This paper presents Gene Projects, a well-tested system developed to integrate all these kinds of solutions.

1. Introduction

Large-scale sequencing projects often target either genomic or cDNA sequences. For genomic sequences, the most used sequencing strategy is the Whole Genome Shotgun - WGS (Venter et al., 1998), where the DNA is broken into fragments that are sequenced randomly. The goal is to assemble these fragments to reconstitute the whole genome sequence or, at least, to obtain an ordered list of large genomic regions with interleaved gaps. In the latter case, we say that it is a draft genome. On the other hand, cDNA sequencing projects aim to look directly at the genes expressed in different conditions. The cDNA is synthesized using the cellular mRNA as template, thus it samples the pool of active genes in a specific cell condition. The fragments of sequenced cDNA are called ESTs (Expressed Sequence Tags), thus, cDNA sequencing projects are often called EST genome projects. The ESTs are often assembled to improve the length and the quality of the ESTs, to obtain the current set of full sequence of cDNAs and to identify expression patterns in different studied conditions.

Genome projects produce a great amount of reads, which are chromatograms representing sequence fragments. This sequencing stage usually lasts for many months. On the other hand, the available tools for retrieving information from these incoming data demand bioinformatics expertise and many computational resources. For this reason, the end user typically has access to this information only after that great part, if not all, of the sequencing had been made. Besides this, small projects almost never can have an own bioinformatics infrastructure.

At the same time, there are different kinds of information related with genome projects that can be produced and utilized during the process of assembly and annotation of the genome. For example, there is a strong relationship between a genome project and the choice of clones to be analyzed in microarrays, as well as the inverse pathway, that is, the selection of sequencing clones from microarrays analyses. To favor this kind of interaction, it is necessary an efficacious search mechanism and a robust analysis mechanism, both able to deal with thousands of clones and to minimize the possibility of human errors. In the other extremity of this process, the annotation of genes and other complex activities need the interaction among several kinds of heterogeneous data, databases and tools. It also needs a special care with the inexistence of a

world consensus for the used nomenclature and the absence of guarantee that all data already annotated are right.

A way to minimize these potential errors is the work of specialists in the different areas involved in the annotation process of each structure. Here, there is another difficulty because these specialists are spread around the world.

In this context of increasing needs for sophisticated data analysis and user helpful interfaces in the genome projects actually carried out in our laboratory, we developed a system called Gene Projects (GP). The GP system can manage small and great genome projects (both, DNA and EST projects) and it allows that annotation occurs in parallel with the sequencing, without the necessity of computer specialists. It is also able to separate analyses by theme, through what we called “projects”. Besides that, it provides interaction among several tools and techniques, facilitates search and annotation of genes, allows access from researchers via Web and can be used by researchers that do not have specialized knowledge in computation or bioinformatics.

2. Projects and Ongoing annotation

The system is called Gene Projects due to the importance of what we called thematic projects (or short, projects). Each user of the system has a specific login and password. With this login he/she can create and manage his/her projects, a flexible and powerful way to make data mining. A project is a structure inside the system where researchers can develop and organize their thematic studies. The user can add reads through several search mechanisms such as BLAST (Altschul et al., 1997) results, read names, keywords, etc. Once the reads are chosen, the user can assemble them, view and edit the assembly results, improve the quality of the contig and enlarge the contig (Saturation Blast), find ORFs, select sub-sequences of a contig (seeds) to future annotation, manage the mapping among the plates of 96 wells used in the sequencing to or from plates of 384 wells used in microarrays experiments, among others. Through these projects it is possible to find, select and send to the annotation, genes of interest for studying genome such as infection related genes or genes that belong to metabolic pathways, for example. One of the advantages of

our system is that all these things can be done during the sequencing stage through the gene annotation interface.

3. System description and architecture

Gene Projects is written in perl (http://www.perl.org) and uses standard, open source modules such as CGI.pm (http://stein.cshl.org/WWW/software/CGI), GD.pm (http://www.boutell.com/gd/) and DBI.pm (http://dbi.perl.org/). The system needs Linux operating system with Web server (i.e. Apache - www.apache.org) and MySQL server (www.mysql.com) installed. As default, Gene Projects needs some programs, databases and directory structure for its execution. Table 1 shows the main required programs.

Figure 1 shows the main characteristics of the system, including services that are being developed and soon will be integrated in Gene Projects. The Web interface allows users spread around the world to access our system. The operations carried out by the users are processed locally in the processing module.

GP has three main integrated functionalities: data mining, annotation and microarrays management. Data mining and annotation are strongly related. Data mining process helps the scientists to analyze great amount of data and thus facilitates the annotation. In the other hand, annotation process (identification and classification of genes) can identify terms to be mining, such as, genes of one specific metabolic pathway.

Microarrays can be built with already known cDNA (cDNA chip) or with random ESTs (blind chip). Gene Projects facilitates the selection of cDNA of interest for the confection of cDNA chips (through data mining and annotation). The selection of ESTs of blind chips is facilitated to the automatic identification of ESTs with a given characteristic of interest, for example, ESTs that presented the same expression pattern. In other words, blind chips can indicate themes for metabolic studies, which can become new *projects* in Gene Projects.

Search tool is one of the most important activities in systems that manage genome projects. Gene Projects has a graphical tool for “Advanced Search” that allows boolean queries, allowing the easy use of queries over the main fields in the databases. Another important functionality of the system is the set of tools that allows the assembly of a set of reads of each *project* and the visualization of the results. This process typically needs expertise of the end user, but it is very simple using the Gene Projects Web interface.

The main advantage of this architecture is that our servers do the great amount of processing, which makes unnecessary for external users to have high performance computers with great amount of memory to perform the kind of operation offered by our system. The system components that demand more CPU time are executed as child processes, running independently from the rest of the system, so, the user can disconnect and, when he reconnects the results can be ready. When there are several processes requisitions to be executed at the same time, they are stored and executed using a queue schema. The results are stored in our database and can be queried at any time.

4. Technological advantages

Our approach is based in perl scripts available via CGI technology. There are some advantages of this approach:

- ✓ Scalability: due to the fact that the internal computational infrastructure of our laboratory is hidden by the Web interface, we have freedom to enact our processing capabilities without changing any user interface, for example, using a distribute infrastructure, such as a computational Grid, to execute the main processes.
- ✓ Availability: all services are available in the Internet and the user needs only a browser to make searches and edit data;
- ✓ Automatic Updates: each time that a user access a GP Web interface he/she sees the last version of the system interface and any updates in internal softwares are transparent to the user;
- ✓ Updated data: whenever an user updates some information, this information is automatically updated in all system and available to the others users;

- ✓ Security: there are some levels of user authentication that determine which kind of information each user can read and/or edit.

5. Practical results

Gene Projects was originally developed to manage data from the *Crinipellis perniciosa* fungus genome (www.lge.ibi.unicamp.br/vassoura) and it has been used in several projects, such as: Coffee genome project (www.lge.ibi.unicamp.br/cafe), Citrus (http://biotecnologia.centrodecitricultura.br) and Eucalyptus genome project (www.lge.ibi.unicamp.br/eucalyptus). More projects and details can be obtained at www.lge.ibi.unicamp.br.

The system was developed to allow users, typically biologists, to make specific searches in the set of sequenced reads. These reads can be filtered for some criteria, such as quality of sequences, percentage of vectors (contaminants), among others. The filtered reads can be processed by several bioinformatics programs and can be compared against a lot of biological databases. All these results are organized and stored inside the projects.

Figure 2 shows some Gene Projects Web interfaces. The Figure 2-A shows the mechanism that users can select and add reads into projects. This search mechanism is composed for:

- ✓ Reads name search: It is useful to analyze specific reads or all reads from one plate (for example, to see the quality of the sequenced plate);
- ✓ Keyword search: It is useful for thematic searches (for example, to find reads related with the product of a given gene);
- ✓ Blast search: It is useful for searches based on similarity of sequences (for example, to find reads which sequence is similar to the sequence of a given gene);
- ✓ Pattern search: It is useful for searches by domain or repeats regions (for example, to find microsatellites, protein domains or transcription factors).

5.1 Assembly

The reads of a given project can be clustered and assembled. CAP3 is used for ESTs projects and phrap for Shotgun projects. The available results of this assembly, showed in the Figure 2-B, are:

- ✓ Assembled fasta sequence;
- ✓ Assembled visualization;
- ✓ Blast results of the assembled sequences against the gene banks NR or NT;
- ✓ Search for ORFs inside the sequences;
- ✓ Assembled sequence reverse complement;
- ✓ Reads that belong to the cluster.

The interest cluster can be submitted to the automatic annotation process, in which the cluster is compared with several databases like GO (http://www.geneontology.org) and NR (http://www.ncbi.nlm.nih.gov/blast/producttable.shtml) and this information is available in the interface showed in the Figure 2-C (annotation interface).

5.2 Saturation Blast

In the typical data mining process, it is common to have keywords searches and searches based on Blast results. These are useful searches, but they cannot find every related read. To improve the quality and increase the number of reads of a given cluster (increasing the size of the cluster) it is necessary to add the right reads to the project. To facilitate this process, we developed a Blast saturation mechanism. This mechanism starts from the original cluster sequence and realizes successive comparisons with the bank of reads, adding the most similar reads and redoing the assembly until to achieve some stop parameter. Typically this saturation runs until the cluster achieves a given length (var2) or when, in a given iteration, there are no more reads with at least a minimum similarity level (var1) to be added to the cluster (see Figure 3).

This tool allows that, independently from the bioinformatics knowledge from the end user, he/she can works with his/her project with the best assembly (following some criteria) using every read sequenced until that moment.

5.3 Annotation

There are three classes of user that interact with GP manual annotation interface in a genome project. The first is the *annotator*, which is the actor that fills in information about the clusters. The *selector* selects interesting clusters for his annotators group and reviews the annotation. The last, the *curator*, is a special kind of user that has the ability to review information filled in by all annotators from all groups. All these classes have tools to register and manage logins and passwords. In the case of personal projects, the owner can chose and change his/her kind of user.

The annotation interface is constituted by eight main sections and has many facilities to save annotator's time and labor. Next, we list the sections; provided facilities and expected user interaction.

- ✓ Classification: here the annotator can view/edit GO terms for the cluster. There is also a direct link from the selected term to Amigo DAG Viewer. A second classification system, defined by the coordinators of the project usually is inserted here to.
- ✓ Identification: in this section, the annotator enters information about the product, phenotype, domain, homologous organism, gene symbol, Enzyme Commission number and Transport Commission number.
- ✓ Visualization: the annotator can view the sequence of the cluster and its reverse complement, the reads that constitute the cluster as well as the assembly of the cluster and its translated sequence in the six frames.
- ✓ Flags: here the annotator set flags to the cluster. There are flags to indicate:
    - o Whether the cluster contains the complete coding sequence of a known gene;
    - o Whether the clone(s) of cluster contains the complete coding sequence of a known gene;
    - o Whether the cluster contains an intron sequence;
    - o Whether the cluster has central role for project's goals;
    - o Whether the cluster has assembling problems such as a frame shift or a significant repeated region, for instance;
    - o Whether the cluster is a contaminant.
- ✓ Pre-processed Blast alignments: here the annotator can view summary data about the

alignments to some databases. Optionally the annotator can visually inspect the alignments by clicking in a link. The list of databases is fully customizable by the project's bioinformatics team in accord to the specific project's requirements.

- ✓ Easy searches: in this section there are links to a set of web Blast interfaces. These Blast searches, unlike the former ones, are dynamically processed. The web Blast interfaces are loaded with custom parameters and the query sequence input field already filled in. There is also an interface to keyword searches in some biological databases. Bioinformatics team customized the list of sites available for both keyword searches and Blast searches.
- ✓ Notepad: in the notepad section there are two input text fields. In the first, the cluster owner can enter personal notes and relevant information about the cluster that was not inputted above. Any annotator can edit the second, called guest notepad,, even if he is not the cluster owner, and it must contain information to help the cluster owner to annotate it.
- ✓ Control: In this region, the annotator can signalize a finished annotation, save the updates and view the annotation historic of cluster. The historic contains all edit operations made in this cluster, including user that made the changes, date and time. The historic has such a level of detail that it can be used to reconstruct the annotation database. The annotator can also reserve the cluster for functional analysis. The system offers an additional functionality to the selector: he can return a cluster that he has selected for his group. For the curator there is a check box that he may use to indicate whether the cluster annotation was reviewed or not.

5.4 Microarray analyses

There is another kind of analysis that can be done with Gene Projects, the microarray analyses, before and/or after microarray experiments. The microarray experiment is divided in two steps: pre-processing and pos-processing. In the pre-processing stage it is necessary to map the clones from plates of 96 wells to plates of 384 wells and, finally, to spot it in the lamina. After this, the microarray starts. The experiment owner defines the criteria of mapping into plates of 96 wells. The mapping from plates of 96 wells to plates of 384 wells follows a rule given by the multi-channel pipette. The experiment owner defines the “spotting” step, but the robot (Flexys – Genomic Solutions – www.flexys.com) generates a file with this information.

In the pos-processing step it is necessary to process the color intensity result of each spot, following these stages: statistical significance normalization (Yang et al, 2002), validation of the colors intensity and gene clustering by expression class (Aittokallio, 2003).

Gene Projects executes these three stages in the pre-processing step:

- ✓ Plates of 96 wells: the arrangement is automatic, the sort criteria is read name alphabetic order. If the user wants, he can make a manual rearrangement;
- ✓ Plates of 384 wells: this stage is automatic following the logic given by the mechanic process.
- ✓ Lamina spotting: the Gene Projects is able to read the robot result file, therefore it generates automatically the rearrangements in the respective grids and spots. Actually, this stage is automatic only when using Flexys – Genomic Solutions robots.

All these stages are registered in the GP interface and there are links to BlastX results. This integration is fundamental to correct interpretation of experimental results.

6. Correlated works

In literature is possible to find many similar programs to Gene Projects, in special after 2002. But, these programs are specific for some situations, for example only microarray analyses or annotation. Because of this, we are showing in table 2 the comparison between GP and some similar programs in function of their characteristics. In this table is possible to see that many programs have a possibility to manage genome projects, but only in large scale. Therefore it is not possible to work in small projects especially during the sequencing stage.

7. Conclusions

Currently, all kinds of information systems must have an especial concern about user-connected issues, such as usability and interface. Systems that deal with genomic projects, by working with large volumes of data, must have an additional concern on the data presentation to users, through graphic visualizations, data summaries and connection among data from heterogeneous sources. Another important characteristic on the genome projects is that the data generation through sequencing is time consuming. Therefore, the systems must make available the already generated

information before the end of the process of sequencing (to allow what we call “ongoing annotation”).

As showed, Gene Projects is a system, available via Web, which has all these concerns. It has been used in real genome projects and has been produced satisfactory results, by integrating, in a transparent way, heterogeneous data and tolls, by extending the functionalities of another system.

Future work

As future work we will improve some GP functionalities. We will take a special care with microarray pos processing, like normalization, statistic significance tests and clustering methods for the identification of expression patterns.

Acknowledgments

The work described in this paper was partially financed by FAPESP, CNPq and CAPES. We acknowledge the researchers that were or are end users of our system and help us to improve the functionalities, interfaces and the general quality of Gene Projects.

References

Aittokallio T, Kurki M, Nevalainen O, Nikula T, West A and Lahesmaa R (2003) Computational Strategies for Analyzing Data in Gene Expression Microarray Experiments. Journal of Bioinformatics and Computational Biology, 1:541-586.

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. Nucleic Acids Research 25:3389-3402.

Berezikov E, Plasterk RH, Cuppen E (2002) GENOTRACE: cDNA-based local GENOme assembly from TRACE archives. Bioinformatics. 18:1396-1397.

Ewing B and Green P (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194.

Hotz-Wagenblatt A, Hankeln T, Ernst P, Glatting KH, Schmidt ER, Suhai S (2003) ESTAnnotator: A tool for high throughput EST annotation. Nucleic Acids Research, 31:3716-3719.

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

Kumar CG, LeDuc R, Gong G, Roinishivili L, Lewin HA, Liu L (2004) ESTIMA, a tool for EST management in a multi-project environment. BMC Bioinformatics, 5:1-10.

Mao C, Cushman JC, May GD, Weller JW (2003) ESTAP-an automated system for the analysis of EST data. Bioinformatics. 19:1720-1722.

Matukumalli LK, Grefenstette JJ, Sonstegard TS, Van Tassell CP (2004) EST-PAGE--managing and analyzing EST data. Bioinformatics, 20:286-288.

Paquola AC, Nishyiama MY Jr, Reis EM, da Silva AM and Verjovski-Almeida S (2003) ESTWeb: bioinformatics services for EST sequencing projects. Bioinformatics. 19:1587-1588

Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M (2004) PartiGene--constructing partial genomes. Bioinformatics, 20:1398-1404.

Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO and Hunkapiller M. (1998) Shotgun sequencing of the human genome. Science. 280:1540-1542.

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

Tables

Table 1 – List of programs required to Gene Projects execution.

Program	Description	Reference/URL
Phred	Base calling and generation of quality values from chromatograms	(Ewing et al., 1998)
Cross_match	Vector screening and generation of FASTA sequences files with masked vector sequences	http://www.phrap.org
Phrap	Clustering and assembling program for shotgun sequences	http://www.phrap.org
CAP3	Clustering and assembling program for ESTs sequences	(Huang et all, 1999)
Local Blast	Sequence alignment	(Altschul et al., 1997)
Fuzznuc	Pattern search	Emboss package (http://emboss.sourceforge.net)

Table 2 – Some surveyed systems and their characteristics.

Characteristics	ESTWeb[1]	ESTAP[2]	GENOTRACE[3]	ESTAnnotator[4]	PartiGene[5]	ESTIMA[6]	EST-PAGE[7]	Gene Projects
i. Web interface	X	X		X	X	X	X	X
ii. Managing ESTs genome projects	X	X		X	X	X	X	X
iii. Managing Shotgun genome projects			X					X
iv. Processing of chromatograms files	X	X	X	X	X		X	X
v. Trimming and filtering sequences	X	X		X	X		X	X
vi. Reads annotation	X	X		X	X		X	X
vii. Clustering and assembly		X	X	X	X			X
viii. Clusters annotation		X	X	X	X			X
ix. Search tools in annotations data	X	X	X		X	X	X	X
x. Microarray analysis								X
xi. Thematic projects environment						X		X
xii. Working in genome projects in progress (sequencing stage)	X	X	X					X
xiii. Data access control (username and password)						X		X

Table references: [1] Paquola et al. 2003; [2]Mao et al., 2003; [3]Berezikov et al., 2002; [4]Hotz-Wagenblatt et al., 2003; [5]Parkinson et al., 2004; [6]Kumar et al., 2004; [7]Matukumalli et al., 2004.

Figures

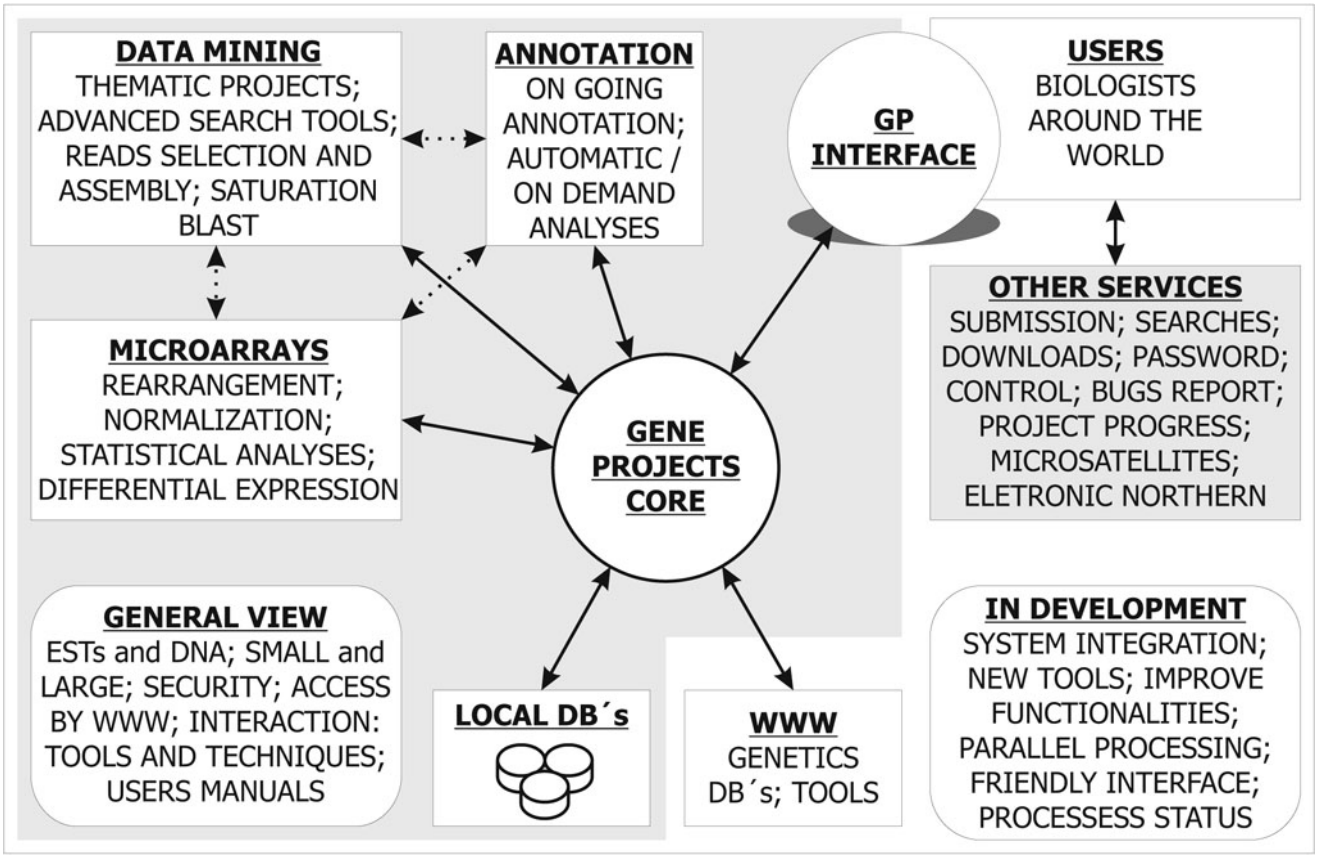


Figure 1 – The Gene Projects architecture. The box on the lower-right corner shows functionalities and features that are being developed and will be soon integrated to Gene Projects. The box on the lower left corner brings a general view on the Gene Projects features.

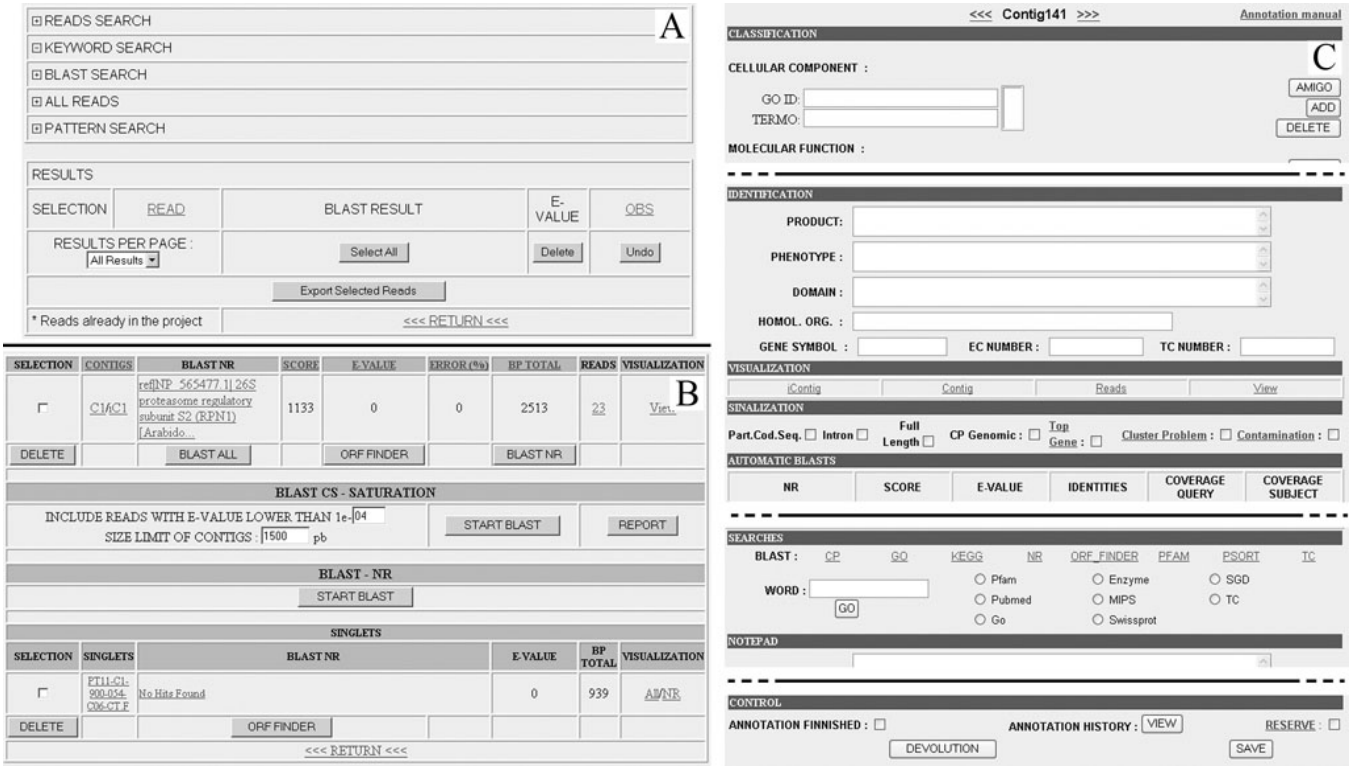


Figure 2 – Flowchart describing the Saturation Blast algorithm.

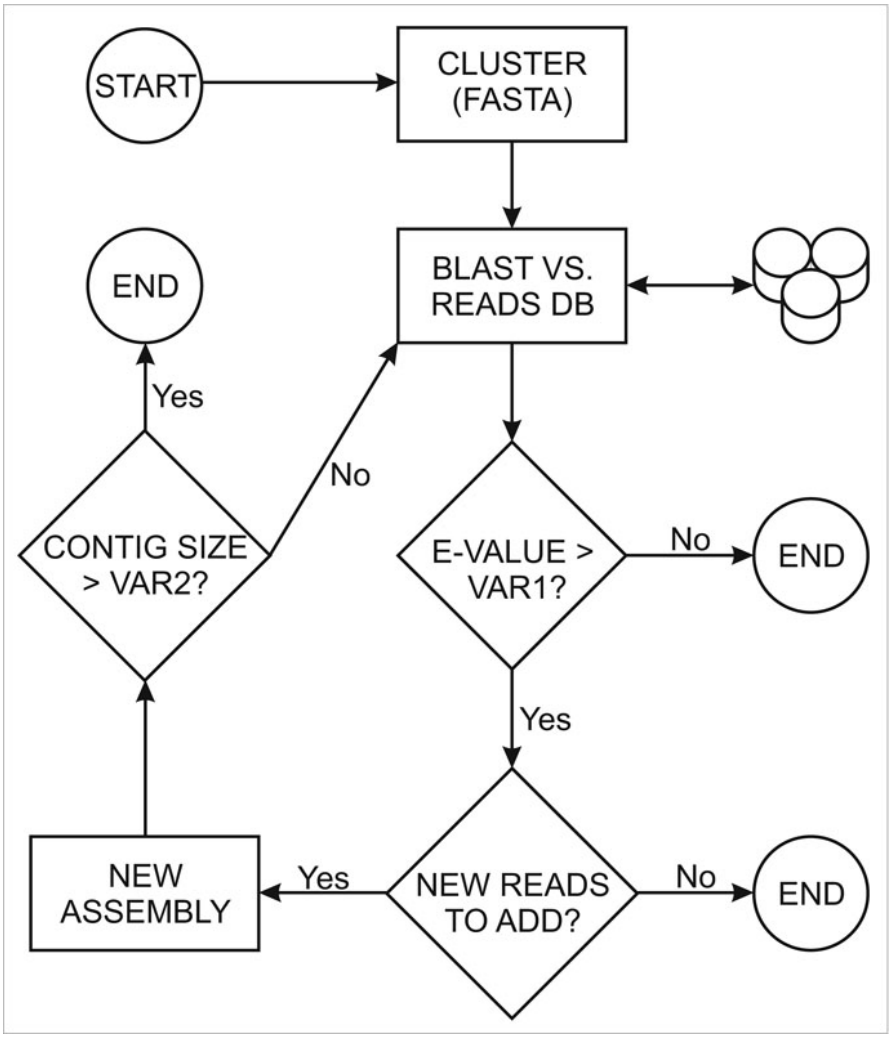


Figure 3 – Saturation Blast schema.

5. CONCLUSÕES

- O genoma mitocondrial de *Moniliophthora perniciosa* é o maior seqüenciado até o momento e o primeiro em que se encontrou um plasmídeo kalilo-like integrado completamente, contendo dois genes que codificam polimerases (*rpo* e *dpoB*) e ainda ORFs hipotéticos. A integração desta estrutura foi confirmada em todas as variedades testadas do biótipo C. A função desta estrutura, sua presença e caracterização em outros biótipos e o motivo de sua integração são alvos importantes para estudos futuros.

- O genoma mitocondrial (mtDNA) de *M. perniciosa* apresenta todos os genes normalmente encontrados em mtDNAs de fungos, tRNAs para todos os aminoácidos e ORFs que apresentam semelhanças com genes típicos, indicando a possibilidade de que codifiquem proteínas ainda desconhecidas. A maioria dos genes encontra-se em sentido horário. Estudos funcionais destes ORFs podem levar ao descobrimento de mecanismos específicos do fungo.

- Não foi detectada sintenia entre o mtDNA de *M. perniciosa* e outros genomas mitocondriais de organismos próximos. Ao contrário de outros organismos, o mtDNA de *M. perniciosa* apresenta poucos íntrons e um grande espaço entre os genes típicos, que contém diversos ORFs hipotéticos que podem vir a ser caracterizados como novos genes. Foi utilizada uma análise estatística (análise de componentes principais) não descrita anteriormente para a análise de uso de códon, que se mostrou muito interessante para apresentar de forma completa a comparação entre o uso de códon dos diferentes ORFs do genoma. Futuramente pode-se buscar descobrir o motivo do tamanho incomum deste mtDNA, o que levou a isto e que vantagens advém deste material gênico adicional.

- O fungo *M. perniciosa* é o primeiro basidiomiceto a ter o fenômeno de senescência descrito, durante a qual a quantidade de DNA mitocondrial é significativamente reduzida. Esta descoberta indica que pode haver uma subdivisão na atual descrição do ciclo de vida do fungo, que inclui quatro fases: esporos; micélio biotrófico; micélio saprotrófico/necrotrófico; e corpo de frutificação. O micélio saprotrófico, na verdade, apresenta duas fases distintas em relação a respiração e produção de energia. Embora plasmídeos kalilo tenham sido relacionados à senescência, não há indícios de que o plasmídeo inserido no genoma

mitocondrial da *M. perniciosa* tenha relação com este processo. O entendimento do mecanismo de senescência é um dos pontos chave na compreensão da interação planta-patógeno, podendo facilitar o combate à doença.

- A principal característica e vantagem do Gene Projects é a análise imediata da informação seqüenciada, mesmo no início do processo de seqüenciamento. O sistema foi e é utilizado e testado em diferentes projetos, permitindo o direcionamento de trabalhos funcionais desde o início do processo de seqüenciamento. Como exemplo, no Projeto Vassoura de Bruxa diversos genes encontrados através do Gene Projects e analisados pelo sistema de anotação foram o início de análises funcionais, como oxidase alternativa (cuja expressão aumenta como resposta a problemas na cadeira respiratória clássica) e proteínas relacionadas à necrose. Novas funções podem ser implementadas para que as buscas sejam cada vez mais completas e rápidas.
- Os sistemas de mineração e anotação podem ser aplicados a projetos de seqüenciamento de genoma ou ESTs, de pequenas pesquisas a grandes redes, com aceso seguro e facilitado via web, permite de forma fácil a realização de operações que exigem normalmente conhecimento específico de bioinformática, e têm como principais características a possibilidade de mineração e anotação durante o seqüenciamento associada a uma interface de anotação desenhada para otimizar a anotação. As próximas fases de trabalho incluem a criação de uma interface de cura e padronização das anotações, a integração de um sistema de anotação metabólica, melhorias na análise de *microarrays* e na navegabilidade do sistema. Melhorias também podem ser realizadas para completar o processo de adaptação automática para novos projetos genoma, visando à automação total do processo via interface web.
- Existe um esforço integrado no Projeto Vassoura de Bruxa buscando a compreensão do fungo e da doença, para encontrar soluções que permitam a volta da maior viabilidade da produção de cacau. A integração das análises bioinformáticas com as pesquisas funcionais do projeto vassoura de bruxa é um bom exemplo a ser aplicado em outros projetos genoma. Deseja-se que as ferramentas desenvolvidas continuem sendo utilizadas nos projetos atuais e em outros futuros para direcionar pesquisas funcionais, sem as quais a informação sobre o metabolismo virtual obtido com o seqüenciamento de genomas não passa de inferência.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Aime, M. C. and Phillips-Mora, W. (2005). "The causal agents of witches' broom and frosty pod rot of cacao (chocolate, *Theobroma cacao*) form a new lineage of Marasmiaceae." Mycologia **97**(5): 1012-22.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Andebrhan, T. and Furtek, D. B. (1994). "Random amplified polymorphic DNA (RAPD) analysis of *Crinipellis perniciosa* isolates from different hosts." Plant Pathology **43**(6): 1020-1027.

Bartley, B. G. D. (1986). Cacao (*Theobroma cacao* L.) breeding for durable resistance in perennial crops. FAO Plant Production and Protection Paper 70. Rome**:** 25–42.

Berezikov, E., Plasterk, R. H. A. and Cuppen, E. (2002). "GENOTRACE: cDNA-based local GENOme assembly from TRACE archives." Bioinformatics **18**(10): 1396-1397.

Bertrand, H. (2000). "Role of mitochondrial DNA in the senescence and hypovirulence of fungi and potential for plant disease control." Annu Rev Phytopathol **38**: 397-422.

Bertrand, H., Chan, B. S. and Griffiths, A. J. (1985). "Insertion of a foreign nucleotide sequence into mitochondrial DNA causes senescence in *Neurospora intermedia*." Cell **41**(3): 877-84.

Bertrand, H., Griffiths, A. J., Court, D. A. and Cheng, C. K. (1986). "An extrachromosomal plasmid is the etiological precursor of kalDNA insertion sequences in the mitochondrial chromosome of senescent *Neurospora*." Cell **47**(5): 829-37.

Chan, B. S., Court, D. A., Vierula, P. J. and Bertrand, H. (1991). "The kalilo linear senescence-inducing plasmid of *Neurospora* is an invertron and encodes DNA and RNA polymerases." Curr Genet **20**(3): 225-37.

Court, D. A. and Bertrand, H. (1992). "Genetic organization and structural features of maranhar, a senescence-inducing linear mitochondrial plasmid of *Neurospora crassa*." Curr Genet **22**(5): 385-97.

Evans, H. C. and Bastos, C. N. (1979). " Uma avaliação do ciclo de vida da vassoura-de-bruxa (*Crinipellis perniciosa*) do cacaueiro." Fitopatologia Brasileira **4**(1): 104.

Ewing, B. and Green, P. (1998a). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-94.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998b). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-85.

Gisi, U., Sierotzki, H., Cook, A. and McCaffery, A. (2002). "Mechanisms influencing the evolution of resistance to Qo inhibitor fungicides." Pest Manag Sci **58**(9): 859-67.

Gordon, D., Abajian, C. and Green, P. (1998). "Consed: A graphical tool for sequence finishing." Genome Research **8**(3): 195-202.

Gordon, D., Desmarais, C. and Green, P. (2001). "Automated finishing with Autofinish." Genome Research **11**(4): 614-625.

Gredilla, R., Grief, J. and Osiewacz, H. D. (2006). "Mitochondrial free radical generation and lifespan control in the fungal aging model *Podospora anserina*." Exp Gerontol **41**(4): 439-447.

Griffith, G. W. and Hedger, J. N. (1994a). "The breeding biology of biotypes of the witches' broom pathogen of cocoa, *Crinipellis perniciosa*." Heredity **72**: 278-289.

Griffith, G. W. and Hedger, J. N. (1994b). "Spatial distribution of mycelia of the liana (L-) biotype of the agaric *Crinipellis perniciosa* (Stahel) Singer in tropical forest." New Phytologist **127**: 243-259.

Griffith, G. W., Nicholson, J., Nenninger, A., Birch, R. N. and Hedger, J. N. (2003). "Witches' brooms and frosty pods: two major pathogens of cacao." New Zealand Journal of Botany **41**(3): 423-435.

Griffiths, A. J. (1992). "Fungal senescence." Annu Rev Genet **26**: 351-72.

Griffiths, A. J., Xiao, Y., Barton, R. and Myers, C. (1992). "Suppression of cytoplasmic senescence in *Neurospora*." Curr Genet **21**(6): 479-84.

Griffiths, A. J. F. (1998). "The kalilo family of fungal plasmids." Botanical Bulletin of Academia Sinica **39**(3): 147-152.

Hedger, J. N., Pickering, V. and Aragundi, J. A. (1987). "Variability of populations of the witches' broom disease of cocoa (*Crinipellis perniciosa*)." Transactions of the British Mycological Society **88**: 533-546.

Hermanns, J., Asseburg, A. and Osiewacz, H. D. (1994). "Evidence for a life span-prolonging effect of a linear plasmid in a longevity mutant of *Podospora anserina*." Mol Gen Genet **243**(3): 297-307.

Hotz-Wagenblatt, A., Hankeln, T., Ernst, P., Glatting, K. H., Schmidt, E. R. and Suhai, S. (2003). "ESTAnnotator: a tool for high throughput EST annotation." Nucleic Acids Research **31**(13): 3716-3719.

Huang, X. and Madan, A. (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-77.

Kerscher, S., Durstewitz, G., Casaregola, S., Gaillardin, C. and Brandt, U. (2001). "The complete mitochondrial genome of *Yarrowia lipolytica*." Comparative and Functional Genomics **2**(2): 80-90.

Kim, E. K., Jeong, J. H., Youn, H. S., Koo, Y. B. and Roe, J. H. (2000). "The terminal protein of a linear mitochondrial plasmid is encoded in the N-terminus of the DNA polymerase gene in white-rot fungus *Pleurotus ostreatus*." Curr Genet **38**(5): 283-90.

Kumar, C. G., LeDuc, R., Gong, G., Roinishivili, L., Lewin, H. A. and Liu, L. (2004). "ESTIMA, a tool for EST management in a multi-project environment." BMC Bioinformatics **5**: 176-185.

Maas, M. F., de Boer, H. J., Debets, A. J. and Hoekstra, R. F. (2004). "The mitochondrial plasmid pAL2-1 reduces calorie restriction mediated life span extension in the filamentous fungus *Podospora anserina*." Fungal Genet Biol **41**(9): 865-71.

Mao, C. H., Cushman, J. C., May, G. D. and Weller, J. W. (2003). "ESTAP - an automated system for the analysis of EST data." Bioinformatics **19**(13): 1720-1722.

Matukumalli, L. K., Grefenstette, J. J., Sonstegard, T. S. and Van Tassell, C. P. (2004). "EST-PAGE - managing and analyzing EST data." Bioinformatics **20**(2): 286-288.

Monteiro-Vitorello, C. B., Bell, J. A., Fulbright, D. W. and Bertrand, H. (1995). "A cytoplasmically transmissible hypovirulence phenotype associated with mitochondrial DNA mutations in the chestnut blight fungus *Cryphonectria parasitica*." Proc Natl Acad Sci U S A **92**(13): 5935-9.

Nakagawa, C. C., Jones, E. P. and Miller, D. L. (1998). "Mitochondrial DNA rearrangements associated with mF plasmid integration and plasmodial longevity in *Physarum polycephalum*." Curr Genet **33**(3): 178-87.

Nakai, R., Sen, K., Kurosawa, S. and Shibai, H. (2000). "Basidiomycetous fungus *Flammulina velutipes* harbors two linear mitochondrial plasmids encoding DNA and RNA polymerases." FEMS Microbiol Lett **190**(1): 99-102.

Orchard, J., Collin, H. A., Hardwick, K. and Isaac, S. (1994). "Changes in morphology and measurement of cytokinin levels during the development of witches-brooms on cocoa." Plant Pathology **43**(1): 65-72.

Osiewacz, H. D. (2002a). "Aging in fungi: role of mitochondria in *Podospora anserina*." Mech Ageing Dev **123**(7): 755-64.

Osiewacz, H. D. (2002b). "Mitochondrial functions and aging." Gene **286**(1): 65-71.

Paquola, A. C. M., Nishyiama, M. Y., Reis, E. M., da Silva, A. M. and Verjovski-Almeida, S. (2003). "ESTWeb: bioinformatics services for EST sequencing projects." Bioinformatics **19**(12): 1587-1588.

Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A. and Blaxter, M. (2004). "PartiGene - constructing partial genomes." Bioinformatics **20**(9): 1398-1404.

Pegler, D. N. (1978). "*Crinipellis perniciosa* (Agaricales)." Kew Bulletin **32**(4): 731-733.

Pereira, J. L., deAlmeida, L. C. C. and Santos, S. M. (1996). "Witches' broom disease of cocoa in Bahia: Attempts at eradication and containment." Crop Protection **15**(8): 743-752.

Purdy, L. H. and Schmidt, R. A. (1996). "Status of cacao witches' broom: Biology, epidemiology, and management." Annual Review of Phytopathology **34**: 573-594.

Resende, M. L. V., Gutemberg, B. A. N., Silva, L. H. C. P., Niella, G. R., Carvalho, G. A., Santiago, D. V. R. and Bezerra, J. L. (2000). "*Crinipellis perniciosa* proveniente de um novo hospedeiro, *Heteropterys acutifolia*, é patogênico a *T. cacao*." Fitopatologia Brasileira **25**(1): 88-91.

Rieck, A., Griffiths, A. J. and Bertrand, H. (1982). "Mitochondrial variants of *Neurospora intermedia* from nature." Can J Genet Cytol **24**(6): 741-59.

Rincones, J., Mazotti, G. D., Griffith, G. W., Pomela, A., Figueira, A., Leal, G. A., Jr., Queiroz, M. V., Pereira, J. F., Azevedo, R. A., Pereira, G. A. and Meinhardt, L. W. (2006). "Genetic variability and chromosome-length polymorphisms of the witches' broom pathogen *Crinipellis perniciosa* from various plant hosts in South America." Mycol Res **110**(Pt 7): 821-32.

Robison, M. M. and Horgen, P. A. (1999). "Widespread distribution of low-copy-number variants of mitochondrial plasmid pEM in the genus *Agaricus*." Fungal Genet Biol **26**(1): 62-70.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences of the United States of America **74**(12): 5463-5467.

Sanogo, S., Pomella, A., Hebbar, R. K., Bailey, B., Costa, J. C. B., Samuels, G. J. and Lumsden, R. D. (2002). "Production and germination of conidia of *Trichoderma stromaticum*, a mycoparasite of *Crinipellis perniciosa* on cacao." Phytopathology **92**(10): 1032-1037.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). "Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray." Science **270**(5235): 467-470.

Silva, S. D. V. M. and Matsuoka, K. (1999). "Histologia da interação *Crinipellis perniciosa* em cacaueiros suscetível e resistente a vassoura-de-bruxa." Fitopatologia Brasileira **24**(1): 54-59.

Singer, R. (1942). "Monograph study of the genera *Crinipellis* and *Chaetochalatus*." Liloa **8**(7): 503.

Sreenivasan, T. N. and Dabydeen, S. (1989). "Modes of penetration of young cocoa leaves by *Crinipellis perniciosa*." Plant Disease **73**(6): 478-481.

Tudzynski, P. and Esser, K. (1979). "Chromosomal and extrachromosomal control of senescence in the ascomycete *Podospora anserina*." Mol Gen Genet **173**(1): 71-84.

7. ANEXOS

7.1. ANEXO A. Artigo "Brazilian coffee genome project: an EST-based genomic resource"

Brazilian coffee genome project: an EST-based genomic resource

Luiz Gonzaga Esteves Vieira*, Alan Carvalho Andrade*, Carlos Augusto Colombo*, Ana Heloneida de Araújo Moraes, Ângela Metha, Angélica Carvalho de Oliveira, Carlos Alberto Labate, Celso Luis Marino, Claúdia de Barros Monteiro-Vitorelloa, Damares de Castro Monte, Éder Giglioti, Edna Teruko Kimura, Eduardo Romano, Eiko Eurya Kuramae, Eliana Gertrudes Macedo Lemos, Elionor Rita Pereira de Almeida, Érika C. Jorge, Érika V. S. Albuquerque, Felipe Rodrigues da Silva, Felipe Vinecky, Haiko Enok Sawazaki, Hamza Fahmi A. Dorry, Helaine Carrer, Ilka Nacif Abreu, João A. N. Batista, João Batista Teixeira, João Paulo Kitajimaa, Karem Guimarães Xavier, Liziane Maria de Lima, Luis Eduardo Aranha de Camargo, Luiz Filipe Protasio Pereira, Luiz Lehmann Coutinho, Manoel Victor Franco Lemos, Marcelo Ribeiro Romanoa, Marcos Antonio Machado, Marcos Mota do Carmo Costa, Maria Fátima Grossi de Sá, Maria Helena S. Goldman, Maria Inês T. Ferro, Maria Laine Penha Tinoco, Mariana C. Oliveira, Marie-Anne Van Sluys, Milton Massao Shimizu, Mirian Perez Maluf, Mirian Therezinha Souza da Eira, Oliveiro Guerreiro Filho, Paulo Arruda, Paulo Mazzafera, Pilar Drummond Sampaio Correa Mariani, Regina L.B.C. de Oliveira, Ricardo Harakava, Silvia Filippi Balbao, Siu Mui Tsai, Sonia Marli Zingaretti di Mauro, Suzana Neiva Santos, Walter José Siqueira, Gustavo Gilson Lacerda Costa, Eduardo Fernandes Formighieri, Marcelo Falsarella Carazzolle*, Gonçalo Amarante Guimarães Pereira*.

*These authors contributed equally to this work.

Braz. J. Plant Physiol., 18(1):95-108, 2006.

MINIREVIEW

Brazilian coffee genome project: an EST-based genomic resource

Luiz Gonzaga Esteves Vieira[1*§], Alan Carvalho Andrade[2*], Carlos Augusto Colombo[3*], Ana Heloneida de Araújo Moraes[2], Ângela Metha[2], Angélica Carvalho de Oliveira[2], Carlos Alberto Labate[4], Celso Luis Marino[8], Claúdia de Barros Monteiro-Vitorello[6a], Damares de Castro Monte[2], Éder Giglioti[9], Edna Teruko Kimura[10], Eduardo Romano[2], Eiko Eurya Kuramae[11], Eliana Gertrudes Macedo Lemos[12], Elionor Rita Pereira de Almeida[2], Érika C. Jorge[5], Érika V. S. Albuquerque[2], Felipe Rodrigues da Silva[2], Felipe Vinecky[2], Haiko Enok Sawazaki[3], Hamza Fahmi A. Dorry[14], Helaine Carrer[7], Ilka Nacif Abreu[15], João A. N. Batista[2], João Batista Teixeira[2], João Paulo Kitajima[17a], Karem Guimarães Xavier[4], Liziane Maria de Lima[2], Luis Eduardo Aranha de Camargo[6], Luiz Filipe Protasio Pereira[18], Luiz Lehmann Coutinho[5], Manoel Victor Franco Lemos[13], Marcelo Ribeiro Romano[4a], Marcos Antonio Machado[19], Marcos Mota do Carmo Costa[2], Maria Fátima Grossi de Sá[2], Maria Helena S. Goldman[20], Maria Inês T. Ferro[12], Maria Laine Penha Tinoco[2], Mariana C. Oliveira[21], Marie-Anne Van Sluys[21], Milton Massao Shimizu[15], Mirian Perez Maluf[22], Mirian Therezinha Souza da Eira[23], Oliveiro Guerreiro Filho[3], Paulo Arruda[24], Paulo Mazzafera[15], Pilar Drummond Sampaio Correa Mariani[16], Regina L.B.C. de Oliveira[25], Ricardo Harakava[26], Silvia Filippi Balbao[15], Siu Mui Tsai[27], Sonia Marli Zingaretti di Mauro[12], Suzana Neiva Santos[2], Walter José Siqueira[3], Gustavo Gilson Lacerda Costa[28], Eduardo Fernandes Formighieri[28], Marcelo Falsarella Carazzolle[28*], Gonçalo Amarante Guimarães Pereira[28*].

**These authors contributed equally to this work.*

[1]Laboratório de Biotecnologia Vegetal (LBI), IAPAR, CP 481, 86001-970, Londrina, PR, Brazil; [2]Núcleo de Biotecnologia-NTBio, Embrapa Recursos Genéticos e Biotecnologia, Parque Estação Biológica, CP 02372, 70770-900, Brasília, DF, Brazil; [3]Instituto Agronômico de Campinas, CP 28, 13001-970, Campinas, SP, Brazil; [4]Departamento de Genética, [5]Departamento de Zootecnia, [6]Departamento de Entomologia, Fitopatologia e Zoologia Agrícola and [7]Departamento de Ciências Biológicas , Escola Superior de Agricultura Luiz de Queiroz, USP, USP, 13418-900, Piracicaba, SP, Brazil; [8]Departamento de Genética, Instituto de Biociências, UNESP, 18618–000, Botucatu SP, Brazil; [9]Centro de Ciências Agrárias, Universidade Federal de São Carlos, 13600-970, Araras, SP, Brazil; [10]Instituto de Ciências Biomédicas, USP, 05508-000, São Paulo, SP, Brazil; [11]Departamento de Defesa Fitossanitária, Faculdade de Ciências Agronômicas, UNESP, CP 237, 18603-970, Botucatu SP, Brazil; [12]Departamento de Tecnologia and [13]Departamento de Biologia Aplicada à Agropecuária, Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal, UNESP, 14884-900, Jaboticabal, SP, Brazil; [14]Departamento de Bioquímica, Instituto de Química, USP, 05513-970, São Paulo, SP, Brazil; [15]Departamento de Fisiologia Vegetal, Instituto de Biologia, [16]Faculdade de Engenharia Química and [17a]Laboratório de Bioinformática, Instituto da Computação, UNICAMP, CP 6109, 13083-970, Campinas, SP, Brazil; [18]Embrapa Café, IAPAR, CP 481, 86001-970, Londrina, PR, Brazil; [19]Centro APTA de Citros Sylvio Moreira, IAC, CP 04, 13490-970, Cordeirópolis SP, Brazil; [20]Departamento de Biologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, USP, 14040-901, Ribeirão Preto, SP, Brazil; [21]Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, 05508-900, São Paulo, SP, Brazil; [22]Embrapa Café, Instituto Agronômico de Campinas, CP 28, 13001-970, Campinas, SP, Brazil; [23]Embrapa Café, Núcleo de Biotecnologia-NTBio, Embrapa Recursos Genéticos e Biotecnologia, Parque Estação Biológica, CP 02372, 70770-900, Brasília, Brazil; [24]Centro de Biologia Molecular e Engenharia Genética, UNICAMP, CP 6010, 13083-970, Campinas, SP, Brazil; [25]Núcleo Integrado de Biotecnologia, Universidade de Mogi das Cruzes, 08780-911, Mogi das Cruzes, SP, Brazil; [26]Centro de Sanidade Vegetal, Instituto Biológico de São Paulo, 04014-002, São Paulo, SP, Brazil; [27]Centro de Energia Nuclear na Agricultura, USP, CP 96, 13400-970, Piracicaba, SP, Brazil; [28]Laboratório de Genômica e Expressão, Instituto de Biologia, UNICAMP, 13083-970, Campinas, SP, Brazil.

Current Address: [4a]Universidade Estadual de Ponta Grossa, Campus Uvaranas, 84030-900, Ponta Grossa, PR, Brazil; [6a]Laboratório Nacional de Computação Científica, Laboratório de Bioinformática, Quitandinha, 25651-075, Petrópolis, RJ, Brazil; [15a]Alellyx Applied Genomics, Rod. Anhanguera, Km 104, 13067-850, Campinas, SP, Brazil;

[1§]Corresponding author: lvieira@iapar.br*

Coffee is one of the most valuable agricultural commodities and ranks second on international trade exchanges. The genus *Coffea* belongs to the Rubiaceae family which includes other important plants. The genus contains about 100 species but commercial production is based only on two species, *Coffea arabica* and *Coffea canephora* that represent about 70 % and 30 % of the total coffee market, respectively. The Brazilian Coffee Genome Project was designed with the objective of making modern genomics resources available to the coffee scientific community, working on different aspects of the coffee production chain. We have single-pass sequenced a total of 214,964 randomly picked clones from 37 cDNA libraries of *C. arabica*, *C. canephora* and *C. racemosa*, representing specific stages of cells and plant development that after trimming resulted in

130,792, 12,381 and 10,566 sequences for each species, respectively. The ESTs clustered into 17,982 clusters and 32,155 singletons. Blast analysis of these sequences revealed that 22 % had no significant matches to sequences in the National Center for Biotechnology Information database (of known or unknown function). The generated coffee EST database resulted in the identification of close to 33,000 different unigenes. Annotated sequencing results have been stored in an online database at http://www.lge.ibi.unicamp.br/cafe. Resources developed in this project provide genetic and genomic tools that may hold the key to the sustainability, competitiveness and future viability of the coffee industry in local and international markets.
Key words: *Coffea*, cDNA, EST, transcriptome.

Projeto Genoma Brasileiro Café: recursos genômicos baseados em ESTs: O café é um dos principais produtos agrícolas, sendo considerado o segundo item em importância do comércio internacional de "commodities". O gênero *Coffea* pertence à família Rubiaceae que também inclui outras plantas importantes. Este gênero contém aproximadamente 100 espécies, mas a produção comercial é baseada somente em duas espécies, *Coffea arabica* e *Coffea canephora*, que representam aproximadamente 70 % e 30 % do mercado total de café, respectivamente. O Projeto Genoma Café Brasileiro foi desenvolvido com o objetivo de disponibilizar os modernos recursos da genômica à comunidade científica e aos diferentes segmentos da cadeia produtiva do café. Para isso, foram seqüenciados 214.964 clones escolhidos aleatoriamente de 37 bibliotecas de cDNA de *C. arabica*, *C. canephora* e *C. racemosa* representando estádios específicos do desenvolvimento de células e de tecidos do cafeeiro, resultando em 130.792, 12.381 e 10.566 seqüências de cada espécie, respectivamente, após processo de trimagem. Os ESTs foram agrupados em 17.982 contigs e em 32.155 singletons. A comparação destas seqüências pelo programa BLAST revelou que 22 % não tiveram nenhuma similaridade significativa às seqüências no banco de dados do National Center for Biotechnology Information (de função conhecida ou desconhecida). A base de dados de ESTs do cafeeiro resultou na identificação de cerca de 33.000 unigenes diferentes. Os resultados de anotação das seqüências foram armazenados em base de dados "online" em http://www.lge.ibi.unicamp.br/cafe. Os recursos desenvolvidos por este projeto disponibilizam ferramentas genéticas e genômicas que podem ser decisivas para a sustentabilidade, a competitividade e a futura viabilidade da agroindústria cafeeira nos mercados interno e externo.
Palavras-chave: *Coffea*, cDNA, EST, transcritoma.

INTRODUCTION

Coffee is an important agricultural commodity produced in more than 60 countries. It generates a turnover of US$10-12 billion per year and ranks second on international trade exchanges, representing a significant source of income to several developing countries in Africa, Asia and Latin America. Brazil, Vietnam and Colombia are responsible for about 50 % of the world-coffee production, and Brazil alone responds for more than one third of the global coffee production and exports. This fact ranks coffee amongst the most important commodities in the Brazilian trade balance.

The genus *Coffea* belongs to the Rubiaceae family, found throughout the tropics, which includes other important plants. About 100 species of the genus *Coffea* have been identified so far (Bridson and Verdcourt, 1988), most of them trees and shrubs growing at low altitudes in the tropical rain forests of Africa and Asia (Sondahl et al., 1992). Commercially, production relies only on two species, *Coffea arabica* L and *Coffea canephora* Pierre ex Froehner, which represent about 70 % and 30 % of the total coffee market, respectively. All known species are diploid (2n=22 chromosomes) and obligate outbreeders with self-incompatibility systems, except for *C. arabica* which is allotetraploid (2n=4x=44) and self-fertile at approximately 90 %. (Charrier and Berthaud, 1985).

C. arabica, providing Arabica coffee, was first described by Linnaeus in 1753.The botanical evidence indicates that the coffee plant *C. arabica* originated on the plateaus of central Ethiopia where it still grows wild. The species *C. arabica* L. is endemic in Southwest Ethiopia and probably originates from a relatively recent cross between *Coffea eugenoides* and *Coffea canephora* (Lashermes et al., 1993) as indicated by chloroplast restriction fragment length polymorphism (RFLP) analyses (Lashermes et al. 1996). The nuclear DNA content of *C. arabica* determined by flow cytometry is 2.4 pg, or 2X=1158 Mb (Arumuganathan et al., 1991). *C. arabica* is the most cultivated species, occupying 75 % of the coffee plantations around the world. The quality of the beverage is potentially excellent, being known in the

trade as mild coffee. Several cultivars have been described for *C. arabica*, but because of the narrow genetic basis of the species, phenotype differences among the cultivars are due mainly to single gene mutations.

The species *C. canephora* is the diploid species most widely cultivated around the world. It is self-sterile and cross-pollinated and consequently displays much more variability than *C. arabica*. *C. canephora* is better adapted to warm and humid equatorial climates and is frequently cultivated in low to medium altitudes. Robusta coffee is grown in West and Central Africa, throughout Southeast Asia and in Brazil, where it is also known as Conilon. The quality of the beverage made from C. *canephora* is generally regarded as inferior to that made of *C. arabica*. However, C. *canephora* is more resistant to adverse conditions than Arabica, particularly to several diseases and pests. Another diploid coffee species originating from Mozambique, *C. racemosa*, is characterized as having low caffeine content, high drought tolerance and resistance to leaf-miner (Clarke and Macrae, 1988), and has been used in breeding programs for introgression of important agronomic traits to *C. arabica* (Guerreiro et al., 1991).

The cultivation of the Arabica coffee began about five hundred years ago in Yemen and reached the southeast of Asia approximately in 1700. In the beginning of the 18th century, progenies of a single plant were taken from Indonesia to Europe and later to America (Chevalier and Dagron, 1928; Carvalho, 1945). Originating from other introductions that took place from Yemen to Brazil, seeds of two different cultivars, Typica and Bourbon, constitute the main genetic basis of all cultivated coffee planted in Brazil and other countries (Krug et al., 1939; Carvalho et al., 1993).

Coffee has long been bred with the view of improving important agronomic characteristics such as flowering, yield, bean size, cup quality, caffeine content and disease and drought resistance. Despite solid efforts, the progress in coffee breeding using conventional approaches has been slow due to many factors such as the narrow genetic basis of cultivated coffee, the lack of genetic markers and efficient screening tools, as well as the long time taken for generation advancement.

The recent development of applied technologies in biology is leading to an enormous production of information in the area of plant genomics, through the sequencing of different organisms. Large-scale sequencing of cDNAs to produce Expressed Sequence Tags (ESTs) and comparing the resulting sequences with public databases has become the method of choice for the rapid and cost-effective generation of data on the coding capacity of genomes and for the potential identification of new genes. For the same reason, several sequencing projects of plant species, such as the Sugar Cane EST Genome Project (SUCEST) accomplished by the ONSA group (Organization for Nucleotide Sequencing and Analysis) (Arruda, 2001) have been carried out in Brazil.

Coffea genomes are large in comparison with the current plant models, Arabidopsis and rice. While the coffee genome may probably have similarities to gene motifs already identified in small-genome plants, the larger genome size of coffee makes it unlikely to anticipate a complete genome-sequencing effort of any species of the genus *Coffea* in the near future, despite the recent increases in DNA-sequencing capacity of modern equipment. Therefore, large-scale discovery, isolation and analysis of gene function in coffee and its relatives must rely on other, less direct methods. The partial sequencing of anonymous cDNA clones (Expressed sequence tags - ESTs) is a rapid and cost-effective method for generating data on the coding capacity of genomes and, for this reason, has become the fastest growing segment of the public DNA databases (Wolfsberg and Landsman, 1997). In plants, the EST approach was initially used for the model species *A. thaliana* (Höfte et al., 1993) and rice (Yamamoto and Sasaki, 1997). Subsequently, a large variety of EST sequences from other species have been deposited in the dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). Recently, an EST database based on sequences from approximately 47,000 cDNA clones and with a special focus on developing seeds of *C. canephora* has also been released (Chenwei et al., 2005).

The Brazilian Coffee Genome Project was designed to develop and deploy useful tools for gene discovery and functional genetic analysis in coffee and related species and to aid in the advance of knowledge on the structure and evolution of the coffee genome. The generated coffee EST database from *C. arabica*, *C. canephora* and *C. racemosa* resulted in the identification of more than 30 thousand different unigenes and will facilitate genetic studies on coffee. This basic information provides a very valuable resource for studies on the biology and physiology of coffee plants that will considerably enhance the isolation and characterization of important agronomic genes for genetic improvement of *Coffea*.

Project organization and goals

The Brazilian Coffee Genome Project was formulated in 2002 through a cooperative agreement signed between the

Brazilian Coffee Research and Development Consortium (CBP&D-Café), a national consortium of 40 public Universities and Research Institutes, the Brazilian Enterprise for Agricultural Research (Embrapa), the São Paulo State Research Support Foundation (FAPESP) and the Permanent Forum for University-Company Relations (UNIEMP). The CBP&D was responsible for the central coordination of the project, but all three institutions supporting this initiative also appointed one project coordinator each with managerial responsibilities in the project aimed at facilitating the maintenance of the information flow from the network of laboratories involved in the Coffee Genome Project.

The initial goal of this project was the development of a large database of ESTs, with a minimum of 200,000 reads and a Unigene set composed of 25,000 genes. Assuming that the number of gene motifs is similar among the angiosperm genomes, this number would theoretically represents about 2/3 of all the gene motifs present in the coffee genomes.

The infrastructure assembled for sequencing was already established by the AEG (Agriculture and Environment Genomes), a network of several laboratories located at different research institutions in São Paulo state, and funded by both FAPESP and Embrapa Recursos Genéticos e Biotecnologia, Brasília. Laboratories from these two groups were also responsible for supervising and coordinating all aspects of cDNA library construction, such as cDNA size selection, cloning, clone-picking and clone library storage, sequencing and sequence submission to the bioinformatics center. Each group was assigned the cloning and sequencing of 100,000 reads.

The Laboratório de Genômica e Expressão (LGE) of the State University at Campinas (http://www.lge.ibi.unicamp.br/) was designated as the central bioinformatics facility to house the Coffee Genome sequence database and coordinate all aspects related to sequence submission, performance and productivity of the sequencing groups, data storage, BLAST analysis and clustering. For safety reasons, a replica of the raw data was transferred to the Embrapa Recursos Genéticos e Biotecnologia's bioinformatics group.

Total cost of the project and the committed institutional efforts was shared among the CBP&D-Café, EMBRAPA and FAPESP in the proportions of 50 %, 25 % and 25 %, respectively. The access of the database is free for six public universities and research institutes linked to FAPESP and for organizations and research institutes that are members of the CBP&D-Café, which in turn grant the opportunity for free access to more than 700 scientists from 40 institutions that develop coffee research in Brazil through collaborative partnerships. Data access restrictions are applied to any other user of the EST database, subject to the approval of the Coffee Genome Project Directive Committee. For this purpose, specific contractual conditions have been established regarding intellectual-propriety rights derived from having access to this information.

cDNA libraries and sequencing

The Instituto Agronômico de Campinas (IAC), which possesses a significant germplasm collection of *Coffea* species, supplied the material for the construction of cDNA libraries covering a wide range of tissues, developmental stages, and plant material submitted to biotic and abiotic stress conditions. The cDNA libraries constructed by the AEG group used plant material from *C. arabica* cv. Mundo Novo and cv. Catuaí, while those constructed at Embrapa Recursos Genéticos e Biotecnologia were made from tissues and organs from *C. arabica* cv. Catuaí (table 1). Also, EST libraries were made from tissues of *C. canephora* and *C. racemosa* lines belonging to the Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (INCAPER) and IAC's collection, respectively.

Total RNA was extracted from coffee tissues at different developmental stages and also submitted to different stress conditions. Poly(A)$^+$ RNA was purified from total RNA using the Oligotex Kit (Quiagen), following the manufacturer's directions. The mRNA purity and integrity were estimated by absorbance at 260/280 nm and agarose gel electrophoresis. cDNA libraries were constructed using the SuperScript Plasmid System and Plasmid Cloning Kit (Invitrogen) with about 1-2 µg poly(A)$^+$ RNA. The efficiency of cDNA synthesis was monitored with radioactive nucleotides. cDNA were size fractionated on a Sepharose CL-2B column. Aliquots of each fraction were eletrophoresed in agarose gel to determine the size range of cDNAs. Fractions containing cDNA larger than 500 pb were ligated into pSPORT1 and pSPORT6 vectors (Invitrogen) at the *SalI-NotI* site. The resulting plasmids were transformed in *E. coli* DH10B or DH5α cells (Invitrogen) by electroporation.

Plasmid DNA was purified using a modified alkaline lysis method (Sambrook et al., 1989). Sequencing reactions were conducted using the ABI BigDye Terminator Sequencing kit (Applied Biosystems). cDNA inserts were sequenced from the 5' end with T7 promoter primer (5'-TAATACGACTCACTATAGGG-3') or M13 Rev in the pSPORT1 vector or

with SP6 primer (5'- ATTTAGGTGACACTATAG-3') in the pSPORT6. Sequencing reaction products were analyzed on ABI 3700 sequencers (Applied Biosystems).

Picking of the clones and storage of stocks were carried out at the Brazilian Clone Collection Center (BCCC) in the case of libraries constructed by the AEG group. For the construction of all libraries, no procedure to eliminate differences in transcript representation was adopted.

Sequencing of coffee EST libraries was carried out by 25 laboratories located at Research Institutes and Universities belonging to the AEG system and at Embrapa Recursos Genéticos e Biotecnologia with 96-lane sequencers (ABI 3700), using standard protocols. Raw sequences and base confidence scores were obtained from chromatogram files using the program Phred (Ewing and Green, 1998; Ewing et al., 1998). Sequences accepted in the project had more than 250 bases with Phred quality ≥ 20. The number of sequences collected for each library was determined by monitoring the redundancy level of produced sequences.

Bioinformatics and database construction

In general, bioinformatics of EST projects includes such services as the organization, storage, integration, and analysis of biological information. The objectives of the Laboratório de Genômica e Expressão (LGE) bioinformatics group were (a) to provide appropriate database methods for the data generated for the sequencing groups in São Paulo and Brasilia; (b) to provide adequate security measures to ensure the integrity of the data; (c) to organize and present the data in such a way that authorized users can readily extract meaningful information from it and (d) to develop user-friendly interfaces to access the core data.

The first bioinformatics objective of the Brazilian Coffee Genome Project was to establish the means by which the various forms of the core data could be stored. The diverse sources of the sequences submitted required personnel that were knowledgeable in the nature of the data, as well as in the collection, manipulation, presentation and sharing of the bioinformatics data. There was also a need for security of the

Table 1. Description of the coffee ESTs libraries

Library code	Tissue/Developmental stage	Number of valid reads
AR1, LP1	Plantlets and leaves treated with araquidonic acid	5664
BP1	Suspension cells treated with acibenzolar-S-methyl	12379
CB1	Suspension cells treated with acibenzolar-S-methyl and brassinoesteroids	10311
CL2	Hypocotyls treated with acibenzolar-S-methyl	11615
CS1	Suspension cells treated with NaCl	10803
EA1, IA1, IA2	Embryogenic calli	9191
EB1	Zygotic embryo (immature fruits)	192
EC1	Embryogenic calli from *Coffea canephora*	8050
EM1, SI3	Germinating seeds (whole seeds and zygotic embryos)	9201
FB1, FB2, FB4	Flower buds in different developmental stages	23036
FR1, FR2	Flower buds + pinhead fruits + fruits at different stages	14779
FR4	Fruits (*Coffea racemosa*)	7967
FV2	Fruits, stages 1,2 and 3 (*Coffea racemosa*)	7195
CA1, IC1, PC1	Non embryogenic calli with and without 2,4 D	12135
LV4, LV5	Young leaves from orthotropic branch	15067
LV8, LV9	Mature leaves from plagiotropic branches	11864
NS1	Roots infected with nematodes	569
PA1	Primary embryogenic calli	2483
RM1	Leaves infected with leaf miner and coffee leaf rust	5567
RT3	Roots	560
RT5	Roots with acibenzolar-S-methyl	2311
RT8	Suspension cells with stressed with aluminum	9119
RX1	Stems infected with *Xylella spp.*	9563
SH1	Leaves from water deficit stresses plants (*Coffea canephora*)	7368
SH2	Water deficit stresses field plants (pool of tissues)	6824
SS1	Well-watered field plants (pool of tissues)	960

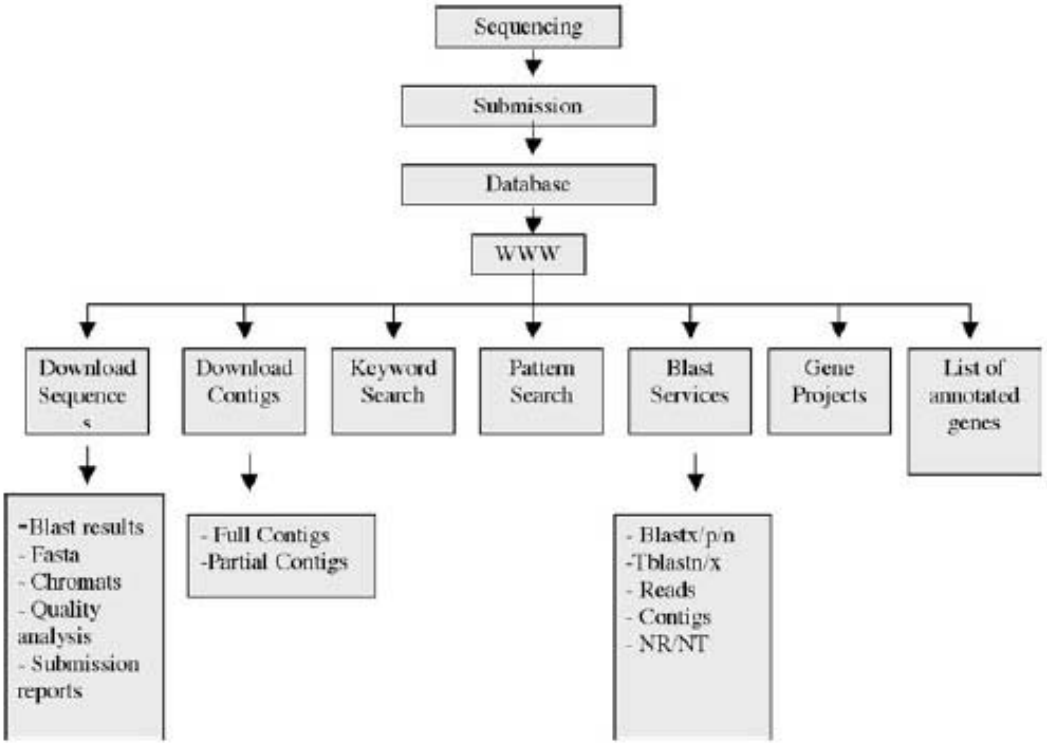


Figure 1. Flow of information and services involved in bioinformatics of the Brazilian Coffee Genome Project.

core data, which requires restricted access and backup, and there was a need for users to be able to access the data on demand.

The second bioinformatics objective was related to the handling of core data to meet the needs of users. For this, a relational database was developed in order to record and readily access the core information. To keep costs to a minimum in a centralized databasing location while making the database amenable to a broad group of users, a MySQL database (http://www.mysql.com/) was preferred for the relational database. EST assembly and the viewing of assemblies, as well as consensus sequences were the prime goals in bioinformatics of the coffee genome project.

Delivery of the information produced by the Brazilian Coffee Genome Project can be retrieved via the internet at the project site (http://www.lge.ibi.unicamp.br/cafe/). The advantage of web-based delivery is that anyone with an internet connection can have access, but this advantage is counterbalanced by the risk of crashes of a single centralized facility and sometimes slow speed of information access. For these reasons, besides LGE, all the core sequence data is also maintained at Embrapa Recursos Genéticos e Biotecnologia bioinformatics group (http://www.cenargen.embrapa.br/biotec/genomacafe/). Also, the possibility of handling the data by two different bioinformatics groups allows the development of derivatives of the core data (e.g., gene specific oligonucleotides, protein sequences, promoters, gene expression data etc) to meet the most frequent needs of users through databases that may be customized to take into consideration the preferences of coffee investigators.

Database analysis

For functional annotation of ESTs and categorization of contigs, the masked (http://www.phrap.com/) and trimmed sequences (Telles and da Silva, 2001) were compared with the protein sequences stored at NCBI databases (National Center for Biotechnology Information), particularly to the NR-(Non-redundant) database (http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.shtml#databases). Also, several off

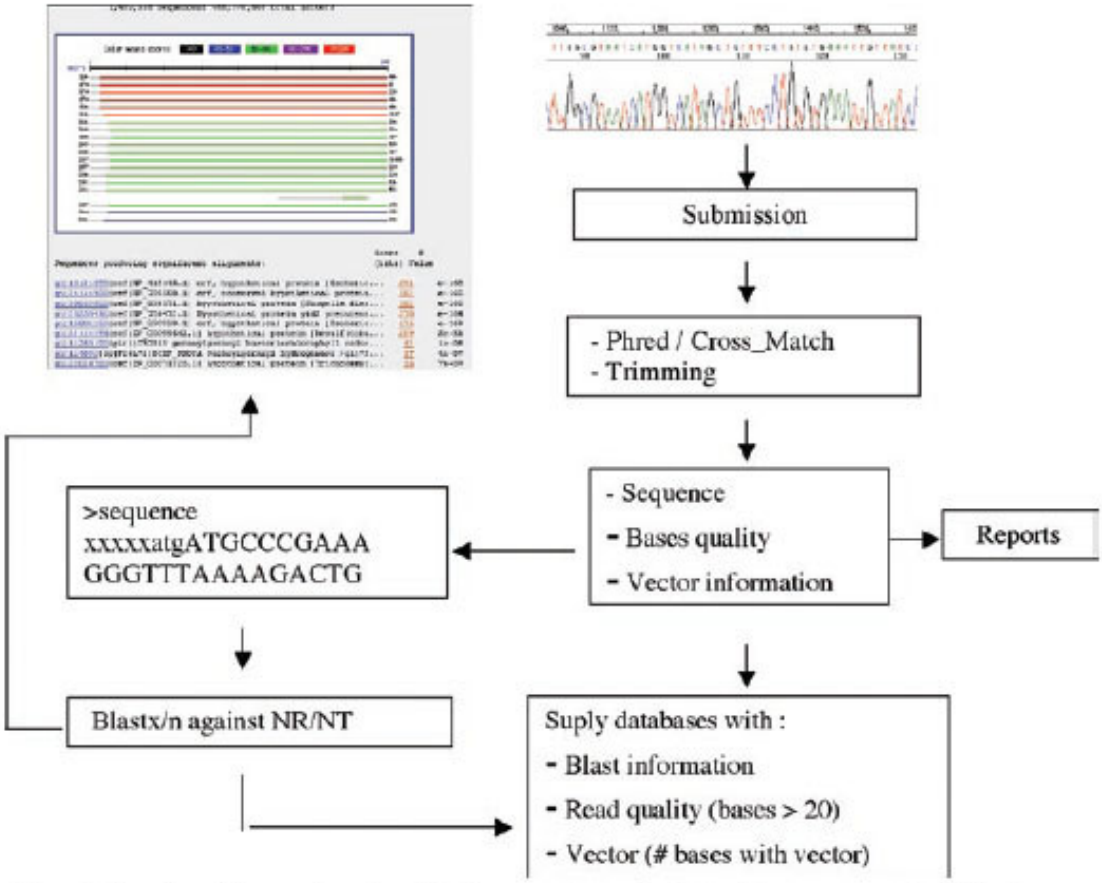


Figure 2. Overview of the procedures for submission, processing and analysis of the sequences submitted by the sequencing laboratories.

the shelf databasing and extraction tools were developed by the LGE team to fulfill many of the initial needs in coffee genomics. The procedures for submitting, processing, storing and analyzing the data are summarized in figure 2.

An overall sequencing efficiency of 70 % was obtained, including failures due to false-positive vector-only clones, short-insert clones, low-intensity or no-labeling reads, low-quality reads etc. The final results of the sequencing of EST libraries done by 26 laboratories produced a total of 214,964 reads, distributed among the three *Coffea* species selected in the project. The quality of the submitted sequences is an important piece of information to validate a database. The majority of the ESTs analyzed at end of the sequencing stage of the Coffee Genome Project had lengths above 500 bp with Phred quality >= 20 (figure 3).

As for any EST project, unwanted sequences are produced such as ribosomal sequences, poly-A fragments, low quality and short sequences, and slippage that all needed to be remove to avoid the introduction of irrelevant information into the EST database. The trimming was carried out with reads from *C. arabica*, *C. canephora* and *C. racemosa* that resulted in 130,792, 12,381 and 10,566 sequences (respectively), with the number of removed sequences summarized in table 2 according to each class.

Clustering and assembly of these ESTs using the CAP3 program (Huang and Madan, 1999) was done separated by species, resulting in 14,886 clusters and 24,426 singletons from *C. arabica*, 2,147clusters and 4,622 singletons for *C. canephora*, and 949 clusters and 3,107 singletons for *C. racemosa* . Close to sixty percent of the 17,982 contigs

and singletons presented a size length between 700 and 900 bp (figure 4). Due to the short-length attributes of part of the ESTs that had been produced, some singlets may have failed to merge into contigs and, therefore, the total number of "unigenes" might be overestimated. Of the contigs in *C. arabica*, the majority (86.7 %) was represented by two to ten ESTs. Due to the small number of clones for *C. canephora* and *C. racemosa* that were produced, the percentage of contigs with a higher number of ESTs in that range was superior to *C. arabica* (97.8 % and 97.5 %, respectively) (figure 5).

Regarding the cDNA libraries from *C. arabica, C. canephora and C. racemosa,* the sequences were analyzed for their similarity with known genes by BLASTX (Altschul et al, 1990) against NR (figure 6), considering 10^{-5} for the E-value as threshold for identity. It is interesting to note that there is a very similar partition between known and unknown sequences (No hit NR) for the three species. Moreover, among the sequences that generated hits, a significant number comprehends cDNA with the complete ORF inside (full length – NR Full).

Close to 11,000 contigs from *C. arabica* (29 % of the dataset) lacked significant similarity to any sequence based

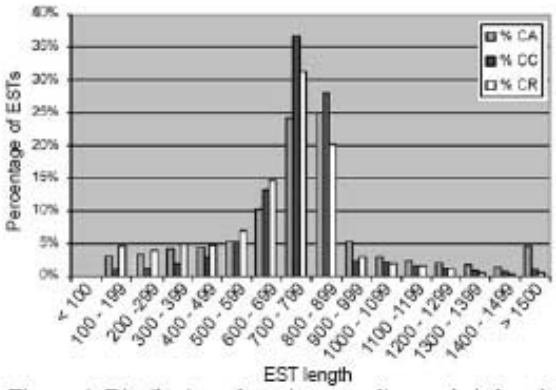


Figure 4. Distribution of contigs according to their length (bp) for each species.CA: *C. arabica*; CC: *C. canephora*; CR: *C. racemosa*

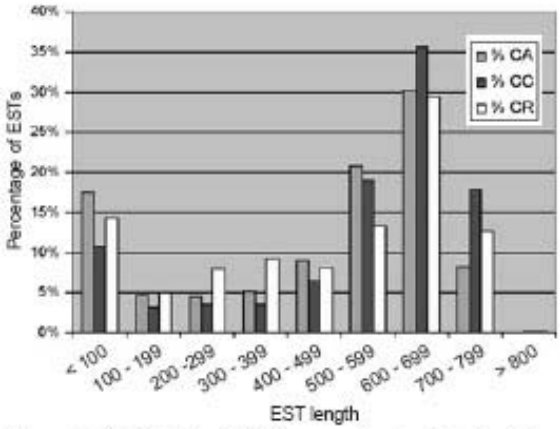


Figure 3. Distribution of ESTs according to their length in the different species. CA: *C. arabica*; CC: *C. canephora*; CR: *C. racemosa.*

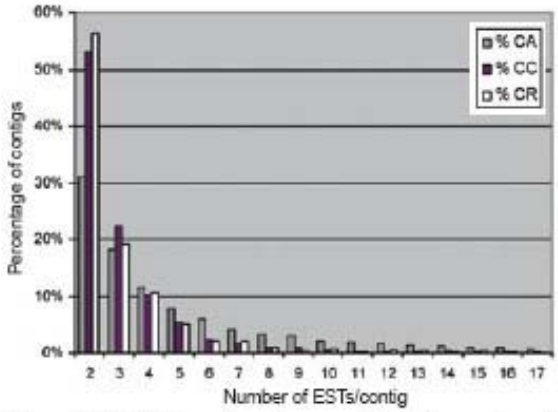


Figure 5. Distribution of contigs according to the number of EST per contig for each *Coffea* species. CA: *C. arabica*; CC: *C. canephora*; CR: *C. racemosa*

Table 2. Distribution of the removed reads by the trimming procedure from libraries of *C. arabica, C. canephora and C. racemosa,* according to classes.

Description	*C. arabica*	*C. canephora*	*C. racemosa*
Ribosomal sequences	1084 (0.56%)	49 (0.31%)	3 (0.04%)
Short sequences	29846 (15.30%)	1655 (10.60%)	1003 (13.23%)
Low quality	4798 (2.46%)	203 (1.30%)	274 (3.62%)
Slippage	23013 (11.79%)	1109 (7.10%)	546 (7.20%)
Poly-A	4077 (2.09%)	213 (1.36%)	409 (5.40%)
Poly-T	1500 (0.77%)	57 (0.37%)	14 (0.18%)

on the three ontological principles of Molecular Function, Biological Process and Cellular Component and broad categories developed for plant gene annotations by the Gene Ontology (GO) Consortium (ftp://ftp.geneontology.org/go/GO_slims/). *C. canephora* and *C. racemosa* had the same percentage of hits by GO (figure 7).

EST-based functional analysis

Coffee breeding, which is carried out through the traditional methods of hybridization and selection of superior progenies, has achieved relative success in satisfying the needs of the coffee industry. Certainly, the value of conventional breeding should not be overlooked, but linked efforts of both molecular techniques and traditional breeding can offer alternatives for making selective breeding more predictable and precise, reducing the time for obtaining new genotypes. Nowadays, the comprehensive examination of an organism that is afforded by functional genomics has changed the way one identifies genes and proteins with potential roles in a particular biological process without any *a priori* knowledge of their function. As with conventional breeding, the main objective is to describe and exploit the genetic diversity that is present in coffee species.

The access to coffee gene sequence information brings new perspectives and approaches to carry out biological research. Genome related databases, as the one made available by the Brazilian Coffee Genome Project, have become an invaluable asset for the scientific community to

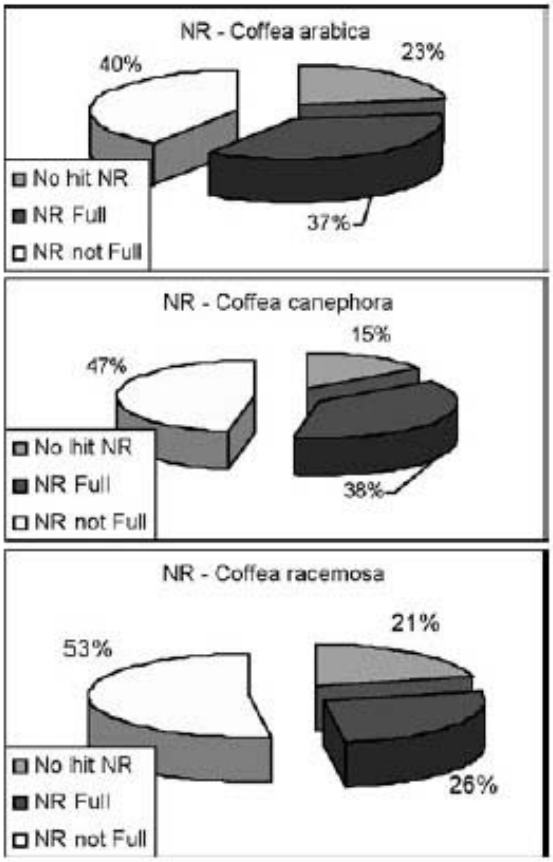


Figure 6. Distribution of the unigenes (contigs plus singletons) according to their comparison against non-redundant protein database (NR at NCBI) by BLASTX considering a threshold of 10^{-5} E-value. No hit NR: no similar sequence has been found in NR; NR Full - coffee sequence with a significantly similar sequence in NR and may encompass the complete ORF.

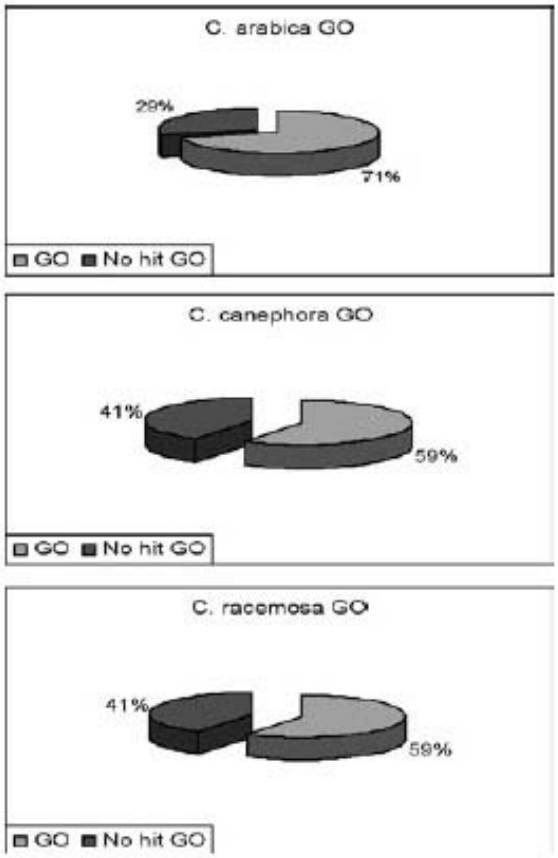


Figure 7. Coffee "unigenes" that had BLASTX matches in Gene Ontology database (GO) with E-value of ≤ $1.0E^{-5}$.

move onto the use of a number of these new technologies. Ultimately, through the use of the coffee EST dataset, genetic markers can be found for breeding programs, coffee genes can be cataloged in association with their location on the genome, the study of gene function and how activity of the gene products fits into complex metabolic pathways can be facilitated, and the regulation of the genes in response to different developmental and environmental stimuli can be examined holistically.

The evaluation of certain characteristics of interest in coffee requires a great deal of time because it can only be carried out on coffee trees after 10-15 years. Particularly, in *C. arabica*, the limited genetic diversity present among elite cultivars planted all over the world is the consequence of few introductions (Pearl et al., 2004). This lack of genetic variability in the gene pool of Arabica coffee limits the potential for germplasm improvement of this species. Therefore, finding new traits that add value to agricultural crops and their products has immense value in the agribusiness.

Due to the knowledge gathered on the coffee genome, these problems can be alleviated by searching for genetic markers closely linked to the candidate genes expressing these characteristics. The detection of such markers permits the screening of large numbers of coffee trees for a gene of interest when the plants are still at early stages of development and may reduce the number of backcrosses required to obtain quality traits (Lashermes et al., 1997). Also, marker-assisted selection for important but complex traits, which are often difficult to select in routine breeding programs, will enhance coffee breeding programs in terms of better-focused problems and save time and resources.

Molecular markers allow for the extension of conventional breeding methods with one important difference, that is the transfer of genetic information in a more precise and controlled manner. In coffee, molecular marker technology has already been implemented in germplasm characterization and management (Sera et al., 2003; Aga et al., 2005; Prakash et al., 2005; Maluf et al., 2005), detecting gene introgression in breeding populations (Prakash et al., 2002; Prakash et al., 2004; Herrera et al., 2004), describing coffee phylogeny with related species (Lashermes et al., 1999;Anthony et al., 2002) and in marker assisted breeding (Bertrand et al., 2001).

Among the molecular markers currently available, the SSRs (Simple Sequence Repeats), or microsatellite, have been extensively used due to their resolution and polymorphism levels. These characteristics make these molecular markers efficient tools for the genetic mapping, linkage studies, genotype identification and conservation of germplasm, pedigree analyses, marker assisted selection, and analysis of DNA libraries for gene cloning (Rufino et al., 2005). In coffee, the SSRs markers are not broadly used due to the limited numbers of primers presently available for this plant. The availability of massive amounts of coffee nucleotide sequence data will certainly offer an alternative to identify microsatellite motifs, which would be much more expensive through conventional laboratory protocols. In the coffee EST data set, a number of SSRs are present in transcripts that can now be readily mapped using existing breeding populations, and such studies are currently underway (Colombo and Caixeta, personal communication). Furthermore, specific genes of interest can be studied for variation within coffee species, allowing their assignment to coffee linkage maps.

Coffee physical maps bridge gaps between genetic maps and gene location. In this way, the availability of coffee BAC libraries will make possible the alignment of physical and genetic maps, bringing along continuity from phenotype to genotype (Noir et al., 2004; Leroy et al., 2005). Furthermore, the combination of EST database and BAC libraries may help to isolate genes through positional cloning.

One of the major objectives of the Brazilian Coffee Genome Project was to provide a tool for creation of transcriptional profiles as they appear in different tissues and as they change in response to development (Gaspari-Pezzopane et al., 2005; Geromel et al., 2005), biotic (Brandalise et al., 2005) and abiotic stresses (Vinecky et al., 2005). To help accomplish these studies, it is necessary to have powerful technologies available that allow the analysis of mRNA transcription patterns of thousands of genes in a single experiment (Kuhn, 2001).

Gene arrays (Lockhart and Winzeler, 2000) hybridized with mRNA populations from a variety of coffee tissues, organs and developmental stages may provide a genome-wide database of the transcriptional changes during plant growth that ultimately determine resistance to pests and diseases, productivity, and quality attributes of the coffee trees and fruits. Using this screening method, solutions for specific agronomic constraints may be found not only through new cultivar development but also by changes in crop management, harvesting, and post-harvest practices. For the construction of arrays, a set of UNIGENE sequences has to be available for use in the analysis of temporal or spatial expression profiles. Recent work to devise a minimal

clone set that represents all transcripts found in the Brazilian Coffee Genome Project was carried out by Sales et al. (2005). In this effort, a single relational database containing close to 33,000 putative transcripts was organized, allowing its use in diverse platforms and languages.

Proteomics as used to identify proteins in complex mixtures is only effective when a sequenced and annotated genome is available or Unigene sets become established (Rounsley et al., 1996). Proteomics is complementary to the ESTs because it also focuses on gene products. Proteomic studies consist of profiling the protein expression levels found in samples derived from different cultivars, tissue types, cultivation or post-harvest conditions in order to understand which proteins may be responsible for a trait of commercial significance, such as pathogens (Andrade et al., 2005), stress tolerance (Vincent et al., 2005) and food quality (Hajduch et al., 2005).

Proteomic characterizations of the coffee genotypes may also be used to validate results derived from DNA arrays and EST studies by verifying protein expression and thereby permit the subsequent coordination of gene transcription with protein expression. Such results can be used to establish baseline protein expression levels, and to identify constitutively expressed proteins that will be used as standards for comparing results derived from different cultivars or crop management conditions.

One of the effective ways to carry out studies on gene function at the morphological, biochemical and physiological level is to establish regulated expression systems of native genes in plants. The cloning of coffee regulatory sequences opens up the possibility of understanding the molecular mechanisms that regulate cellular/developmental processes and production of coffee metabolites at the biochemical and molecular levels, and provides the possibility of using regulatory elements to manipulate expression of entire metabolic pathways.

At the moment, only a few regulatory sequences for some coffee genes (Aldwinckle and Gaitan, 2002, 2004; Marraccini et al., 1999, 2003; Satyanarayana et al., 2005) have been identified. One of the most effective ways to obtain clones for promoter analysis of genes is from large insert genomic libraries. The construction of BAC libraries (Noir et al., 2004; Leroy et al., 2005) in addition to the already available EST sequences may greatly speed up the process of identification and isolation of important genetic control elements in coffee (promoters, enhancers, silencers etc). A highly efficient transformation system in coffee is an important complementary technology for evaluating promoter function.

Production of genetically modified plants is one of the techniques that opens new perspectives to coffee improvement, allowing the fast incorporation of desirable characteristics into elite cultivars. Despite the fact that the discussions on plant transformation are mainly centered on the commercial applications, for the scientific community, transgenic plants are important tools to study various aspects of plant sciences (Pereira, 2000). The enormous amounts of DNA sequence information available in the coffee EST data set opens up new experimental opportunities for functional genomic analysis.

Although genetic transformation procedures for coffee have been established (Hatanaka et al., 1999; Leroy et al., 2000; Ribas et al., 2005), the current technology has serious limitations, including low efficiency and throughput, which is still a key limitation for the widespread use of this technology (see: Genetic Transformation of Coffee, Ribas et al., in this issue). Successful genetic transformation of coffee is still limited to characters controlled by major genes and to transgenic plants that have been produced for insect resistance (Leroy et al., 2000), low caffeine content in seeds (Ogita et al., 2003) and herbicide resistance (Ribas et al., submitted). Based on the current public understanding of this technology, characteristics with low variability in the *Coffea* gene pool or of great appeal to consumers, such as delayed fruit ripening, resistance to pests and diseases (e.g., coffee borer, nematodes, coffee berry disease, leaf rust, *Xylella*), tolerance to abiotic stress and enhanced health benefits such as disease-fighting compounds, are the main candidates for academic work in future years.

CONCLUSION

The Brazilian Coffee Genome Project briefly presented here provides the genomic tools required for applied research to address the various constraints associated with the economic production of the coffee industry, mainly regarding the development of new cultivars. When combined with progress made in the development of *in vitro* technologies required for genetic transformation, data made available by the Coffee Genome Project may place the development of coffee cultivars on the future research agenda through the use of these new genetic technologies.

It is our belief that the Coffee EST database will not be limited to cultivar development applications, but will make a decisive contribution to other applied supplementary applica-

tions such as transcriptional profiling and proteomic analyses, leading to a better understanding of the way plants cope with biotic and abiotic stresses. Practical problems faced by the coffee agribusiness, represented by farmers, roasters, processors, exporters and specialty coffee associations, such as control of pre- and post-harvest physiological factors involved in quality, disease and pests control, management of plant response to water deficit, and elevated production costs can be partially overcome by integrated efforts of genomics research and breeding. Also, improvements through coffee genomic research may result in increased consumption and better health value of the beverage through new value-added products derived from coffee (e.g., nutraceutics, oils and flavors).

Finally, with the use of the coffee EST set it may be predicted that the integration of gene discovery, marker development and gene deployment become routine practices in Brazilian coffee research programs. Currently, genome annotation is being carried out by different institutions of the CBP&D-Café to improve the information in the database of the Coffee Genome Project. Annotating EST records will allow the coffee scientific community to use EST databases as an opportunity for gene discovery. Further efforts by the Coffee Genome Project bioinformatics groups may include assembly of ESTs to form Unigene sequences, complete gene sequences, gene specific oligonucleotides, alignment of gene sequences with related genes from other organisms, grouping of genes according to expression pattern and function, genetic linkage maps and physical maps.

Acknowledgments: The authors thank to the following researchers and technicians who contributed to the sequencing effort: A. Dalben, A.L.A. Beraldo, A.R de Oliveira, A.S. Zanca, A.S. Castro, D. Truffi, E.A Amaral da Silva, E.A.N. Pedrinho, E.S. Ferro, E.L. da Silveira, F.S. Prada, G.H. Goldman, J.C. Setúbal, J.P. Piazza, K.M. Borges, K.M. Brito, L.B.D. Labuto, M.M. Zerillo, M.A.C. da Silva, M.C. Oliveira, C.R. Borges Neto, R.L.B.C. Oliveira, R. Padovani, Z.A.R. Souza. This project was sponsored by Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café (CBP&D-Café), Empresa Brasileira de Pesquisa Agropecuária (Embrapa) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252: 1651-1656.

Aldwinckle SH, Gaitan AL (2002) Constitutive and inducible promoters from coffee plants. [US Patent N° 6,441,273].

Aldwinckle SH, Gaitan AL, (2005) Constitutive α-tubulin promoter from coffee plants and uses thereof. [US Patent N° 6903247].

Aga E, Bekele E, Bryngelsson T (2005) Inter-simple sequence repeat (ISSR) variation in forest coffee trees (*Coffea arabica* L.) populations from Ethiopia. Genetica 124:213-221.

Altschul SF, Gish W, Miller W, Myers EW, Lipman, DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-10.

Andrade AE, Albuquerque EVS, Grossi de Sá MF, Carneiro RMDG, Metha, A (2005) Expressão diferencial de proteínas em raízes de *Coffea canephora* infectadas com o nematóide endoparasita *Meloidogyne paranaensis*. In: Anais do IV Simpósio de Pesquisa dos Cafés do Brasil. Londrina, Brasil. Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café, CD-ROM.

Anthony F, Combes MC, Astorga C, Bertrand B, Graziosi G, Lashermes, P (2002) The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. Theor. Appl. Genet. 104:894-900.

Arruda P (2001) Sugarcane transcriptome: a landmark in plant genomics in the tropics. In: Arruda P (ed), Special volume on Sugarcane Transcriptome. Genet. Mol. Biol. 24:1–296.

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

Bertrand B, Anthony F, Lashermes P (2001) Breeding for resistance to Meloidogyne exigua in Coffea arabica by introgression of resistance genes of Coffea canephora. Plant Pathol. 50:637-643.

Brandalise M, Maluf M.P, Guerreiro Filho O,Gonçalves W, Maia IG (2005) Caracterização de genes com expressão tecido especifica em raízes e folhas de *Coffea arabica*. In: Anais do IV Simpósio de Pesquisa dos Cafés do Brasil. Londrina, Brasil,Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café, CD-ROM.

Bridson DM, Verdcourt B (1988) Flora of tropical East Africa: Rubiaceae. (Part 2). Cape Town: Iziko Museums of Cape Town, pp.415-747.

Charrier A, Berthaud J (1985) Botanical classification of coffee. In: Clifford MN, Wilson KC (eds), Coffee: botany, biochemistry and production of beans and beverage, pp.13-47. Croom Helm, London, Sydney.

Chenwei, L, Mueller, LA, Mc Carthy, J, Crouzillat, D, Pétiard, V, Tanksley, SD (2005). Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts, Theor. Appl. Genet. 112:114-130.

Chevalier A, Dagron M (1928) Recherches historiques sur les débuts de la culture du caféier en Amérique. Communications et Actes de Académie des Sciences Coloniales, Paris.

Carvalho A, Fazuoli LC (1993) O melhoramento de plantas no Instituto Agronômico. In: Furlani AMC, Viégas GP (eds), Café. pp.29-76. Campinas, Brasil

Carvalho A (1945) Distribuição geográfica e classificação botânica do gênero *Coffea* com referência especial à espécie Arabica. Bol. Superint. Serv. Café. 21:174-180.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. 8: 186–194.

Ewing B, Hillier L, Wend MC, Green P (1998) Basecalling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. 8:175–185.

Gaspari-Pezzopane C, Maluf MP, Pinto FO (2005) Expressão gênica diferencial em frutos de *Coffea arabica* L. em diferentes estádios de desenvolvimento e maturação. In: Anais do IV Simpósio de Pesquisa dos Cafés do Brasil. Londrina, Brasil, Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café, CD-ROM.

Geromel C, Ferreira LP, Cavalari AA, Pereira LFP, Vieira LGE, Leroy T, Mazzafera P, Marraccini, P (2005) Metabolismo de açúcares durante o desenvolvimento de frutos de café. In: Anais do IV Simpósio de Pesquisa dos Cafés do Brasil. Londrina, Brasil, Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café, CD-ROM.

Guerreiro Filho O, Medina Filho HP, Carvalho A (1991) Fontes de resistência ao bicho-mineiro. *Perileucoptera coffeella* em *Coffea* sp. Bragantia 50:45-55.

Hajduch M, Ganapathy A, Stein JW, Thelen JJ (2005) A Systematic Proteomic Study of Seed Filling in Soybean. Establishment of High-Resolution Two-Dimensional Reference Maps, Expression Profiles, and an Interactive Proteome Database Plant Physiol. 137:1397-1419.

Hatanaka T, Choi YE, Kusano T, Sano H (1999) Transgenic plants of *Coffea canephora* from embryogenic callus via *Agrobacterium tumefaciens*-mediated transformation. Plant Cell Rep. 19:106-110.

Herrera JC, Combes MC, Cortina H, Lashermes P (2004) Factors influencing gene introgression into the allotetraploid Coffea arabica L. from its diploid relatives. Genome 47: 1053-1060.

Höfte H, Desprez T, Amselem J, Chiapello H, Caboche M, Moisan A, Jourjon MF, Charpenteau JL, Berthomieu P, Guerrier D, Giraudat J, Quigley F, Thomas F, Yu DY, Mache R, Raynal M, Cooke R, Grellet F, Delseny M, Parmentier Y, Marcillac G, Gigot C, Fleck J, Philipps G, Axelos M, Bardet C, Tremousaygue D, Lescure B (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. Plant J. 4:1051-1061.

Huang X, Madan A (1999) CAP3: A DNA assembly program. Genome Res. 9:868-877.

Khun E (2001) From library screening to microarray technology: Strategies to determine gene expression profiles and to identify differentially regulated genes in plants. Ann. Bot. 87:139-155.

Krug CA, Mendes JET, Carvalho A (1938) Taxonomia de *Coffea arabica* L. Descrição das variedades e formas encontradas no Estado de São Paulo. Boletim Técnico do Instituto Agronômico. Campinas, Brasil, 62:1-57.

Lashermes P, Cros J, Marmey P, Charrier A (1993) Use of random amplified DNA markers to analyze genetic variability and relationships of *Coffea* species. Genet. Res. Crop Evol. 40:91–99.

Lashermes P, Cros J, Combes MC, Trouslot P, Anthony F, Hamon S, Charrier A (1996) Inheritance and restriction fragment length polymorphism of chloroplast DNA in the genus *Coffea* L. Theor. Appl. Genet. 93:626-632.

Lashermes P, Combes MC, Trouslot P, Charrier A (1997) Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. Theor. Appl. Genet. 94:947-955.

Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A (1999) Molecular characterisation and origin of the *Coffea arabica* L. genome. Mol. Gen. Genet. 261:259-266.

Leroy T, Henry AM, Royer M, Altosar I, Frutos R, Duris D, Philippe R (2000) Genetically modified coffee plants expressing the *Bacillus thuringiensis* cry1Ac gene for resistance to leaf miner. Plant Cell Rep. 19:382-389.

Leroy T, Marraccini, P, Dufour, M, Montagnon, C, Lashermes, P., Sabau, X., Ferreira L.P., Jourdan, I., Pot, D., Andrade A. C., Glaszmann, J.C., Vieira, L.G. E. and Piffanelli P. (2005). Construction and characterization of a *Coffea canephora* BAC library to study the organization of sucrose biosynthesis genes. Theor. Appl. Genet. 111:1032–1041.

Lockhart DJ, Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature, 405:827–836.

Maluf MP, Silvestrini M, Ruggiero LCM, Guerreiro Filho O, Colombo CA (2005) Genetic diversity of cultivated *Coffea arabica* inbred lines assessed by RAPD, AFLP and SSR marker systems. Sci. Agric. 62:366-373.

Marraccini P, Deshayes A, Petiard V, Rogers WJ (1999) Molecular cloning of the complete 11S seed storage protein gene of *Coffea arabica* and promoter analysis in transgenic tobacco plants. Plant Physiol. Biochem. 37:273-282.

Marraccini P, Courjault C, Caillet V, Lausanne F, Lepage B, Rogers WJ, Tessereau S, Deshayes A (2003) Rubisco small subunit of *Coffea arabica*: cDNA sequence, gene cloning and promoter analysis in transgenic tobacco plants. Plant Physiol. Biochem. 41:17-25.

Noir S, Patheyron S, Combes MC, Lashermes P, Chalhoub B (2004) Construction and characterization of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica* L.). Theor. Appl. Genet. 109:225-230.

Ogita S, Uejuji H, Yamaguchi Y, Koizumi N, Sano H (2003) RNA interference: Producing decaffeinated coffee plants. Nature 423:823.

Pereira A (2000) A transgenic perspective on plant functional genomics. Transgenic Res. 9:245–260.

Pearl HM, Nagai C, Moore PH, Steiger DL, Osgood RV, Ming R (2004) Construction of a genetic map for arabica coffee. Theor. Appl. Genet. 108:829–835.

Prakash NS, Combes MC, Somanna N, Lashermes P (2002) AFLP analysis of introgression in coffee cultivars (*Coffea*

arabica L.) derived from a natural interspecific hybrid. Euphytica 124:265-271.

Prakash NS, Marques DV, Varzea VMP, Silva MC, Combes MC, Lashermes P (2004) Introgression molecular analysis of a leaf rust resistance gene from Coffea liberica into C. arabica L. Theor. Appl. Genet. 109:1311-1317.

Prakash N, Combes MC, Dussert S, Naveen S, Lashermes P (2005) Analysis of genetic diversity in Indian robusta coffee genepool (*Coffea canephora*) in comparison with a representative core collection using SSRs and AFLPs. Genet. Resour. Crop Evol. 52:333-343.

Ribas AF, Kobayashi AK, Pereira LFP, Vieira, LGE (2005) Genetic transformation of *Coffea canephora* P. by particle bombardment. Biol. Plant. 49:493-497.

Ribas, A.F., Kobayashi, A.K., Pereira, L.F.P. and Vieira, L.G.E. (2006). Production of herbicide-resistant coffee plants (*Coffea canephora* P.) via *Agrobacterium tumefaciens*-mediated transformation. Brazil. Arch. Biol. Technol. (accepted for publication).

Rounsley SD, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JG, Kerlavage AR (1996) The construction of *Arabidopsis* expressed sequence tag assemblies. Plant Physiol. 112:1177-1183.

Rufino RJN, Caixeta ET, Zambolim EM, Pena GF, Almeida RF, Alavarenga SM, Zambolim L, Sakaiyama NS (2005) Microsatellite markers for coffee tree. In: Anais do IV Simpósio de Pesquisa dos Cafés do Brasil. Londrina, Brasil.Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café. CD-ROM.

Sales RMOB, Andrade AC, da Silva FR (2005) Determinação do Unigene do Projeto Genoma Café. In: Anais do IV Simpósio de Pesquisa dos Cafés do Brasil. Londrina, Brasil Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café. CD-ROM.

Satyanarayana KV, Kumar V, Chandrashekar A, Ravishankar GA (2005) Isolation of promoter for N-methyltransferase gene associated with caffeine biosynthesis in *Coffea canephora*. J. Biotech. 119:20–25.

Sera T, Ruas PM, Ruas CD, Diniz LEC, Carvalho VD, Rampim L, Ruas EA, Silveira SR (2003) Genetic polymorphism among 14 elite Coffea arabica L. cultivars using RAPD markers associated with restriction digestion. Genet. Mol. Biol. 1:59-64.

Telles GP, da Silva FR (2001) Trimming and clustering sugarcane ESTs. Gen. Mol. Biol. 24:17-23.

Vincent D, Lapierre C, Pollet B, Cornic G, Negroni L, Zivy M (2005) Water deficits affect caffeate O-methyltransferase, lignification, and related enzymes in maize leaves. A proteomic investigation. Plant Physiol. 137: 949-960.

Vinecky F, Brito KM, da Silva FR, Andrade AC (2005) Análise *in silico* de genes potencialmente envolvidos na resposta aos estresses abióticos presentes na base de dados do Genoma Café. In: Anais do IV Simpósio de Pesquisa dos Cafés do Brasil. Londrina, Brasil. Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café. CD-ROM.

Yamamoto K, Sasaki TL (1997). Large-scale EST sequencing in rice. Plant Mol. Biol. 35:135-144.

Wolfsberg TG, Landsman D (1997) A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res. 25:1626-1632.

7.2. ANEXO B. Produção didática: 3 Quickguides para *European Molecular Biology network* (EMBnet):

- ✓ ***A Quick Guide BLAST*;**
- ✓ ***A Quick Guide Phrap*;**
- ✓ ***A Quick Guide Phred*.**

-D [integer] DB genetic code (def = 1).
-M [string] matrix (def = BLOSUM62).
-T [T/F] produces HTML output (def = F).
-U [T/F] uses lower case filtering (def = T) Obs.: T = any lower-case letter in input FASTA file should be masked.

Position Specific Iterated BLAST

PSI-BLAST is a variant of blast that searches a query against a database using a position-specific scoring matrix created by PSI-BLAST. First run **blastpgp** to create and save a position-specific scoring matrix, then run **blastpgp** again to search iteratively with the previously saved matrix. e.g.,

```
blastpgp -i ff.chd -d yeast -C ff.chd.ckp
blastpgp -i ff.chd -d nr -j 3 -R ff.chd.ckp
```

Selected blastpgp arguments for PSI-BLAST:

-j [integer] maximum number of iterations (def = 1).
-h [number] E-value threshold for including sequences in the score matrix model (def = 0.001).
-C [file out] stores the query and frequency count ratio matrix in a file (opt).
-Q [file out] output file for PSI-BLAST matrix in ASCII (opt).
-R [file in] restarts from a file stored previously with -C.
-B [file in] input alignment for restart.

Pattern-Hit Initiated BLAST

PHI-BLAST is a search program that combines the matching of regular expressions with local alignments surrounding the match. E.g.,:

```
blastpgp -i query.file -k pattern.file -p patseedp
```

Select blastpgp arguments for PHI-BLAST:

-i [file in] input sequence file in FASTA format.
-k [file in] pattern (syntax follows the PROSITE conventions).
-p [string] usage mode (def = blastpgp). Obs: use 'patseedp', if pattern occurs only once, and 'seedp', if it occurs more than once per protein.

Obs.: You can integrate a PSI-BLAST search after the PHI-BLAST search, using the argument "-j". E.g.,

```
blastpgp -i query -k pattern -p patseedp -j 2
```

Mega BLAST

Mega BLAST uses a greedy algorithm optimized for aligning sequences that differ slightly as a result of sequencing or other similar «errors». When a larger word size is used, it is up to 10 times faster than more common sequence similarity programs. It is also able to efficiently handle much longer DNA sequences than the blastn program.

Selected megablast arguments:

-D [integer] type of megablast output (def = 0 = alignment endpoints and score; 1 = all ungapped segments endpoints; 2 = traditional BLAST output; 3 = tab-delimited one line format).
-M [integer] maximal total length of queries for a single search (def = 20000000).
-f [T/F] shows full Ids in the output (def = F, only GIs or accessions).
-p [real] identity percentage cut off (def = 0).
-s [integer] minimal hit score to report (def = 0).

To compare two sequences

bl2seq performs a pairwise comparison between two sequences.

Selected bl2seq arguments:

-i [file in] first sequence.
-j [file in] second sequence.
-p [string] program name as in blastall; def = blastp).
-o [T/F] alignment output (def = stdout).
-G [integer] cost to open a gap (def = 0; zero invokes default behavior).
-E [integer] cost to extend a gap (def = 0; zero invokes default behavior).
-W [integer] wordsize (def = 0; zero invokes default behavior).
-M [string] matrix (def = BLOSUM62).
-F [string] filters query sequence (def = T).
-e [real] expectation value E (def = 10.0).
-T [T/F] produces HTML. (def = F).

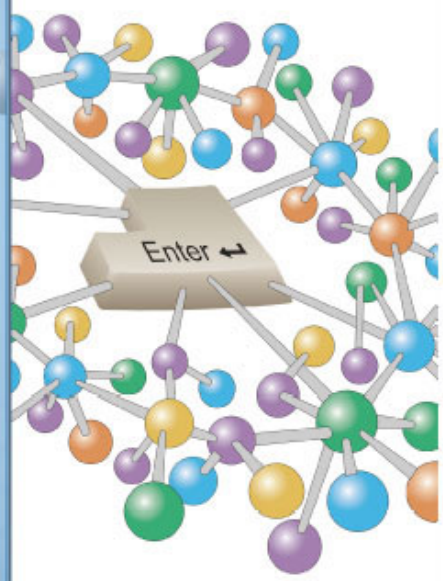
This document was written and designed by Eduardo Fernandes Formighieri with the help of Marcos Renato R. Araújo, Marcelo Falsarella Carazzolle and Gonçalo A. Guimarães Pereira from the Brazilian EMBnet node and distributed by the P&PR Publications Committee of EMBnet.

EMBnet – European Molecular Biology network – is a network of bioinformatics support centers situated primarily in Europe. Most countries have a national node, which can provide training courses and other forms of help for users of bioinformatics software.

http://www.embnet.org/

A Quick Guide to NCBI Blast
First edition © 2004

EMBnet



A Quick Guide
BLAST

A Quick Guide to the NCBI Blast

http://www.ncbi.nlm.nih.gov/blast

Blast (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is a sequence comparison algorithm optimized for speed and used to search sequence databases for optimal local alignments to a query. The NCBI implementation was established by the National Center for Biotechnology Information. The program can be used through the NCBI site or can be installed locally (stand alone blast).

This guide doesn't replace the entire documentation for Blast but can be used as a reference.

Where to start?

For beginners we suggest to first read the documentation of the Blast related to similarity searching (see link below). Other useful pages are available by following the links at the top of this page.

E.g., the glossary and the tutorials:

`http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/similarity.html`

Program selection – web interface options

BLASTN – used to search nucleotide databases with a nucleotide query sequence.

MEGABLAST – a version of BLAST specially designed to efficiently find very similar sequences in a database.

Discontiguous MEGABLAST – a version of MEGABLAST used to identify similar but not identical nucleotide sequences.

Search for short nearly exact matches – used to search for primer or short nucleotide motifs in nucleotide sequences or short peptides in protein sequences.

BLASTP – used to search protein databases with a protein query sequence.

PSI-BLAST (Position-Specific Iterated BLAST) – used to search protein databases with increased sensitivity potentially locating distant homologies. A position-specific scoring matrix is created after each iteration using the selected results from the previous search.

PHI-BLAST (Pattern-Hit Iterated BLAST) – a version similar to PSI-BLAST, but including a user-defined pattern limiting the output to sequences matching the pattern. The patterns must follow the pattern syntax conventions from PROSITE.

BLASTX – makes a six-frame nucleotide query search against a protein database, finding proteins similar to those encoded by the query. Useful when the reading frame of the query is unknown or when it contains errors that may lead to frame shifts.

TBLASTN – makes a protein query search against a dynamically translated nucleotide database. Useful when searching for a specific protein against an unannotated nucleotide database, like HTGs or ESTs databases.

TBLASTX – searches all six-frame query translations against all six-frame database translations. Effectively performs a more sensitive blastp search without doing manual translations.

CDD-Search (Conserved Domain Database Search) – used to identify conserved protein domains.

CDART (Conserved Domain Architecture Retrieval Tool) – explores the domain architectures of proteins.

Blast 2 sequences – direct comparison of two sequences.

VecScreen – screens DNA sequence queries for vector contamination using a database of known vectors.

Main databases (available at NCBI)

Protein *nr* (non-redundant + PDB + SwissProt + PIR + PRF); *swissprot* (latest major release of the SWISS-PROT); *pat* (proteins from patent division of GenBank); *month* (new data released in the last 30 days); *pdb* (3-dimensional structure records from Protein Data Bank).

Nucleotide *nr* (GenBank + EMBL + DDBJ + some PDB); *est* (GenBank + EMBL + DDBJ from EST division); *pat* (nucleotides from patent division); *pdb* (3-dimensional structure records); *month* (new data released in the last 30 days); *chromosome* (complete genomes and chromosomes); *est human* (human subset of EST); *est mouse* (mouse subset of EST); *est others* (subset of EST other than human or mouse); *gss* (Genome Survey Sequence); *htgs* (Unfinished High Throughput Genomic Sequences); *alu_repeats* (select Alu repeats from REPBASE); *dbsts* (STS division + EMBL + DDBJ); *wgs* (assemblies of whole genome shotgun sequences).

LOCAL BLAST INSTRUCTIONS

Format source databases

formatdb formats protein or nucleotide source databases before they can be searched by blastall, blastpgp or megablast. The source database may be in either FASTA or ANS.1 format.

Selected formatdb arguments:

-t [string]	title for database (opt).
-i [file in]	input file for formatting.
-l [file out]	logfile name (opt; def = formatdb.log).
-p [T/F]	type of file (opt; T = protein (def), F = nucleotide).
-o [T/F]	parse options (opt; T = parse SeqID and create indexes; F = no parse, no indexes (def)). Obs.: the first word on the fasta definition line should be a unique identifier (SeqID).
-v [integer]	size of the volume in millions of letters (opt, def = 0). Obs.: This option breaks up large FASTA files into 'volumes' (each with a maximum size of 2 billion characters). I.e.: -v 2000.
-n [string]	base name for BLAST files (opt).

Fasta from databases

fastacmd retrieves FASTA formatted sequences from a BLAST database, if it was formatted using the '-o' option.

Selected fastacmd arguments:

-d [string]	database (def = nr).
-s [string]	search string.
-i [string]	input file with GIs/accessions/locuses for batch retrieval (opt).
-l [integer]	line length for sequence (def = 80, opt).

Stand-alone blast

blastall performs all five flavors of blast comparison.

Selected blastall arguments:

-p [string]	program name (input should be one of "blastp", "blastn", "blastx", tblastn" or "tblastx").
-d [string]	database (def = nr). Obs.: Multiple database names will be accepted if quoted. E.g., -d "nr est".
-i [file in]	query file (def = stdin). Obs.: Query should be in FASTA format. If multiple FASTA entries are in the input file, all queries will be searched.
-e [real]	expectation value threshold (def = 10.0).
-o [file out]	BLAST report output file (opt; def = stdout).
-F [string]	filter query sequence (def = T). Obs.: T = DUST for blastn or SEG for others, and F = no filtering. To change SEG options, use: `-F "S 10 1.0 1.5"`, where 10 = window value, 1.0 = low cut and 1.5 = high cut. For coiled-coil filter: `-F "C 28 40.0 32"`, where 28 = window, 40.0 = cut off and 32 = linker. To use both SEG and coiled-coil: `-F "C;S"`.
-b [integer]	number of alignments (def = 250).
-v [integer]	number of one-line description (def = 500).
-Q [integer]	query genetic code (def = 1).

And the flags are:

–tags	tags selected lines in the Phrap output.
–screen	masks phrap–inferred vectors and chimeric segments in the *.ace* file.
–old_ace	creates an old style format *.ace* file.
–new_ace	creates an new style format *.ace* file.
–ace	same as *new_ace*.
–view	creates *.view* file for *phrapview*.
–print_extraneous_matches	shows non–local matches between contigs.

Other *miscellaneous* options are:

–max_subclone_size	maximum subclone size (to check forward–reverse pairs). (Default = 5000)
–confirm_length	minimum alignment size to confirm an alignment. (Default = 8)
–confirm_score	minimum alignment score to confirm an alignment. (Default = 30)

And the flags are:

–retain_duplicates retain exact duplicate reads.

Phrap output

There are three types of output for Phrap: *standard output*, *standard error* and *files*. The *standard output* and *standard error* are the messages shown on the console during the execution of Phrap. These are independent and can be independently redirected to a file (i.e. using '*> phrap.out*' to redirect standard output into a file name *phrap.out* and '*2> phrap.err*' to redirect standard error into a file named *phrap.err*). The standard output gives a summary of the final assembly, including data anomalies and possible sites of assembly error. The information it provides includes:

(i) Contigs and corresponding reads;
(ii) Read alignments and qualities against contigs;
(iii) Matching regions within and between contigs
(iv) Suspect reads (probable deletions or chimeras)
(v) Possible assembly errors
(vi) Forward/reverse consistency checks.

The standard error basically shows execution errors and warnings, and the point reached in the execution of Phred. The output files keep the name from the input file, but with suffixes added according to file type as follows:

`.contigs`	FASTA file containing contig sequences.
`.contigs.qual`	corresponding qualities file for *.contigs*.
`.singlets`	FASTA file containing singlets (unassembled reads).
`.log`	diagnostic only.
`.problems`	diagnostic only.
`.ace`	input for *Consed* (needs *ace*, *new_ace* or *old_ace* flags).
`.view`	input for *phrapview* (needs *view* flag).

Phrap vector screening

Before running Phrap, vector sequences should be masked or removed, since they may interfere with the assembly. The easiest way to do this is using the program *cross_match*: just create a FASTA file containing all the vector sequences you want to screen for and execute *cross_match* with both screen and sequence files with option *–screen*. E.g., for a FASTA sequence file named *seq_file* and a FASTA vector file named *vector*:

```
cross_match seq_file vector -screen
```

This causes *cross_match* to create a file named *seq_file.screen* containing all sequences from *seq_file* where all bases matching sequences in *vector* are replaced by 'X', and ignored by Phrap during assembly.

Phred/Phrap integration

The easiest way to execute both Phred and Phrap is using the *phredPhrap* script available in the Consed package. To proceed, create a directory to store your assembly (i.e., '*assemble*'), create your vector sequences file and name it *vector.seq* and put it inside this directory. Then, create three directories in it: '*chromat_dir*', '*edit_dir*' and '*phd_dir*'. Put all your trace files (or correspondent symbolic links) in '*chromat_dir*' and just execute *phredPhrap* under '*edit_dir*'. After execution, all PHD files and assembly results will be available in '*phd_dir*' and '*edit_dir*', respectively.

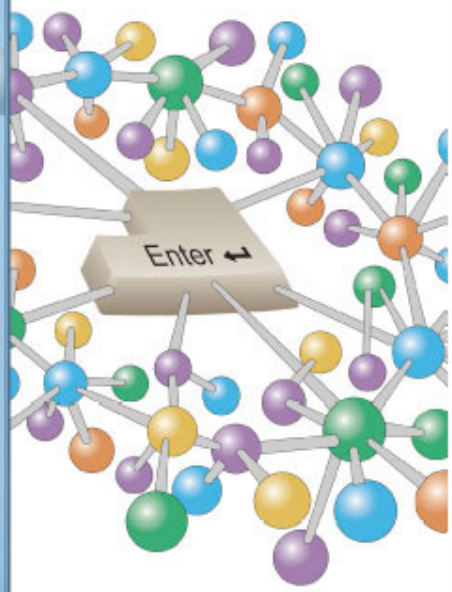
This document was written and designed by Marcos Renato R. Araujo with the help of Eduardo F. Fornighieri, Marcelo Falsarella Carazzolle and Gonçalo A. Guimarães Pereira from the Brazilian EMBnet node, and distributed by the P&PR Publications Committee of EMBnet.

EMBnet - European Molecular Biology network - is a network of bioinformatics support centers situated primarily in Europe. Most countries have a national node which can provide training courses and other forms of help for users of bioinformatics software.

http://www.embnet.org/

A Quick Guide To Phrap
First edition © 2004

EMBnet



A Quick Guide
Phrap

A Quick Guide To Phrap

http://www.phrap.org

Phrap is part of the phred/phrap package that is designed to analyze and assemble sequences from large genomes. It was developed by Phil Green and Brent Ewing at the Department of Molecular Biotechnology, Washington University, and is in the public domain for academic purposes. Phred processes trace data files, generating relevant information used by phrap to assemble shotgun DNA sequences. Phrap is also designed for stand alone use.

This guide doesn't replace for the entire documentation for phrap, but can be used as reference for those who want to use the whole potential of the program. For additional information, please read the phrap documentation.

Phrap command line

The Phrap command line has only one obligatory parameter: the FASTA file containing the sequences to be assembled, but there are several modifiers that can be applied. A common Phrap command line is:

```
phrap seq_file -penalty -9 -ace > phrap.out
```

where *seq_file* is the FASTA file containing the sequences, "*-penalty*" a modifier of the penalty for base mismatch, "*-ace*" a modifier generating the file *seq_file.ace* and the standard output is redirected to *phrap.out* file.

Phrap also checks for a corresponding quality file suffixed with *.qual* (*seq_file.qual* in the example above). In the quality file, each position in the FASTA file corresponds to a quality value between 0 and 97 separated by spaces. The special quality values 98 and 99 are used in visual inspection (manual edition) to indicate inaccurate bases that must be ignored (98) and highly accurate bases, being used to break false joins made by Phrap (99). Every entry in a quality file must match an entry in the corresponding FASTA file.
The easiest way to generate both the FASTA and its corresponding quality file is with *phd2fasta* (available in the Consed package) over PHD files generated by Phred.

Phrap naming convention

In addition to the sequence and quality data, Phrap needs to know three things for each read:

(i) The subclone or other template from which the read is derived;
(ii) Read orientation (forward or reverse) within the subclone;
(iii) The chemistry used to generate the read;

This information is obtained from the sequence name (the string between the '>' symbol and the first space in the header) using the St. Louis naming convention: the portion of the read name up to the first '.' (or whatever value used in *subclone_delim* option – see below) identifies the subclone and the first letter following it indicates the orientation of the read within the subclone and its chemistry, as follows:

	Orientation	Strand	Chemistry
s	Fwd	Single	Dye primer
f	Fwd	Double	Dye primer
r	Rev	Double	Dye primer
x	Fwd	Single	Std. dye terminator
z	Fwd	Double	Std. dye terminator
y	Rev	Double	Std. dye terminator
l	Fwd	Single	Big dye terminator
b	Fwd	Double	Big dye terminator
g	Rev	Double	Big dye terminator

And other special codes:

	Feature
t	T7 cDNA
p	SP6 cDNA
e	T3 cDNA
d	Special cases
c	Consensus pieces
a	Assembly pieces

Phrap command line options

The Phrap command line options are divided into two main groups: *optional modifiers* and *flags*. The *optional modifiers* are used to change default values in Phrap (require a value), and *flags* are used to enable or disable specific features. Some redundant modifiers are not shown here; for a complete list, please read the Phrap documentation.

The *scoring* options change the way Phrap calculates the score for an alignment between sequences, referring to values used in the comparison matrix (matching residues are always rewarded by +1). They are:

–penalty	penalty for base mismatch. Diminish to increase identity. (Default = –2)
–gap_init	penalty to initiate a gap. Decrease to reduce number of potential gaps. (Default = *penalty*–2)
–gap_ext	penalty to extend a gap. Decrease to reduce length of gaps. (Default = *penalty*–1)

And the flags are:

–raw	don't penalize low–complexity regions.

The *banded search* options refer to the banded Smith–Waterman algorithm used by Phrap to realize sequence comparisons, referring to the length of initial perfectly identical subsequences (matches) used to start the comparisons. They are:

–maxmatch	maximum match length. Decrease under lack of memory at cost of speed. (Default = 30)
–minmatch	minimum match length. Increase for speed at cost of stringency. (Default = 14)

The *filtering* options cause Phrap to remove selected matches from the assembly. They are:

–minscore	minimum score for alignment. Increase for stringency. (Default = 30)
–vector_bound	potential vector bases at beginning of each sequence. (Default = 80)

The *input interpretation* options change the way Phrap processes input data. They are:

–subclone_delim	delimitates subclone name. (Default = '.')
–group_delim	delimitates groupname (check *preassemble* flag below). (Default = '_')
–trim_start	bases to remove at start. (Default = 0)

The *assembly* options change the way Phrap merges and assembles sequences. They are:

–forcelevel	stringency relaxation (0 to 10). (Default = 0)
–maxgap	maximum gap size for unmatched region when merging contigs. (Default = 30)

And the flags are:

–revise_greedy	shatters and reattachs pre-contigs on low quality regions to increase overall score. Enable to correct some types of misassemblies.
–shatter_greedy	same as above, without reattaching.
–preassemble	preassembles reads within groups. The group names are defined by the first part of the sequence name prior to *group_delim*.
–force_high	ignores edited high-quality discrepancies.

The *consensus construction* options directly affect the algorithm that creates the weighted directed graph used to determine consensus. Each node in this graph is a subsequence segment. Higher values reduce memory usage but decrease the accuracy of the consensus sequence found. They are:

–node_seg	minimum segment size to define a node. (Default = 8)
–node_space	minimum space between segments in a single sequence to define a node. (Default = 4)

The *output* options change the generation of files after the execution of Phrap. They are:

–qual_show	minimum quality to show sequence in uppercase. (Default = 20)

processing a single file.

–log creates a log file named "phred.log" in the current directory.

–v <n> increases verbosity of output by *n*.

–tags tags common output to facilitate parsing.

–h displays a brief help screen.

–doc displays full phred documentation.

–V displays phred version.

Phred processing options

The processing options affect the way phred works and are designed for experimented users. The available options are:

–nocall disables phred base calling algorithm, setting output sequence to that defined in the chromatogram.

–nonorm disables phred trace normalization. If enabled, the chromatogram is read "as is" instead of normalizing the peaks by the medium value of each nucleotide signal.

–nosplit disables phred compressed peak splitting. If enabled, merged "C" and "G" peaks in signal are not split.

–nocmpqv disables phred compressed peak values recognition. If enabled, qualities of "CC" and "GG" merged peaks aren't reduced.

–ceilqv <value> specifies value as maximum phred quality, without normalization.

–beg_pwd<position> sets the position where phred should start predicting peaks.

–exit_nomatch stops the execution if primer ID is not defined on Phred Parameter File.

–process_nomatch if primer ID is not defined in Phred Parameter File, searches for special "__no_matching_string__" entry on Phred Parameter File and uses it to identify the chromatogram type. If not defined, stops the execution.

The following options refer to trimming capabilities. These options are useful to locate high quality regions on reads and when cutting off low quality regions in them. The options are:

–trim <sequence> finds chromatogram's highest quality region. If "sequence" is defined, phred uses it as restriction enzyme recognition sequence and trims bases off the start of the current sequence before proceeding.

–trim_alt<sequence> same as above, but using "Maximum Score Subsequence" algorithm to locate the highest quality region (recommended).

–trim_cutoff<value> Sets trimming error probability for the previous option, needed for "Maximum Score Subsequence" algorithm. The default value is 0.05.

–trim_fasta trims sequence and quality values written to FASTA files.

–trim_scf trims sequence and quality values written to SCF files.

–trim_phd trims sequence and quality values written to PHD files.

–trim_out trims sequence and quality values written to FASTA, SCF and PHD files.

Phred quality determination

To determine the quality values, phred analyzes the four base traces on each chromatogram, calculates all functions that fit each peak and the best global function fitting all peaks. Based on these, phred determines the error probability on each called base and uses it to calculate the base quality using the following formula:

$$Q = -10 \times \log_{10}(P)$$

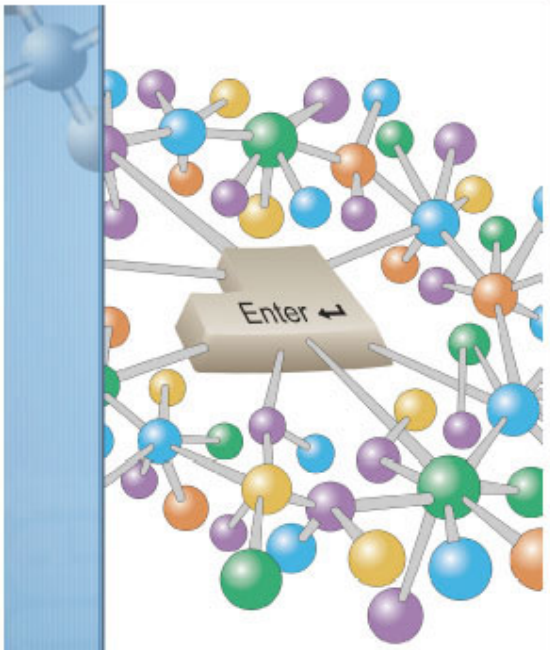
Where Q denotes base quality and P denotes the probability of that base being wrong. For instance, with $P = 0.01$ (one error per 100 bases) the value of Q is 20 and with $P = 0.001$ (one error per 1,000 bases) the value of Q is 30. Note that the default value for P on *–trim_cutoff* option is 0.05, meaning that Q ~13 will be used as the background level value.

This document was written and designed by Marcos Renato R. Araujo with the help of Marcelo Falsarella Carazzolle and Eduardo Fernandes Formighieri from the Brazilian EMBnet node and distributed by the P&PR Publications Committee of EMBnet.

EMBnet - European Molecular Biology network - is a network of bioinformatics support centers situated primarily in Europe. Most countries have a national node which can provide training courses and other forms of help for users of bioinformatics software.

http://www.embnet.org/

A Quick Guide To Phred
Second edition © 2004



EMBnet

A Quick Guide
Phred

A Quick Guide To Phred

http://www.phrap.org

Phred is part of the phred/phrap package that is designed to analyze and assemble sequences from large genomes. It was developed by Phil Green and Brent Ewing at the Department of Molecular Biotechnology, Washington University, and is in the public domain for academic purposes. Phred processes trace data files, generating relevant information used by phrap to assemble shotgun DNA sequences. Phred is also designed for stand alone use.

This guide doesn't replace the entire documentation for phred, but can be used as reference for those who want to use the whole potential of program. For additional information, please read the phred documentation.

Source data

The main source data are trace files. Trace files are also known as chromatograms and consist of computational representations of data provided by sequencing machines:

Phred recognizes three formats of chromatograms, depending on the sequencing machine used: SCF, ABI and processed ESD.

Phred Parameter File

The first step to use phred is to set up the chemistry parameter file. This file comes with the phred/phrap package under the name *phredpar.dat* and can be edited by the user to add new calibrations when needed. It defines which calibration phred should use to process a chromatogram, allowing phred to resolve certain motif problems. The file format is:

```
begin chem_list
"primer ID"    chemistry    dye    machine
"primer ID"    chemistry    dye    machine
.
.
end chem_list
```

Where:

primer ID dye identification. Should match "primer ID" string in chromatogram.

chemistry identifies reaction chemistry. Options are "*primer*", "*terminator*" or "*unknown*".

dye identifies dye type. Options are "*rhodamine*", "*d-rhodamine*", "*big-dye*", "*energy-transfer*", "*bodipy*" or "*unknown*".

machine identifies the machine used. Current machines recognized by phred (version 0.020425.c) are "*ABI_373_377*", "*ABI_3100*", "*ABI_3700*", "*Beckman_CEQ_2000*", "*MolDyn_MegaBACE*" and "*LI-COR_4000*".

If you're not sure which "primer ID" to use, don't worry: phred tells you what to do.

This file must be located in a common place accessible by all users allowed to use phred and its location must be indicated with the PHRED_PARAMETER_FILE environment variable.

For example, in Unix machines

```
export PHRED_PARAMETER_FILE=/usr/local/etc/phredpar.dat
```

and Windows NT machines:

```
set PHRED_PARAMETER_FILE=\usr\local\etc\phredpar.dat
```

Phred input parameters

The input parameters inform phred where trace files are. There are two ways to do that:

–id <directory> directory containing all chromatograms.

–if <text file> text file containing the full path of each chromatogram.

Phred output parameters

The output parameters for phred define in which format the results should be produced. There are several output parameter classes: *base calling*, *quality*, *SCF*, *PHD* and *poly*. Note that these options can be combined to produce several output files in distinct formats.

The base calling options class refers to DNA sequences. The options are:

–st <fasta|xbap> output file format. Recognized formats are FASTA and XBAP (default FASTA).

–s: creates files in the local directory using chromatogram file names and appending ".seq" to them.

–sd <directory> creates files in the directory indicated using chromatogram file names and appending ".seq" to them.

–sa <file> creates a single file containing the results for all processed chromatograms.

The quality options are used to generate quality information for the produced DNA sequences. The options are:

–qt <fasta|xbap|mix> output file format for quality values. Should be "fasta" for trimmed FASTA format, "xbap" for XBAP format or "mix" for untrimmed FASTA format (default FASTA). Check the "Phred processing options" below for trimming capabilities.

–q creates quality files in FASTA format in the local directory using chromatogram file names and appending ".qual" to them.

–qd <directory> creates files in FASTA format in the directory indicated using chromatogram file names and appending ".qual" to them.

–qa <file> creates a single quality file containing the results for all chromatograms processed.

–qr <file> creates a histogram file with the number of high quality bases for each chromatogram.

The SCF options are used to generate SCF files based on trace data. These files can be used by programs that do not recognize the ABI or ESD formats. The options are:

–c creates SCF files in the local directory using chromatogram file names.

–cd <directory> creates SCF files in the directory indicated using chromatogram file names.

–cp <1|2> number of bytes used to represent a single trace data value.

–cv <1|2|3> writes SCF output in SCF format 1, 2 or 3.

–cs ensures that the largest trace value represents the highest value in the SCF output.

The PHD options are used to generate PHD files based on trace data. These files are in human readable format and contain information about base calling and quality for each chromatogram. The options are:

–p creates PHD files in the local directory using chromatogram file names and appending ".phd.1" to them.

–pd <directory> creates PHD files in the directory indicated and appending ".phd.1" to them.

The poly options are used to generate poly files. These files are used to detect polymorphic bases, storing information concerning the values for each peak represented on a chromatogram. The options are:

–d creates poly files in the local directory using chromatogram file names and appending ".poly" to them.

–dd <directory> creates poly files in the directory indicated and appending ".poly" to them.

Other miscellaneous output options:

–raw <name> writes "name" in the header of the output file instead of the current file name when

7.3. ANEXO C. Detalhes do sistema de anotação semi-automática.

O sistema de bioinformática para genomas do LGE inclui desde a submissão de seqüências e a realização de diversas análises automáticas até o controle de produção dos diferentes laboratórios de cada projeto, listas de discussão, análises em grande escala e a integração com análise de microarranjos (*microarrays*). A página de entrada do Projeto Vassoura de Bruxa é mostrada na figura 1.

Figura 1 – Página inicial do Projeto Vassoura de Bruxa.

Neste capítulo será discutida uma parte deste trabalho, voltada à anotação de genes em genomas seqüenciados, *drafts* (ou rascunhos, seqüenciamentos incompletos de genomas) ou completos através do sistema de anotação do LGE. Detalhes do sistema de mineração e de anotação podem ser encontrados nos anexos C e D.

7.3.1. **Análises iniciais**

Durante o processo de submissão seqüências são realizadas ou disparadas várias análises que servirão de base para os sistemas de mineração e anotação. Existem algumas diferenças de acordo com o tipo de projeto (ESTs ou genômico), mas de modo geral ocorrem os passos descritos abaixo. O armazenamento de dados ocorre no decorrer das análises.

- Verificação de padrão de nomenclatura e de existência de submissão anterior da mesma placa (ferramenta de edição de nomenclatura na figura 2);
- Verificação de qualidade;
- Trimagem – marcação de regiões de vetores, contaminantes, caudas poli-A etc;
- Análises de Blast – o programa utilizado (normalmente BlastX e BlastN) e o banco contra o qual a procura é feita variam de acordo com cada caso;
- Análise de redundância (para ESTs);
- Geração e envio de relatórios
- Armazenamento das informações em bancos de dados e diretórios.

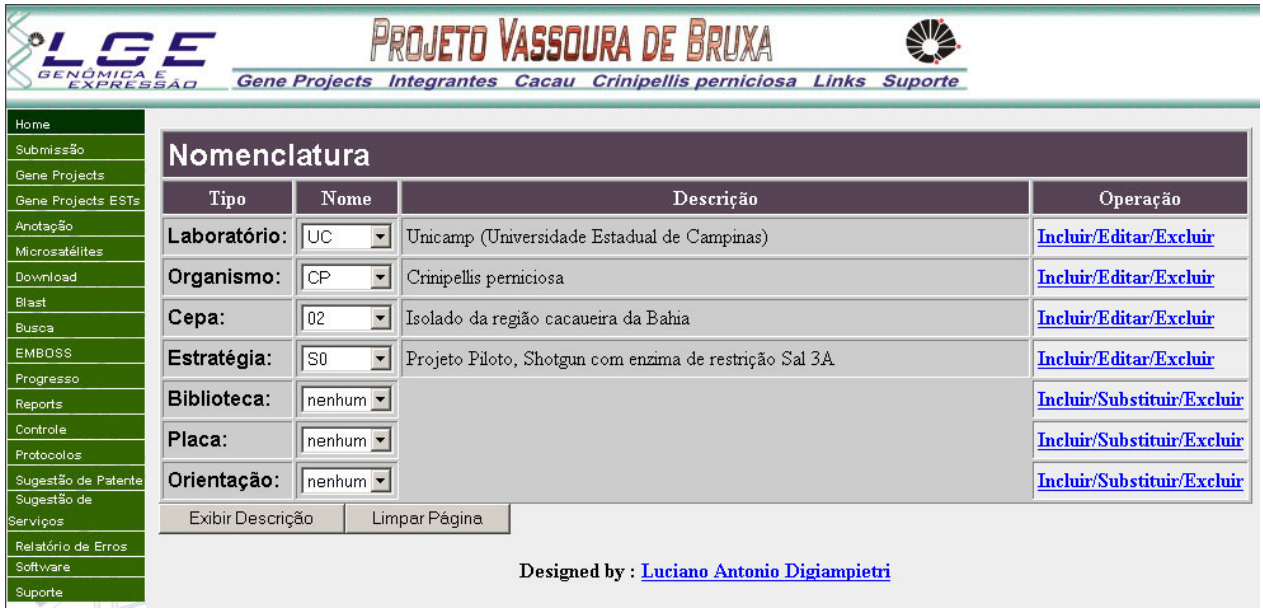


Figura 2 – Ferramenta de edição de nomenclatura.

7.3.2. **Sistema de anotação**

O principal programa da interface de anotação de clusters chama-se anotacao.cgi e é um programa desenvolvido na linguagem Perl e usa módulos de código aberto como CGI.pm (http://stein.cshL.org/WWW/CGI), GD.pm (http://www.boutell.com/gd/) e DBI.pm (http://dbi.perl.org). A lista de ferramentas e material relacionado á anotação é mostrada na fig. 3.

Figura 3 – Menu de ferramentas de anotação do Projeto Vassoura de Bruxa.

O programa anotacao.cgi é disponibilizado através da tecnologia CGI. Esta escolha foi feita por favorecer a escalabilidade, disponibilidade e compatibilidade do sistema. Além disso, as atualizações nos bancos de seqüências usados pelo sistema precisam ser realizadas apenas em nossos servidores.

Existem três classes de usuários que interagem com a interface de anotação. Primeiramente, o *anotador*, que é o usuário que coleta e submete informações sobre os clusters. Segundo, o selecionador, que é o usuário que seleciona clusters interessantes para o seu grupo de anotação e revisa a anotação dos anotadores de seu grupo. Finalmente, o *curador*, que é um tipo especial de

usuário que tem permissão para revisar e alterar as informações entradas no sistema por qualquer anotador de qualquer grupo. Há ferramentas on-line para cadastrar e gerenciar usuários e senhas destas três classes.

O fluxo de anotação começa com a seleção de clusters. Com o objetivo de ajudar o selecionador na escolha de seus clusters, a ferramenta disponibiliza buscas baseadas em nomes de arquivos, laboratório de seqüenciamento, estratégia de seqüenciamento, orientação dos *reads*, organismo, biblioteca, prato, posição e palavras-chave. A fig. 4 mostra um detalhe da interface inicial desta ferramenta e a fig. 5 parte da página de resultados de uma busca.

Figura 4 – Detalhe da ferramenta de seleção de clusters.

Clusters List

Legenda:

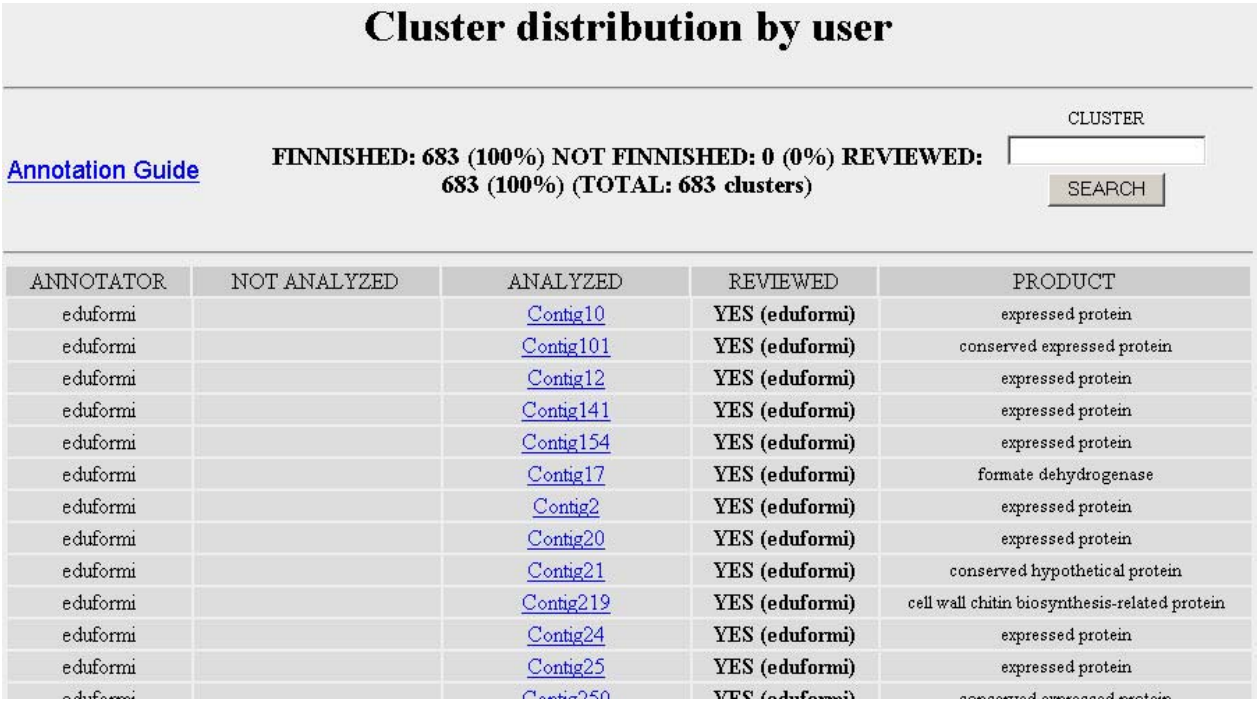
Anotação automática | Anotação finalizada | Seqüência selecionada

Select	Contig	Gene Name	Gene Symbol	Blast Result	E-value
	Contig165			gi|2245560|gb|AAB63450.1| cytochrome c oxidase subunit II [Homo sapiens]	2e-06
	Contig187	Cytochrome b-245 heavy chain	cybb	gi|66848844|gb|EAL89173.1| NADPH oxidase (NoxA), putative [Aspergillus fumigatus Af293] gi|70992725|ref|XP_751211.1| NADPH oxidase NoxA [Aspergillus fumigatus Af293]	2e-61
	Contig190	Glucose oxidase precursor	gox	gi|66846943|gb|EAL87274.1| GMC oxidoreductase [Aspergillus fumigatus Af293] gi|70988923|ref|XP_749312.1| GMC oxidoreductase [Aspergillus fumigatus Af293]	2e-20
	Contig242	NADPH oxidase	noxA	gi|83767498|dbj|BAE57637.1| unnamed protein product [Aspergillus oryzae]	9e-17
				gi|71018881|ref|XP_759671.1|	

Figura 5 – Parte da página de resultados de busca de clusters.

Há também uma ferramenta de Web Blast que possibilita ao usuário fazer buscas no banco de clusters para encontrar seqüências do banco similares a uma seqüência de consulta. Os resultados da busca incluem, para cada cluster, seu melhor hit de alinhamento contra o banco NR, anotação automática com termos do *Gene Ontololgy*, o grupo que possui o cluster e quando o cluster foi modificado pela última vez. O selecionador navega nos resultados da busca e pode escolher para seu grupo clusters de interesse que ainda não foram selecionados.

Uma vez que a fase de seleção começou, a anotação dos clusters pode começar. Quando um anotador entra no sistema com seu nome de usuário e senha, ele pode ver uma tabela com todos os clusters associados a seu grupo e se ele já foi anotado ou não (Fig. 6). Então, o anotador pode selecionar um cluster e clicar sobre o nome do cluster para acessar a interface de anotação. A página também traz ferramenta de busca e o status da anotação (número de clusters finalizados, não finalizados, revisados e total de clusters).



Cluster distribution by user

Annotation Guide

FINNISHED: 683 (100%) NOT FINNISHED: 0 (0%) REVIEWED: 683 (100%) (TOTAL: 683 clusters)

CLUSTER SEARCH

ANNOTATOR	NOT ANALYZED	ANALYZED	REVIEWED	PRODUCT
eduformi		Contig10	YES (eduformi)	expressed protein
eduformi		Contig101	YES (eduformi)	conserved expressed protein
eduformi		Contig12	YES (eduformi)	expressed protein
eduformi		Contig141	YES (eduformi)	expressed protein
eduformi		Contig154	YES (eduformi)	expressed protein
eduformi		Contig17	YES (eduformi)	formate dehydrogenase
eduformi		Contig2	YES (eduformi)	expressed protein
eduformi		Contig20	YES (eduformi)	expressed protein
eduformi		Contig21	YES (eduformi)	conserved hypothetical protein
eduformi		Contig219	YES (eduformi)	cell wall chitin biosynthesis-related protein
eduformi		Contig24	YES (eduformi)	expressed protein
eduformi		Contig25	YES (eduformi)	expressed protein

Figura 6 – Parte da página de lista de clusters atribuídos a um selecionador.

A interface de anotação é constituída por oitos seções principais e tem vários facilitadores para poupar tempo e trabalho do anotador. Durante o decorrer deste projeto, a interface de anotação foi totalmente remodelada de acordo com a demanda dos usuários. Os e-mails, telefones e formulários pra contato foram utilizados com freqüência para notificar erros na ferramenta, solicitar novas funcionalidades e dar sugestões. A ferramenta já foi bastante utilizada em diferentes projetos genoma, já podendo ser considerada estável.

A seguir uma descrição breve da interface de anotação:

- Cabeçalho: traz links do projeto, manuais, setas de navegação e o nome do cluster (fig. 7).
- Identificação: nesta seção o anotador entra com informações sobre o produto, fenótipo, domínio, organismo homólogo, símbolo do gene, número EC (*Enzyme Comission number*) e número TC (*Transport Comission number*) (fig. 7).
- Classificação: aqui o anotador pode ver ou editar termos *Gene Ontology* que descrevem o cluster. Existe também um link direto do termo selecionado para o visualizador de

ontologias AMIGO (http://www.amigo.org). Opcionalmente, os coordenadores do projeto podem inserir nesta seção um segundo sistema de classificação (fig. 7).

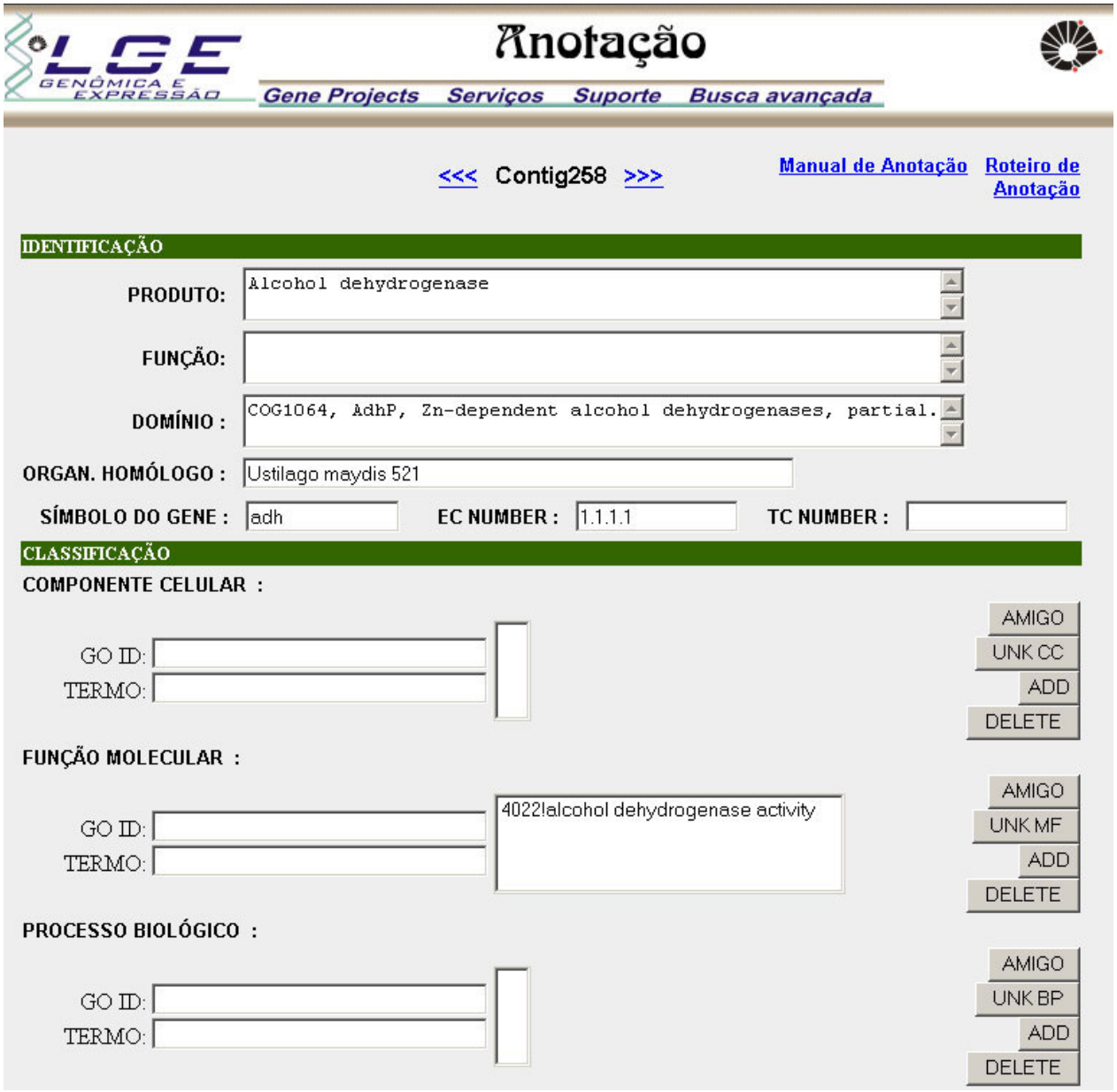


Figura 7 – Parte da interface de anotação – cabeçalho, identificação e classificação.

- Visualização: o anotador pode ver a seqüência do cluster e o complemento reverso da seqüência, os *reads* que constituem o cluster, a montagem do cluster e a seqüência do cluster traduzida em todos os seis quadros de leitura (fig. 8).

- Sinalização: aqui o anotador pode atribuir marcadores para o cluster. Existem marcadores para indicar se um cluster contém a seqüência codante completa de um gene conhecido; se algum *read* que constitui o cluster contém a seqüência codante completa de um gene conhecido; se o cluster tem problemas de montagem tais como um deslocamento no quadro de leitura ou uma região de repetições significativa, por exemplo; se o cluster é um contaminante (fig. 8).
- Alinhamentos Blast pré-processados: aqui o anotador pode ver um sumário sobre os alinhamentos contra alguns bancos de seqüências. Opcionalmente, o anotador pode clicar em um link para visualizar os alinhamentos. A lista de alinhamentos pré-processados disponíveis é totalmente configurável pela equipe de bioinformática do projeto de acordo com as necessidades específicas do projeto (fig. 8).

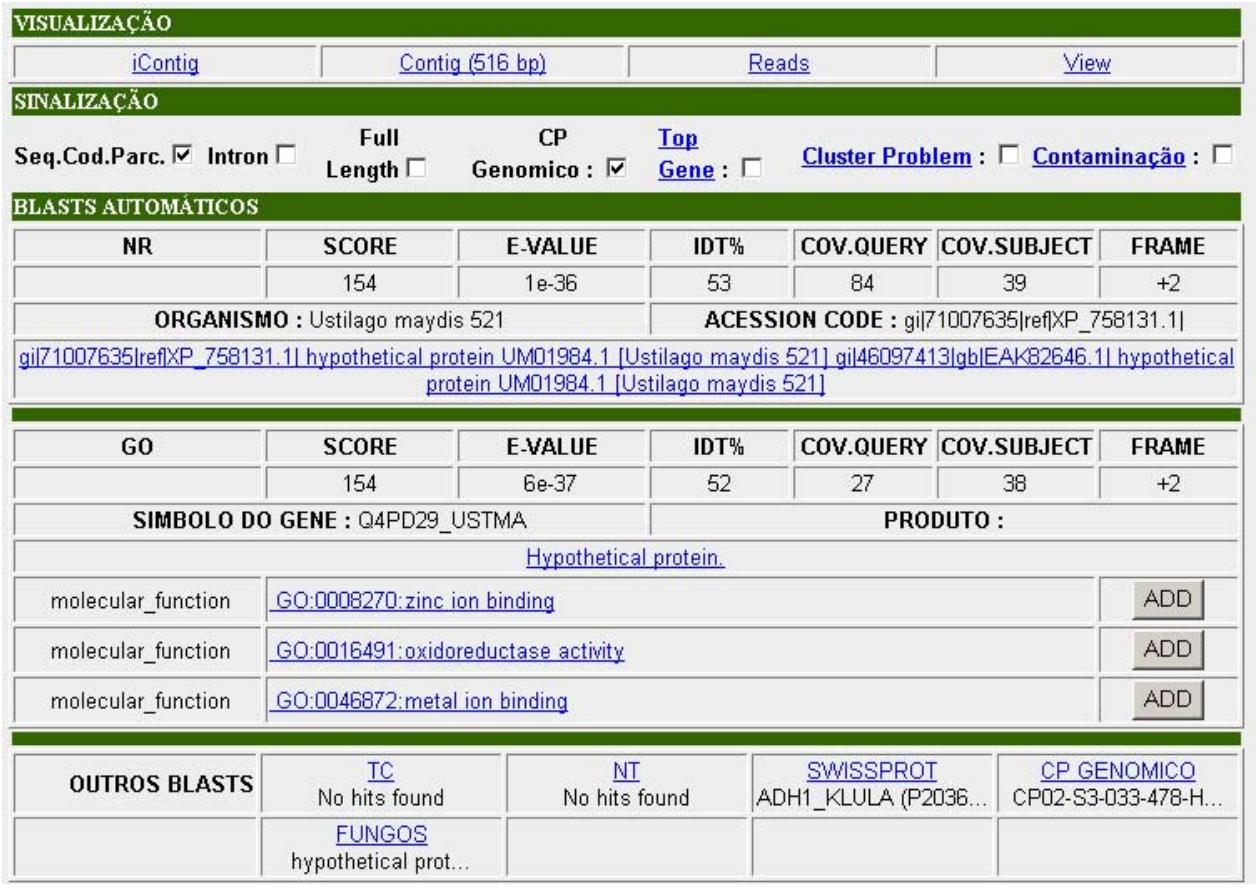


Figura 8 – Parte da interface de anotação – visualização, sinalização e resultados de blast.

- Buscas facilitadas: nesta seção há links para um conjunto de interfaces de *web blast*. Estas buscas blast, diferentemente das anteriores, são dinamicamente processadas. O sistema carrega as interfaces de *web blast* com parâmetros padrões e com a seqüência de consulta já preenchida. Existe também uma interface para buscas por palavras-chave em algumas bases de dados biológicas. A lista de sites disponíveis tanto para buscas por palavras-chave como para buscas blast são configuráveis pela equipe de bioinformática (fig. 9).
- Bloco de notas: na seção do bloco de notas há dois campos de entrada de texto. No primeiro, o dono do cluster pode entrar com notas pessoais e informação relevante sobre o cluster que ainda não tenha sido descrita em nenhuma seção anterior. A segunda, chamada *guest notepad*, pode ser editada por qualquer anotador, mesmo que ele não seja o dono do cluster, e deve conter informação para ajudar o dono do cluster a anotá-lo (fig. 9).
- Controle: nesta seção, o anotador pode sinalizar que a anotação está terminada, salvar as atualizações ou ver um histórico de anotação do cluster. O histórico contém todas as operações de edição feitas no cluster, inclusive o usuário que fez as edições, data e hora. O histórico tem um nível de detalhe tal que ele pode ser usado para reconstruir o banco de anotação. O selecionador pode também reservar o cluster para pesquisa funcional ou devolver um cluster selecionado previamente. Para o curador, existe um marcador para sinalizar se a anotação do cluster já foi revisada (fig. 9).

Figura 9 – Parte da interface de anotação – buscas facilitadas, bloco de notas e controle.

Adicionalmente foram desenvolvidos scripts de reconstrução do banco de dados a partir da tabela de histórico de anotação, o que representa um nível a mais de segurança para os dados do projeto além dos convencionais *backups*. Esta tabela de Histórico de anotação registra triplas de (usuário, data, operação) de modo a poder recuperar todo o banco de dados de anotação em caso de falhas em alguma tabela.

O tempo de anotação foi bastante reduzido com a implementação de atalhos, resultados de Blast pré-processados, ferramentas de tradução on-line e uma visualização dos clusters com marcação das qualidades das bases em escalas de cinza e marcação em vermelho de discrepâncias em regiões de alta qualidade. A verificação manual de clusters com similaridade muito fraca nos resultados de blast teve seu tempo reduzido drasticamente com o novo sistema. Em alguns genomas a anotação destes clusters, antes das mudanças, consumia a maior parte do precioso tempo dos especialistas.

Além desta interface foi desenvolvido um módulo de busca avançada que permite que o usuário localize, pelos dados de anotação, clusters de seu interesse. Essa busca implementa uma lógica booleana simplificada, onde o usuário pode realizar a operação E (AND) e a operação OU (OR) simultaneamente em todos os campos, assim como negar (NOT) alguns campos de busca específicos. O número de resultados por página é fornecido pelo usuário. Nos resultados da busca, é possível encontrar um breve sumário da anotação de cada cluster.

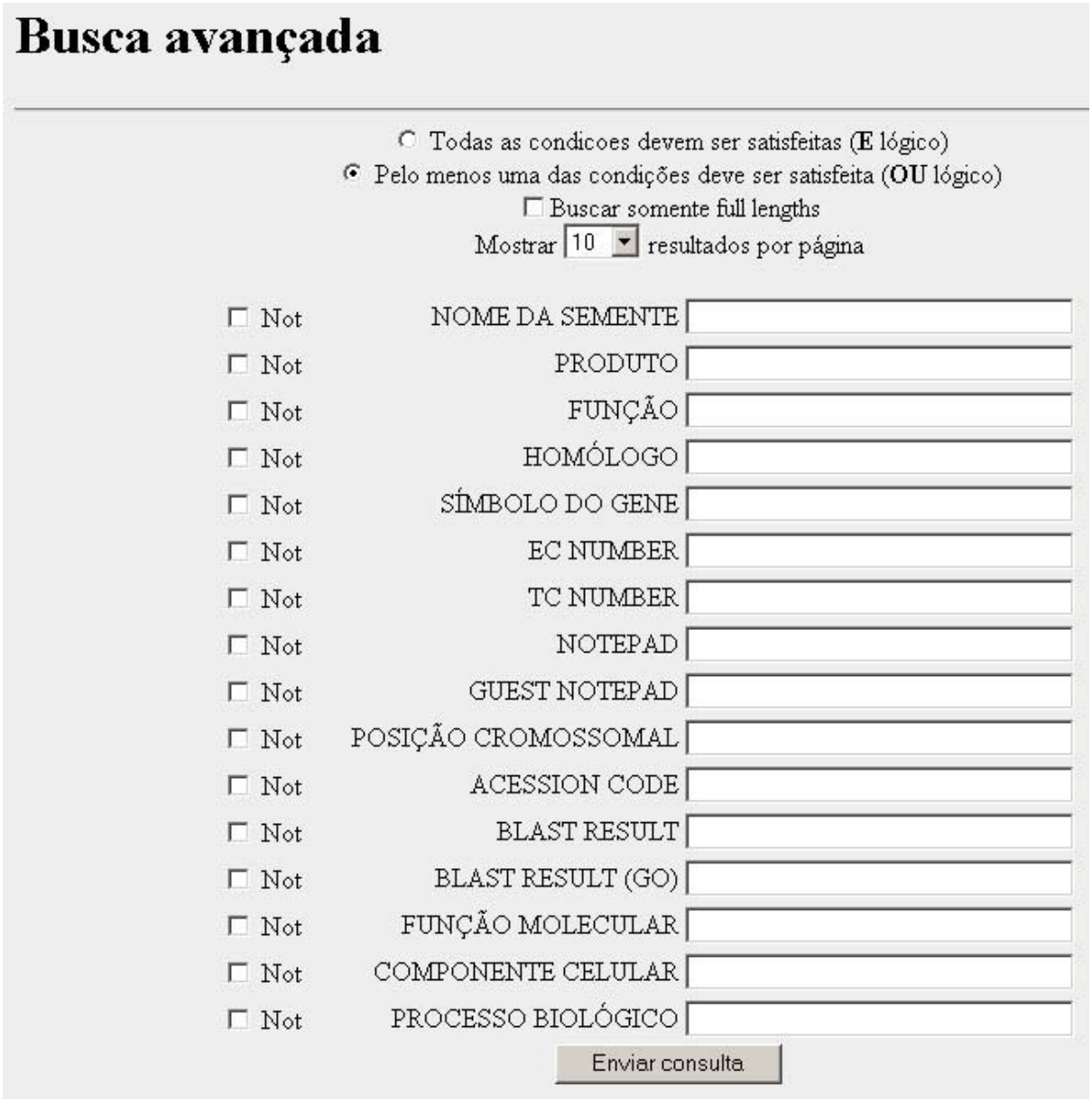


Figura 10 – Interface da ferramenta de busca avançada.

A interação com pessoas de outros laboratórios propiciada principalmente pelos projetos genoma através de projetos, reuniões, treinamentos e cursos foi muito proveitosa (o anexo G apresenta uma lista de cursos e palestras). A localização do Laboratório de Bioinformática dentro do Laboratório de Genômica e Expressão, com contato direto com pesquisadores “de bancada” nas reuniões semanais e no dia-a-dia permitiu, conjuntamente com a interação através dos projetos, o desenvolvimento de ferramentas com grande aplicabilidade, constantemente exigidas e melhoradas. Esta demanda constante gera o desenvolvimento apressado em alguns casos, exigindo planejamento e adaptações posteriores, entretanto a ótima aceitação das ferramentas demonstra que o grupo do Laboratório de Bioinformática do LGE já conta com uma boa estrutura de programas e desenvolvimento, e atingiu maturidade para utilizar menos tempo com serviços e mais com desenvolvimento e publicações.

7.4. ANEXO D. Produção didática: Manual de utilização do programa Gene Projects para o Sistema de mineração e anotação do Laboratório de Genômica e Expressão – LGE/UNICAMP. Utilizado nos projetos Genoma Vassoura de Bruxa, Café, Genolyptus, Camarão, entre outros.

Manual do Gene Projects

1. Imprescindível

início

É muito importante ler as dicas iniciais. Lá se encontram informações que irão prevenir erros e poupar seu tempo. Boa leitura e bom trabalho.

início

2. Dicas iniciais

início

- Para seleção de informações no Word, é possível fazer uma seleção independente de linhas utilizando a tecla “alt” enquanto se move o mouse. Interessante para selecionar reads de um resultado de blast, por exemplo;
- Sempre utilize os links “VOLTAR” da própria página. Se utilizar o botão do navegador, o programa pode se confundir e gerar erro;
- Os campos de observações e os filtros têm como objetivo facilitar o seu trabalho;
- O programa é constantemente melhorado, mas já possui recursos suficientes para facilitar a busca de genes;
- Reporte falhas ou sugestões. São sempre bem vindas;
- Convenções – para descrever uma seqüência de ações, convencionamos o seguinte significado para:
  - <> - apertar a tecla “Enter”;
  - <ctrl> + ação – executar ação enquanto mantém a tecla ctrl apertada. O mesmo vale para <alt> e <shift>;
  - ação1, ação2, ação3 – seqüência de ações a ser executada ;
  - Por “ação” entenda-se apertar uma tecla ou clicar com o mouse, por ex.;
    Cópia de texto no wordpad para novo arquivo:
    “<ctrl> + T, <ctrl> + C, novo documento, <ctrl> + V”
    – são 4 ações separadas por vírgulas. (1) manter a tecla “ctrl” apertada e apertar a tecla “T”, (2) manter a tecla “ctrl” apertada e apertar a tecla “C”, (3) abrir um novo documento, (4) manter a tecla “ctrl” apertada e apertar a tecla “V”.

início

3. Login

início

O primeiro passo para trabalhar com o Gene Project é cadastrar um login e uma senha. Na página de entrada você pode cadastrar um novo usuário. Para o e-mail que fornecer serão enviados avisos de término de processo.

Confirmado o cadastro, o login pode ser feito na página principal. Não se esqueça de guardar bem sua senha.

início

4. Criar novo projeto

início

O primeiro passo é criar um novo projeto, a partir do link correspondente. São três campos para preencher. No primeiro, 'Título', não pode haver espaços ou caracteres especiais. Se quiser utilizar mais de uma palavra, utilize o caractere "_" entre elas. Ex.: "titulo_do_projeto".

A seguir, o campo 'Descrição'. Aqui se pode escrever em detalhes o que é seu projeto, ou do que ele trata. O campo permite espaços, acentos etc.

Finalmente cadastre o e-mail, para onde serão enviadas mensagens do programa e de outros usuários. 'Incluir'.

O programa voltará à tela dos seus projetos. Na coluna 'Projeto' estão links para as páginas de cada projeto criado. Na coluna 'Descrição', links para alterar a descrição de cada projeto. Ainda data da última modificação e botões de seleção para apagar um projeto.

início

5. Novos reads

(reads search | keyword search | blast search | all reads | pattern search)

início

A partir da página dos seus projetos, entre no projeto em que quer trabalhar. Clique no link 'ACRESCENTAR NOVOS READS'.

Existem dois quadros na página. O primeiro apresenta as diferentes formas de acrescentar reads: reads search, keyword search, blast search e all reads. O segundo, opções de ação a serem realizadas com os reads listados.

Vamos tratar de cada forma de busca de reads separadamente. Em todos os casos, existem duas opções que exigem atenção: Filtros e Observação. Em 'FiILTROS' existem dois campos para filtrar a qualidade dos reads a serem inseridos, sendo possível determinar a porcentagem máxima de vetores e a porcentagem mínima de qualidade acima de 20. O campo 'OBSERVAÇÃO' é muito importante para a organização do projeto, pois permite associar os reads à forma como foram inseridos. Por exemplo, uma palavra chave.

5.1. Reads Search

(5. Novos reads | início)

O primeiro é o 'reads search', que realiza busca pelo nome de cada read. Clicando na figura à esquerda do título (⊞) você verá as opções desta busca. Neste caso existem duas formas principais de busca: por nome de arquivo ou por seleção. Para 'BUSCA POR NOME', você deve usar uma lista de nomes, sendo um por linha, ou utilizar coringas. Pode ainda escolher um arquivo texto que contenha a lista de reads.

Coringas? Sim, coringas. Eles são excelentes para facilitar nossa vida. Funcionam da seguinte forma: O " ? " substitui um caractere qualquer, e o " * " significa quaisquer caracteres daqui para diante.

Por exemplo: para buscar CP02-S2-032-338-F10-UE.F e CP02-S2-032-338-F10-UE.R, os dois lados do mesmo clone, pode-se pedir: "CP02-S2-032-338-F10-UE.?". Para buscar todos os reads de CP feitos no lab EU, pode-se pedir: "CP*-UE*". Significa: todos os reads que comecem com "CP", tenham qualquer número ou qualidade de caracteres até que se encontre um texto "-UE", e tenham qualquer número ou qualidade de caracteres após este "-UE".

Para 'BUSCA POR SELEÇÃO', basta clicar nos itens de cada coluna para selecionar o que interessa. Pode-se fazer múltipla seleção, com as teclas "ctrl" e "shift". Para selecionar dois itens da mesma coluna: "clique1, <ctrl> + clique2". Para selecionar um trecho: "clique1 (no início do trecho), <shift> + clique2 (no fim do trecho)". Na coluna que não houver nenhuma seleção, serão utilizados todos os valores.

Após efetuar a busca, aparecerão os resultados da mesma. Você deve selecionar os reads que interessam e exportar os reads selecionados. Após selecionar todos os reads que deseja, pelos diferentes processos, tecle em 'voltar' para chegar à página principal do projeto.

5.2. Keyword Search

(5. Novos reads | início)

Esta opção de busca por palavras nos resultados de blast provavelmente vai ser a primeira a ser utilizada na criação de um projeto. A busca por

palavras permite a utilização de operadores booleanos (“ ”, AND, OR e AND NOT – sempre em letras maiúsculas), e exige que as palavras a serem procuradas estejam sempre em letras minúsculas.

Operadores:

- “ ” – para delimitar expressões, como *“homo sapiens”*;
- AND – para trazer mais de uma palavra, como *sapiens AND human*, que busca resultados com as duas palavras, em qualquer ordem;
- OR – para trazer uma ou outra palavra. *sapiens OR human* traz resultados que contenham a palavra sapiens ou a palavra human;
- AND NOT – para não retornar resultados que contenham a palavra. No caso *sapiens AND NOT human*, a busca trará os resultados que contém a palavra sapiens, mas que não contenham a palavra human;
- ( ) – para delimitar a ação de outros operadores. Ex. *sapiens AND NOT (human OR mouse)*.

A opção ‘MATCH WHOLE WORD’ restringe a procura à palavra exata. Existem ainda as opções ‘READS NEW’ e ‘READS BELOW’, que ainda não estão ativas.

Após efetuar a busca, aparecerão os resultados da mesma. Você deve selecionar os reads que interessam e exportar os reads selecionados. Após selecionar todos os reads que deseja, pelos diferentes processos, tecle em ‘voltar’ para chegar à página principal do projeto.

5.3. Blast Search

(5. Novos reads | início)

Busca através de comparação de seqüência com banco de dados da Cp através do Blast. Basta escolher o programa e o banco de dados (‘database’), colocar a seqüência em formato fasta no campo correspondente, colocar a palavra para a observação e executar a procura (‘SEARCH’). Você também pode localizar um arquivo texto ao invés de colar a seqüência.

O formato fasta inclui um título e uma seqüência. O título é sempre precedido por >, e a seqüência deve ter a linha quebrada após cerca de 50 bases (este valor pode variar, mas não deve ser muito grande). Por ex.:

```
>nome do gene
cagctagcgttatactgtgcagtcgatcgatctattccgttgcgatctat
atatatatatctgcgtcgagtcgcgtagtgatatatatcgagcgcgatat
atatcgcgatctacgatgctgatcagtgtacgatcgatgctagcagatcg
atcgactcgatgatcgatcgacgctgactagctgactacgctagctagct
atatcgcgatctacgatgctgatcagtgtacgatcgatgctagcagatcg
```

```
cgtcgagtcgcgtagtgatatat
```

Após efetuar a busca, aparecerão os resultados da mesma. Você deve selecionar os reads que interessam e exportar os reads selecionados. Após selecionar todos os reads que deseja, pelos diferentes processos, tecle em 'voltar' para chegar à página principal do projeto.

A seqüência fasta que você inserir receberá o nome "query", e cada seqüência que alinhar com ela e aparecer no resultado da busca será chamada de "subject". No site do NCBI existe uma excelente documentação, mas daremos uma breve explicação sobre os principais programas BLAST:

Query	BD	Compara	Programa
nt	nt	nt	Blastn
nt (traduzido)	aa	aa	Blastx
aa	aa	aa	Blastp
aa	nt (traduzido)	aa	Tblastn
nt (traduzido)	nt (traduzido)	aa	Tbaslx

- Blastn – compara nucleotídeos contra BD de nucleotídeos;
- Blastx – compara seis fases de leitura da seqüência de nucleotídeos (query) contra BD de proteínas;
- Blastp – compara aminoácidos contra BD de proteínas;
- Tblastn – compara seqüência de aminoácidos contra BD de nucleotídeos, traduzido dinamicamente nas seis fases de leitura;
- Tblastx – compara nucleotídeos traduzidos nas seis fases contra BD de nucleotídeos também traduzido nas seis fases.

As principais utilizações destes programas no Gene projects são:

- Quando temos um ORF definido em nossa seqüência, podemos utilizar o blastp;
- Para saber se existe algum gene que codifique proteína no nosso trecho, utiliza-se o blastx;
- Para buscar um gene conhecido no nosso BD de contigs, utilizamos o tblastn;
- Quando queremos achar genes em nosso trecho que possam não ter sido ainda descritos, mas que existam em seqüências de nucleotídeos depositadas, utilizamos o tblastx;
- E finalmente, para comparações de genes de RNA (como o ribossomal), e ainda para comparações entre organismos próximos, utilizamos o blastn.

Na interpretação dos resultados de Blast é preciso levar em consideração alguns fatores.

Comentaremos os principais: E value, *identities* e *score*.

Identities, ou identidade é a relação entre o número de *matchs*, ou letras similares, com o tamanho do alinhamento. *Score* é a pontuação do programa para o alinhamento. Porém, estes dois valores não podem ser analisados individualmente. É preciso utilizar como parâmetro principal de seleção o E-value, que é o valor estatístico da probabilidade de o alinhamento acontecer por acaso. Quanto menor este valor, menor a chance de que o alinhamento tenha acontecido por acaso, e, portanto, maior a chance de que este mostre uma similaridade real.

Existem três formatos possíveis para o valor de E-value: 0.0; 3e-35; e 0.14. O formato 3e-35 significa 3 x 10^{-35}, e portanto é muito mais próximo de 0.0 (zero) do que 0.14. O quanto o alinhamento pode ser significativo depende também de outras informações, mas normalmente são utilizados valores de E-value menores que 1e-5 (por exemplo: 1e-6, 1e-7, 1e-8 etc.). Valores de E-value maior que 1e-5, principalmente maiores que 0.001 ou 1e-3, são insignificantes estatisticamente para indicar uma possível homologia. Isto não significa necessariamente que exista homologia, pois podemos estar utilizando uma seqüência parcial, de baixa qualidade ou ainda com um gene ainda não descrito, mas significa que estatisticamente não existe "boa" similaridade com os genes comparados.

5.4. All Reads

(5. Novos reads | início)

Busca com base apenas em valores de filtro. Esta ferramenta apresenta todos os reads, dentro dos parâmetros do filtro, com os respectivos resultados de blast. Pode ser utilizada para um panorama geral do projeto, para buscas de genes (numa ordenação por resultado de blast, por exemplo) etc.

início

5.5. Pattern search

(5. Novos reads | início)

Busca de padrões nas seqüências. Interessante para busca de domínios ou trechos específicos, como primers. Existem alguns recursos adicionais para esta busca.

- [AT] – letra A ou letra T;
- (3) – três repetições da letra anterior;

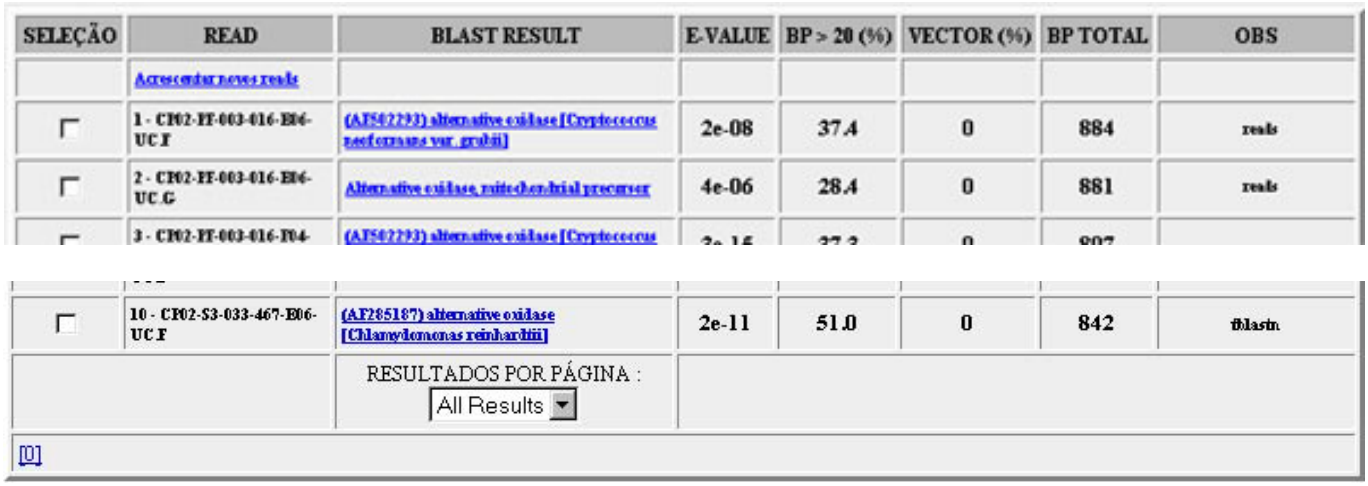
- (1,4) – uma a quatro repetições da letra anterior;
- {G} – qualquer letra menos G;
- N – qualquer letra;
  Ex.: A(3)N(10,15)[G,C] – 3 As, de 10 a 15 letras quaisquer, 1 G ou 1 C.

início

6. Projeto

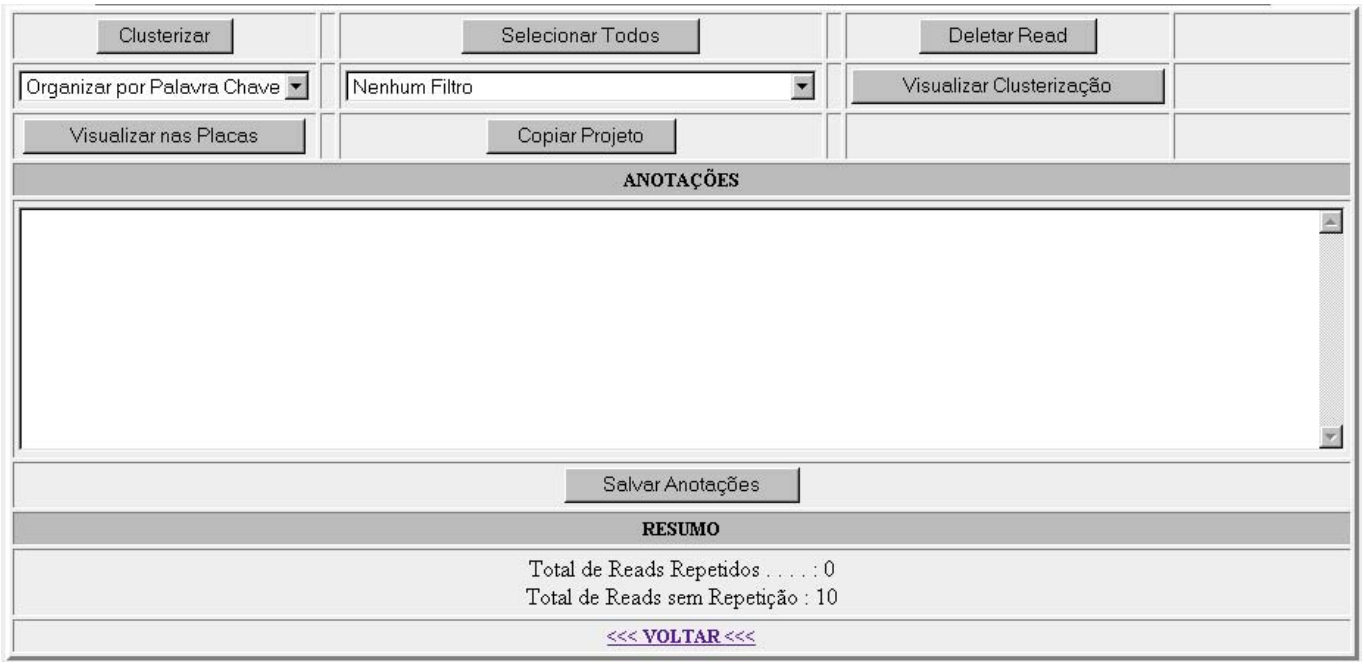
início

Agora que já foram acrescentados os reads de interesse, voltamos à página do projeto para os próximos passos. Esta página apresenta dois quadros. O primeiro contém informações sobre os reads selecionados para o projeto.



SELEÇÃO	READ	BLAST RESULT	E-VALUE	BP > 20 (%)	VECTOR (%)	BP TOTAL	OBS
	Acrescentar novos reads						
☐	1 - CP02-FF-003-016-E06-UC.F	(AF502793) alternative oxidase [Cryptococcus neoformans var. grubii]	2e-08	37.4	0	884	reads
☐	2 - CP02-FF-003-016-E06-UC.G	Alternative oxidase, mitochondrial precursor	4e-06	28.4	0	881	reads
☐	3 - CP02-FF-003-016-F04-	(AF502793) alternative oxidase [Cryptococcus	3e-15	37.3	0	807	
☐	10 - CP02-S3-033-467-E06-UC.F	(AF285187) alternative oxidase [Chlamydomonas reinhardtii]	2e-11	51.0	0	842	tblastn
		RESULTADOS POR PÁGINA : All Results					
[0]							

A coluna ‘SELEÇÃO’ permite selecionar os reads que sofrerão a ação de algum botão. ‘READS’ mostra os nomes de cada read selecinado. ‘BLAST RESULT’ contém o nome do primeiro hit do blastx automático, com link para a análise do blast. ‘E-VALUE’ mostra o valor de E-value do melhor resultado do blast. ‘BP > 20 (%)’ contém a porcentagem de bases do read com qualidade phred acima de 20. ‘VECTOR’ mostra a porcentagem de vetores trimados do read. ‘BP TOTAL’ apresenta o tamanho total do read, independente de qualidade ou trimagem. Finalmente, o campo ‘OBS’ mostra uma palavra que pode ser utilizada para controle pessoal de forma/motivo de inserção de read no projeto. Por exemplo, um read pode ser selecionado por palavra chave, outro por blast, outro por padrão etc. O quadro de baixo apresenta possibilidades de ação a serem realizadas com os reads.



O botão 'CLUSTERIZAR' vai executar uma montagem (Phrap ou CAP3) com os reads selecionados. Não esqueça de selecionar os reads antes de clicar no botão 'CLUSTERIZAR'. Quando esta montagem estiver terminada, o programa enviará um e-mail avisando. Não adianta clicar mais de uma vez. Isto apenas iniciará novos processos, o que tornará a chegada de respostas mais lenta. Após a chegada do e-mail, a montagem poderá ser visualizada por intermédio do botão 'VISUALIZAR CLUSTERIZAÇÃO'. Detalharemos os resultados da montagem no item "Clusterização".

'SELECIONAR TODOS' seleciona todos os reads do quadro 1. 'DELETAR READ' irá retirar do projeto os reads selecionados. A caixa abaixo do botão 'CLUSTERIZAR', que mostra 'ORGANIZAR POR PALAVRA CHAVE' possui algumas opções de organização dos reads, para facilitar sua seleção. A outra caixa de seleção, que mostra 'NENHUM FILTRO' apresenta opções de filtro também para facilitar a seleção de reads. 'COPIAR PROJETO' é um botão que permite seja realizada uma cópia do projeto, para o próprio usuário ou outro.

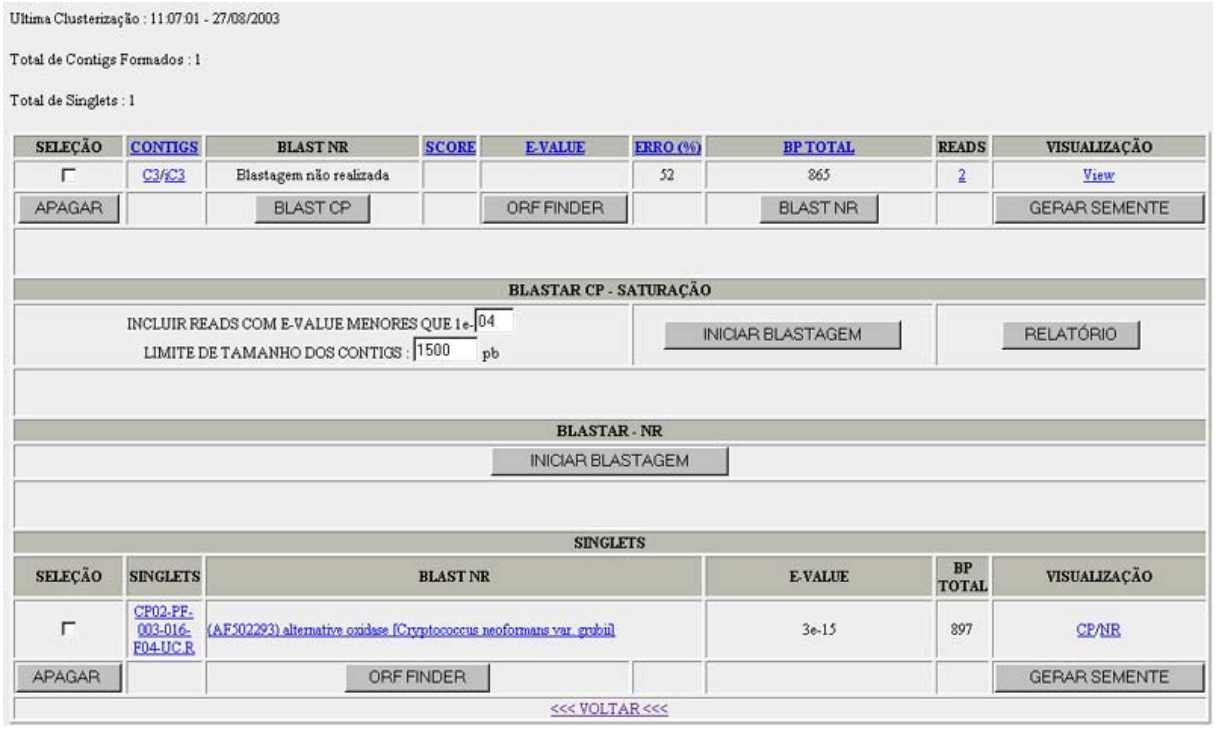
'VISUALIZAR NAS PLACAS' mostra as posições dos reads nas placas e é utilizado para facilitar a localização rápida de reads. O campo 'ANOTAÇÕES' permite que sejam copiados ou escritos textos referentes ao projeto. Sempre que for modificado, deve-se clicar em 'SALVAR ANOTAÇÕES'. No fim da página estão os dados sobre repetição de reads e o link '<<<<VOLTAR<<<<'. Estes links devem sempre ser utilizado para movimentação entre as páginas do projeto, ao invés dos botões do navegador, que o programa pode não reconhecer, gerando erros.

início

7. Clusterização

início

Após o término da clusterização clica-se no botão 'VISUALIZAR CLUSTERIZAÇÃO', para entrada em uma tela como a da figura abaixo:



Ultima Clusterização : 11:07:01 - 27/08/2003

Total de Contigs Formados : 1

Total de Singlets : 1

SELEÇÃO	CONTIGS	BLAST NR	SCORE	E-VALUE	ERRO (%)	BP TOTAL	READS	VISUALIZAÇÃO
☐	C3/iC3	Blastagem não realizada			52	865	2	View
APAGAR		BLAST CP		ORF FINDER		BLAST NR		GERAR SEMENTE

BLASTAR CP - SATURAÇÃO

INCLUIR READS COM E-VALUE MENORES QUE 1e-04

LIMITE DE TAMANHO DOS CONTIGS : 1500 pb

INICIAR BLASTAGEM | RELATÓRIO

BLASTAR - NR

INICIAR BLASTAGEM

SINGLETS

SELEÇÃO	SINGLETS	BLAST NR	E-VALUE	BP TOTAL	VISUALIZAÇÃO
☐	CP02-PF-003-016-F04-UC.R	(AF502293) alternative oxidase [Cryptococcus neoformans var. grubii]	3e-15	897	CP/NR
APAGAR		ORF FINDER			GERAR SEMENTE

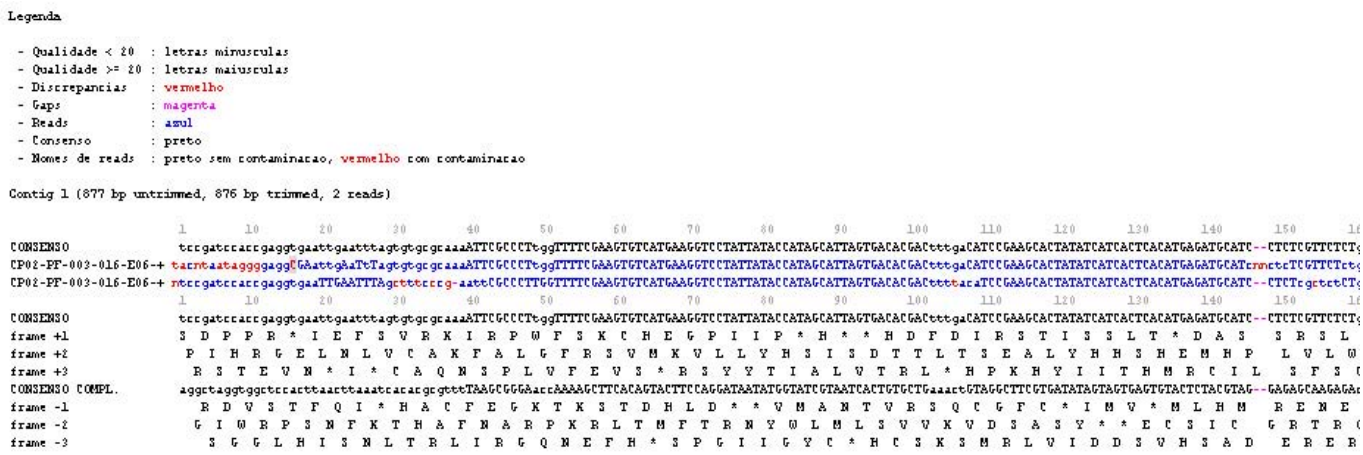
<<< VOLTAR <<<

A parte superior da tabela apresenta os contigs formados e algumas características destes (um contig por linha). Na coluna 'CONTIGS' são apresentados o nome do contig e dois links para a seqüência deste em formato fasta: 'C3' leva à seqüência "normal" do contig 3, e 'iC3' leva ao complemento reverso desta. A coluna 'BLAST NR' mostra os resultados de blastx dos contigs. Estes dados aparecerão depois que o botão 'INICIAR BLASTAGEM' for acionado, e o término do processo é avisado por e-mail. SE nenhum read for selecionado, este botão realiza o blast de todos os contigs do projeto. Caso contrário "blasta" os selecionados.

As colunas 'SCORE' e 'E-VALUE' apresentam os respectivos valores para o primeiro resultado do 'BLAST NR' de cada contig. O valor da coluna 'ERRO' mostra o número de prováveis erros da montagem a cada 10.000 bases. No exemplo, seriam 52 prováveis erros a cada 10.000 bases. Como o contig tem 865 bases, são prováveis 4,5 erros no contig. 'BP TOTAL' mostra o tamanho do

contig montado, e ‘READS’ o número de reads que fazem parte deste contig. Neste número existe um link que abre uma janela com a lista dos reads.

A última coluna, ‘VISUALIZAÇÃO’, traz um link (‘view’) que mostra a montagem do contig, como no exemplo da figura abaixo:



As legendas explicam os detalhes, mas é importante destacar que bases em desacordo com o consenso aparecem em vermelho, e as letras maiúsculas indicam qualidade phred acima de 20. Abaixo da montagem, são mostrados os 6 quadros de leitura.

Abaixo dos contigs existe uma linha com diversos botões:

Todas as ações destes botões serão executadas no contig selecionado acima destas. ‘APAGAR’ apaga o contig da montagem; ‘BLAST CP’ é o blast do contig contra o organismo do projeto, no caso, Crinipellis perniciosa; ‘ORF FINDER’ executa este programa para localização de ORFs no contig; ‘BLAST NR’ permite blast contra BD NR; e o botão ‘GERAR SEMENTE’ será detalhado no item ‘Semente’.

Uma ferramenta muito interessante e útil é o Blast saturação:

Após escolhido o contig de interesse, para verificar se existem outros reads que possam melhorar a qualidade ou estender o contig, o procedimento normal é blastar o contig contra os reads, verificar se existem reads novos, acrescentar estes reads novos à montagem, remontar (clusterizar), verificar novamente com o blast se com o novo tamanho existem outros reads que possam aumentar o seu contig, e assim ciclicamente, até atingir o tamanho pretendido ou não conseguir mais novidades.

Sim, isto dá muito trabalho, e por isso foi desenvolvido o 'Blast saturação', que faz isto tudo sozinho. Basta definir o tamanho pretendido, e o limite de similaridade do blast e deixar o programa trabalhar por você, clicando em 'INICIAR BLASTAGEM'.

Quando o processo terminar, o programa envia um e-mail avisando. Se preferir acompanhar o progresso, basta clicar em 'RELATÓRIO' e ir atualizando a página. São mostradas todas as etapas com detalhes até a finalização do processo.

No final da página existe uma tabela semelhante à primeira, mas com singlets ao invés de contigs.

'SINGLETS' mostra os nomes dos reads que não entraram nos contigs, com link para a seqüência em formato fasta do mesmo. A coluna 'BLAST NR' mostra o melhor resultado de blastx (contra o BD nr) executado durante o processo de submissão, com o E value correspondente a este alinhamento na coluna 'E-VALUE'. Tamanho do read em 'BP TOTAL'. A coluna 'VISUALIZAÇÃO' mostra dois links: 'CP/NR'. 'CP' (ou o código do organismo do projeto) mostra o resultado de blastn do read contra o BD de reads do projeto. 'NR' mostra o resultado de blastx contra o BD nr.

Os botões 'APAGAR', 'ORF FINDER' e 'GERAR SEMENTE' funcionam como explicado para os contigs, para o read selecionado.

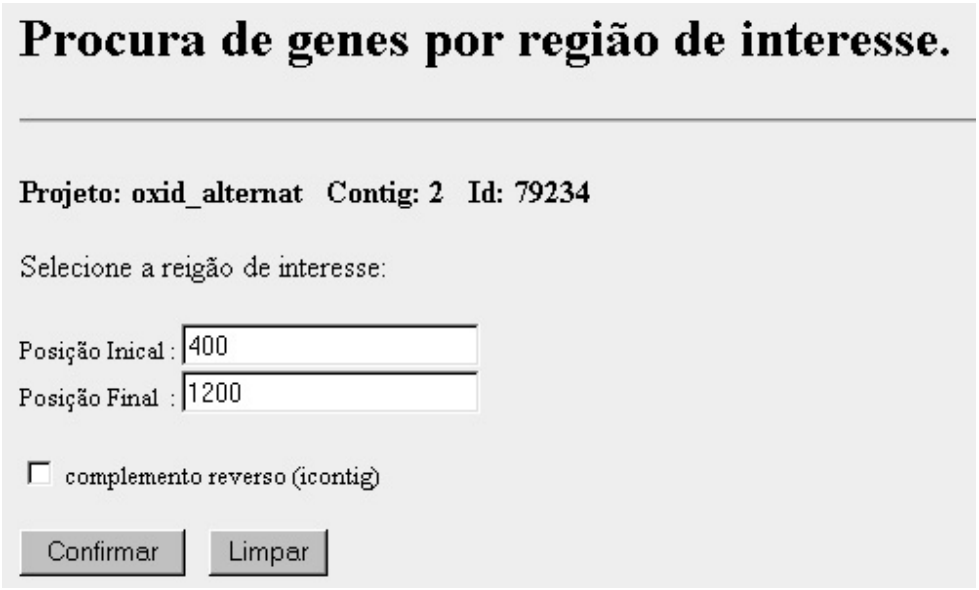
início

8. Semente

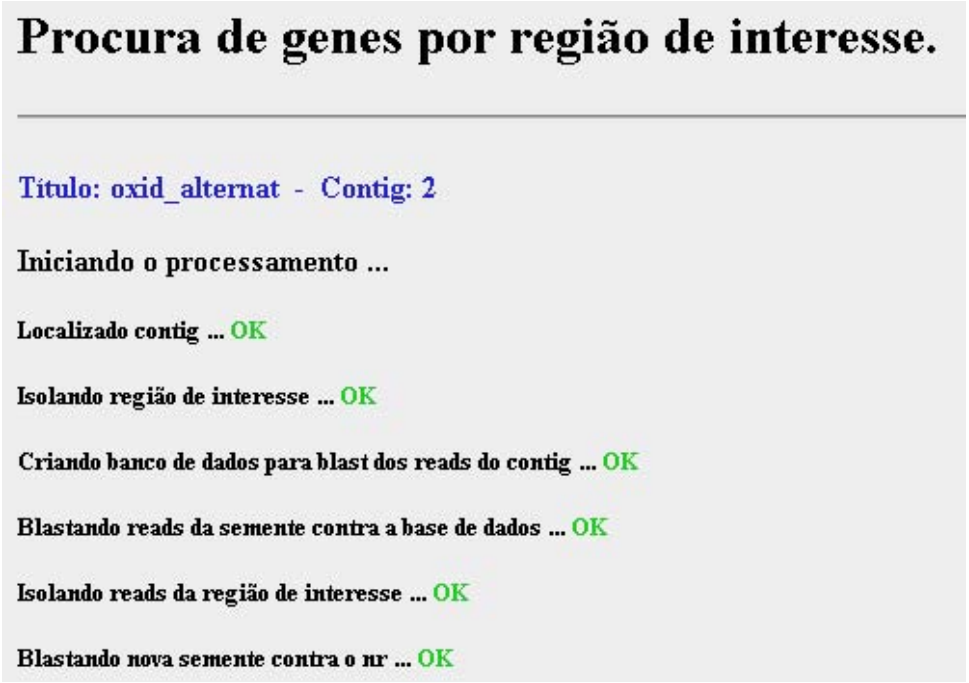
início

A criação de semente é o processo que precede a anotação. Depois do processo de clusterização, para localizar regiões de interesse, é preciso preparar estas regiões para a anotação. A semente é o trecho de DNA que será anotado. Pode conter um gene inteiro, parcial, com ou sem região promotora.

Depois de selecionado o contig de interesse, clica-se no botão 'GERAR SEMENTE'. A primeira tela será como a da figura abaixo:



Utilizando blast, ORF Finder ou outras ferramentas, pode-se determinar que região do contig contém o gene de interesse. A partir desta região, será iniciada a geração da semente. Caso o gene esteja invertido em relação á montagem do contig, e se for de seu interesse, basta clicar em 'COMPLEMENTO REVERSO (ICONTIG)' para capturar a seqüência complementar.



Procura de genes por região de interesse.

Título: oxid_alternat - Contig: 2

Iniciando o processamento ...

Localizado contig ... OK

Isolando região de interesse ... OK

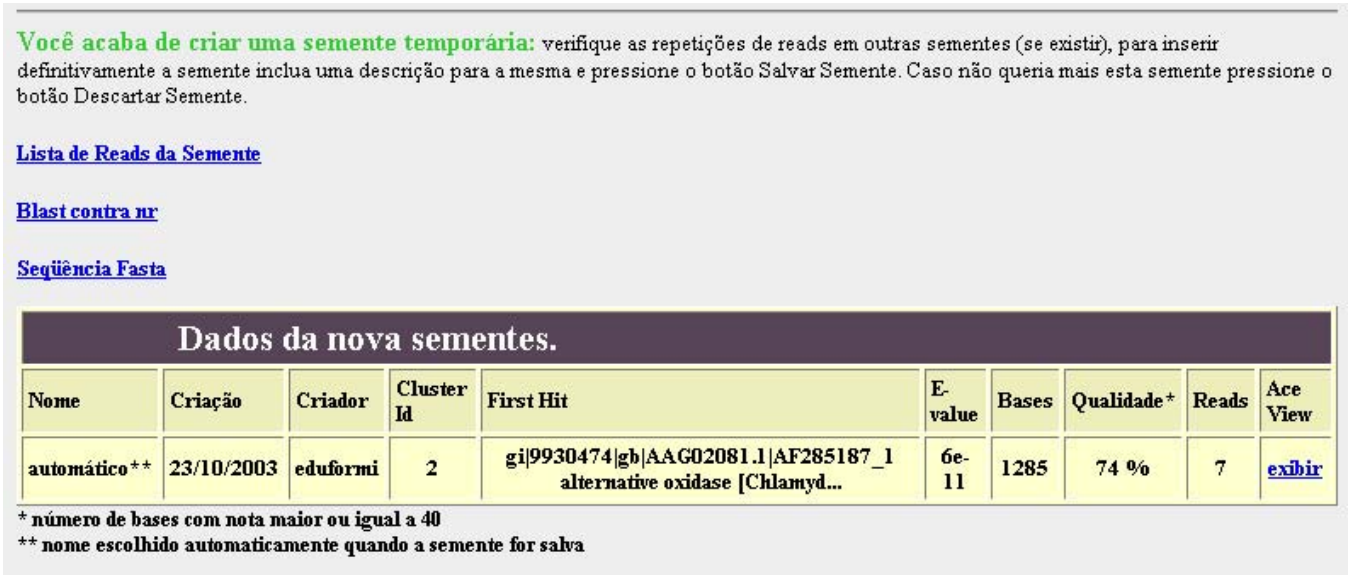
Criando banco de dados para blast dos reads do contig ... OK

Blastando reads da semente contra a base de dados ... OK

Isolando reads da região de interesse ... OK

Blastando nova semente contra o nr ... OK

O programa irá executar alguns itens, gerar uma semente temporária e uma tabela de informações:



Você acaba de criar uma semente temporária: verifique as repetições de reads em outras sementes (se existir), para inserir definitivamente a semente inclua uma descrição para a mesma e pressione o botão Salvar Semente. Caso não queria mais esta semente pressione o botão Descartar Semente.

Lista de Reads da Semente

Blast contra nr

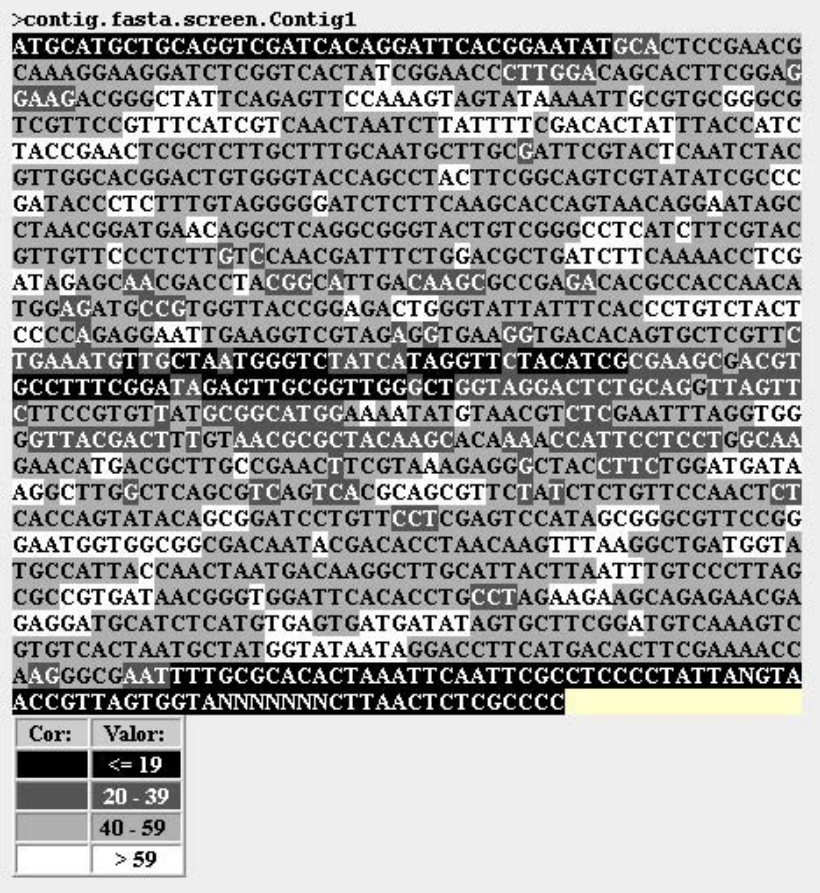
Seqüência Fasta

Dados da nova sementes.

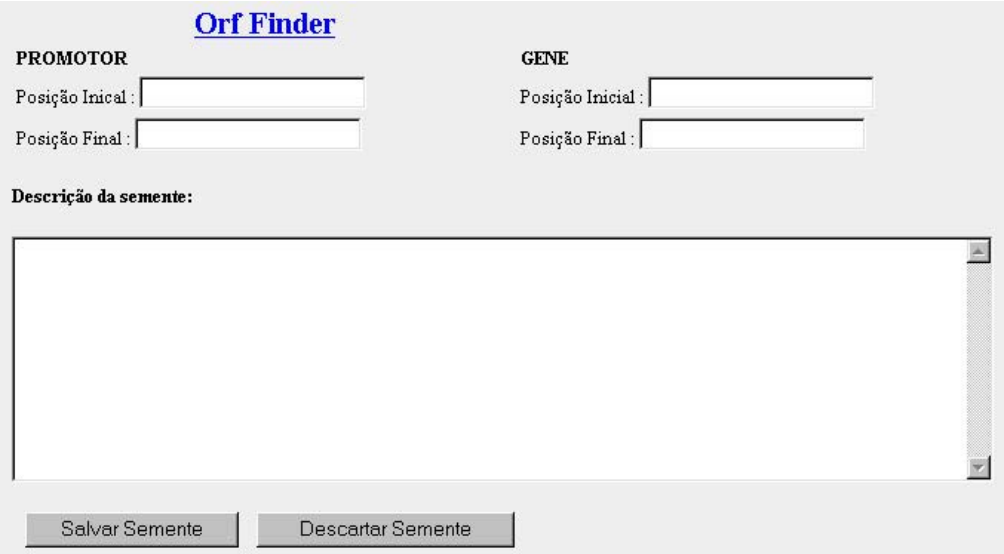
Nome	Criação	Criador	Cluster Id	First Hit	E-value	Bases	Qualidade*	Reads	Ace View
automático**	23/10/2003	eduformi	2	gi|9930474|gb|AAG02081.1|AF285187_1 alternative oxidase [Chlamyd...	6e-11	1285	74 %	7	exibir

* número de bases com nota maior ou igual a 40
** nome escolhido automaticamente quando a semente for salva

Além da tabela com as informações da semente, são apresentados links para conferir as informações referentes a esta. A lista de reads que compõem a semente, o resultado de blastx contra BD nr e a seqüência fasta com valores de qualidade, como na figura abaixo.



O final da página vai apresentar as informações finais para cadastro ou descarte da semente. Caso não exista sobreposição (regiões já salvas como sementes serão acusadas), a semente pode ser salva. Basta preencher as informações de região de promotor e gene, e a descrição da semente. O link 'ORF FINDER' permite que sejam conferidas as posições.



Após clicado o botão 'SALVAR SEMENTE', é aberta uma página com a confirmação e um link para a 'LISTA DE SEMENTES'.

Para acessar esta lista, é necessário que seu login tenha permissão de anotador. O sistema de anotação varia com cada projeto, e o manual de anotação será desenvolvido em separado.

Obrigado por ter lido o manual e ter dado sentido ao nosso trabalho ☺

início

9. Dicas de utilização

início

As dicas foram distribuídas nas explicações de cada item, mas é importante citar que a forma mais simples de começar um projeto sobre determinado tema é realizar uma busca por palavra chave (keyword search) ou um tBLASTn da proteína pretendida (de preferência de organismo próximo) e acrescentar os reads no projeto. Outra opção é a busca nos resultados de blast, ou utilizar uma combinação dos métodos apresentados.

Depois que conseguir, da forma que preferir, acrescentar alguns reads no projeto, execute o blast saturação colocando como limite o tamanho do gene ou um pouco mais. Se ainda não conseguiu o tamanho completo do gene, tente aumentar o tamanho pretendido e rode novamente o blast saturação.

início

10. Sobre o programa

início

O principal objetivo do Gene Projects é permitir a "*on going annotation*", ou anotação enquanto o genoma está sendo seqüenciado. O sistema está sendo complementado para englobar as principais ferramentas do Laboratório de Genômica e Expressão (LGE/UNICAMP), objetivando a automatização total na criação de um sistema de bioinformática para projetos genoma, desde os pequenos casos, de pesquisas individuais, até programas de seqüenciamento completo de organismos.

Foi idealizado pelo Prof. Dr. Gonçalo A. G. Pereira, e o principal responsável pelo seu desenvolvimento é Marcelo F. Carazzolle. Time de desenvolvimento: Luciano A. Digiampietri, Marcos R. R. Araújo e Eduardo F. Formighieri. Os demais integrantes da bioinformática também começaram ou irão começar a trabalhar neste sistema (Eliezer e Danilo).

Manual redigido por Eduardo F. Formighieri (eduformi@lge.ibi.unicamp.br).

Versão do Gene Projects: 1.3.0.0
Versão do Manual: 1.0.0.2

Sugestões, críticas e contato através do e-mail
suporte@lge.ibi.unicamp.br.

início

7.5. ANEXO E. Produção didática: Manual e roteiro de anotação para o Sistema de anotação do Laboratório de Genômica e Expressão – LGE/UNICAMP. Utilizado nos projetos Genoma Vassoura de Bruxa, Café, Genolyptus, Camarão, entre outros.

Manual Avançado de Anotação – Camarão

| Personagens e senhas | Dicas iniciais | Processo de anotação | Interface de anotação | Como anotar | Sobre o sistema de anotação LGE |

11. Definição de personagens e senhas

início

Curadores – No caso da anotação dos ESTs de camarão os curadores ainda não foram definidos, mas futuramente irão padronizar as informações anotadas.

Selecionadores – São os coordenadores de cada equipe de anotação, podendo haver mais de uma equipe por instituição. Eles terão o poder de selecionar inicialmente o material de interesse para a anotação (*sequences control*), que será distribuído entre os anotadores ligados a cada selecionador. Após a seleção dos clusters de interesse, os clusters não selecionados serão distribuídos. Cada selecionador será responsável por uma cota de seqüências para anotar, e pela qualidade da anotação do seu grupo de anotadores. O controle pode ser feito pela página de anotação específica de cada selecionador.

Anotadores – São os pesquisadores que pertencem a uma das instituições conveniadas, que farão a anotação das seqüências segundo as normas gerais do projeto e as específicas de cada equipe. Cada anotador possui seu login e acesso à página de anotação do selecionador responsável. Existem ferramentas para a troca (trocar) de senha e para o caso de esquecimento da mesma (enviar).

Senhas – Todos os personagens têm acesso ao sistema através de login e senha geral do coordenador de cada instituição. Para a anotação e o controle desta pelos selecionadores, cada pesquisador (anotador ou selecionador) terá um login próprio e único, com permissões de acordo com sua função.

início

12. Dicas iniciais

início

- Sempre utilize os links das próprias páginas de serviços, como o "VOLTAR". Se utilizar o botão do navegador, o programa pode se confundir e gerar erro;
- Reporte falhas ou sugestões. São sempre bem vindas.
  - Utilize para isto o e-mail: suporte@lge.ibi.unicamp.br.
- Dúvidas sobre o processo de anotação devem ser enviadas para:
  - eduformi@lge.ibi.unicamp.br.
- Utilize o roteiro de anotação.
- Para descrever uma seqüência de ações, convencionamos o seguinte significado para:
  - <> – apertar a tecla "Enter";

- <ctrl> + ação – executar ação enquanto mantém a tecla *ctrl* apertada. O mesmo vale para <alt> e <shift>;
- ação1, ação2, ação3 – seqüência de ações a ser executada;
- Por "ação" entenda-se apertar uma tecla ou clicar com o mouse, por ex.; Cópia de texto no wordpad para novo arquivo:
  "<ctrl> + T, <ctrl> + C, novo documento, <ctrl> + V"
  – são 4 ações separadas por vírgulas. (1) manter a tecla "ctrl" apertada e apertar a tecla "T" – selecionar tudo, (2) manter a tecla "ctrl" apertada e apertar a tecla "C" – copiar, (3) abrir um novo documento, (4) manter a tecla "ctrl" apertada e apertar a tecla "V" – colar.

início

13. Geral do processo de anotação

(Cadastro de login | Seleção de seqüências | Cotas | Formas de busca | Análise dos resultados)
início

13.1. Cadastro de login/senha

Serão criados logins e senhas para todos os selecionadores que serão enviados por e-mail individuais.

De posse destas informações, cada selecionador criará login para cada um de seus pesquisadores (anotadores) através do serviço específico (http://bioinfo05.ibi.unicamp.br/camarao/services/atuselecionador.php). A senha será criada automaticamente e enviada para o e-mail do anotador, que poderá alterá-la através de serviço específico.

Para cadastrar os anotadores basta preencher os campos 'NOME' (sempre completo), 'LOGIN' e 'E-MAIL'. Estes estão automaticamente relacionados ao selecionador que os cadastrou. Por questão de segurança é necessário que cada anotador cadastrado seja bloqueado e depois desbloqueado para que o mesmo tenha acesso aos serviços.

Na coluna de gerenciamento existem 4 opções:

 - Editar.

 - Bloquear anotador.

 - Desbloquear anotador.

 - Excluir anotador (abre nova figura de confirmação).

 - Confirma exclusão de anotador (clicar no lixo para confirmar).

13.2. Seleção de seqüências

13.2.1. Cotas

(Processo de Anotação | início)

Cada selecionador poderá manter sob seus cuidados cerca de 300 seqüências por vez, tendo um prazo determinado para finalizar a anotação destas (definido por comunicados da coordenação de anotação). Após o prazo, a contar da data de seleção, a seqüência não finalizada volta automaticamente a fazer parte do grupo das que podem ser escolhidas para anotação pelos selecionadores. O limite de 300 seqüências não significa limite por processo de seleção, e sim o máximo de seqüências que um selecionador pode manter ao mesmo tempo. Assim que uma seqüência é finalizada, outra pode ser selecionada.

O selecionador poderá também desistir de determinada seqüência retirando-a de seu grupo e assim permitir que outros pesquisadores possam ter acesso à mesma e ao mesmo tempo para que possa escolher outra seqüência para seu grupo de trabalho. Isto vale para trocas de seqüências entre grupos ou desistência de sequencias escolhidas de forma errada. Na fase final as seqüências restantes serão distribuídas, e o botão de devolução não deverá mais ser utilizado. Casos especiais podem ser discutidos diretamente com a coordenação de anotação.

13.2.2. Formas de busca

(Processo de Anotação | início)

Ao acessar a página de seleção de clusters, o selecionador terá diferentes opções para localização dos clusters de interesse: busca por palavras chave, busca por blast, busca por seqüências e ainda visualizar todos os clusters. Basta clicar no ⊞ para abrir o respectivo quadro de busca, e em ⊟ para fechá-lo. A análise dos resultados das buscas será discutida após a descrição de todas as formas de busca.

Busca por palavras – Pode ser feita em todos os campos citados ou apenas em alguns, sendo feita esta escolha através das caixas de seleção (☐ e ☑). Depois de digitada a palavra e definidos os campos em que será feita a busca, define-se o número de clusters a serem mostrados a cada página de resultados. Para executar a busca clicar no botão 'SEARCH'.

Busca por BLAST – Buscas através do programa NCBI-BLAST contra o Banco de Dados das seqüências de ESTs do projeto. É possível utilizar os programas blastn (nt *vs.* nt), tblastn (aa *vs.* nt trad) e tblastx (nt trad *vs.* nt trad). A seqüência *query*, em formato FASTA, pode ser colada no campo ou localizada através do botão de procura. Define-se o número de clusters a serem mostrados em cada página e para executar a busca clica-se em 'SEARCH'.

Busca por seqüências – Pode ser realizada por uma listagem com o nome dos reads ou contigs (onde podem ser utilizados caracteres curingas (* e/ou ?)), digitada na caixa ou localizada em arquivo, ou através de seleção diretamente nas caixas da

nomenclatura cadastrada. Para executar a busca, clicar em 'SEARCH' depois de definir o número de resultados por página.

Visualizar todas as seqüências – Mostra todas as seqüências, com diferentes possibilidades de ordenação. Mesma opção de definição de quantidade de resultados a serem visualizados por página.

13.2.3. Análise dos resultados da busca

(Processo de Anotação | início)

Todas as formas de busca trazem os resultados no mesmo padrão de página. Existem três cores possíveis para as informações de cada seqüência:

 – apenas anotação automática realizada;
 – reservado para anotação; e
 – anotação manual finalizada.

Esta página contém uma tabela com os seguintes campos:

Seleção – Campo para seleção de seqüências a anotar.
Seqüência – Nome da seqüência (contig ou read). O link no nome de cada seqüência abre a página de anotação desta.
Produto – Nome completo do produto do gene. Ex.: cytochrome oxidase, subunit 3.
Símbolo – Abreviação do nome do gene, o equivalente ao "gene name" do genbank. Ex.: cox3.
Primeiro hit BLAST – Descrição da seqüência com o melhor hit da comparação do BLAST. O link na descrição abre a página do resultado do blast.
E-value – Valor de expectativa. Equivalente a probabilidade de o alinhamento ocorrer por acaso. Quanto mais próximo de zero, maior a chance de o alinhamento ter significado biológico.
Função Molecular – Descrição dos termos GO da genealogia de função molecular. O link deste campo abre a página do referido termo GO na ferramenta computacional AMIGO.
Componente Celular – Descrição dos termos GO da genealogia de componente celular. O link deste campo abre a página do referido termo GO na ferramenta computacional AMIGO.
Processo Biológico – descrição dos termos GO da genealogia de processo biológico. O link deste campo abre a página do referido termo GO na ferramenta computacional AMIGO.
Selecionador – Login do selecionador responsável pela anotação da seqüência.
Anotador – Login do último anotador que salvou alguma modificação na seqüência. Todas as modificações feitas por qualquer anotador são registradas no histórico de anotação. Este controle existe para facilitar a determinação de quais

seqüências já começaram a ser anotadas, e a forma como a anotação será distribuída no grupo depende da preferência de trabalho de cada selecionador.

Última atualização – Data da ultima modificação salva na página de anotação.

Depois de escolhidas as seqüências, basta clicar no botão 'ENVIAR SEQÜÊNCIAS' para que as mesmas sejam cadastradas na página de anotação do selecionador, até o limite de 300 seqüências. Após o período definido para esta seleção, todos os slecionadores receberão clusters aleatórios para completar 300 seqüências.

início

14. Interface de anotação

(Cabeçalho | Navegação | GO | Categoria | Identificação | Visualização | Sinalização | BLAST | Buscas facilitadas | Anotações)

início

As informações sobre a anotação são detalhadas no roteiro de anotação. O objetivo deste manual é mais ensinar a utilizar a interface de anotação, e dar alguma base em assuntos essenciais.

14.1. Cabeçalho

O cabeçalho está em um frame separado, e estará sempre disponível para acesso. Nele se encontram links para outras páginas do projeto, dos quais destacamos:

Gene Projects – Acesso ao programa desenvolvido no LGE para análise aprofundada de clusters ou assunto de interesse. Ligado diretamente aos dados do projeto.

Serviços – Abre a página de serviços do Projeto camarão no LGE.

Suporte – Envio de e-mail para grupo de suporte do LGE.

Busca avançada – Busca nos bancos de dados de ESTs do camarão no LGE. Possui opções de seleção dos diferentes campos para delimitar e especificar a busca.

14.2. Navegação

(Interface de anotação | início)

A barra de navegação possui no centro o nome da seqüência, e nas extremidades dois links de cada lado.

O ">>" abre a página de anotação da próxima seqüência do selecionador que ainda não foi anotada. O ">>>" abre a próxima seqüência, independentemente de esta ter sido anotada ou não. "<<" e "<<<" funcionam com a mesma lógica para as seqüências anteriores, considerando ordenação alfabética.

Também possui os links para manual e roteiro de anotação.

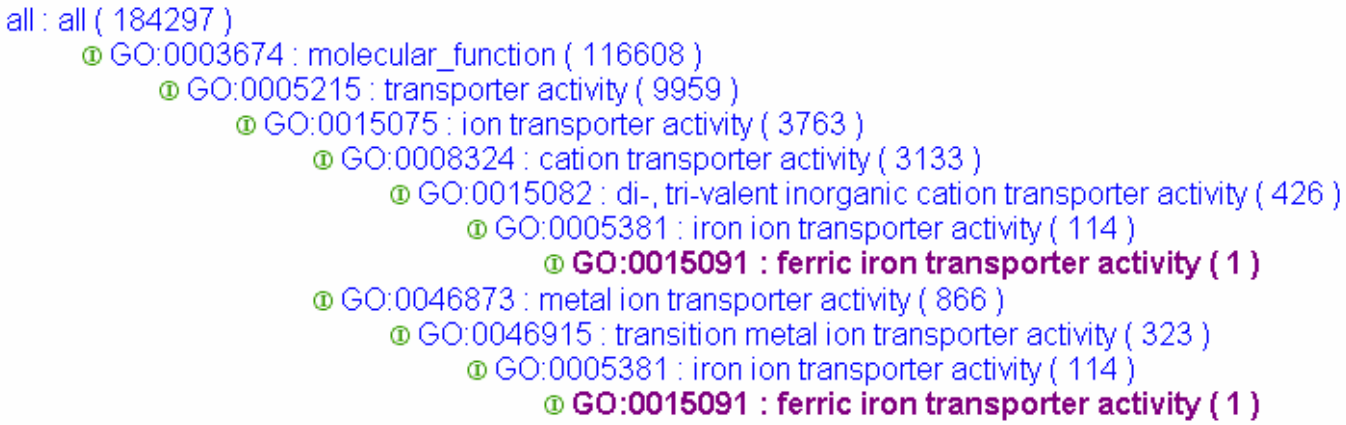
14.3. Classificação - Gene ontology

(Interface de anotação | início)

No campo 'CLASSIFICAÇÃO' voc pode classificar cada cluster segundo sistema do Gene Ontology (GO).

Mais baixo na interface de anotação existe um resultado automático de blast contra BD de seqüências associadas ao GO, que será discutido posteriormente. Este resultado pode ajudar na classificação quanto ao GO. Para uma análise atualizada existe mais abaixo, em 'BUSCAS FACILITADAS', a opção de busca por palavra chave ou por blast (GO) no GO.

Depois de determinados os termos GO relacionados à seqüência, preenche-se os dados do GO ID e do nome do termo nos campos apropriados. Utilize preferencialmente copiar/colar para evitar erros de digitação. Em caso de dúvida em relação aos termos mais específicos, seleciona-se um termo mais geral, ou seja, mais próximo à raiz da árvore. Exemplo:



Neste caso, a ontologia de função molecular está aberta até o último nível. Se não se tem certeza sobre o íon transportado, pode-se anotar uma das categorias acima, como 5381, 15082 ou 8324. Quanto mais próximo da raiz da árvore, menos específica a classificação, mas em casos de dúvida o melhor é classificar no nível em que se pode ter maior grau de certeza.

14.4. Identificação

(Interface de anotação | início)

Nome do gene – Descrição por extenso do produto do gene, incluindo informação sobre ser subunidade e qual. Ex.: cytochrome c oxidase, subunit 3.

Função – Inserir informações sobre a função do produto do gene. Pode-se simplesmente colar a informação do UNIPROT ou do NCBI-Gene homólogo.

Domínio – Descrição do(s) domínio(s) encontrado(s) na seqüência. Utilizar nomenclatura completa, incluindo identificador (o padrão pode ser o do PFAM). Em caso de similaridade parcial, indicar no final da descrição do domínio. Ex.: "COG0071, IbpA, Molecular chaperone (small heat shock protein), incomplete."

Organismo homólogo – Nome científico completo do organismo referente ao melhor resultado do blastx vs. nr.

Símbolo do gene – Nome curto do gene, o equivalente ao termo "gene name" das entradas do genbank. Ex.: cox3.

Número EC – Número de classificação enzimática.

Número TC – Número de classificação de proteínas de transporte.

14.5. Visualização

(Interface de anotação | início)

iContig – Complemento reverso da seqüência em formato fasta.

Contig – Seqüência em formato fasta.

Reads – Lista de reads que compõem a seqüência.

View – Visualização da montagem.

14.6. Sinalização

(Interface de anotação | início)

Seq. Cod. Parc. – Deve ser marcado quando a região do alinhamento não inclui todo o gene similar, potencialmente homólogo. Diz respeito à região seqüenciada

Intron – indicador da presença de intron.

Full Length – indicador da presença de toda região codante no clone. Deve ser marcado quando a região do alinhamento possui a região codante do início do gene similar.

CP genômico – Indicador da existência de seqüência similar no conjunto de DNA genômico seqüenciado.

Top gene – Indicador de que o gene é importante de acordo com os objetivos do projeto e deve ser estudado com cuidado, e/ou reservado para futuras patentes.

Cluster problem – Indicador de problemas na seqüência, como frameshift ou consenso com regiões de qualidade muito baixa.

Contaminação – Indicador de contaminante, a ser eliminado do conjunto de dados a serem anotados no projeto.

14.7. BLAST automático

(Interface de anotação | início)

14.7.1. BLAST vs. nr

Informações retiradas automaticamente do resultado de blastx contra nr.

Score – Pontuação do primeiro alinhamento do blastx.

E-value – Valor de expectativa do primeiro alinhamento do blastx.

Identidade – Valor em porcentagem do número de acertos (*matchs*) em relação ao tamanho total do alinhamento.

Coverage query – Porcentagem da divisão do tamanho do 'query' em relação ao tamanho do alinhamento.

Coverage subject – Porcentagem da divisão do tamanho do 'subject' em relação ao tamanho do alinhamento.

Organismo – Nome científico do organismo do primeiro alinhamento do blastx.

Acession code – Código de entrada no banco de dados genéticos da seqüência do primeiro alinhamento do blastx.

Descrição – Campo com o nome inteiro da entrada do primeiro alinhamento do blastx, que também é um link para o resultado do blastx.

14.7.2. BLAST vs. GO

Informações retiradas automaticamente do resultado de blastx contra banco de dados com seqüências relacionadas a termos GO.

Score – Pontuação do primeiro alinhamento do blastx.

E-value – Valor de expectativa do primeiro alinhamento do blastx.

Identidade – Valor em porcentagem do número de acertos (*matchs*) em relação ao tamanho total do alinhamento.

Coverage query – Porcentagem da divisão do tamanho do 'query' em relação ao tamanho do alinhamento.

Coverage subject – Porcentagem da divisão do tamanho do 'subject' em relação ao tamanho do alinhamento.

Símbolo do gene – Símbolo do gene.

Produto – Descrição do nome do produto do gene.

Descrição – Campo com o nome inteiro da entrada do primeiro alinhamento do blastx. Também é um link para o resultado do blastx contra o BD relacionado aos termos GO.

Molecular_function, Cellular_component e Biological_process – descrição dos termos GO com similaridade. O link de cada termo abre a janela de descrição do termo no programa AMIGO.

14.7.3. Outros Blasts

Links para resultados de blasts estáticos: Swiss prot, NT e TC.

14.8. Buscas facilitadas

(Interface de anotação | início)

14.8.1. Buscas atualizadas por BLAST e ORFFinder com nucleotídeos

Links que abrem páginas de BLAST contra bancos de dados específicos na tela de seleção de parâmetros, já com a seqüência fasta específica preenchida. Verificar os parâmetros em cada

caso e executar a busca atualizada. O link para o programa OrfFinder permite a localização de ORFs. Cuidado ao interpretar ORFs em ESTs...

BD GO – Seqüências do Gene Ontology.
BD Kegg – Seqüências do KEGG.
BD Nr – Non redundant, do NCBI.
ORF Finder – busca de ORFs para definição de início e/ou final da região codante quando for do interesse do anotador.
BD TC – Seqüências relacionadas a classificação de proteínas de transporte.

14.8.2. Buscas por CD_Search com aminoácidos

Escolha o frame que deseja e execute a ferramenta.

CD_Search – Busca por domínios.
Fasta aas – Seqüência fasta traduzida no frame escolhido.

14.8.3. Buscas por palavra chave

Buscas por palavra chave em diferentes locais. Seleciona-se o local onde se quer fazer a busca da palavra digitada e clica-se no botão 'CONSULTAR'. Descrição concisa dos bancos de dados:

Enzyme – Conjunto de informações relativas à nomenclatura de enzinas.
GO – Vocabulário de produtos de genes dinâmico controlado com potencial de ser aplicado a todos os organismos.
Pfam – Coleção de alinhamentos múltiplos de seqüências e de HMMs cobrindo os mais comuns domínios de proteínas.
Pubmed – Serviço da NLM incluindo mais de 15 milhões de citações para artigos relacionados à ciência da vida.
Swissprot – Banco de dados de seqüências de proteínas curado - alto nível de anotação, mínimo de redundância e alto de integração com outros BDs.
Famílias de transporte (TC) – Sistema de classificação para proteínas de transporte de membrana conhecido como sistema TC.
SGD – Saccharomyces Genome Database

14.9. Bloco de notas

(Interface de anotação | início)

Notepad – Notas do próprio anotador sobre a seqüência anotada. São atualizadas, como toda a página, pelo botão 'ATUALIZAR'. Pode ser utilizado para inserir informações importantes sobre a anotação que não possam ser colocadas em outros campos.
Guest notepad – Notas de qualquer anotador, sobre a anotação da seqüência. Estas informações são salvas através do botão 'GUEST UPDATE'.

14.10. Controle

(Interface de anotação | início)

Anotação terminada – Selecionar após o término da anotação manual.

Histórico da anotação – O botão 'VIEW' mostra todo o histórico de modificações da anotação.

Reserva – Selecionar para reservar a seqüência para trabalho funcional.

Atualizar – O botão 'ATUALIZAR' aparece somente para o grupo de anotadores do selecionador. Deve ser clicado para salvar as alterações feitas na página. O recomendado é que as alterações sejam salvas logo após serem feitas.

Devolução – retira o cluster do grupo do selecionador. Só deve ser utilizado pelo selecionador, em fase de seleção de clusters.

Revisado – para selecionador clicar quando revisar a anotação de sua equipe.

15. Como anotar

(O que é anotar | Identificar proteína | Classificar | Detalhes | Anotação metabólica)

início

15.1. O que é anotar uma seqüência?

O roteiro de anotação se encontra em arquivo separado, mas serão tratados aqui os conceitos básicos e alguns tópicos com mais detalhes.

Para começar, o que é esta seqüência? Cada uma das seqüências a serem anotadas é resultado da montagem de todos os ESTs seqüenciados no Projeto Genoma Camarão, com diferentes bibliotecas de ESTs (as estratégias presentes no projeto podem ser vistas na página de serviços do projeto, em 'SUBMISSÃO' / 'NOMENCLATURA') após a trimagem de contaminantes, vetores etc.

Portanto cada seqüência a ser anotada é um trecho de cDNA seqüenciado. Como passou por um processo de montagem, pode ser um read (singlet) ou um contig (grupo de reads alinhados formando um trecho normalmente maior e com melhor qualidade do que um read sozinho).

Por que anotá-lo? Além das metas de anotação, é claro... A anotação é a fase mais importante de um projeto genoma. É através da anotação que encontraremos as proteínas expressas em diferentes situações, poderemos inferir sobre vias metabólicas presentes, alvos específicos para pesquisas funcionais e futuras patentes e produtos. Enfim, uma boa anotação facilita a continuidade da pesquisa, a redação de artigos e a continuidade da genômica no Brasil ☺.

Mas então o que é anotar? Anotar é determinar a função do produto de um gene. Como trabalhamos neste projeto com ESTs, tudo o que temos foi expresso (não precisamos nos preocupar em determinar ORFs). A missão é identificar a função de cada uma. Mas como?

Para uma anotação bem feita precisaremos passar por quatro fases principais: a identificação da proteína, sua classificação, o detalhamento da anotação, e a anotação metabólica. Este processo de anotação será detalhado no curso de anotação, mas trataremos brevemente os principais pontos da anotação a seguir.

15.2. Identificar a proteína

(Como anotar | início)

Esta é a parte mais importante da anotação. Mais ou menos trabalhosa dependendo do caso, mas essencial para que as próximas fases sejam úteis. Ou seja, não adianta classificar perfeitamente o gene errado. Esta é a principal fase onde a anotação automática necessita da confirmação humana.

Para identificar uma seqüência o primeiro passo é fazer uma comparação desta seqüência com BDs de seqüências já anotadas. A melhor ferramenta para uma primeira análise é o BLASTX contra o Swissprot e contra o nr (BD *non redundant* de proteínas do NCBI). Para quem não conhece o BLAST, o 'quadro 1' traz uma descrição breve de suas principais ferramentas.

Quadro 1 – Introdução ao NCBI-BLAST

Basic Local Alignment Search Tool (BLAST) - Programas de busca de similaridade projetados para explorar todas as bases de dados disponíveis de seqüências de proteínas ou DNA. A seqüência a ser utilizada na busca deve

sempre estar no formato fasta (podendo estar sem a linha de título).

O formato FASTA inclui um título e uma seqüência. O título é sempre precedido por '>' e finalizado por uma quebra de linha (ou seja, um <Enter> no final da linha). A seqüência deve ter a linha quebrada após cerca de 50 bases (este valor pode variar, mas não deve ser muito grande). Por ex.:

```
>nome do gene
cagctagcgttatactgtgcagtcgatcgatctattccgttgcgatctat
atatatatatctgcgtcgagtcgcgtagtgatatatatcgagcgcgatat
atatcgcgatctacgatgctgatcagtgtacgatcgatgctagcagatcg
atatatatatctgcgtcgagtcgcgtagtgatatatatcgagcgcgatat
atatcgcgatctacgatgctgatcagtgtacgatcgatgctagcagatcg
atcgactcgatgatcgatcgacgctgactagctgactacgctagctagct
atcgactcgatgatcgatcgacgctgactagctgactacgctagctagct
atatcgcgatctacgatgctgatcagtgtacgatcgatgctagcagatcg
cgtcgagtcgcgtagtgatatat
```

A seqüência fasta que você inserir receberá o nome "query", e cada seqüência que alinhar com ela e aparecer no resultado da busca será chamada de "subject". No site do NCBI existe uma excelente documentação, mas daremos uma breve explicação sobre os principais programas BLAST:

Query	BD	Compara	Programa
nt	nt	nt	Blastn
nt (traduzido)	aa	aa	Blastx
aa	aa	aa	Blastp
aa	nt (traduzido)	aa	Tblastn
nt (traduzido)	nt (traduzido)	aa	Tbaslx

Blastn – compara nucleotídeos contra BD de nucleotídeos;

Blastx – compara seis fases de leitura da seqüência de nucleotídeos (query) contra BD de proteínas;

Blastp – compara aminoácidos contra BD de proteínas;

Tblastn – compara seqüência de aminoácidos contra BD de nucleotídeos, traduzido dinamicamente nas seis fases de leitura;

Tblastx – compara nucleotídeos traduzidos nas seis fases contra BD de nucleotídeos também traduzido nas seis fases.

rpretação dos resultados de Blast é preciso levar em consideração alguns fatores. Comentaremos os principais: E-value, *identities, positives* e *score*.

Identities, ou identidade é a relação entre o número de *matchs*, ou letras similares, com o tamanho do alinhamento. *Positives* traz a porcentagem de matchs de semelhança química. Além dos matchs iguais, os sinais + representam aminoácidos diferentes, mas semelhantes. *Score* é a pontuação do programa para o alinhamento. Porém, estes valores não podem ser analisados individualmente. É preciso utilizar como parâmetro principal de seleção o *E-value*, que é o valor estatístico da probabilidade de o alinhamento acontecer por acaso. Quanto menor este valor, menor a chance de que o alinhamento tenha acontecido por acaso, e portanto, maior a chance de que este mostre uma similaridade real, com significado biológico.

Existem três formatos possíveis para o valor de E-value: 0.0; 3e-35; e 0.14. O formato 3e-35 significa 3 x 10^{-35}, e portanto é muito mais próximo de 0.0 (zero) do que 0.14. O quanto o alinhamento pode ser significativo depende também de outras informações, mas normalmente são utilizados valores de E-value menores que 1e-5 (por exemplo: 1e-6, 1e-7, 1e-8 etc.). Valores de E-value maiores que 1e-5, principalmente maiores que 0.001 ou 1e-3, podemos considerar como pouco significativos estatisticamente para indicar uma possível homologia. Isto não significa necessariamente que não exista homologia, pois podemos estar utilizando uma seqüência parcial, de baixa qualidade ou ainda com um gene ainda não descrito, mas significa que estatisticamente não existe "boa" similaridade com os genes comparados. É preciso analisar diversos fatores para inferir a homologia, e a restringência escolhida depende do anotador.

Filtro de região de baixa complexidade (*filter – low complexity*). Utilizado para evitar alinhamentos em regiões de baixa complexidade, como caudas poliA, o *default* é que esteja acionado. Para uma análise mais cuidadosa, desmarcar esta opção pode trazer informações mais completas sobre o alinhamento.

O resultado do blast pode ser um ótimo indicador da identidade do gene, mas não basta olhar o primeiro hit. É preciso analisar os principais resultados que tenham bons E-values e verificar se existe concordância nestes. Em alguns casos é preciso verificar se os diferentes nomes encontrados não se referem na verdade ao mesmo gene.

Para confirmar a identidade, podemos utilizar outras ferramentas, como as buscas facilitadas. Por blast para a identificação, e por palavras para padronizar nomenclatura e definir alguns detalhes, como nome e símbolo corretos.

15.3. Classificar a proteína

(Como anotar | início)

As informações obtidas na identificação ajudarão na classificação do gene no sistema adotado pelo projeto.

15.3.1. Pelo GO

Para a classificação pelo GO o primeiro passo é verificar a anotação automática do GO, no 'Blast automático'. Mais do que olhar os resultados automáticos, é importante analisar o a tela de resultado do blast (botão 'BLAST GO'). Com base nas informações de identificação e do Blast GO, pode-se fazer a classificação buscando a melhor descrição possível. Um bom local para encontrar termos GO é a descrição do homólogo no uniprot (através do link do resultado do swissprot, verificar a calssificação do iProClass, no final da página).

15.4. Detalhes importantes

(Como anotar | início)

É importante que a anotação seja bem feita desde o começo, para evitar desperdício de tempo. Uma anotação mal feita pode gerar ou frustrar expectativas futuras, e mesmo que não o faça, sempre vai custar tempo na correção. Dois minutos gastos na confirmação de um número EC ou no nome correto do gene podem facilitar muito as análises futuras, da categoria de anotação e mesmo na identificação de genes importantes para que se atinja o objetivo do projeto inteiro.

Nomes oficiais podem ser obtidos através de Swissprot, Entrez – gene, Enzyme etc. Sempre que for definido o EC number é possível buscar o nome oficial no Enzyme. Para proteínas relacionadas a transporte, preencha o campo de número TC, caso a busca específica traga bons resultados. Sempre descreva detalhes e observações no campo 'NOTEPAD', e se for um visitante, no 'GUEST NOTEPAD'. Todas as interações com o sistema de anotação serão registradas.

Finalmente, não se esqueça da 'SINALIZAÇÃO'. Avise o sistema sobre importância ou problemas nos genes. É muito mais fácil detectar isto quando se está anotando o mesmo. Sempre atualize os dados depois de qualquer modificação, e finalize a anotação do gene depois que tiver preenchido todas as informações importantes que conseguir. Depois de algum tempo a anotação se torna mais intuitiva e rápida, mas é preciso fazer uma boa anotação desde o princípio para que se automatize o processo correto, e a anotação seja sempre bem feita.

15.5. Anotação metabólica

(Como anotar | início)

Será realizada posteriormente, provavelmente concomitantemente com o porcesso de mineração. Quando forem iniciadas as análises de vias metabólicas, se tornará clara a importância das informações complementares ao nome do gene. Dica: para busca de uma proteína que não foi localizada inicialmente, a forma mais fácil é comparar por tblasn uma proteína homóloga de um organismo próximo contra o BD do projeto.

16. Sobre o sistema de anotação LGE

início

O sistema de anotação do LGE foi idealizado pelo grupo de Bioinformática do LGE, sob orientação e supervisão do Prof. Dr. Gonçalo A. G. Pereira. Começou a partir da estrutura inicial do projeto *Xylella fastidiosa*, e foi adaptado e melhorado para contemplar outros projetos genoma, como Vassoura de Bruxa, Genolyptus, Humano – Câncer Cabeça e Pescoço entre outros, até as atuais adaptações específicas para os Projeto Genoma Vassoura de bruxa, Café e Camarão. O time atual de bioinformatas do LGE conta com: Marcelo F. Carazzolle, Eduardo F. Formighieri, Gustavo G. Lacerda Costa, Lucas P. Parizzi, Taís S. Herig e novos alunos que estão chegando.

Manual redigido por Eduardo F. Formighieri (eduformi@lge.ibi.unicamp.br).

Versão da interface de anotação: 1.6.0.0
Versão do Manual: 1.3.0.0

Sugestões, críticas e contato através do e-mail: suporte@lge.ibi.unicamp.br.

início

Roteiro para anotação do Genoma Camarão

[Identificação e sinalização | Classificação | Padrão de nomenclatura | Dinâmica de anotação]

Laurival Vilas Boas[1] & Eduardo Fernandes Formighieri[2]
(Adaptado de texto inicial de: Laurival Vilas Boas, Elizabete K. Takahashi, Luiz Filipe Pereira)
[1]IB/UFBA (lavboas@ufba.br) ; [2]LGE/IB/UNICAMP (eduformi@lge.ibi.unicamp.br)

Salvar no Favoritos:

- ✓ http://www.lge.ibi.unicamp.br/camarao
- ✓ http://psort.nibb.ac.jp/form.html (Predição de localização celular)
- ✓ http://www.cbs.dtu.dk/services/NetPGene (Predição de introns)

[início]

I. IDENTIFICAÇÃO E SINALIZAÇÃO

1. Entrar na página de anotação: http://www.lge.ibi.unicamp.br/camarao
   1.1. No menu lateral: Anotação/Interface de anotaçao. Entrar com login/senha de acesso web.
   1.2. Na página lista de clusters, entrar com login/senha de anotador.
2. Selecionar cluster, e na página de anotaçao do contig, fazer as seguintes análises para definir a identidade do cluster e algumas informações para sinalizar:
3. BLASTX (clicar no link com o nome do gene).
   3.1. Verificar se o Organismo do primeiro hit confere com o encontrado na página. Corrigir a entrada de "Organismo homólogo", se necessário.
   3.2. Verificar possibilidade de contaminação

Caso os resultados do BlastX apresentem similaridade somente com organismos procarióticos, provavelmente o cluster seja uma contaminação. Para confirmar, verificar o BlastN. Se for realmente um contaminante (hit forte com blastn), marcar o campo apropriado e anotar somente o produto (verificação pelo blastn mesmo). Abrir a montagem do contig (No view – colocar a lista de reads no notepad) e escrever justificativa.

Ex. de justificativa: "Provável contaminação. Seqüência similar a diferentes bactérias sem correspondentes em eucariotos nos resultados de blasts."

Se não for contaminante, seguir com a identificação do produto.

   3.3. No resultado do BlastX
      A. Observar melhor hit e comparar com os subseqüentes.
      B. Confirmar o frame correto dos alinhamentos
4. CD-SEARCH – Fazer busca facilitada utilizando o frame do blastx
   4.1. Verificar E-values, copiar o melhor domínio (na dúvida, utilize o pfam) (ex: *pfam00639, Rotamase, PPIC-type PPIASE domain.*) e colar no campo domínio.
   4.2. Verificar se domínios com nomenclatura diferentes (pfam, Smart e COG) correspondem ao mesmo domínio. Neste caso dar preferência ao pfam, sem necessidade de colocar os outros.
   4.3. Caso haja mais de um domínio e estes não estejam relacionados (ou seja, não são versões da mesma coisa em BDs diferentes), copiar todos um representante de cada tipo e colar no campo de domínios em linha s distintas, separados com ponto e vírgula (;) e ponto após o último domínio. Ex:
      A. pfam00639, Rotamase, PPIC-type PPIASE domain;
         Cd00171, *fulanase*, full domain.
   4.4. Quando o domínio estiver incompleto, descrever o domínio e colocar entre parêntese a palavra incomplete. Ex:
      A. pfam00639, Rotamase, PPIC-type PPIASE domain (incomplete).
   4.5. Note que um mesmo domínio pode estar presente em proteínas com funções distintas, portanto o domínio sozinho nem sempre pode definir a proteína.
5. SWISSPROT – utilizar blast automático SWISSPROT
   5.1. Observar o melhor hit e clicar no link do mesmo que remeterá ao UniProt.

5.2. Comparar os principais resultados de BLASTX, CD-SEARCH e SWISSPROT. Se a informação é suficientemente clara, e coerente entre as diferentes fontes, copiar o nome do produto do gene para o campo "PRODUTO".
5.3. Copiar as informações do campo "Function" do UNIPROT para o campo "FUNÇÃO" da interface de anotação.
5.4. Caso a informação não seja suficiente para a descrição da função, utilizar análises adicionais, como blast TC, KEGG, PSORT etc. para a tomada de decisão. Da mesma forma, utilize as buscas por palavra chave para definir a melhor nomenclatura nos casos de nomes diversos. O ideal é utilizar o nome oficial do UNIPROT (quando for possível definí-lo).

6. EC NUMBER
   6.1. Pode ser encontrado principalmente através de informação do UNIPROT, NCBI-Gene, busca no KEGG e busca no ENZYME.
   6.2. Utilizar a classificação do ENZYME é particularmente interessante, para definir em que nível é possível afirmar a classificação do gene. Ou seja, se não tem certeza da especificidade da função, utilize uma classe acima da classificação.
   6.3. Para colocar a informação na interface de anotação, utilize sempre quatro (4) campos separados por pontos, definindo a classe de acordo com o que consegue afirmar com segurança. Ex:
      A. 3.17.1.15 (EC number completo) ou 3.7.1.-
   6.4. Muito cuidado ao utilizar o EC Number, pois um mesmo EC # pode ser utilizado para várias subunidades, ou mesmo proteínas semelhantes. Ou seja, o EC por si não identifica um gene, embora identifique a função. Busque encontrar a função correspondendo ao seu gene, mas não utilize um EC number encontrado para definir quem é seu gene.
7. SÍMBOLO
   7.1. Preferencialmente encontrado no UNIPROT. Também existe boa definição de símbolos no NCBI-Gene (chamado lá de "gene name").
   7.2. Quando não se conseguiu definir com boa precisão o produto do gene, não colocar nada no campo símbolo.
   7.3. Nem sempre existe consenso claro sobre o símbolo. Um símbolo que só existe num organismo tem grande chance de ser apenas a nomenclatura do ORF utilizada no dado genoma, não se trata do símbolo que queremos descrever aqui.
      A. Ex. clássico do que não deve ser utilizado são os genes de *A. thaliana*, que seguem nomenclatura no padrão: At4g38740.
   7.4. Utilizar preferencialmente o padrão de três (3) letras, sem números posteriores, que normalmente representam a seqüência dos genes no genoma, que não corresponde necessariamente à seqüencia do nosso genoma. Exceção feita a genes com nomenclatura típica, como os genes relacionados à cadeia respiratória: cox1, cox2, cox3, nad4, nad5 etc.
8. INFORMAÇÕES COMPLEMENTARES – para serem colocadas no NOTEPAD
   8.1. Utilizando as informações do UNIPROT, buscar as informações do campo Comments, e copiar itens relevantes como: Catalytic activity; Enzyme Regulation; Subcellular location; Tissue Specificity.
   8.2. O número de acesso no UNIPROT.
9. TC NUMBER
   9.1. Avaliar o resultado do blast TC e preencher o campo TC NUMBER, se preciso.
10. OUTROS SINALIZADORES
   10.1. Seq. Cod. Parc. – Seqüência codificadora parcial
      A. Deve ser marcado quando a região do alinhamento não inclui todo o gene similar, potencialmente homólogo. Diz respeito à região seqüenciada.
      B. Se faltar uma pequena região no início e/ou no fim do gene, e havendo bases suficientes no cluster para cobrir a região faltante do gene, não é necessário marcar o cluster como seqüência codante parcial.
   10.2. Intron – indicador da presença de intron
      A. Deve ser marcado quando houver indício da presença de intron no cluster. Isto pode ser observado nos resultados de blast.
   10.3. Full Length – o clone possui toda região codante

A. Deve ser marcado quando a região do alinhamento possui a região codante do início do gene similar (potencialmente homólogo). Isto porque o método de obtenção dos ESTs utilizado sempre inclui a região terminal do cDNA (utilizado primer poli-T).

B. Se faltar uma pequena região no início do gene, e havendo bases suficientes no cluster para cobrir a região faltante do gene, o cluster pode ser considerado Full Length.

10.4. Top Gene

A. Só deve ser marcado quando houver certeza de que o gene é importante para os objetivos do projeto. Quando encontrar genes relevantes, comunicar seu selecionador.

10.5. Cluster problem

A. Indicador de montagem parecendo errada, frameshift, quimeras etc.

[início]

II – CLASSIFICAÇÃO

11. GO
    - 11.1. A classificação GO deve ser coerente com a identificação do produto do gene.
    - 11.2. Observar o Blast GO e a classificação GO do gene homólogo no UNIPROT e no NCBI-Gene.
    - 11.3. Esta informação deverá ser utilizada como base para a determinação do nível de classificação GO que você consegue afirmar para o cluster que está sendo anotado. Na dúvida, utilize uma classe mais geral, um nível acima.
    - 11.4. Para verificar a árvore, utilize o link no próprio termo da página, colocado automaticamente. Caso não tenha havido nenhum hit, utilize a ferramenta AmiGO: http://www.godatabase.org/cgi-bin/amigo/go.cgi para percorrer as árvores de classificação ou fazer buscas por palavra, e definir as classes a serem anotadas.
    - 11.5. Procure sempre preencher ao menos um termo em cada um dos três (3) campos (componente biológico, função molecular e processo biológico). Se o processo/função/componente for desconhecido (exemplo de todos os casos de "Expressed protein" e "Conserved expressed protein"), marque o termo GO apropriado segundo informação abaixo:
        - A. GO:0008372 cellular component unknown
        - B. GO:0005554 molecular function unknown
        - C. GO:0000004 biological process unknown
    - 11.6. Nos casos em que não consegue se decidir sobre a classificação, mas conseguiu definir a identidade do gene, deixe o referido campo em branco, e explique no notepad a sua dúvida.
    - 11.7. Para preencher a informação de termo GO na página, sempre utilize copiar/colar. Copie o termo inteiro (por ex.: GO:0005634:nucleus) e cole nos campos correspondentes, deixando no campo "GO ID" o trecho do ID (ex.: GO:0005634 ou 5634) e no campo "TERMO" o nome exato do termo (ex.: nucleus).

[início]

III – PADRÃO DE NOMENCLATURA

12. PRODUTO – nome do produto do gene
    - 12.1. amino transferase – quando encontrar similaridade com o gene correspondente dentro dos critérios discutidos anteriormente. Preferência para nome oficial, segundo UNIPROT.
    - 12.2. conserved expressed protein – quando é similar a outras proteínas hipotéticas (expressed protein, hypothetical protein, unknow protein, etc.)
    - 12.3. expressed protein – quando não encontra nenhuma similaridade nos bancos de dados (no hits found).
    - 12.4. Sempre começar em letra minúscula e sem ponto final.
13. SÍMBOLO
    - 13.1. Em princípio, três letras minúsculas sem número seqüencial. Nos casos de nomes clássicos bem definidos, como atp6, atp8, cox3, utilizar o nome contendo o número. Na dúvida, utilize apenas as 3 letras.
        - A. Ex. de símbolos: roc, cox2, fur.

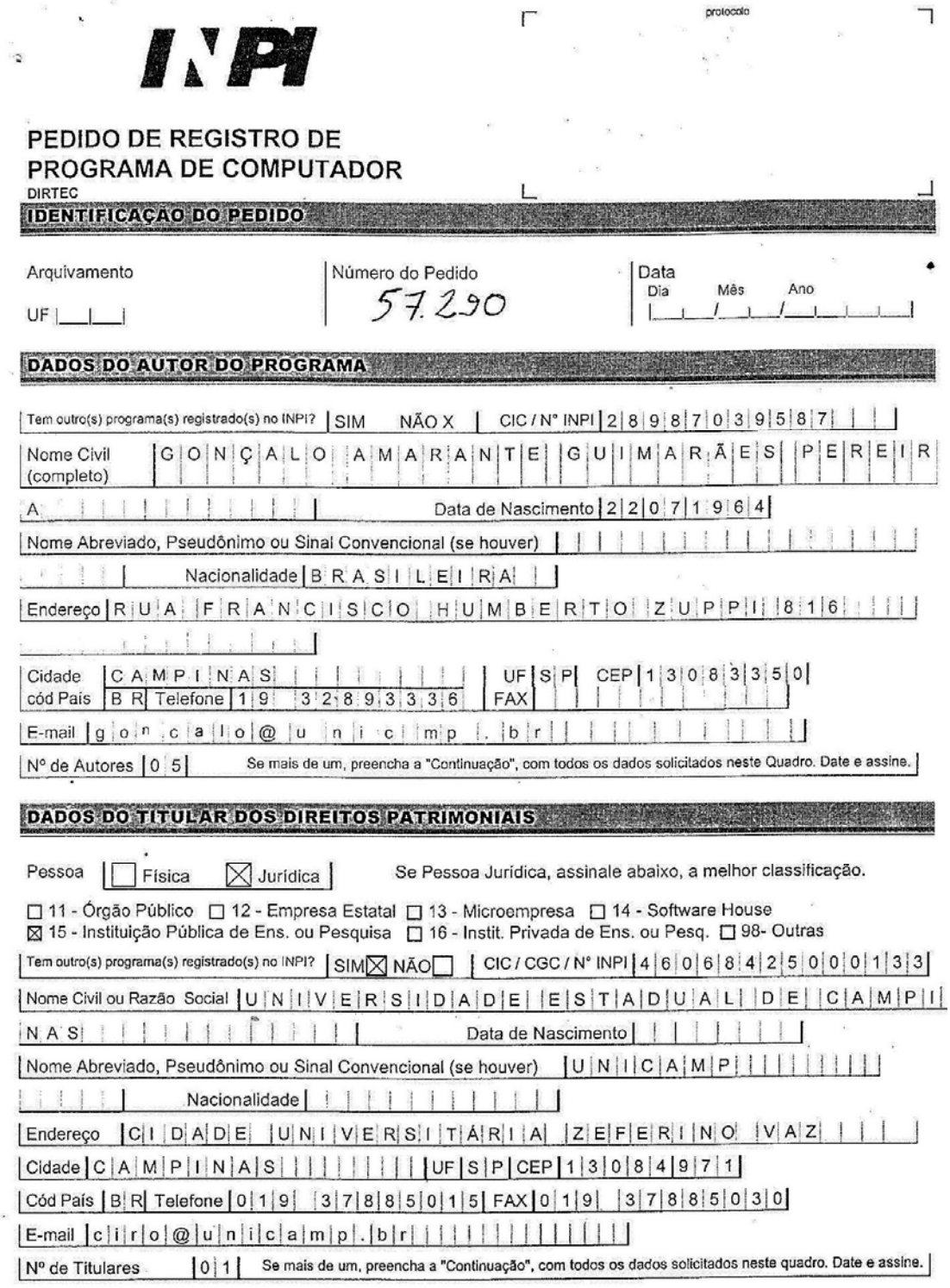
[início]

IV – SUGESTÃO DE DINÂMICA DE ANOTAÇÃO

- ✓ Abrir a interface de anotação do cluster que irá anotar.
- ✓ Abrir a página do *blastX* estático.
  - ✓ Verificar contaminação.
    - ✓ Se for um contaminante, preencher os campos devidos e finalizar.
  - ✓ Observar os nomes dos genes localizados e verificar a coerência dos dados.
  - ✓ Verificar e corrigir se necessário o campo do Organismo Homólogo.
  - ✓ Verificar o *frame* do gene.
- ✓ Fazer análise de CD-Search
  - ✓ Analisar o alinhamento dos domínios encontrados e preencher o campo.
- ✓ Na página da estrutura do Cluster (*View*) verificar:
  - ✓ A qualidade da montagem ou do singlet;
  - ✓ Os tipos de *reads* presentes no *contig*
    - ✓ segundo a nomenclatura – espécie, cepa, estratégia etc.
- ✓ Abrir a página do blast *SwissProt* e observar os nomes dos genes localizados.
- ✓ Baseando-se na coerência do resultado dos *blasts*, no resultado da presença dos domínios e do Blast SwissProt, tomar a decisão quanto à identificação do gene.
- ✓ Inserir a *função* descrita para o produto do gene.
- ✓ Para a definição do *EC Number*, utilizar os dados do swiss prot, NCBI-Gene, busca no *ENZYME* e o blast vs KEGG.
- ✓ Preencha o campo de símbolo do gene utilizando de preferência aqueles descritos no *SWISSPROT*. Sempre que utilizar o *SWISSPROT* para anotar, coloque no *notepad* o número da entrada do homólogo.
  - ✓ O símbolo pode ser localizado ainda em outras bases como *EC number* e *NCBI*.
- ✓ Quando o gene indicar função no transporte de moléculas, fazer a pesquisa na base do *TC number* e lançar no campo apropriado.
  - ✓ Verificar o Blast automático contra TCDB. Em caso de macth significativo, analisar o blast e a coerência com os demais dados. Completar a anotação.
- ✓ Finalmente fazer o levantamento do GOs. Para isto considerar os GOs automáticos, aqueles presentes na descrição do homólogo, *SWISSPROT* e outras bases e preencher a classificação.
- ✓ Sempre que necessário preencher os itens no campo de sinalização.
- ✓ Sempre que for possível selecionar os campos específicos na categoria.
- ✓ Finalizar a anotação.

[início]

7.6. ANEXO F. Registro de patente do programa Gene Projects.



INPI

protocolo

PEDIDO DE REGISTRO DE PROGRAMA DE COMPUTADOR

DIRTEC

IDENTIFICAÇÃO DO PEDIDO

Arquivamento
UF

Número do Pedido: 57.290

Data: Dia / Mês / Ano

DADOS DO AUTOR DO PROGRAMA

Tem outro(s) programa(s) registrado(s) no INPI? SIM NÃO X | CIC / Nº INPI 2 8 9 8 7 0 3 9 5 8 7

Nome Civil (completo): GONÇALO AMARANTE GUIMARÃES PEREIRA

Data de Nascimento: 2 2 0 7 1 9 6 4

Nome Abreviado, Pseudônimo ou Sinal Convencional (se houver)

Nacionalidade: BRASILEIRA

Endereço: RUA FRANCISCO HUMBERTO ZUPPI 816

Cidade: CAMPINAS | UF: SP | CEP: 13083350

Cód País: BR | Telefone: 19 3289 3336 | FAX

E-mail: gonc alo@ unic mp . br

Nº de Autores: 05 — Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

DADOS DO TITULAR DOS DIREITOS PATRIMONIAIS

Pessoa: ☐ Física ☒ Jurídica — Se Pessoa Jurídica, assinale abaixo, a melhor classificação.

☐ 11 - Órgão Público ☐ 12 - Empresa Estatal ☐ 13 - Microempresa ☐ 14 - Software House
☒ 15 - Instituição Pública de Ens. ou Pesquisa ☐ 16 - Instit. Privada de Ens. ou Pesq. ☐ 98- Outras

Tem outro(s) programa(s) registrado(s) no INPI? SIM ☒ NÃO ☐ | CIC / CGC / Nº INPI 4 6 0 6 8 4 2 5 0 0 0 1 3 3

Nome Civil ou Razão Social: UNIVERSIDADE ESTADUAL DE CAMPINAS

Data de Nascimento

Nome Abreviado, Pseudônimo ou Sinal Convencional (se houver): UNICAMP

Nacionalidade

Endereço: CIDADE UNIVERSITÁRIA ZEFERINO VAZ

Cidade: CAMPINAS | UF: SP | CEP: 13084971

Cód País: BR | Telefone: 019 3788 5015 | FAX: 019 3788 5030

E-mail: cirro@ unicamp . br

Nº de Titulares: 01 — Se mais de um, preencha a "Continuação", com todos os dados solicitados neste quadro. Date e assine.

DADOS DO PROGRAMA

Título: GENE PROJECTS

Data de Criação do Programa: 22092003

Linguagens: PERL

Modificação Tecnológica ou Derivação? SIM ☐ NÃO ☒

Título do Programa Original:

Registro composto por outra natureza de ordem intelectual? SIM ☐ NÃO ☒ Se SIM, assinale a(s) natureza(s) abaixo

☐ Literária ☐ Musical ☐ Artes Plásticas ☐ Áudio-Visual ☐ Arquitetura ☐ Engenharia

Classificação do Campo de Aplicação: BL-02 CO-02 IF-01 IF-07

Classificação do Tipo de Programa: GI-01 GI-02 GI-03 TC-01

DOCUMENTOS ANEXADOS

Código	*Quant.*	*Nome*	*Código*	*Quant.*	*Nome*
01	01	Guia de Recolhimento	05	__	Contrato de Trabalho/Prestação de Serviço
02	01	Procuração	06	10	Envelope
03	05	Termo de Cessão	99	04	Outros(especificar) DADOS DOS OUTROS 04 AUTORES
04	__	Termo de Autorização para Modificações Tec. ou Derivações			

DECLARAÇÕES

DECLARO, PARA TODOS OS FINS DE DIREITO:

A) que estou ciente de **TODAS AS RECOMENDAÇÕES** constantes do "Manual do Usuário de Registro de Programas de Computador", **ESPECIALMENTE NO QUE TANGE AO TÍTULO E AOS DOCUMENTOS DO PROGRAMA**, bem como da legislação pertinente ao assunto, constante dos anexos "A" ; "B" ; "C" ; "E" e "F", do referido Manual;

B) que se deixar de solicitar a prorrogação do sigilo, nos casos necessários, estarei desistindo desse caráter de guarda dos documentos de programa do presente depósito, na forma do art. 4º, § 3º, da Lei 7646, de 18 de dezembro de 1987;

C) que, se devido à qualidade do papel ou à qualidade gráfica dos documentos sigilosos anexos ao presente, houver deterioração ou perda de seu conteúdo, nenhuma responsabilidade caberá ao INPI, desde que mantida a inviolabilidade dos invólucros (ressalvadas as hipóteses de serem abertos por ordem judicial ou motivo de força maior;

D) que em caso de perda do SIGILO ou dos documentos, por culpa exclusiva do INPI, a indenização por perdas e danos, porventura cabível, estará limitada a 20 (vinte) salários mínimos;

E) que devo manter guardado, em segurança e inviolado, o COMPARTIMENTO "3" do invólucro especial para depósito, que é restituído pelo INPI, para fins de recomposição do arquivo do Instituto, no caso de sua destruição total ou parcial por algum tipo de sinistro;

F) que deverei manter endereço atualizado junto ao Serviço de Registro de Programas de Computador, a fim de garantir o recebimento das comunicações relativas ao andamento do meu pedido/registro, ressalvando o INPI de qualquer responsabilidade decorrente da não observação deste preceito.

DADOS DO PROCURADOR

Código do Procurador:

Nome: MARIA CRISTINA VAILM LOURENÇO GOMES

UF: SP Telefone: 019 3788 4771

DECLARO, SOB AS PENAS DA LEI, SEREM VERDADEIRAS AS INFORMAÇÕES PRESTADAS

Local/Data **Campinas,SP, 08.01.2004** Assinatura/Carimbo

Maria Cristina Vailm Lourenço Gomes
Procuradora de Universidade Subchefe
Matrícula n.º 193861
OAB/SP n.º 99243-B

REGISTRO DE PROGRAMA DE COMPUTADOR - CONTINUAÇÃO

Utilize este ANEXO, em quantas folhas forem necessárias, para complementar as informações dos formulários "Pedido de Registro de Programa de Computador" e "Folha de Petição" (DIRTEC).

DADOS DOS OUTROS AUTORES DO PROGRAMA

Tem outro programa registrado no INPI – NÃO

CIC nº 190.397.088-19

Data de nascimento: 22.07.64

Nome civil (completo) – EDUARDO FERNANDES FORMIGHIERI

Nome abreviado – NÃO TEM

Nacionalidade - BRASILEIRA

Endereço – RUA DA FAZENDA, 155, CASA 26, CONDOMÍNIO BAMBUS

Cidade – SUMARÉ

UF – SP

CEP – 13175-658

Cód. País – BR

Telefone – 0 XX 19 9611.4957

FAX – NÃO TEM

E-Mail : eduformi@lge.ibi.unicamp.br

Nº DE AUTORES: 05

Campinas, SP, 08 DE JANEIRO DE 2004.

x

Maria Cristina Vallim Lourenço Gomes
Procuradora de Universidade Subchefe
Matrícula n.º 193861
OAB/SP n.º 99243-B

7.7. ANEXO G. Participações e colaborações atuais em manuscritos em diferentes fases de redação.

7.7.1. Evaluation of the Genome Sequence of *Moniliophthora perniciosa,* the Causative Agent of Cacao´s Witche´s Broom Disease (sendo redigido)

Jorge M. C. Mondego, Marcelo F. Carrazzolle, Eduardo F. Formighieri, Gustavo G. C. Lacerda, Ramon O. Vidal, Lucas P. Parizzi, Johana Rincones, Ricardo A. Tiburcio, Leandra M. Scarpari, Maricene Sabha, Carolina Cotomacci, Taís S. Herig, Dirce M. Carraro, Aristóteles Góes-Neto, Luciana V. Barbosa, Mariza S. Gonçalves, José P. Moura Neto, Karina Gramacho, Lyndel W. Meinhardt, Julio Cezar M. Cascardo, Gonçalo A. G. Pereira*.

7.7.2. Differential gene expression of the biotrophic and saprotrophic mycelia of the Witches' broom hemibiotrophic pathogen *Moniliophthora perniciosa* (sendo redigido)

Johana Rincones, Leandra M. Scarpari, Marcelo F. Carazzolle, Robson Dias, Joan G. Barau, Eduardo F. Formighieri, Bruno V. Oliveira, Jorge Mondego, Julio Cascardo, Ricardo A. Azevedo, Lyndel W. Meinhardt, and Gonçalo A.G. Pereira.

Abstract: With the object of acquiring a better understanding of the molecular mechanisms involved in the fungal pathogenesis during the development of Witches' broom disease of cacao, we performed DNA microarrays and EST analyses of the biotrophic and saprotrophic stages of *M. perniciosa*. For the EST analysis, three unsubtracted full-length cDNAs libraries were constructed: glucose- and cacao-induced saprotrophic mycelia and cacao-induced biotrophic mycelia. 3650 accepted reads of all four libraries were clusterized producing 1427 unigenes. 534 of these unigenes (36%) showed significant similarity (E-value ≤ 1E-05) with protein sequences in public databases. The DNA microarray analysis was performed for 2,304 fragments obtained from the genomic DNA libraries of the Witches' Broom Genome Project, and which were

selected based on their sequence similarity to pathogenicity genes of other pathogens. These fragments were hybridized with probes prepared from total RNA of the biotrophic and saprotrophic mycelia of *M. perniciosa*, induced with cacao extracts. The induced genes in the saprotrophic phase mycelia were mainly involved in oxidative phosphorilation and production/degradation of reactive oxygen species (ROS), in addition to toxins with antifungal properties and enzymes involved in lignin degradation. The analysis of genes expressed during the biotrophic phase also included several involved in ROS production and genes showing significant similarity to proteins involved in the Hypersensitive Response of plants, such as Prohibitins and Stomatins. Additionally, numerous proteases and peptidases were found to be expressed in the biotrophic mycelia, probably involved in pathogen defense against the plant defense systems. Differential expression of selected genes was confirmed by reverse northern and Real Time-PCR analyses. These results contribute to the overall understanding of the molecular mechanisms underlying the plant-pathogen interactions governing witches' broom disease and has aided in the identification of possible drug targets. This work is part of the on-going *M. perniciosa* Genome Project.

7.7.3. An Alternative Oxidase is Preferentially Expressed in the Biotrophic Mycelium of the Hemibiotrophic Fungal Pathogen *Moniliophthora perniciosa*, causal agent of Witches' Broom Disease of Cocoa (sendo redigido)

Daniela P. T. Thomazella, Johana Rincones, Odalys García, Elzira Saviani, Ione Salgado, Ricardo A. Tiburcio, Eduardo F. Formighieri, Lyndel W. Meinhardt, Gonçalo A. G. Pereira.

Abstract: The hemibiotrophic fungus *Moniliophthora perniciosa* is the causal agent of witches' broom disease of cocoa. Fungicides containing Cytochrome-dependent Respiratory Chain (CRC) inhibitors have proved unsuccessful to control the disease. The *M. perniciosa* Genome Project has led to the identification of a putative gene homologue of an alternative oxidase protein (AOX) and its activity could account for the fungal resistance against these inhibitors. This work aimed to study the activity and transcriptional regulation of the AOX from *M. perniciosa*. AOX expression during the biotrophic and saprotrophic stages of the fungus was analyzed through

Reverse Transcription-PCR. Additionally, AOX activity was verified by measuring the rate of oxygen consumption in the presence of inhibitors of CRC and AOX (Strobilurin and SHAM, respectively). Although detected in both stages, AOX expression was significantly higher in the biotrophic phase. Moreover, Strobilurin inhibited oxygen consumption of the saprotrophic mycelium by 70%, indicating that CRC is the predominant respiratory pathway during this stage. In contrast, respiratory activity of the biotrophic mycelium was insensitive to Strobilurin and highly inhibited by SHAM, with AOX activity accounting for about 90% of the oxygen consumption. Plants are known to produce large amounts of Reactive Oxygen Species (ROS) and Nitric Oxide (NO) (a CRC inhibitor) in response to pathogen invasion. During Witches' Broom disease, this situation would correspond to the biotrophic phase of *M. perniciosa*. Therefore, AOX, which is a NO insensitive protein that is able to stall ROS production, could contribute to pathogen resistance to the plant defense mechanisms

7.7.4. Multidrug resistance-associated-like proteins (MRP) from ATP-binding cassette (ABC) transporters induced in citrus plants (submetido à "Genetics and Molecular Biology")

Alexandre Morais do Amaral, Juliana Cristina Baptista, Eduardo F. Formighieri, Daniel Saito, Edenilson Rabello, Adriane N. de Souza, Maria Estela Silva-Stenico, and Siu Mui Tsai.

Abstract: The multidrug resistance-associated proteins (MRP), a subfamily of the ATP-binding cassettes (ABC) transporters, represent one of the most numerous transporters in all living organisms. However, there is no report on these transporters in citrus plants to date. On another hand, many of their substrates have been identified in other plants, including both organic and inorganic compounds and MRP proteins have been shown to play a role in plant detoxification processes as well as in stress response. Here, to obtain genetic information on citrus, we constructed a cDNA library from various citrus species and tissues and identified the genes expressed through the construction of expressed sequence tags (EST) libraries. We used cDNA sequencing and alignment-based annotation of genomic sequences to identify putative MRP genes. Although the search found 488 ESTs, which were assembled into 62 contiguous

sequences, we have identified only two proteins highly homologous to members of the MRP subfamily. One of these clusters was found to be expressed in various citrus species (*Citrus aurantium, C. aurantifolia, C. reticulata, C. sinensis, C. latifolia* and *Poncirus trifoliata*) and tissues (leaf, fruit, seed, bark, and root). Interestingly, the other cluster was constructed only with cDNA from leaves of three species (*C. reticulata, C. aurantifolia,* and *C. latifolia*), indicating a more restricted expression, based on organ and species. Both MRP-like clusters contained two copies each of the nucleotide-binding domains and one copy of transmembrane domain, which are located in the C-terminal. The best protein homology was found with AtMRP10 and AtMRP6, to which no functional role has been identified so far. The characterization of these clones may indicate in citrus important roles in detoxification and stress response as well as in growth and developmental processes. To our knowledge, this is the first report on MRP-like proteins in citrus species.

7.7.5. **Pleitropic drug resistance (PDR)-like proteins induced in citrus** (submetido à "Genetics and Molecular Biology")

Alexandre Morais do Amaral, Daniel Saito, Eduardo F. Formighieri, Edenilson Rabello, Adriane N. de Souza, Maria Estela Stenico, and Siu Mui Tsai.

Abstract: Pleiotropic drug resistance (PDR) proteins, a subfamily of the ATP-binding cassette (ABC) transporters, have been recently shown to play a role in plant defense against biotic and abiotic stresses. However, nothing is known about their expression in citrus. In order to investigate the occurrence of PDR homologues in citrus species, we have surveyed the EST sequences from different tissues and conditions of the citrus expressed sequence tags (CitEST) database, through sequence similarity search analyses and inspection for characteristic PDR domains. Multiple sequence alignments and prediction of transmembrane topology were additionally performed. The identification of seven putative proteins showing characteristic PDR features in tissues under stress,conditions indicate a potential correlation between PDRs and defense mechanisms in citrus. To our knowledge, this is first report of PDR gene expression in citrus.

7.7.6. Structure of Genetic Diversity among Common Bean (*Phaseolus vulgaris* L.) Varieties of Mesoamerican and Andean Origins Using New Developed Microsatellite Markers (submetido à "Genetics Resources and Crops Evolution")

Luciana Lasry Benchimol*, Tatiana de Campos, Sérgio Augusto Morais Carbonell, Carlos Augusto Colombo, Alisson Fernando Chioratto, Eduardo Fernandes Formighieri, Lígia Regina Lima Gouvêa and Anete Pereira de Souza.

Abstract: A common bean genomic library was constructed using the 'IAC-UNA' variety enriched for (CT) and (GT) for microsatellite motifs. From 1,209 sequenced clones, 714 showed microsatellites distributed over 471 simple and 243 compound motifs. GA/CT and GT/CA were the most frequent motifs found among these sequences. A total of 123 microsatellites has been characterized. Out of these, 87 were polymorphic (73.7%), 33 monomorphic (26.8%), and 3 (2.4%) did not amplify at all. In a sample of 20 common bean materials selected from the Agronomic Institute Germplasm Bank, the number of alleles per locus varied 2 to 9, with an average of 2.82. The polymorphic information content (PIC) of each marker varied from 0.05 to 0.83, with a 0.45 average value. Cluster and principal coordinate analysis of the microsatellite data were consistent with the original assignment of the germplasm accessions into the Andean and Mesoamerican gene pools of common bean. Low polymorphism levels detected could be associated with the domestication process. These microsatellites could be a valuable resource for the bean community because of their use as new markers for genetic studies.

7.7.7. Characterization of twenty dinucleotide microsatellite loci for common bean (*Phaseolus vulgaris* L.) (sendo redigido)

Tatiana de Campos, Luciana Lasry Benchimol, Sérgio Augusto Moraes Carbonell, Alisson Fernando Chioratto, Eduardo Fernandes Formighieri, and Anete Pereira de Souza.

Abstract: Common bean (*Phaseolus vulgaris* L.) is an important legume food in the world. We identified and characterized twenty microsatellite loci in common bean. An enriched library was constructed and screened for two microsatellite repeat sequences (CT and GT). Microsatellite PCR-amplification was tested for 14 accessions of the IAC Germplasm Bank, including Andean and Mesoamerican gene pools. The allele number ranged from one to three, and the polimorphism information content (PIC) between 0.14 and 0.65. These polymorphic loci can be used for genetic and QTL mapping, marker-assisted selection and germplasm characterization in common bean.

7.7.8. **Development, characterization and comparative analysis of polymorphism at common bean-SSR loci isolated from genic and genomic sources** (submetido à "Genome")

Luiz R. Hanai, Tatiana de Campos, Luis E.A. Camargo, Luciana L. Benchimol, Anete P. de Souza, Maeli Melotto, Sérgio A.M. Carbonell, Alisson F. Chioratto, Luciano Consoli, Eduardo F. Formighieri, Marcos V.B.M. Siqueira, Siu M. Tsai and Maria L.C. Vieira.

Abstract: Microsatellites or SSR (Single Sequence Repeats) have been used to construct and integrate genetic maps in crop species, including *Phaseolus vulgaris*. In the present study, three cDNA libraries generated by the Bean EST project (http://lgm.esalq.usp.br/BEST/) comprising a unigene collection of 3,126 sequences and a genomic microsatellite-enriched library were analyzed for the presence of SSR. A total of 219 EST were found to carry 240 SSR (named EST-SSR), whereas 714 genomic sequences contained 471 SSR (named genomic-SSR). A subset of 80 SSR, 40 EST-SSR and 40 genomic-SSR, were evaluated for molecular polymorphism in 23 genotypes of cultivated beans from the Mesoamerican and Andean genetic pools, including Brazilian cultivars, and two related species. Of the common bean genotypes, 31 EST-SSR loci were polymorphic, yielding 2-12 alleles compared with 26 polymorphic genomic-SSR, accounting for 2-7 alleles. Cluster analysis from data using both genic and genomic SSR revealed a clear separation between Andean and Mesoamerican beans. The usefulness of these loci for distinguishing bean genotypes and genetic mapping is discussed.

7.8. ANEXO H. Cursos e palestras durante o período de trabalho no Laboratório de Bioinformática do LGE/IB/UNICAMP.

(informação do menu cursos da página http://www.lge.ibi.unicamp.br/bioinfo)

2006		
Como el proyecto genoma modifico la investigacion en Brasil: de *Xylella fastidiosa* a Genolyptus	The 2nd International Seminar on Genomics,Proteomics,Bioinformatics and Systems Biology, Universidad del Cauca, Popayán (Cauca) – Colombia	Eduardo F. Formighieri
Genomica y Bioinformatica Aplicada a un Caso de Escoba de Bruja en el Cacao	The 2nd International Seminar on Genomics,Proteomics,Bioinformatics and Systems Biology, Universidad del Cauca, Popayán (Cauca) – Colombia	Eduardo F. Formighieri
Mini Curso de anotação de ESTs de *Crinipellis perniciosa*	LGE - UNICAMP	Eduardo F. Formighieri
Mini Curso de anotação de ESTs de Camarão	Universidade Federal de São Carlos	Eduardo F. Formighieri, Marcelo F. Carazzolle

2005		
I Curso de Bioinformática	Universidade Federal de Lavras, Lavras, MG.	Eduardo F. Formighieri, Marcelo F. Carazzolle
Palestra Bioinformática do Projeto Vassoura de Bruxa	Workshop da Vassoura de Bruxa, Universidade Estadual de Santa Cruz, Campus Soane Nazaré de Andrade, UESC, Ilhéus, BA.	Eduardo F. Formighieri
Curso Ferramentas de Bioinformática	XXIII Congresso Brasileiro de Microbiologia, Sociedade Brasileira de Microbiologia, Mendes Convention Center, Santos, SP.	Eduardo F. Formighieri
Palestra e treinamento sobre anotação de genes	Intensivo para Anotação de Genes do Genoma Café, Embrapa Café, Brasília, DF.	Eduardo F. Formighieri
Palestra Projetos Genoma	Curso Introdução a Ferramentas de Bioinformática, FIOCRUZ - Centro de Pesquisa René Rachou, Belo Horizonte, MG.	Eduardo F. Formighieri

2004		
Bioinformática: Palestra e treinamento prático	Disciplina de graduação Biodiversidade e conservação: um enfoque molecular, CENA/USP, Piracicaba, SP.	Eduardo F. Formighieri
Bioinformática: Palestra e treinamento prático	Disciplina de pós-graduação Ecologia Experimental de Microrganismos, CENA/USP, Piracicaba, SP.	Eduardo F. Formighieri
Bioinformática e Anotação	Disciplina de pós-graduação em genética e Biologia Molecular, Instituto de Biologia/UNICAMP, Campinas, SP.	Eduardo F. Formighieri
Bioinformática e Anotação do Projeto Genoma Vassoura de Bruxa: Palestras e treinamento prático	Laboratório de Biologia Molecular e Propagação de Plantas, FESP/UEMG, Passos, MG.	Eduardo F. Formighieri, Marcelo F. Carazzolle
Aplicación de Herramientas Bioinformáticas para el Análisis Avanzado de Secuencias Genómicas	CABBIO/CBAB (Centro Brasileiro Argentino de Biotecnologia), Instituto de Biotecnologia, Universidad Nacional de Colombia, Bogotá, Colômbia.	Eduardo F. Formighieri
Mini-curso da genômica à biotecnologia	IV Semana de Química da UNICAMP, All Química/CAEQ e Sociedade Brasileira de Química, Campinas, SP.	Eduardo F. Formighieri
Mini-curso teórico-prático: Bioinformática	V Semana da Biologia, Curso de Ciências Biológicas EFOA/CEUFE, Alfenas, MG.	Eduardo F. Formighieri