Universidade de Brasília
Instituto de Biologia
Departamento de Biologia Celular

Rodrigo Theodoro Rocha

Genômica comparativa de cepas de *Aspergillus terreus* visando a produção de lovastatina

Brasília
2017

Rodrigo Theodoro Rocha

Genômica comparativa de cepas de *Aspergillus terreus* visando a produção de lovastatina

Dissertação apresentada ao Departamento de Biologia Celular do Instituto de Biologia da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Biologia Molecular com ênfase em Biologia Computacional.

Área de Concentração: Bioinformática
Orientador(a): Prof. Dr. Georgios Joannis Pappas Júnior
Co-orientador(a): Prof. Dr. Nádia Skorupa Parachin

Brasília

2017

Dedico este trabalho a meus familiares, amigos e amada que sempre me apoiaram nos momentos alegres e conturbados.

Agradecimentos

Sendo os cientistas seres humanos sujeitos a rotineiras experiências sociais, como separar o subjetivismo da pesquisa científica? Responderia que certamente deve-se aplicar o método científico. O método científico, se aplicado corretamente, é o menos enviesado a preceitos de cunho religioso, social e subjetivo. Porém, antes de aplicá-lo as experiências empíricas dos cientistas e estudiosos se destacam. Elas influenciam na condução do pensamento subjetivo que, consequentemente, poderá levar à formulação da hipótese científica a ser testada. Os fluxos de pensamentos subjeitos nascem das diversas interações do cientista com o mundo. Portanto, gostaria de agradecer a todos e qualquer coisa que indiretamente (seja via e-mails, vídeo aulas, posts ou artigos) e diretamente (familiares, professores e amigos) moldaram minha forma de pensar. Não desejo excluir ninguém que algum dia passou na minha vida da lista de agradecimentos porém algumas pessoas merecem mais créditos.

Gostaria de agradecer meu avô pelas excelentes discussões e crescimento intelectual durante toda a minha vida e a minha querida avó pelo incrível coração. A vida de vocês é um espelho para todos. Além disso, agradeço a meus pais pela educação e carinho e ao meu irmão pelas confidências e diversões. Não poderia esquecer mais um integrante da casa nesta lista. Incluo meu cachorro pela companhia persistente nas noites de trabalho ao computador.

A vida só tem sentido quando é compartilhada. Então, partindo desta máxima gostaria de agradecer imensamente pelos *post-it* deixados pela minha namorada Alessa Bembom com as melhores frases coladas nos mais diversos cantos do meu PC. É tão mais fácil quando você sabe que tem alguém para contar no início, durante e fim do dia. Logo, te agradeço imensamente por tudo. Incluo nos agradecimentos todos meus amigos, especialmente o

Gustavo Coelho e Marcus Renan, com quem posso contar e sempre me influenciam - positivamente - nas decisões diárias.

Para finalizar, agradeço a todos do laboratório de Bioinfo e aos envolvidos diretamente neste trabalho. Especialmente à Flávia Mulinari e Kelly Assis pelos experimentos e paciência comigo, ao Wendell Pereira no qual tenho uma dívida inestimável pelas revisões e correções textuais. Minha dívida e privilégio é ainda maior com meu orientador Georgios Pappas no qual as envolventes discussões científicas e notório saber tornaram a conclusão deste trabalho possível. Sou também muito grato ao convite da Nádia Parachin para trabalhar com os dados apresentados nesta dissertação.

"*Tenho um amigo artista plástico, e (...) ele tem uma opinião com a qual não concordo muito. Ele segura uma flor e diz, "Veja como é bonita", e eu concordo. Aí ele diz: "Sabe, como artista consigo ver como a flor é bonita, mas você, como cientista, quer dissecar tudo, e a beleza torna-se uma coisa sem graça". Acho que ele é meio maluco. Primeiro, a beleza que ele vê está disponível para todos, e para mim também, eu creio. Embora eu não seja tão refinado esteticamente quanto ele, posso apreciar a beleza de uma flor. Ao mesmo tempo, vejo na flor muito mais do que ele. Consigo imaginar as células dela, as ações complicadas lá dentro que também tem sua beleza. Quer dizer, não existe beleza só na dimensão de um centímetro; também há beleza em dimensões menores, na estrutura interna, nos processos (...). O fato de que as cores da flor evoluíram para atrair insetos que irão polinizá-las (...) significa que insetos podem (...) ser atraídos pela beleza como nós somos (...) Todo os tipos de questões interessantes – objetos do interesse da ciência – só aumentam a beleza e o mistério de uma flor. Ciência só acrescenta. Não entendo como alguém possa achar que subtrai.*"

Richard Feynman, The pleasure of finding things out

"*Não é no espaço que devo procurar a minha dignidade, mas na direção do meu pensamento. Não deverei tê-la mais se possuir mundos. Pela amplidão, o universo me envolve e me traga como um átomo; pelo pensamento eu compreendo o mundo.*"

Blaise Pascal, Pensées

Resumo

Metabólitos secundários (MS) são moléculas heterogêneas de baixo peso molecular e, alternativamente aos metabólitos primários, não estão diretamente envolvidas no crescimento do organismo que as produz. Entre os microrganismos produtores de MS destacam-se os fungos filamentosos do gênero *Aspergillus*, abrangendo diversas espécies com estilos de vida variados e produzindo inúmeros compostos de importância para os homens, animais e plantas. A espécie *A. terreus* é produtora de diversos MS, entre eles destaca-se a lovastatina, fármaco da classe das estatinas, que são mundialmente utilizadas para a redução dos níveis de colesterol.

Visando a bioprospecção de cepas de *A. terreus* produtoras de lovastatina utilizamos a genômica comparativa para detalhar a estrutura e variabilidade dos genes responsáveis por sua biossíntese. Oito cepas de *A. terreus* isoladas no Brasil foram submetidas a sequenciamento de segunda geração utilizando a plataforma Illumina. Os dados resultantes foram mapeamentos contra o genoma de referência da espécie (cepa NIH 2624) e observou-se uma conservação de 86% de todo genoma entre as cepas. No entanto, grandes regiões com tamanho maior que 10 kb apresentaram cobertura anômala, e posteriormente verificou-se que se tratava de variantes estruturais (grandes indels) nos genomas das cepas, inclusive no agrupamento gênico de biossíntese (BCG) de lovastatina. As variantes estruturais dentro deste loco foram validadas experimentalmente via ensaios de PCR e a ausência de genes essenciais à biossíntese da lovastatina explica o fenótipo não produtor em algumas cepas.

A observação variabilidade genômica entre as cepas motivou o desenvolvimento de uma nova metodologia para detecção de BCGs em geral. Esta baseia-se na estrutura de grafos de Bruijn coloridos e pode ser aplicada diretamente nos dados brutos de sequenciamento, sem necessitar montagens genômicas de alta qualidade. Com esta abordagem foi possível

identificar os limites gênicos dos agrupamentos de biossíntese dos metabólitos acetilaranotina e terretonina.

As análises de genômica comparativa neste estudo apontam as relevantes diferenças na composição gênica de indivíduos da mesma espécie, as quais podem ser correlacionadas com fenótipos de interesse biotecnológico. Ademais, ressaltam a complexa história evolutiva dos fungos e a plasticidade de seus genomas. Assim como acontece em muitos procariotos, os genomas dos fungos filamentosos devem ser representados por um pan-genoma.

Abstract

Secondary metabolites (SM) are a heterogeneous class of low molecular weight compounds not directly involved in the growth of the producing organism. Among the SM producing microorganisms stands out the genus *Aspergillus*, a diverse group of filamentous fungi of medical and biotechnological relevance. The species *A. terreus* is known to produce several metabolites, noticeably lovastatin, belonging to the statins class with worldwide application as cholesterol lowering drugs.

Aiming at the bioprospection of lovastatin producing strains we employed a comparative genomics study to pinpoint the structure and variability of the genes involved in its biosynthesis. Eight *A. terreus* strains, isolated in different locations, were sequenced by second generation genome sequencing platform Illumina. The resulting reads were mapped against the reference genome of *A. terreus* (strain NIH 2624) unveiling a 86% genome-wide conservation between the strains. However, large blocks spanning over 10 kb showed anomalous mapping coverage depth, and further analyses showed that these were structural variants (large indels) occurring in the genome of the strains. Strikingly, some strains exhibited structural variations in the lovastatin biosynthetic gene cluster (BCG), further validated by PCR assays, which offers a plausible explanation for the non-producing phenotype observed in some strains.

The observation of strain-specific genome variation prompted the development of a new BCG detection methodology based on colored de Bruijn graphs and directly applied to raw sequencing data without the necessity of a high quality reference genome. This approach uncovered the presence and the gene boundaries of several BCGs in our study, such as the biosynthetic clusters for the metabolites acetilranotin and terretonin.

The comparative genomic analyses in this study highlight the gene composition dif-

ferences among individuals of the same species and the correlation with biotechnological relevant phenotypes. Moreover, underlies the complex fungal evolutive pathways and the plasticity of microbial genomes.

Lista de Figuras

1.1 Diagrama dos domínios essenciais (em laranja) e opcionais (em azul) detectados nas enzimas policetídeos sintases (PKSs) e peptídeos sintases não-ribossômicos (NRPSs), além da identificação dos módulos com função iterativa. 37

1.2 Via metabólica de biossíntese do metabólito secundário lovastatina em *A. terreus* 46

1.3 Agrupamento gênico de biossíntese (BCG) de lovastatina em *A. terreus* ATCC 20542 47

3.1 Metodologia curvas tipo Hilbert. 59

4.1 Quantificação do metabólito secundário lovastatina no sobrenadante de 8 indivíduos após cultivo das cepas de acordo com metodologia A.3. 63

4.1 Espectro de cobertura do mapeamento das leituras-curtas (*short reads*) das oito cepas de *Aspergillus terreus* contra o genoma de referência da espécie (NIH) 66

4.2 Distribuição das regiões genômicas com cobertura de mapeamento $\leq$2x. . . 67

4.3 Construção do grafo de ordem parcial para representação de múltiplos alinhamentos. 69

4.4 Esquema representando o peso em cada resta em grafos POA. 71

4.5 Esquema representativo da heurística proposta para desenhar primers em estudos de resequenciamento. 73

4.6 Análise dos resultados da PCR. 76

4.7 Variantes únicas nas regiões gênicas do agrupamento de biossíntese de lovastatina da cepa ATCC 20542. 79
4.8 Circos das cepas ATCC 20542, BU35, BU27 e BU33. 85
4.9 Circos das cepas U9, U10, U22 e U26. 86
4.10 Múltiplo Alinhamento das oito cepas do estudo contra o genoma de referência NIH. 89
4.11 Predição de metabólitos secundários nas montagens. 91
4.12 Representação esquemática da metodologia proposta para detectar agrupamentos de biossíntese de metabólito secundário (BCGs) a partir da estrutura de grafos colorida. 93
4.13 Comparando a metodologia de Bruijn colorida contra a ferramenta tradicional para predição de agrupamentos gênicos de biossíntese de metabólitos secundários antiSMASH. 95
4.14 Comparação das predições dos genes responsáveis pela síntese de terretonina em *Aspergillus terreus*. 99
4.15 Comparação das predições dos genes responsáveis pela biossíntese de lovastatina em *Aspergillus terreus*. 100
4.16 Comparação das predições dos genes responsáveis pela síntese de lovastatina em *Aspergillus terreus*. 103

Lista de Tabelas

1.1 Tabela comparando os principais atributos de cada plataforma de sequenciamento a partir dos dados compilados por Goodwin et al. (2016). A sigla PE significa *paired-end*, pb - pares de base, indel - pequenas inserções e deleções na sequência. Os prefixos K, M e G são as siglas para as gradezas científicas 10^3, 10^6 e 10^9 respectivamente. 27

3.1 Identificadores das cepas usadas neste estudo, identificadores oficiais da micoteca da UFPE 53

3.2 Plataformas de sequenciamento utilizadas para o resequenciamento genômico das cepas 54

4.1 Métricas avaliadas após a etapa de sequenciamento. Na nomenclatura das amostras as extensões .1 e .2 referem as leituras resultantes do sequenciamento *paired-end* 64

4.2 Tabela sumarizando os resultados pós filtragem dos dados brutos. 65

4.3 Tabela com as informações dos primers desenhados nesse estudo. 75

4.4 Contagem dos polimorfismos únicos de sequência (SNPs). O termo het refere-se às variantes heterozigotas, classificadas em RA quando possui um alelo igual referência e outro alternativo e AA onde os dois alelos são alternativos à referência e divergentes entre si. O termo hom refere-se às variantes homozigotas alternativas (hom AA) e homozigoto referência (hom RR). . . 77

4.5 Sumário das métricas avaliadas na montagem *de novo*, predição dos genes e mapeamento. 82

4.6 Estatística descritiva dos blocos sintênicos usando como referência a cepa NIH 2624. 87
4.7 Genes únicos ao genoma de referência *A. terreus* NIH 2624 em comparação às oito cepas usadas no estudo. 101

B.1 Supercontigs da montagem de referência NIH 2624. 147
B.2 Lista de metabólitos secundários (SMs) conhecidos preditos pela ferramenta antiSMASH (Weber et al., 2015) nas montagens *de novo* das oito cepas. . . 148

Lista de Abreviações

ACP Proteína Carreadora de Acil

antiSMASH antibiotics and Secondary Metabolite Analysis SHell

AT Aciltransferase

BCG Agrupamento de Biossíntese de Metabólito Secundário

BNS Blocos Não Sintênicos

BS Blocos Sintênicos

CON Condensação

DH Desidratase

DMAT Triptofano Dimetilalil Sintase

DML Dihidromonacolina L

ER Enoil-redutase

FAS Ácidos Graxos Sintetases

Gb Giga Bases

GRAS Generally Recognized As Safe

HC Curva Hilbert

HGT Horizontal Gene Transfer

HMG 3-hidroxi-3-metil-glutaril

HMGR 3-hidroxi-3-metil-glutaril-CoA redutase

HMM Modelo Markoviano Oculto

HPLC Cromatografia Líquida de Alta Eficiência

INDELs Pequenas Inserções e Deleções de Sequência

KR Cetoredutase

KS Cetosintase

LDKS Lovastatina Dicetato Sintase

LNKS lovastatina nonacetato sintase

MIBiG Minimum Information about a Biosynthetic Gene cluster

Mp Milhão de bases

MS Metabólito Secundário

MSA Alinhamento Múltiplo de Sequência

MT Metil Transferase

NGS Next Generation Sequencing

NRP Peptídeos não Ribossômicos

NRPS Peptídeo Sintases Não-Ribossômicos

OLC Overlap-layout-consensus

ORF Open Reading Frame

PCR Polimerase Chain Reaction

PK Policetídeos

PKS Policetídeos Sintases

PO-MSA Grafo Parcial de Alinhamento Múltiplo

SAM S-Adenosil Metionina

SMURF Secondary Metabolite Unknown Regions Finder

SNPs Single Nucleotide Polymorphism

SSR Repetições de Sequências Simples

TE tioesterase

VE Variante Estrutural

Sumário

1. Introdução 25

1.1 Introdução à Genômica 25

1.1.1 Montagem de genomas 28

1.1.2 Estudos de resequenciamento 30

1.2 Genomas fúngicos: estrutura e relevância biotecnológica 32

1.2.1 Metabólitos Secundários 35

1.2.1.1 Policetídeos 35

1.2.1.2 Peptídeos não ribossômicos 35

1.2.1.3 Alcaloides 36

1.2.1.4 Terpenos 36

1.2.2 Biossíntese de metabólitos secundários 37

1.2.2.1 Identificação bioquímica 37

1.2.2.2 Genômica para a descoberta de metabólitos secundários . 39

1.2.3 Predição de metabólitos secundários *in silico* 41

1.2.4 Histórico da lovastatina 44

1.2.4.1 Bases Bioquímicas e Moleculares da Lovastatina 46

1.2.4.2 LovB e LovC 46

1.2.4.3 LovF e LovD 48

1.2.4.4 LovA 49

1.2.4.5 LovG (ORF5) 49

1.2.4.6 Outros genes do agrupamento 49

2. Objetivos 51
2.1 Objetivo geral 51
2.2 Objetivos específicos 51

3. Métodos 53
3.1 Seleção dos Isolados 53
3.1.1 Sequenciamento de DNA 54
3.1.2 Genoma de referência *A. terreus* NIH 2624 54
3.2 Mapeamento dos dados de resequenciamento 55
3.2.1 Mapeamento das leituras contra genoma referência NIH 2624 55
3.2.2 Detecção e Genotipagem das variantes genéticas 55
3.3 Montagem *de novo* dos genomas 56
3.3.1 Avaliando a completude dos genomas montados 56
3.3.2 Anotações genômicas 57
3.3.3 Predição de metabólitos secundários nas montagens *de novo* 58
3.3.4 Alinhamento Múltiplo dos Genomas 58
3.3.5 Visualização com curva tipo Hilbert 58
3.3.6 Comparando a região genômica de produção de lovastatina entre as cepas 60
3.3.7 Reproducibilidade das etapas 60
3.3.8 Teste de significância entre múltiplos conjuntos 60

4. Resultados 63
4.1 Quantificação dos níveis de lovastatina 63
4.2 Sequenciamento das cepas de *A. terreus* 64
4.3 Mapeamento das leituras-curtas contra o genoma de referência 65
4.4 Desenho de primers para o agrupamento de biossíntese de lovastatina . . . 67
4.4.1 Desenvolvendo algoritmos baseados em PO-MSA para desenho de primers 68
4.4.2 Validação dos primers desenhados por PO-MSA 74
4.4.3 Análise global de polimorfismo de sequências 77
4.4.4 Análise local de polimorfismo de sequências no agrupamento de biossíntese de lovastatina da ATCC 20542 78

4.5 Análises baseadas nas montagens *de novo* 79
4.5.1 ATCC 20542 80
4.5.2 U26: Avaliando a montagem genômica 81
4.5.3 BU35, BU33, BU27, U9, U10 e U22 81
4.5.4 O agrupamento de biossíntese de lovastatina 83
4.6 Alinhamento Múltiplo dos Genomas 84
4.7 Predição de agrupamentos de genes biossintéticos 90
4.7.1 Predição de BCGs com ferramentas tradicionais 90
4.7.2 Metodologia para detecção de BCGs usando estruturas de grafo de Bruijn multicolorido 92
4.7.3 Identificação de loci referência putativamente envolvidos em vias de biossíntese de MSs em *Aspergillus terreus* 96
4.7.4 Identificando o agrupamento gênico de biossíntese de acetilaranotina 98
4.7.5 Identificando o agrupamento gênico de biossíntese de terretonina . . 98
4.7.6 Identificando o agrupamento gênico de biossíntese de lovastatina . . 99
4.7.7 Consenso dos pan-genes da cepa NIH 2624 100

5. Discussão 105
5.1 O resequenciamento de cepas de *A. terreus* 106
5.1.1 Metodologia para detecção de anomalias de mapeamento 107
5.1.2 Mapeamento no BCG de lovastatina 108
5.1.3 Montagem genômica *de novo* das cepas 110
5.1.3.1 Qualidade das montagens *de novo* 111
5.1.4 Exploração de agrupamentos de metabólitos secundários 112

6. Conclusão 119

Referências 121

Apêndice 141

A. Metodologias Adicionais 143
A.1 Cultivo das cepas de *A. terreus* 143

A.2 Extração de DNA Genômico 144
A.3 Extração e Quantificação de lovastatina no sobrenadante 145

B. Tabelas adicionais 147

Capítulo 1

Introdução

1.1 Introdução à Genômica

A evolução tecnológica precursora que iniciou a nova área de investigação biológica, a genômica, foi concretizada em 1995 com a publicação do genoma completo da bactéria *Haemophilus influenzae*, contendo 1.8 milhão de bases (Mp). O feito foi extremamente complexo em razão da limitação técnica imposta pelo método Sanger de sequenciamento (Sanger et al., 1977). O método Sanger, único método de sequenciamento existente na época, consiste na fragmentação aleatoria do genoma formando fragmentos de sequências com poucos milhares de bases, estes são clonados em plasmídeos que são sequenciados gerando algumas centenas de bases efetivamente decodificadas (Hutchison, 2007).

Não obstante, a partir deste feito formou-se os alicerces para a genômica, que objetiva a obtenção e caracterização funcional de sequências genômicas dos organismos. O genoma representaria, em primeira instância, o catálogo de informações fundamentais para o desenvolvimento e manutenção de biomoléculas (RNAs, proteínas) abrangendo o dogma central da Biologia. De posse destas informações acreditava-se que delimitar a relação direta entre genótipo e fenótipo seria elementar.

Atualmente, temos ciência da enorme complexidade dos seres vivos, e que a simples descrição dos componentes (ex. genes) não é suficiente, na maioria das situações, para se elucidar o funcionamento dos sistemas biológicos. Ao mesmo tempo, pode-se argumentar que o conhecimento dos genomas é condição necessária para o entendimento dos sistemas biológicos. Neste contexto, o século XXI iniciou a era genômica da Biologia, onde diversos organismos modelo tiveram seus genomas decifrados, culminando com o próprio genoma humano em 2001 (Lander, 2011). No começo da era genômica, todos os genomas

elucidados foram obtidos por grandes consórcios internacionais utilizando a tecnologia de Sanger, em projetos milionários que consumiam anos de trabalho de dezenas a centenas de pesquisadores.

No ano de 2005 uma publicação descrevendo uma metodologia alternativa (Margulies et al., 2005) ao método de Sanger inaugurou a era das novas tecnologias de sequenciamento de DNA, ou "Next Generation Sequencing"(NGS) (Reuter et al., 2015). As novas abordagens tecnológicas para sequenciamento de DNA diferem do método tradicional de Sanger pela química do processo e pelo processamento simultâneo de um grande número de amostras. Enquanto os sequenciadores automáticos por capilares, que utilizam o método de Sanger, geram até 384 sequências por corrida, este número sobe para centenas de milhares ou milhões com as novas abordagens tecnológicas.

Empresas como Illumina lançaram comercialmente sequenciadores de DNA com tecnologias NGS a partir de 2008, os quais rapidamente foram adotados pela comunidade científica em função do aumento de ordens de magnitude na capacidade de sequenciamento com um efeito inverso no preço por base. Entretanto, apesar da grande quantidade de bases produzidas existe uma limitação marcante em relação à tecnologia Sanger: enquanto esta consegue gerar sequências de até 700 bases por fragmento de DNA lido, as tecnologias NGS utilizadas na última década restringem o sequenciamento a fragmentos curtos de nucleotídeos, também chamados de leituras-curtas (*short reads*) (Goodwin et al., 2016). Embora o avanço tecnológico das plataformas NGS com o passar dos anos contribuiu substancialmente para o incremento do tamanho das sequências geradas sem aumentar a taxa de erro (Kircher e Kelso, 2010), aumento da quantidade de sequências geradas (*throughput*) e, indiscutível redução dos custos, a plataforma mais utilizada mundialmente - Illumina - ainda limita-se ao sequenciamento máximo de 300 nucleotídeos por fragmento, obtido, pela versão mais recente (versão 3) do sequenciador Illumina MiSeq (tabela 1.1).

Recentemente, uma nova linhagem de sequenciadores conhecidos como plataformas de terceira geração surgiu. Estas novas plataformas de sequenciamento diferenciam-se, principalmente, pela capacidade de produzir leituras longas (*long reads*), apesar da desvantagem de altas taxas de erros na imputação das bases sequenciadas (Wang et al., 2015). As leituras-longas são resultado de novas abordagens de sequenciamento na qual sequenciam-se moléculas únicas de DNA em contrapartida ao *pool* necessário pelas químicas de sequenciamento-por-síntese e sequenciamento por ligação utilizados pelas plataformas de NGS Illumina e

SOLiD, respectivamente (Metzker, 2010).

Tabela 1.1 - Tabela comparando os principais atributos de cada plataforma de sequenciamento a partir dos dados compilados por Goodwin et al. (2016). A sigla PE significa *paired-end*, pb - pares de base, indel - pequenas inserções e deleções na sequência. Os prefixos K, M e G são as siglas para as gradezas científicas 10^3, 10^6 e 10^9 respectivamente.

Plataforma	Tamanho da leitura (pb)	Capacidade	Leituras	Tempo da corrida	Perfil do Erro	Preço por Giga Bases (US$)
Illumina MiSeq v2	250 (PE)	7.5-8.5 Gb	24-30 M (PE)	39 h	0.1%, substituições	$142
Illumina MiSeq v3	300 (PE)	13.2-15 Gb	44-50 M (PE	26 h		$110
Illumina HiSeq2500	150 (PE)	75-90 Gb	600 M (PE)	40 h		$ 40
Pacific BioSciences RS II (PacBio)	∼20 Kb	500 Mb-1G	∼55.000	4 h	13%, indel	$1000
Oxford Nanopore MK 1 MinION	até 200Kb	até 1.5Gb	>100.000	até 48 h	12%, indel	$750

Entre as plataformas de sequenciamento de leituras longas destacam-se o PacBio (Eid et al., 2009) e o MinION da companhia Oxford Nanopore (Clarke et al., 2009). Apesar da taxa de erro dessas plataformas, cerca de 10%, elas são capazes de gerar sequências na faixa de 55.000 a mais de 100.000 pares de bases (Goodwin et al., 2016). Enquanto as plataformas de leituras curtas garantem mais confiança nas bases, as de leituras longas permitem a amostragem de grandes extensões genômicas contínuas. Publicações recentes demonstram que o uso conjunto das abordagens, gerando leituras curtas e longas, são suficientes para montar genomas *de novo* com qualidade comparável aos genomas montados pela convencional tecnologia Sanger (Utturkar et al., 2014).

A redução no custo de sequenciamento de DNA teve uma queda brusca nos anos recentes, porém, este ainda é fator limitante na quantidade de sequências geradas por um experimento e, com isso, no impacto das respostas biológicas obtidas (Sims et al., 2014).

Em experimentos que envolvem a produção de sequências de nucleotídeos para responder hipóteses científicas os pesquisadores devem elaborar o desenho experimental de modo a otimizar a obtenção das respostas biológicas em questão e o custo do experimento. Inevitavelmente, algumas perguntas requerem mais dados para respondê-las e, consequentemente, maior é o custo. Por exemplo, se a questão proposta é analisar a fundo as bases moleculares que determinam o fenótipo de resistência à antibióticos de uma bactéria, sendo que esta não possui genoma de referência publicado, o desenho experimental requer a produção de grande quantidade de dados referentes ao DNA genômico desse indivíduo para garantir a correta montagem *de novo* a nível nucleotídico. Por outro lado, se o foco da pergunta é a

caracterização das variantes genéticas de polimorfismos únicos de sequência (SNPs) entre indivíduos de uma população - com genoma de referência publicado - o desenho experimental altera-se a fim de garantir maximização da quantidade de indivíduos sequenciados e, consequentemente, aumentar a probabilidade de amostrar alelos raros na população.

1.1.1 Montagem de genomas

A metodologia de sequenciamento ideal seria aquela que, sem erros, conseguisse determinar as bases do começo ao fim de um cromossomo. E, de forma serial, dos próximos cromossomos até obter o genoma completo do indivíduo sequenciado. Com isso, todas as variantes polimórficas de genomas diplóides ou poliplóides seriam identificadas, pois saberíamos exatamente a posição e identidade da base determinada, e as longas sequências de nucleotídeos idênticas ou quase idênticas que repetem-se ao longo dos genomas - os elementos repetitivos - poderiam ser atribuídas, de forma precisa, às posições genômicas.

Muitas limitações impostas pela metodologia de mapeamento podem ser resolvidas através da montagem *de novo* dos genomas utilizando as leituras curtas (Chaisson et al., 2015). A abordagem algorítmica usada na montagem de genomas com dados de leituras curtas diferencia-se da abordagem usada anteriormente quando as leituras eram provenientes da tecnologia Sanger (Miller et al., 2010). Anteriormente o algoritmo de montagem, conhecido como "overlap-layout-consensus"(OLC), baseava-se na sobreposição das leituras para posicioná-las e, posterior elucidação da sequência consenso dado as informações das sobreposições (Myers et al., 2000). Com a introdução das tecnologias de leituras-curtas, a abordagem dos algoritmos foi alterada. Os algoritmos modernos de montagem não usam toda a extensão das leituras mas baseiam-se na clivagem destas ao longo de janelas sobrepostas contendo k nucleotídeos, conhecidas como k-mers (Goodwin et al., 2016). Portanto, para representar um genoma através dos seus k-mers forma-se um grafo cujos nós são o prefixo e sufixo dos k-mers. O prefixo e sufixo representam respectivamente as k-1 primeiras bases do k-mer e as k-1 últimas. Liga-se, por exemplo, uma aresta direta entre o nó A e B quando existe um k-mer com prefixo igual a A e um sufixo igual a B. Repetindo essa operação para os n k-mers do conjunto de todos os *k-mers* descendentes das leituras-curtas constrói-se um grafo direto (Compeau et al., 2011), conhecido como grafo de Bruijn, em homenagem ao matemático holandês que o propôs para solucionar o seguinte problema: encontrar a menor superpalavra circular que contêm todas as possíveis subpalavras de ta-

manho k (*k-mers*) dado um alfabeto. Isso equivale a trilhar um caminho que visite todas as arestas do grafo uma única vez. A estrutura grafo de Bruijn foi aplicada nas montagens de genomas *de novo* com a finalidade de substituir a clássica abordagem OLC (Pevzner et al., 2001).

Alguns fatores fundamentais impactam a qualidade das montagens *de novo*: a cobertura de sequenciamento, a taxa de erro e tamanho da leitura resultante da metodologia de sequenciamento e a complexidade das sequências repetitivas existente no genoma a ser montado (Alkan et al., 2011).

A cobertura teórica ou esperada refere-se ao número médio de vezes que espera-se sequenciar uma base nucleotídica dado uma quantidade de sequências geradas com certo tamanho, pressupondo-se que as sequências geradas são uniformemente distribuídas ao longo de um genoma, de acordo com o modelo matemático de Lander e Waterman (Lander e Waterman, 1988). Na prática, a obtenção de coberturas uniformes ao longo de toda extensão do genoma não é o usual devido a dificuldades técnicas intrínsecas das plataformas de sequenciamento (Goodwin et al., 2016). No caso dos sequenciadores da plataforma Illumina, a preparação da biblioteca é um processo de múltiplas etapas na qual a última envolve amplificação via *Polimerase Chain Reaction* (PCR) antes de inserir as amostras no sequenciador (Kozarewa et al., 2009). Segmentos com baixa complexidade de sequência, ou seja, que possuem alto desvio na porcentagem de conteúdo GC são menos amostrados (Liu et al., 2012; Chen et al., 2013).

O viés na etapa de preparação da amostra a ser sequenciada somado aos erros impostos pelas plataformas de sequenciamento na determinação da base nucleotídica contribuem para a amostragem desigual dos fragmentos genômicos sequenciados (Goodwin et al., 2016).

Portanto, em detrimento ao termo cobertura teórica proposto por Lander e Waterman (1988) usa-se o termo cobertura empírica. O último termo refere-se ao número de vezes reais que determinada base do genoma referência é amostrada observando-se as sequências mapeadas com alto grau de confiança contra o genoma [1] (Sims et al., 2014).

Apesar da boa qualidade das montagem genômicas com 8-10X de cobertura usando a

[1] Em estudos NGS e, no decorrer dessa dissertação, a palavra cobertura será usada como sinônimo de cobertura empírica. Ou seja, a cobertura média esperada de nucleotídeos por sequências mapeadas numa referência (loco, genoma ou fragmento de genoma). Quando o alvo do mapeamento não for claro, este será escrito explicitamente.

tecnologia Sanger, as montagens equiparáveis usando dados de NGS requerem, aproximadamente, 70X de cobertura (Sims et al., 2014). Entre as consequências da baixa cobertura de sequenciamento está a errônea interpretação biológica das análises subsequentes. Isto é, analisando-se resultados de NGS com baixa cobertura não é possível determinar se a ausência de um gene codificador de proteína, ou uma quebra no quadro de leitura de uma ORF, representam deficiências na montagem devido não amostragem da sequência ou eventos evolucionários reais de perda de DNA. Outro acontecimento que pode levar a interpretações biológicas errôneas, principalmente na busca de variantes genéticas, é a propagação de erros inerentes às plataformas de sequenciamento (Sims et al., 2014).

O tamanho das leituras também impacta negativamente as montagens genômicas, e a disponibilidade de apenas sequências curtas impõe que a montagem resultante seja fragmentada em diversos *contigs*, independente da cobertura de sequenciamento (Alkan et al., 2011; Birney, 2011; van Dijk et al., 2014). A complexidade dos elementos repetitivos presentes no genoma é fator que, em ocasiões cujo tamanho da sequência repetitiva é maior que o tamanho da leituras-curtas, impossibilita a montagem das sequências ainda que a taxa de cobertura seja altíssima resultando em montagens fragmentadas ou quiméricas (Treangen e Salzberg, 2011).

1.1.2 Estudos de resequenciamento

As tecnologias de sequências curtas (NGS), especificamente a tecnologia Illumina domina o mercado de sequenciamento de DNA, devido à alta capacidade de geração de dados em poucos dias e o relativo baixo custo (Reuter et al., 2015). Apesar da grande quantidade de sequências produzidas, as montagens genômicas usando dados de NGS tem deficiências devido, principalmente, dois fatores intrínsecos à tecnologia. O primeiro refere-se ao viés no conteúdo GC e o segundo fator ao tamanho da leitura que dificulta a alocação de regiões repetitivas. Entretanto, as tecnologias NGS despontam com amplo nicho de aplicações na área de genômica comparativa, especialmente, na busca de variantes genéticas (Alkan et al., 2011; Sims et al., 2014).

Portanto, a tarefa de buscar polimorfismos entre indivíduos cujo genoma de referência da espécie encontra-se montado foi ampliada com as tecnologias NGS. A busca de variantes entre os genomas não necessita da montagem da sequência genômica bastando a ancoragem das leituras curtas no genoma de referência, surgindo com isso as metodolo-

gias de alinhamento escalonadas para grande quantidade de dados (Trapnell e Salzberg, 2009). Para tanto, foram desenvolvidos novos programas, denominados "mapeadores de genomas" que conseguem encontrar rapidamente os pontos em um genoma com sequência disponível, o genoma de referência, onde se ancoram milhões de pequenas sequências do genoma desconhecido, nas posições conservadas (Sims et al., 2014).

A medida que elucidou-se os genomas de diversas espécies, houveram tentativas de catalogar as variações genéticas entre indivíduos da mesma espécie, utilizando a estratégia acima, também chamada de resequenciamento genômico. O "1000 genomes project"(Durbin et al., 2010), onde se catalogou a variação genética em diversas populações humanas por resequenciamento é um exemplo marcante desta abordagem.

As plataformas de NGS propiciaram uma explosão na quantidade de estudos de resequenciamento de DNA cujo objetivo principal é a detecção das variantes genéticas entre indivíduos (Chaisson et al., 2014), acelerando a taxa de caracterização das variantes genéticas presentes em populações mostrando-se eficiente na determinação dos polimorfismos únicos de sequência (SNPs) responsáveis por doenças genéticas (Koboldt et al., 2014), imputação do grau de diversidade em populações (McVean et al., 2012) e estudos relacionados à descoberta de variantes somáticas (Chang et al., 2013). Contudo, a sensitividade da metodologia de resequenciamento é reduzida na identificação de pequenas inserções e deleções no genoma (INDELs) e detecção de grandes variantes estruturas (VEs) (Chaisson et al., 2014; Ross et al., 2013). Iqbal et al. (2012) listaram três limitações principais que podem ser obstáculos nas análises de dados oriundos de resequenciamento de DNA.

A primeira limitação refere-se ao requisito mínimo da existência do genoma de referência montado e devidamente anotado da espécie estudada no qual as leituras-curtas serão mapeadas. A segunda limitação torna-se aparente quando amostras sequenciadas contêm extensões de sequências genômicas ausentes ou divergentes da referência (Holtgrewe et al., 2015; Chaisson et al., 2015). Nesse caso, as leituras-curtas não serão mapeadas (Iqbal et al., 2012). A terceira limitação impacta negativamente a qualidade dos mapeamentos é decorrente da imputação das variantes, inflando o número de falso-positivos devido mapeamento a regiões parálogas (Iqbal et al., 2012). Observa-se a terceira limitação em genomas com regiões altamente variáveis (Holcomb et al., 2011; Huang et al., 2014) ou com densidade elevada de sequências repetidas em série (*tandem repeats*) (Trapnell e Salzberg, 2009; Treangen e Salzberg, 2011).

1.2 Genomas fúngicos: estrutura e relevância biotecnológica

Existem espécies com características intrínsecas ao seus genomas que acentuam o grau de impacto das limitações impostas pelas leituras curtas, dentre as quais destacamos as espécies pertencentes ao reino Fungi. Ainda que o número de genomas de fungos montados e disponíveis publicamente vêm aumentando ao longo dos últimos anos devido ao incentivo de grandes consórcios - como o projeto 1000 genomas fúngicos (Grigoriev et al., 2014; Sharma, 2015) - a quantidade de genomas disponíveis desse reino ainda é relativamente escassa (Grigoriev, 2013). Portanto, a ausência de genomas de referência e má qualidade das anotações dos genomas fúngicos (Sharma, 2015; Abbas et al., 2014) acentuam a primeira limitação. Outro fator característico dos genomas fúngicos e possivelmente o mais impactante na etapa de mapeamento, em especial nos genomas das espécies pertencentes ao gênero *Aspergillus*, são a elevada incidência de genes cepa específicos (Nierman et al., 2005; Rokas et al., 2007; Fedorova et al., 2008). Isso corrobora para acentuar a fração de leituras curtas não mapeadas contra o genoma de referência (Iqbal et al., 2012).

Pode-se afirmar que o ponto de partida para o entendimento dos genomas fúngicos e de conceitos de biologia envolvendo organismos eucarióticos foi o sequenciamento da levedura *Saccharomyces cerevisiae* cepa S288C (Goffeau et al., 1996). Subsequentes sequenciamentos de espécies correlatas demonstraram adicional importância de análises comparativas na identificação de mecanismos moleculares que envolvem a evolução dos genomas. Apesar do número considerável de genomas montados de espécies do subfilo *Saccharomycotina* a taxa de sequenciamento das outras famílias fúngicas não acompanhou o mesmo ritmo (Sharma, 2015). A maior parte dos genomas fúngicos publicamente disponíveis, quando não pertencentes ao gênero *Saccharomyces*, eram de espécies envolvidas com a saúde humana como *Candida albicans* (Jones et al., 2004), *Aspergillus fumigatus* (Pain et al., 2004), *Cryptococcus neoformans* (Loftus, 2005). Somente na última década grandes consórcios formados pelos institutos *Broad* e o *US Department of Energy Joint Genome Institute* (JGI) propuseram amostrar a diversidade do Reino dos Fungos através de projetos como o 1000 genomas fúngicos (Sharma, 2015). Essas iniciativas permitem a cada dia incrementar a coleção de informações desse reino caracterizado por espécies com diferentes estilos de vida.

O gênero *Aspergillus* claramente exemplifica a diversidade no padrão de vida dos fun-

gos e sua importância direta para os seres humanos. Descrito aproximadamente há 300 anos pelo padre e botânico Antonio Micheli que nomeou o gênero dado a similaridade da sua estrutura formadora de esporos com o instrumento chamado aspergillum, usado para dispersar água benta (Gibbons e Rokas, 2013). No clássico tratado do gênero, Raper e Fennel (1965) chegou a estimativa de 250 espécies compondo este gênero. Entretanto, 50 novas espécies foram adicionadas ao gênero nesse século devido otimização de metodologias de classificação taxonômica e uso de sistemáticas envolvendo sequências genômicas (Geiser et al., 2007). Dentre as espécies pertencentes a esse gênero algumas merecem importância adicional dado a correlação direta com os interesses humanos. Os fungos da espécie *A. fumigatus* são importantes patógenos humanos responsáveis pelo maior número de mortes e segundo maior número de infecções causadas por fungos (Gibbons e Rokas, 2013), tendo seu genoma elucidado em 2004, o primeiro do gênero devido sua importância médica (Pain et al., 2004). De patógeno humano a peste agricultural a espécie *A. flavus* contamina diversas culturas com a potente aflatoxina, causando prejuízos na produção agrícola e algumas poucas mortes humanas por ano (Yu et al., 2011). O nicho de atuação do gênero foi demonstrado pela espécie *A. sydowii* que infectou comunidades de corais Caribenhos ameaçando todo um ecossistema (Rypien et al., 2008). Mudando o foco da importância patogênica está a espécie *A. niger* produtora de diversas moléculas base usadas na indústria e atestado como segura para animais e plantas através da certificação GRAS (*Generally Recognized As Safe*) emitida pelo órgão governamental norte-americano *Food And Drug Administration* (FDA), responsável pela liberação de alimentos, cosméticos e medicamentos para utilização humana (Pel et al., 2007). As espécies *A. oryzae*, *A. sojae* e *A. kawachii* asseguram a produção de diversas bebidas e molhos típicos dos países do extremo leste como o sake, molho de soja e a bebida típica japonesa shochu (Machida et al., 2008). A importância para a genética é atestada pela espécie modelo *A. nidulans* (Galagan, 2005) e o valor farmacêutico do gênero confiado à espécie *A. terreus* que produz a molécula lovastatina usada mundialmente para reduzir os níveis de colesterol humano (Greenspan e Yudkovitz, 1985a; Treiber et al., 1989).

Depois de uma década após o sequenciamento da primeira espécie, alguns bancos de dados especializados no gênero Aspergillus foram criados para armazenar e disponibilizar publicamente as informações à medida que outras espécies de Aspergillus são sequenciadas. Entre os bancos destacam-se o repositório *Central Aspergillus Resource* (CADRE) (Mabey

Gilsenan et al., 2012), o portal *Aspergillus Genome Database* (AspDB) (Cerqueira et al., 2014) e embora não especializado no gênero o portal de *MycoCosm* (Grigoriev et al., 2014) armazena diversos genomas fúngicos.

A alta diversidade no gênero *Aspergillus* foi demonstrada nas análises comparativas das sequências proteicas das espécies "muito próximas"*A. fumigatus* e *A. fischerianus*. Rokas et al. (2007) identificou grau de similaridade tão discrepante quanto os observados entre homens e peixes. Entretanto, contrariamente ao esperado, a estrutura dos genomas é extremamente conversada entre as espécies do gênero Aspergillus sendo que todos os genomas das espécies do gênero avaliados até a data são constituídos por 8 cromossomos e têm tamanho total variando entre 28 até 40 milhões de pares de base (Mpb) (Gibbons e Rokas, 2013). A tecnologia de sequenciamento Sanger ainda é a mais usada pelos grandes consórcios nas etapas de montagem de genomas completos, embora a montagem de algumas espécies de Aspergillus usaram a combinação de leituras geradas pela tecnologia Sanger e 454.

Outra característica constatada no genoma das espécies de *Aspergillus* é a falta de associação entre o estilo de vida e afinidade evolucionária. Por exemplo, *A. oryzae*, espécie domesticada de *A. flavus*, compartilham 99.5% de identidade entre seus genomas (Rokas et al., 2007). Apesar do alto grau de sintenia entre os genomas, *A. oryzae* é usado em processos de fermentação e mantêm a designação de microrganismo seguro (GRAS) enquanto *A. flavus* produz a carcinogênica molécula aflatoxina (Gibbons et al., 2012). Reciprocamente, as espécies *Aspergillus fumigatus*, *A. flavus* e *A. terreus* comumente caracterizadas como patogênicas para o homem não agrupam-se na árvore filogenética proposta para o gênero (Peterson, 2008). A ausência de conexões ligando estilo de vida e afinidade evolucionária entre as espécies de *Aspergillus* é atribuída à capacidade de biossíntese de metabólitos secundários (MS) específicos para cada espécie (Gibbons et al., 2012). Por exemplo, aproximadamente, 80% dos genes são compartilhados entre *A. fumigatus*, *A. clavus* e *A. fischerianus*. Entretanto, quando analisam-se os genes classificados como essenciais para a biosíntese de metabólitos secundários somente 30% deles são conservados entre as espécies (Fedorova et al., 2008).

1.2.1 Metabólitos Secundários

Metabólitos secundários (MS) são moléculas heterogêneas de baixo peso molecular e, alternativamente aos metabólitos primários, não estão diretamente envolvidas no crescimento do microrganismo que as produz. As ações dos metabólitos secundários são diversas no contexto do próprio organismo, e pesquisadores reportaram diversas atividades heterólogas, como ações antibacterianas, antifúngicas, antitumorais, citotóxicas, teratogênicas e mutagênicas (Bennett e Bentley, 1989).

A biosíntese dos MSs envolve a condensação repetitiva de estruturas precursoras padrões oriundas do metabolismo primário como aminoácidos, cadeias de carbono simples como malonil/acetil, isoprenos e seus derivados, de forma semelhante ao anabolismo de ácidos graxos (Brakhage, 2013; Demain, 2014). Reportam-se 4 classes principais de enzimas envolvidas na produção de metabólitos secundários, o que fornece uma classificação para os próprios MSs (Keller et al., 2005a), detalhada a seguir.

1.2.1.1 Policetídeos

Os policetídeos são os metabólitos secundários mais abundantes nos fungos e pertencem à esta classe o carcinógeno aflatoxina (Payne e Brown, 1998), a clinicamente usada para reduzir os níveis de colesterol lovastatina (Kennedy, 1999) e o pigmento amarelado do fungo *A. nidulans* naftopirona (Fujii et al., 2001). Policetídeos são sintetizados por enzimas multimodulares chamadas policetídeos sintases (PKS), em fungos, as PKS são iterativas em contrapartida à ação das PKS de bactérias. Os módulos das PKS iterativas são reutilizados na formação de um único MS enquanto as PKS modulares produzem MSs através da ação catalítica sequencial de seus módulos (Keller et al., 2005a).

1.2.1.2 Peptídeos não ribossômicos

As enzimas da categoria peptídeo sintases não-ribossômicos (NRPS) são as responsáveis pela produção da classe de MSs denominada peptídeos não ribossômicos que englobam importantes moléculas de interesse humano como os antibióticos penincilina e cefalosporina, os imunossupressores ciclosporina e a gliotoxina (Brakhage, 2013).

1.2.1.3 Alcaloides

Já a classe de metabólitos secundários alcaloides, normalmente derivados de triptofano ou do pirofosfato de dimetilalilo (intermediário da via de mevalonato), são biossintetizados por enzimas triptofano dimetilalil sintase (DMAT) (Keller et al., 2005a). Pertencem à essa classe os alcaloides de ergot, conjunto de MSs produzidos pelo fungo *Claviceps*, identificados por cientistas do século XX para aplicação em diversos distúrbios neurológicos, embora historiadores sustentam a utilização prévia desses alcaloides na Grécia antiga para induzir alucinações durante rituais holísticos (Bennett e Bentley, 1999).

1.2.1.4 Terpenos

Outra classe de metabólitos secundários, os terpenos, é bem conhecida em plantas apesar dos fungos produzirem terpenos importantes como, por exemplo, as giberelinas, carotenoides e tricotecenos (Keller et al., 2005a).

A existência de moléculas híbridas (PKS-NRPS) com domínios conjuntos classificados como policetídeos e peptídeos não-ribossômicos adiciona mais um nível para a diversidade de metabólitos secundários encontrados na natureza (Brakhage, 2013).

O mecanismo de biossíntese dos metabólitos secundários para as classes PKS, NRPS e PKS-NRPS envolvem, principalmente, a ação de enzimas multimodulares. As policetídeos sintases (PKSs) e peptídeos sintases não-ribossômicos (NRPSs) catalizam a formação do metabólito através da utilização iterativa dos seus módulos (Keller et al., 2005a). Para realizar essa função as PKS típicas possuem no mínimo 3 módulos essenciais (figura 1.1): domínio aciltransferase (AT) permitindo seleção da subunidade padrão e transferência para o domínio proteína carreadora de acil (ACP) que carregará covalentemente o bloco formado. A medida que este é estendido iterativamente pela ação do domínio cetosintase (KS) pelas inúmeras reações de condensação descarboxilativa da unidade extensora ao bloco (usualmente malonil-CoA ou metilmalonil-CoA) (Brakhage, 2013). Reciprocamente, as NRPS também são constituídas de, no mínimo, 3 módulos: o domínio de adenilação responsável pela ativação do aminoácido, o domínio proteína carreadora de peptídeo para ligação covalente do aminoácido ativado e o domínio de condensação que catalisa a formação da ligação peptídica (figura 1.1). Tipicamente, as PKS e NRPS possuem módulos

extras, além dos 3 essenciais, que realizam modificações nos intermediários da formação do bloco como módulos com funções desidratase, enoil-redutase, metil-transferase e epimerização (Brakhage, 2013). Enzimas auxiliares às multimodulares personalizam o policetídeo ou peptídeo nascente das PKS ou NRPS, respectivamente, conferindo funcionalidades estruturais ao produto final através de, por exemplo, ação de oxidases.

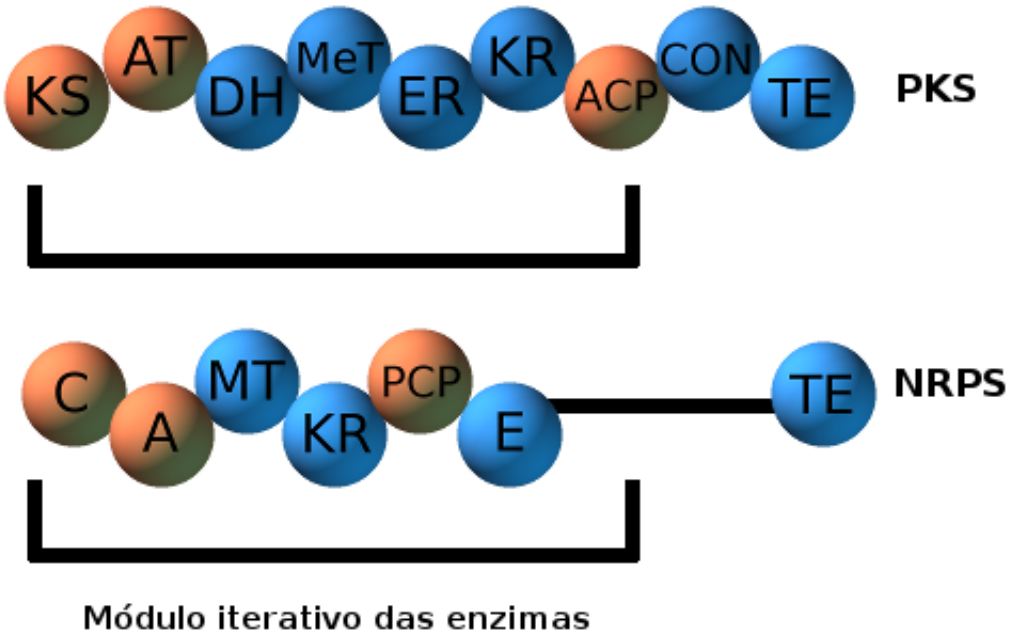

Figura 1.1: Diagrama dos domínios essenciais (em laranja) e opcionais (em azul) detectados nas enzimas policetídeos sintases (PKSs) e peptídeos sintases não-ribossômicos (NRPSs), além da identificação dos módulos com função iterativa.

1.2.2 Biossíntese de metabólitos secundários

1.2.2.1 Identificação bioquímica

As primeiras identificações de metabólitos secundários baseavam-se no teste de diversos compostos purificados ou pertencentes a extratos celulares contra alvos a fim de determinar possíveis ações (Bennett e Bentley, 1989). Caso algum composto identificado exibisse ação relevante, este era caracterizado quimicamente através de esforços intensos e custosos (Bentley, 1999). Como no caso da descoberta da compactina que culminou no *screening* de 3.600 cepas fúngicas à procura de potenciais inibidores da enzima chave na produção de colesterol HMG-CoA redutase (Endo et al., 1976). Atualmente a estratégia mais utili-

zada para caracterizar os genes responsáveis pela biossíntese de MSs é a análise do perfil metabólico via cromatografia líquida de alta eficiência (HPLC) ou espectrômetro de massa de cada cepa knock-out resultante da deleção de genes possivelmente ligados à produção do MS (Guo et al., 2012, 2013; Andersen et al., 2013; Yin et al., 2016).

Apesar da descoberta de alguns compostos que inibem HMG-CoA redutase entre as cepas testadas por Endo et al. (1985), possivelmente a biossíntese de outros putativos inibidores estavam desestimuladas. Sabe-se hoje em dia da conexão entre estímulos externos e produção de MSs pelos fungos, gerando um grande problema para a identificação dos MSs, pois deve-se estabelecer condições fisiológicas propicias para que os microrganismos ativem sua produção (Brakhage, 2013; Cacho et al., 2015; Fischer et al., 2016). Isso envolve otimização das condições de cultivo, como temperatura e composição do meio de cultura. Os metabólitos secundários que não são sintetizados, isto é, encontram-se inativos em determinadas condições são denominados crípticos.

Entre as estratégias para ativar a biossíntese de MSs crípticos sem otimizar meios de cultura estão a super-expressão de genes regulatórios (Weber, 2014a). Os genes necessários para a biossíntese de um metabólito, normalmente, são controlados por reguladores codificados por duas classes de genes. Os genes que regulam somente um MS são classificados como reguladores específicos enquanto os reguladores globais ativam ou inibem tanto a produção de MS quanto de moléculas pertencentes ao metabolismo primário (Brakhage, 2013). Genes que codificam os reguladores específicos, normalmente, tem o motivo de ligação ao DNA tipo *Zinc finger* (Fischer et al., 2016), e sua expressão ativa a transcrição coordenada do conjunto gênico responsável pela produção do metabólito secundário (Weber, 2014a). Além dos reguladores específicos, a produção de MSs está atrelada a resposta de diversos sinais ambientais. A ativação do fator de transcrição AreA é induzida pela disponibilidade de nitrogênio no meio de cultivo e, além de afetar a fisiologia do organismo também ativa a produção de diversos metabólitos secundários (Fischer et al., 2016). Cepas de *Aspergillus nidulans* mutantes *knock-out* da proteína nuclear LaeA apresentaram decréscimo na produção de diversos MSs (Bok e Keller, 2004). Como LaeA possui domínio metiltransferase sugeriu-se que ela impacta a expressão dos genes responsáveis pela produção de MSs via modificações das estruturas da cromatina (Fischer et al., 2016).

Portanto, a troca das sequências promotoras de genes reguladores específicos por promotores fortes e induzíveis, por exemplo, promotor do gene álcool desidrogenase *alcA*, é

uma das estratégias usada rotineiramente para ativar a expressão de genes reponsáveis pela produção de MSs e, consequentemente, identificar novos metabólitos secundários em *Aspergillus sp.* (Bergmann et al., 2007; Maiya et al., 2006; Fischer et al., 2016). Entretanto, para resultado satisfatório na super expressão dos reguladores específicos faz-se necessário, primeiramente, definir a sequência deste gene e, em cerca de 50% dos casos, os MSs não possuem regulador específico conhecido (Unkles et al., 2014). Para superar este obstáculo pode-se também expressar de forma heteróloga os genes responsáveis pela biossíntese de um MS. Neste caso, mais uma vez, a sequência dos genes essencias à produção do MS deve ser conhecida. Com isso, diversas metodologias computacionais que consideram as características genômicas dos fungos foram propostas ultimamente para identificar os genes responsáveis pela produção de metabólitos secundários (Keller et al., 2005a; Sanchez et al., 2012; Cacho et al., 2015).

1.2.2.2 Genômica para a descoberta de metabólitos secundários

Em geral, os genes responsáveis pela biosíntese de um metabólito secundário, tanto em bactérias quanto em fungos, estão localizados no genoma em grupos (*clusters*) onde o gene que codifica a enzima chave para síntese do metabólito, por exemplo o gene PKS ou NRPS, encontra-se adjacente aos genes codificadores das enzimas auxiliares (oxidases, desidratases, tioesterases), das proteínas de transporte e do regulador específico do metabólito (Martín e Gil, 1984). Esses grupos são chamados de agrupamentos de biossíntese de metabólitos secundários (*Biosynthesis Clustered Genes* - BCG). O tamanho dos BCGs diferem mas, em média, estendem-se por mais de 10.000 pares de base no genoma (Bentley, 1999; Sanchez et al., 2012; Hoffmeister e Keller, 2007; Keller et al., 2005a) e, não distribuem-se uniformemente nos cromossomos, pelo contrário, estima-se que têm prevalência por regiões subteloméricas (Palmer e Keller, 2010). Como dito anteriormente, duas classes de reguladores parecem ativar ou inibir a transcrição coordenada dos genes na maioria dos BCGs: o regulador global e o específico (local). Diferentemente do regulador específico, os genes que codificam os reguladores globais localizam-se fora dos BCGs e também regulam diversos genes do metabolismo primário. Já os reguladores locais parecem ser metabólito específicos, ou seja, co-regulam os genes necessários a síntese de determinado MS e, como mencionado, encontram-se inseridos no BCG que regulam (Keller et al., 2005a).

Anteriormente à era genômica, a caracterização dos genes responsáveis pela produção

de MS baseava-se na utilização de sondas de DNA contendo as sequências de domínios conservados de enzimas-chave na síntese, como as PKS e NRPS, para capturar as sequências margeadoras e, subsequente, sequenciamento destas porções de DNA (Hendrickson et al., 1999). A informação genética era, então, utilizada para modificar o microrganismo com a finalidade de aumentar a produção do metabólito secundário em questão, expressá-lo heterologamente em microrganismos modelo, como *Aspergillus nidulans* ou *Saccharomyces cereviseae*, ou modificar a estrutura química do metabólito secundário para gerar novos compostos ativos (Askenazi et al., 2003; Sorensen et al., 2003).

Com o advento da genômica, o foco na identificação dos metabólitos secundários foi alterado (Sanchez et al., 2012). Estratégias alternativas baseadas nas informações contidas em todo o genoma são utilizadas em contraste às buscas experimentais baseadas em fenótipos mensuráveis por meio de testes bioquímicos ou por estudos de genética com sondas. A disponibilidade de sequências genômicas permite uma busca computacional pelos agrupamentos de genes envolvidos na biossíntese de metabólitos secundários (BCGs), em uma espécie de "descoberta reversa"das enzimas que contribuem para a linha de produção dos compostos. Entre as inúmeras vantagens desta abordagem destacam-se a não necessidade de cultivo do microrganismo, que atualmente restringe o espectro de organismos estudados; a não necessidade de otimizar as condições fisiológicas (composição do meio e do processo de crescimento) para garantir detecção da biossíntese de metabólitos; e potencialização do processo experimental de triagem pois a informação genômica pode contribuir para caracterização em detrimento aos processos custosos e laboriosos de caracterização de amostras aleatórias de cepas.

Em conjunto à aplicação das diversas "ômicas" propiciada, principalmente, pela era genômica nos fungos estabeleceu-se a metabolômica secundária. Ou seja, a busca passou a ser feita predizendo-se os BCGs putativos contidos no genoma de espécies produtoras e associando-se o locus genômico do BCG com o metabólito biossintetizado. Além disso, descobriu-se que a quantidade de BCGs putativos encontrados nos genomas é bem maior do que o número de metabólitos secundários conhecidos (Bergmann et al., 2007; Hertweck, 2009). Isto é, diversos agrupamentos, denominados órfãos, de metabólitos ainda não estão associados aos seus respectivos compostos produzidos. Para se ter uma ideia, somente 50% dos MSs foram associados aos seus respectivos agrupamentos gênicos na espécie modelo *Aspergillus nidulans* (Fischer et al., 2016).

A identificação dos MSs ultimamente é guiada pelas hipóteses produzidas pela exploração dos dados genômicos criando-se a oportunidade de identificar o conjunto de metabólitos produzidos pelo microrganismo estudado. Traçando o perfil de produção de MSs para uma cepa. Esta abordagem representa um novo paradigma de descoberta, pois os genomas podem revelar a presença de uma maquinaria de produção de metabólitos, mesmo sem que se tenha a identificação química destes. Esta exploração ampla e sem viés promete expandir o universo de novos compostos bioativos, e ressalta a importância das análises computacionais para identificação de BCGs em novas sequências genômicas (Keller et al., 2005b; Weber, 2014a).

1.2.3 Predição de metabólitos secundários in silico

A maioria dos programas desenvolvidos para predizer metabólitos secundários *in silico* empregam o conhecimento sobre a arquitetura dos domínios das enzimas-chave na biossíntese do metabólito. Isto é, utilizam conjuntos de sequências de enzimas-chave previamente caracterizadas experimentalmente para gerar perfis de Modelos Markovianos Ocultos (*hidden markov models*-HMM) que podem ser usados para buscar genes envolvidos com a biossíntese de MSs e inferência da classe do metabólito secundário (Khaldi et al., 2010). Analogamente, a busca de genes e, subsequente caracterização do restante dos genes codificadores de enzimas acessórias dos agrupamentos, pode ser realizada via busca por homologia entre o genoma estudado e o repositório contendo sequências de BCGs conhecidas MIBiG (Medema et al., 2015). A maioria das ferramentas desenvolvidas para predição de metabólitos secundários baseadas na interpretação das arquiteturas dos domínios ou homologia com BCGs conhecidos usam as ferramentas HMMer3 (Eddy, 1998) ou BLAST (Altschul et al., 1990) para auxiliar a predição. Com os resultados destas ferramentas, os programas de detecção de MS adicionam informações como a estrutura química do metabólito predito, os limites do agrupamento biossintético ou os genes auxiliares presentes no agrupamento e suas funções (Weber, 2014a).

Algumas ferramentas foram desenvolvidas para predição de classes específicas de MS, como a SEARCHPKS (Yadav et al., 2003) que identificava somente PKS e, em 2004, integrada no sistema NRPS-PKS (Ansari et al., 2004) adicionou a predição de NRPS e, posteriormente, incluiu um módulo para predição baseada na homologia da estrutura 3D dos domínios modulares (Anand et al., 2010). Enquanto a ferramenta NP.search, além de

predizer metabólitos das classes PKS e NRPS, incluiu a funcionalidade de detectar híbridos PKS/NRPS e a tentativa de elucidação da estrutura química do MS predito (Li et al., 2009). Nas enzimas NRPS a composição do sítio ativo do domínio de adenilação determina qual aminoácido será precursor para a formação do peptídeo não-ribossômico. A análise do sítio ativo foi incorporada nos programas PKS/NRPS, NP.search para melhorar a determinação do produto sintetizado pelas NRPS (Weber, 2014a). Ademais, para aumentar a eficiência na predição *in silico* de qual aminoácido será o substrato para a NRPS, a classificação baseada em Máquinas de Vetores Suporte (*Suport Vector Machines* - SVMs) foi usada pelos programas NRPSpredictor, NRPSpredictor2 e ACS pipeline (Rausch, 2005; Rottig et al., 2011; Röttig et al., 2010).

Outras ferramentas desenvolvidas não limitam-se somente à predição das enzimas-chave responsáveis pela biossíntese do metabólito secundário ou do substrato precursor. Elas estendem as predições de enzimas-chave e além disso incluem na análise o papel dos outros genes pertencentes aos agrupamentos de biossíntese de metabólitos secundários (BCG) (Weber, 2014a). Com isso determinam os limites da localização genômica dos BCG e facilitam as anotações de grandes análises genômicas. Nesse sentido a ferramenta web *Secondary Metabolite Unknown Regions Finder* (SMURF) (Khaldi et al., 2010) foi desenvolvida para predizer 4 classes de metabólitos secundários em fungos: policetídeos (PK), peptídeos não-ribossômicos (NRP), alcalóides, híbridos PK-NRP e alcalóides indólicos. A ferramenta não é capaz de predizer metabólitos da classe terpeno devido alto grau de variabilidade entre as enzimas-chave, terpeno ciclases, responsáveis pela produção desse MS (Keller et al., 2005a). A metodologia proposta pelo SMURF é dependente da busca de domínios conservados presentes no conjunto proteico da espécie analisada usando modelos markovianos ocultos (HMMs). Após a busca, uma proteína é considerada PKS se possui ao menos um domínio acil-transferase (AT), um domínio C-terminal beta-cetoacil sintetase C-terminal (KS-C) e um domínio N-terminal beta-cetoacil sintetase (KS-N). Enzimas NRPS são identificadas com pelo menos um domínio adenilação (A), um domínio tiolase (PCP) e um domínio condensação (C). Enzimas híbridas PKS-NRPS são aquelas que possuem ao menos um domínio de cada conjunto de domínios PKS e NRPS citados. Os desenvolvedores do SMURF aumentaram a permissividade na busca de MSs incluindo duas classificações: tipo-PKS e tipo-NRPS. Essas são definidas quando possuem 2 domínios entre os 3 requeridos para predição de PKS e NRPS, respectivamente. Já as enzimas-chave na biossíntese

de alcalóides são preditas pela presença do domínio triptofano dimetilalil sintase (DMAT) (Khaldi et al., 2010; Weber, 2014a).

A ferramenta *antibiotics and Secondary Metabolite Analysis SHell* (antiSMASH) (Medema et al., 2011; Weber et al., 2015) utilizada na identificação e análise de agrupamentos de biossíntese de metabólitos secundários ao longo dos genomas é sem dúvida a mais funcional e, consequentemente, mais empregada ultimamente. Ela incorpora diversos algoritmos e módulos propostos anteriormente para predizer os genes que codificam enzimas-chave na biossíntese de MS (Medema et al., 2011). Isso a torna um compêndio de ferramentas para análise de metabólitos secundários com diversas opções oferecidas para a busca de BCGs e visualização dos resultados via página web interativa. Todas as classes majoritárias de MS são preditas pela ferramenta antiSMASH, incluindo a posição e função dos genes responsáveis pelas enzimas auxiliares, e putativa estrutura química do metabólito sintetizado pelo agrupamento (Weber, 2014a). Além disso, os módulos *ClusterBlast* e *SubCluster Blast* possibilitam a busca por homologia da BCG predita contra banco de dados contendo BCGs conhecidos como o MiBIG (Medema et al., 2015). A versão 3.0 da ferramenta antiSMASH incorporou o módulo *ClusterFinder* (Cimermancic et al., 2014) que têm a capacidade de predizer BCGs pertencentes a classes desconhecidas através de modelo HMM baseado na frequência de domínios PFAM dentro e fora da putativa região de BCG. O conceito explorado pelo *ClusterFinder* é a suposição de que mesmo BCGs que codifiquem classes desconhecidas de metabólitos utilizam, também, o mesmo conjunto de enzimas auxiliares (oxidorredutases, metiltransferases) para a formação do produto (Weber et al., 2015).

Sendo assim, a maioria das ferramentas desenvolvidas para predizer agrupamentos de biossíntese de metabólitos secundários são baseadas na busca dos genes chave na produção via perfil HMM de domínios conservados das enzimas PKS, NRPS ou busca via ferramentas de homologia com agrupamentos conhecidos (Weber, 2014a). Essas ferramentas apresentam dificuldades em identificar BCGs não convencionais, ou seja, que sintetizam metabólitos de classes diferentes às majoritárias PKS, NRPS, terpeno ou DMAT (Andersen et al., 2013; Takeda et al., 2014). Outra dificuldade é a correta identificação dos limites da BCG, isto é, a correta estipulação dos genes membros pertencentes ao agrupamento (Umemura et al., 2013; Inglis et al., 2013). As ferramentas de predição SMURF e antiSMASH tendem a superestimar o número de genes auxiliares que margeiam os genes codificadores das enzimas-chave pois, a identificação deles é também via domínios conservados e, além

disso, alguns genes auxiliares não têm domínio conhecido (Umemura et al., 2015). Algumas ferramentas foram desenvolvidas para superar essas dificuldades e predizer BCGs via metodologias independente de motivos de sequência (*motifs*).

Algumas ferramentas para detecção de BCGs a partir de RNA-Seq foram propostas (Andersen et al., 2013; Umemura et al., 2013). Os algoritmos usados por essas ferramentas baseiam-se, principalmente, em um fato característico das BCGs: a co-regulação dos seus genes membros. Com isso, o transcritoma de duas condições diferentes é analisado para detectar BCGs (Weber, 2014a). Quando as condições de indução de determinado metabólito são conhecidas pode-se comparar a condição ausência *versus* produção do metabólito para identificar o BCG responsável pela sua biossíntese. Como diversas BCGs encontram-se silenciadas até que determinadas condições e sinais ambientais sejam impostos, a maior dificuldade nas detecções via RNA-Seq é estipular o meio de cultura (nutrientes, fonte de carbono) ou condição ideal (pH, temperatura) que garanta a expressão diferencial dos genes das BCGs inativas (Umemura et al., 2015).

Análises de sintenia entre os genomas de *A. oryzae* contra *A. nidulans* e *A. fumigatus* revelaram que blocos não sintênicos (BNS) estão distribuídos em mosaicos ao longo dos genomas. Curiosamente, constatou-se que blocos não sintênicos, ocupando 25% do genoma de *A. oryzae*, estão enriquecidos com genes envolvidos no metabolismo secundário do fungo. Outra característica dos BNS é a grande porção de genes com função desconhecida em comparação aos blocos sintênicos (BS) (Inglis et al., 2013). Tais ilhas genômicas podem ser de considerável importância na detecção de novos metabólitos secundários e já foram notadas em diversas outras espécies de fungos filamentosos (Fedorova et al., 2008). Takeda et al. (2014) propôs a detecção de BCGs identificando-se genes ortólogos nos BNS observados entre a comparação de genomas. Isto é, como nas abordagens com RNA-Seq, a detecção não é baseada nos motivos de sequência das enzimas-chave. Com isso, comparando-se 10 genomas do gênero *Aspergillus* foi possível identificar agrupamentos de biossíntese de metabólitos secundários conhecidos, predizendo seus limites sem grande desvio (Takeda et al., 2014).

1.2.4 Histórico da lovastatina

No final dos anos 70, pesquisadores da empresa farmacêutica Merck encarregados em descobrir moléculas capazes de reduzir os níveis de colesterol sanguíneo isolaram o com-

posto mevinolina da cultura de *Aspergillus terreus*, mais tarde o composto foi renomeado lovastatina (Endo, 2010) . A incumbência em buscar moléculas com esta capacidade estava relacionada à elevada taxa de morte entre pessoas que continham altas concentrações de colesterol sanguíneo. Embora confirmada a ação da lovastatina em reduzir os níveis de colesterol, o composto requer prescrição precoce por parte do médico e uso rotineiro pelo paciente para garantir eficácia no tratamento. Em países como o Brasil, o Sistema Único de Saúde é responsável pela aquisição e distribuição deste medicamento para pessoas carentes, sendo que entre os períodos de 2008 a 2009, o gasto público com derivados de lovastatina (estatinas) foi de cerca de R$ 92 milhões (Brasil, 2009). De tal forma, o Ministério da Saúde em 2010, incluiu os medicamentos da classe estatina na lista de produtos estratégicos para garantir a saúde pública. Com isso, garantir efetiva produtividade da molécula lovastatina é fator chave para reduzir os gastos públicos e melhorar as condições de vida já que segundo dados da OMS, divulgados em 2010, a hipercolesterolemia é responsável por 2,6 milhões de mortes por ano (OMS, 2010). Além disso, as taxas de mortalidade são muito mais elevadas em países de baixa e média renda (OMS, 2011).

Portanto, a seguir detalhou-se o que há de conhecido sobre a biossíntese de lovastatina em *Aspergillus terreus*.

A prospecção em busca de compostos bioativos que reduzem os níveis de colesterol iniciou-se no início dos anos 70 onde Endo e seu grupo identificaram a molécula compactina em extratos de *Penicillium citrinum*. Esta molécula é capaz de inibir a enzima 3-hidroxi-3-metil-glutaril-CoA redutase (HMGR) que converte HMG-CoA em ácido mevalônico na via de produção de colesterol. A busca por outras moléculas inibidoras de HMG-CoA redutase não cessaram e, pesquisadores do grupo Merck Sharp, identificaram em culturas do fungo *Aspergillus terreus* a lovastatina (Alberts et al., 1980). Esses compostos também foram posteriormente identificados em outras culturas de fungos, como em *Monascurs ruber* e *Penicillium brevicomaptum* (Endo, 1979). Tais inibidores, responsáveis por mimetizar o substrato da HMGR, foram coletivamente denominados estatinas.

Atualmente, as estatinas pertencem ao grupo de medicamentos mais prescritos anualmente no mundo, movimentando um mercado de bilhões de dólares (Zamosky, 2012).

1.2.4.1 Bases Bioquímicas e Moleculares da Lovastatina

O metabolito lovastatina é formado pela incorporação consecutiva de nove unidades de acetato na cadeia principal, isto é, um nonacetato, mais a adição de um dicetato via esterificação ao grupo hidroxil na posição 8 da molécula (Greenspan e Yudkovitz, 1985b; Treiber et al., 1989; Endo et al., 1985). A adição de moléculas de oxigênio ao composto ocorre após a ação da enzima policetato sintase (PKS), através da participação da enzima citocromo P450 (CP450). A única diferença entre a compactina e a lovastatina é a adição do grupo metil advindo da molécula S-adenosil metionina (SAM) ao carbono 6 (Mulder et al., 2015).

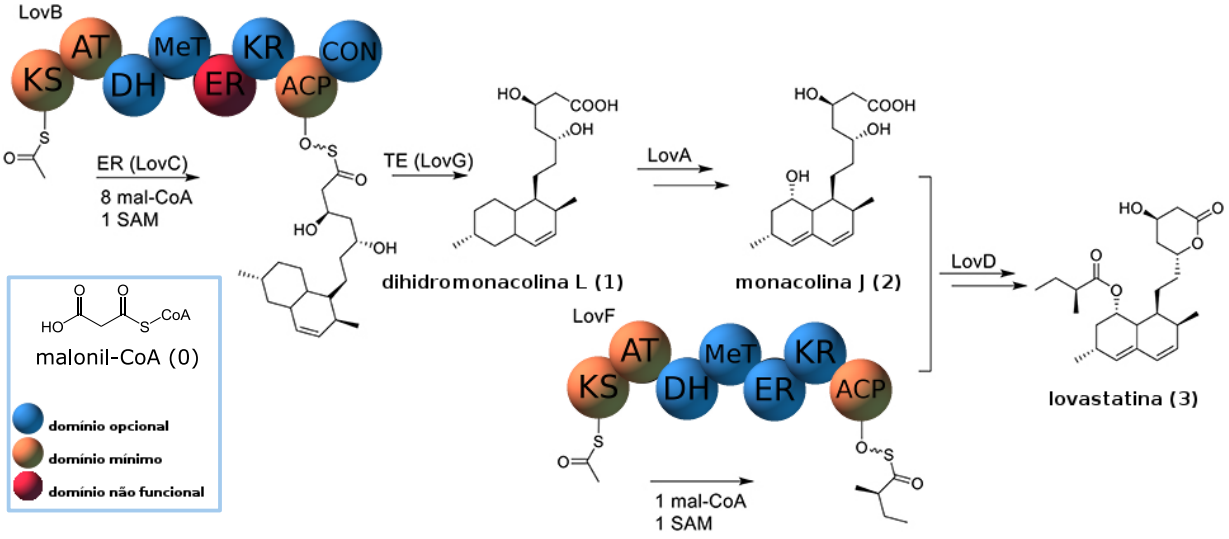

Figura 1.2: Via de biossíntese do metabólito secundário lovastanina em *A. terreus* sumarizada. A figura representa os genes essenciais da via, em parêntese, com ênfase nas duas policetídeos sintases (PKS) lovB e lovF cujos domínios são mostrados. As esferas em laranja representam domínios essenciais, isto é, aqueles com sítios ativos característicos da família de enzimas policetídeo sintase (PKS). Em azul, os domínios opcionais encontrados nas PKS. E, as esferas vermelho, os domínios identificados, porém, não funcionais na enzima. Como mostrado na figura, dihidromonacolina L (1) é produzida pela lovB a partir do precursor malonil-CoA (0) e uma molécula S-adenosil metionina (SAM) e da ação em *cis* do domínio enoil redutase da lovC. O desacoplamento da molécula (1) e lovB é feito através da ação tioesterase (TE) da lovG. Em seguida, o ácido dihidromonacolina L é oxidado duas vezes pela lovA formando ácido monacolina J (2). A porção 2-metil-butirato da lovastatina é sintetizado pela lovF e ligado covalentemente ao composto (2) pela transesterase lovD formando a lovastatina (3).

1.2.4.2 LovB e LovC

Estudos com cepas de *A. terreus* mutantes não produtoras de lovastatina levaram à identificação de uma nova proteína de ∼250kDa, denominada lovastatina nonacetato sintase (LNKS) ou comumente chamada lovB pois é codificada pelo gene lovB, a qual

possibilitou o sequenciamento e a identificação do bloco gênico (figura 1.3) que flanqueia o loco do gene lovB (Kennedy, 1999).

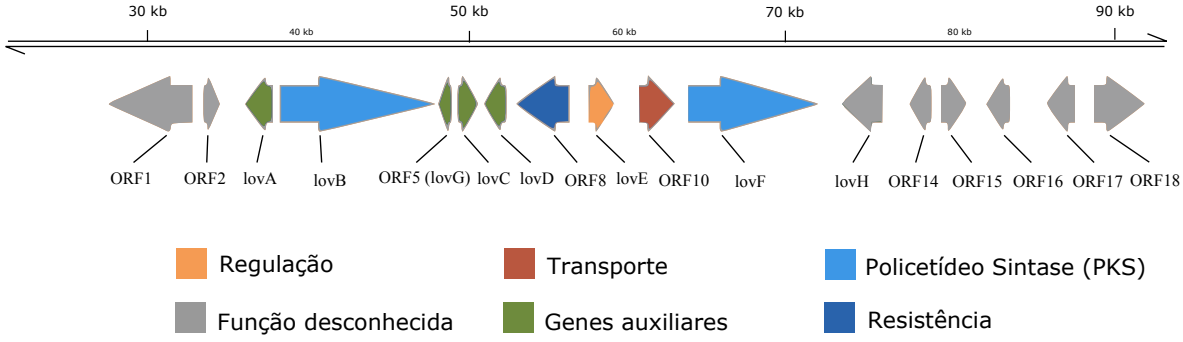

Figura 1.3: Agrupamento gênico de biossíntese (BCG) de lovastatina e regiões que o flanqueiam foram sequenciadas e anotadas por Kennedy (1999). Os genes essenciais que participam diretamente da biossíntese de produção de lovastatina são mostrados as PKS (azul) e os genes auxiliares (verde). Os genes ORF8, lovE e ORF10 participam indiretamente da biossíntese e estão, possivelmente, envolvidos em eventos de resistência, regulação e transporte da lovastatina.

Com a finalidade de caracterizar funcionalmente a lovB e, entender, o seu papel na síntese de lovastatina, foi feita a expressão heteróloga desse gene em *Aspergillus nidulans* (Kennedy, 1999). Os clones transformantes que expressavam lovB foram submetidos a fermentações e o sobrenadante analisado em busca da molécula dihidromonacolina L (DML). Essa molécula é o produto esperado da catálise-enzimática da LNKS - ou lovB (figura 1.2 A, composto (1)). Apesar de não identificarem o composto DML no sobrenadante, encontraram dois novos compostos. Analisando-os, estabeleceu-se que eles pertenciam à classe das pironas e que, provavelmente, eram produtos da interrupção abrupta no funcionamento da policetídeo sintase LNKS (lovB). Atribuiu-se a formação desses produtos intermediários à não funcionalidade do domínio enoil-redutase (ER) da lovB (figura 1.2). Como a porção ER da lovB não é funcional e o processo de formação de policetatos é iterativo, existe uma etapa na qual o ER é requisitado para catálise de intermediários de DML porém, a sua ausência, acarreta a interrupção da enzima lovB e liberação de pironas. Com isso, postulou-se que alguma enzima com domínio ER funcional é fundamental no processo de elongamento da cadeia principal da lovastatina. Posteriormente, identificaram que o gene LovC produzia uma proteína de 363 aminoácidos com sequência homóloga ao domínio ER de outras PKs (figura 1.2). Quando co-expressaram LovB e LovC em *Aspergillus nidulans* houve formação do produto DML. Isso representou a primeira evidência de um domínio não funcional sendo complementado pela ação de outra enzima endógena

(Rangaswamy et al., 1998; Guenzi et al., 1998). Com concentrações puras da enzima lovB e controle de substrato, Ma et al. (2009) reconstituíram as etapas iterativas desta PKS. Assim, estabeleceu-se que a via biossintética de lovastatina começa a partir da adição de nove unidades de acetato, oriundas do composto malonil-CoA (figura 1.2). Quando não há suplementação ao sistema do cofator NADPH, há interrupção da LovB e liberação de uma pirona. Com adição do NADPH, o domínio cetoredutase (KR) fica ativo e a LovB é capaz de incorporar 3 malonatos até a etapa de adição do grupo metil onde só ocorre com a presença da molécula SAM e, posteriormente, a enzima LovC complementando a ação do domínio não funcional ER da LovB termina o processo de síntese de DML.

Surpreendentemente, quando todos os cofatores foram adicionados ao sistema nenhum produto foi identificado. Isso ocorreu, pois, a ausência do domínio tioesterase (TE) na LovB, o qual, encontra-se presente em enzimas FAS (ácidos graxos sintetases) que possuem alta similaridade com a lovB. Portanto, o produto DML se encontrava ligado à enzima lovB e, com a introdução de uma base forte (KOH) foi possível liberar o produto no sobrenadante.

Resumindo os principais pontos acima, referentes à caracterização funcional através da expressão heteróloga de LovB e LovC em *A. nidulans* podemos citar: (1) a importância da presença dos cofatores NADPH e substratos SAM e malonil-CoA para a elaboração do produto final de lovB: a dihidroximonocolina L (DML). (2) A enzima endógena lovC é essencial ao processo complementando o domínio não funcional ER da lovB e sendo necessária para a adição do grupamento metil ao C6 (Ma et al., 2009; Ames et al., 2012).

1.2.4.3 LovF e LovD

Análises do bloco gênico que flanqueia a lovB mostrou a presença de uma segunda PKS, denominada lovastatina dicetato sintase (LDKS), também chamada LovF. Seu papel é formar a cadeia lateral 2(S)-metilbutirato da lovastatina. A sequência de 2.532 aminoácidos da LovF deu suporte à hipótese de que trata-se de uma PKS compartilhando os domínios catalíticos KS, AT, DH, MT, ER, KR e ACP da LovB, mas com o domínio CON ausente e o domínio ER funcional. Com isso, a LovF não requer a ação de uma enoil redutase endógena complementar para realizar com êxito sua ação catalítica. Note que, analogamente à LovB, essa enzima não possui o domínio TE, significando que o produto gerado por ela encontra-se ligado ao seu domínio ACP até que ocorra a transesterificação dele ao composto monacolina

J pela ação da enzima endógena codificada pelo gene LovD (Xie et al., 2006). O gene LovD possui homologia com outras esterases, incluindo β-lactamases, carboxipeptidades, e lipases.

1.2.4.4 LovA

O produto da ação em conjunto da LovB e LovC é a molécula dihidroximonacolina L (DML), o qual deve ser oxidada gerando a monacolina J. Essa molécula então, estará quimicamente susceptível a receber o dicetato produzido pela LovF através da catálise da reação de transacetilação pela LovG. Sendo assim a busca por oxidases dentro do bloco gênico da biosíntese de lovastatina reportou dois putativos genes CP450: LovA e ORF17. Foi proposto que ao menos um desses estaria atuando na oxidação da DML em monacolina J e estudos mutacionais mostraram que cepas deficientes em LovA produziam somente o composto DML implicando, assim, na ação da LovA nessa reação (Sorensen et al., 2003; Barriuso et al., 2011). A função real da ORF17 continua desconhecida.

1.2.4.5 LovG (ORF5)

Sabe-se que a LovB necessita da ação conjunta da LovC para produzir corretamente o produto DML e que, após o uso de uma base forte (KOH) é possível liberar esse produto no meio. Entretanto, para elucidar o real agente da liberação do DML em *A. terreus* Xu et al. (2013) examinaram o bloco biossintético à procura de genes que possuíssem o domínio tioesterease (TE). Após genômica comparativa com o fungo *Penincillium citrinum*, produtor de compactina e *Monascus pilosus*, de lovastatina, identificou-se que entre o gene codificador da PKS e o gene que complementava o ER não funcional, existe um gene que revelou homologia à família das esterases. Além do mais, rupturas nesse gene, denominado LovG (sinônimo: ORF5), levaram ao declínio na produção de lovastatina.

1.2.4.6 Outros genes do agrupamento

Após a elucidação dos genes-chave na produção do metabólito lovastatina, diversos estudos expressaram de forma heteróloga, na levedura *Saccharomyces cerevisiae*, os genes LovB, LovC, LovD e LovA, com isso, conseguiram obter, sem a adição de nenhum substrato exógeno a molécula monacolina J. Sendo assim, os outros genes pertencentes ao BCG da lovastatina atuam como putativo fator de transcrição (LovE), suposto gene codificador

da HMG-CoA redutase envolvida na resistência do hospedeiro produtor de lovastatina (ORF8) e putativa bomba de efluxo (ORF10) (Dietrich e Vederas, 2014).

Capítulo 2

Objetivos

2.1 Objetivo geral

Desenvolvimento de ferramentas computacionais para análise do sequenciamento genômico de oito cepas de *Aspergillus terreus* investigando, em nível molecular, as diferenças na biossíntese do metabólito secundário lovastatina.

2.2 Objetivos específicos

I) Analisar os mapeamentos das leituras-curtas

- Validar experimentalmente regiões com cobertura de sequenciamento anômalas

II) Identificar variantes genéticas

- Busca de polimorfismos de base única e grandes blocos de inserção/deleção

III) Montar genomas *de novo*

- Comparar múltiplos *softwares* de montagem variando inúmeros parâmetros
- Comparar as montagens resultantes contra o agrupamento de biossíntese de lovastatina descrito na literatura

IV) Investigar os agrupamentos de biossíntese de metabólitos secundários (BCGs) presentes nos genomas das cepas

- Desenvolver novos algoritmos para detecção de BCGs

Capítulo 3

Métodos

3.1 Seleção dos Isolados

Ao todo oito cepas de *Aspergillus terreus* foram selecionadas de duas micotecas diferentes para este estudo. Seguem os identificadores usados no estudo e a micoteca hospedeira das cepas: ATCC 20542 (*American Type Culture Collection*[1]) e as cepas BU35, BU27, U22, U9, U10, U22 e U26 da micoteca da Universidade Federal de Pernambuco (UFPE)[2]. Todas as cepas selecionadas foram classificadas como *Aspergillus terreus* pelas respectivas micotecas. Ressalta-se que os identificadores usados no estudo não são os mesmos usados pelas micotecas, com exceção da cepa ATCC 20542. Na tabela 3.1 encontram-se os identificadores usados no estudo e os identificadores na micoteca de origem.

Tabela 3.1 - Identificadores das cepas usadas neste estudo, identificadores oficiais da micoteca da UFPE

Identificador	Id. Micoteca	Localidade	Substrato
ATCC 20542	ATCC 20542	Japão	Solo
BU35	URM 5961	Pernambuco	Solo de caatinga
BU33	URM 5650	-	Torta de Girassol
BU27	URM 5256	-	Rizosfera de Croton sp. (Euphorbiaceae)
U9	URM 224	Pernambuco	-
U10	URM 1876	Amapá	Solo
U26	URM 5254	-	Rizosfera de Cereus sp. (Cactaceae)
U22	URM 5061	Pernambuco	Solo

[1] `https://www.atcc.org/products/all/20542.aspx`

[2] `https://www.ufpe.br/micoteca/nova/home.php`

3.1.1 Sequenciamento de DNA

As bibliotecas genômicas das oito cepas usadas neste estudo foram preparadas de acordo com o protocolo Nextera DNA, de acordo com as especificações do fabricante (Illumina). As amostras referentes às cepas ATCC 20542 e U22 foram sequenciadas através da plataforma *Illumina HiSeq 2500* protocolo *paired-end* de 150 pb por leitura (2x150 pb) e 300 pb de tamanho de inserto. As referentes às amostras BU35, BU27, BU33, U9, U10, U26 foram sequenciadas com *Illumina MiSeq 2000* cujo protocolo emprega *paired-end* com 250 pb por leitura (2x250 pb) e 300 pares de base de tamanho do inserto (tabela 3.2).

Tabela 3.2 - Plataformas de sequenciamento utilizadas para o resequenciamento genômico das cepas

Cepa	Plataforma de sequenciamento
ATCC 20542	HiSeq 2500
BU35	MiSeq 2000
BU33	MiSeq 2000
BU27	MiSeq 2000
U9	MiSeq 2000
U10	MiSeq 2000
U26	MiSeq 2000
U22	HiSeq 2500

3.1.2 Genoma de referência *A. terreus* NIH 2624

O genoma de referência do fungo filamentoso *Aspergillus terreus* foi sequenciado, em 2006, pelo Broad Institute (EUA) e financiado pela organização norte-americana Instituto Nacional de Alergia e Doenças Infecciosas (em inglês, sigla *NIH*) (terreus Broad, 2006). A amostra de DNA genômico sequenciada pertence à cepa NIH 2624, isolada de um paciente imunocomprometido acometido por aspergilose no Reino Unido (Guo e Wang, 2014). A cobertura sequenciada foi de $11X$ resultando em 26 *supercontigs* após montagem pelo Broad Institute (tabela B). A versão usada neste trabalho consiste no genoma *A. terreus* NIH 2624, versão 29 da base de dados *Ensembl Fungi Genomes* (Kersey et al., 2016).

3.2 Mapeamento dos dados de resequenciamento

3.2.1 Mapeamento das leituras contra genoma referência NIH 2624

As leituras oriundas do sequenciamento foram submetidas à etapa na qual checou-se o controle de qualidade do sequenciamento. Nessa etapa o adaptador de sequenciamento inserido durante a preparação das amostras foi removido. Uma janela móvel de tamanho 15 foi perpassada nas leituras-curtas. As janelas com média de qualidade de base menor que 5 foram excluídas. As bases com qualidade (phred) 3 ou menor nos dois extremos das leituras-curtas também foram removidas. Essa etapa foi realizada com o software Trimmomatic (Bolger et al., 2014) (versão 0.30) com os parâmetros `LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 HEADCROP:15`. As leituras resultantes desta etapa foram mapeadas contra o genoma NIH 2624 utilizando-se o software BWA-mem (Li e Durbin, 2009) com parâmetros padrão.

3.2.2 Detecção e Genotipagem das variantes genéticas

As variantes foram identificadas e recalibradas usando o programa Genome Analysis Toolkit (GATK) (McKenna et al., 2010). O GATK usa um modelo de erro adaptativo baseado em variantes conhecidas para diferenciar variantes verdadeiras de artefatos inseridos pela metodologia de sequenciamento. Como até o momento inexistem anotações sobre os sítios reais de variantes em *A. terreus* e, para assegurar a utilização do modelo de erro no GATK, foram identificados sítios de polimorfismos entre as oito cepas e o genoma de referência NIH 2624 através da execução do GATK sem a etapa da modelagem do erro e, as variantes resultantes desta etapa com qualidade de genótipo (GQ) maior que 90 foram usadas como os sítios "reais"a segunda execução do GATK. A segunda execução também seguiu as boas práticas do GATK (Van der Auwera et al., 2013). Os polimorfismos únicos de sequência (SNPs) identificadas foram filtradas com a finalidade de reduzir o número de falso positivos. Para tanto, utilizou-se a ferramenta VariantFiltration do pacote GATK com parâmetros `DP < 16.0 || DP > 44.0`, indicando que foram consideradas variantes com cobertura mínima de 16X e máxima 44X para evitar a identificação de variantes em regiões repetitivas e de falsos positivos. Esses valores são o primeiro e terceiro quartil, respectivamente, da distribuição de cobertura das leituras-curtas mapeadas com atributo mapeamento único. Além disso adicionou-se os parâmetros `QD < 2.0 || FS > 60.0 ||`

`MQ` $<$ `58.0` $\|$ `MQRankSum` $<$ `-12.5` cujo argumento QD significa a qualidade da variante normalizada pela cobertura do alelo não homozigoto da referência, isso evita que variantes em regiões de alta cobertura recebam o índice qualidade (QUAL) desproporcionalmente altas pois cada leitura contribui um pouco para a qualidade inflando esse parâmetro. O parâmetro `FisherStrand (FS)` é uma métrica para evitar o viés entre a fita senso e anti-senso. `MQ` refere-se à média da qualidade entre as leituras das amostras e `MQRankSum` compara a qualidade de mapeamento das leituras suportando o alelo de referência contra as suportando o alternativo.

3.3 Montagem de novo dos genomas

As leituras-curtas das oito cepas foram usadas como entrada para a ferramenta SPAdes de montagem de genomas *de novo* (Bankevich et al., 2012). Os parâmetros foram os padrão, com a especificação adicional dos seguintes tamanhos de k-mer: `-k 21,33,55,77,99,127`. Com a finalidade de testar a influência da ferramenta SPAdes na qualidade das montagens usamos três outros programas para montar a cepa *A. terreus* U26: SOAPdeNovo (Luo et al., 2012), IDBA (Peng et al., 2010) e Abyss (Simpson et al., 2009). As métricas resultantes dos quatro montadores usados na cepa U26 foram obtidos através da ferramenta Quast (Gurevich et al., 2013). Considerou-se o número de contigs nas montagens, tamanho do maior contig e N50 como métricas para comparar a qualidade das montagens entre as quatro ferramentas.

3.3.1 Avaliando a completude dos genomas montados

Para obter uma métrica alternativa aos tradicionais N50 ou L50 (Miller et al., 2010) relacionados ao estado e qualidade da montagem *de novo* foi utilizada a ferramenta Fungal Genes Mapping Project (FGMP) versão 1.0 (Cisse e Stajich, 2016). A metodologia usada pela ferramenta estima a qualidade do genoma montado checando a presença de marcadores. Os marcadores são 593 genes ultra conservados em 40 espécies de fungos analisadas e 172 segmentos de genoma ultra conservados em 10 espécies de fungos, abrangendo os principais clados da árvore filogenética do reino fúngico (Cisse e Stajich, 2016). A proposição em utilizar marcadores para testar a qualidade das montagens *de novo* de eucariotos não é recente. A ferramenta CEGMA (Parra et al., 2007) foi proposta para essa tarefa. Apesar

da possibilidade de aplicar CEGMA na avaliação da montagem *de novo* de genomas de fungos (Moore et al., 2016), o principal ponto que deve ser considerado é que os marcadores eucarióticos desta ferramenta são generalistas, pois são baseados em somente 6 espécies modelo eucarióticas (Cisse e Stajich, 2016).

3.3.2 Anotações genômicas

As predições gênicas a partir das montagens *de novo* das oito cepas foram realizadas com a ferramenta CodingQuarry (versão 2.0) (Testa et al., 2015) usando as anotações dos transcritos montados a partir dos experimentos de RNA-Seq do fungo *A. terreus* cepa LYT10 publicamente disponíveis (Qingdao, 2015). A principal motivação para o uso dessa ferramenta, em detrimento ao usual preditor gênico de organismos eucarióticos AUGUSTUS (Stanke et al., 2004), originalmente usado para predição no genoma de referência *A. terreus* NIH 2624, é a otimização para aplicação em fungos. A ferramenta CodingQuarry baseia-se em duas etapas para predizer a posição dos genes, éxons, íntrons e regiões UTR. A primeira usa o conjunto de transcritos montados do organismo pelas ferramentas Cufflinks/TopHat para treinar os parâmetros da cadeia oculta de Markov generalizada (GHMM) usada para predizer os estados que caracterizaram os limites gênicos. Além disso, o modelo GHMM treinado é aplicado para determinar os limites gênicos da sequência codificadora para cada gene predito a partir dos transcritos contidos nos dados experimentais do RNA-Seq. A segunda etapa, baseia-se na predição gênica usando a montagem do genoma corroborada pelas informações resultantes da etapa anterior. Nessa etapa também é realizada a predição *ab initio* de genes contidos em regiões na qual o conjunto de transcritos não foi capaz de inferir corretamente os limites gênicos devido baixa cobertura ou condições experimentais do RNA-Seq. Todas as etapas descartam os genes que não obedecem algumas características inerentes aos genes de fungos: íntrons curtos, poucas isoformas de transcritos e UTR compartilhado entre genes (Kupfer et al., 2004; Galagan et al., 2005; McGuire et al., 2008). A última característica é uma consequência da alta densidade gênica em regiões do genoma de fungos. Algumas ferramentas como AUGUSTUS interpretam curtas distâncias intergênicas como íntrons levando à predição de genes fusionados (Testa et al., 2015).

3.3.3 Predição de metabólitos secundários nas montagens de novo

A posição genômica dos agrupamentos de biossíntese de metabólitos secundários (BCGs), os genes que os compõe e o putativo metabólito biossintetizado foram preditos através da ferramenta AntiSMASH (versão 3.0) (Weber et al., 2015). As opções usadas na predição das BCGs compreendem a busca por BCGs conhecidas depositadas na base de dados *Minimum Information about a Biosynthetic Gene cluster* (MIBiG) pelo módulo ClusterBlast e BCGs preditas *ab initio*. As predições de BCGs resultantes do programa antiSMASH foram classificadas em 7 principais grupos: alcalóides, NRPS, PKS, híbridos PKS-NRPS, terpenos, sideróforos e outros. Considerou-se BCGs classificadas como "outros" aquelas que produzem putativamente metabólitos secundários de classes alternativas às 6 outras classes usuais. A classe "outros" é resultado da predição de BCGs pelo módulo ClusterFinder do software AntiSMASH v3.0 através das regras de classificação definidas por (Cimermancic et al., 2014). Novas classes referem-se aos metabólitos preditos que não se encaixam em nenhuma outra das 6 categorias propostas usualmente (PKS, NRPS, híbrido PKS-NRPS, alcalóides, sideróforos, terpenos).

3.3.4 Alinhamento Múltiplo dos Genomas

Os genomas montados *de novo* das 8 cepas foram alinhados contra genoma de referência NIH utilizando a ferramenta Mugsy (Angiuoli e Salzberg, 2011). O pacote MafFilter Dutheil et al. (2014) foi usado para processamento do arquivo maf resultante contendo os alinhamento para computar e explorar o grau conservação entre as cepas.

3.3.5 Visualização com curva tipo Hilbert

Curvas tipo Hilbert (HC) são gráficos que objetivam mapear as coordenadas dos espaços de 1 dimensão (1-D) em 2 dimensões (2-D) maximizando o aproveitamento do plano utilizado. Uma característica marcante das curvas tipo Hilbert é o fato do mapeamento entre dois pontos próximos no espaço 1-D permanecerem próximos no 2-D, por exemplo, as coordenadas da base 10 e 100 do cromossomo 1 quando transformado em HC mantêm a proximidade no espaço 2-D. Essa característica permitiu a utilização das HCs na área da Genômica, na qual, Anders (2009) explorou para a caracterização de picos produzidos pela metodologia ChIP-Seq ao longo do genoma. Entretanto, uma desvantagem da HC é a im-

possibilidade de realizar o mapeamento reverso, ou seja, relacionar uma posição no gráfico HC a uma coordenada no espaço 1D. Essa desvantagem é acentuada quando o interesse é obter as posições absolutas, não influenciando inferências relacionadas a visões gerais como homogeneidade, espaçamento, localização de agrupamentos Anders (2009).

As curvas tipo Hilbert, fig. 3.1, são construídas de maneira recursiva, onde uma linha é dobrada sobre si formando um segmento perpendicular. O nível da curva refere-se ao número de segmentos que o gráfico HC conterá. Seja a curva HC com nível n, o número de segmentos s formados no espaço 2-D é $s = 4^n - 1$. Na figura 3.1 é mostrada a curva tipo Hilbert resultante para $n = 1, 2, 3, 4$ com cada segmento pintado em vermelho. Note que à medida que o nível da curva aumenta, o tamanho da curva torna-se maior e a curva dobra-se mais densamente. A direção na qual a curva é construída é indicada pelas setas nas curvas. Imagine uma longa fita de DNA esticada. Pode-se dobrar ela sobre si formando uma curva Hilbert. Com isso é possível representar grandes genomas em um espaço bidimensional ao invés de um espaço linear. Quando mapeamos as, aproximadas, 30×10^6 (Mpb) do genoma de *Aspergillus terreus* no espaço hilbertiano cada segmento da curva tem resolução de 7.326 e 1.831 bases para HC com níveis 6 e 7, respectivamente.

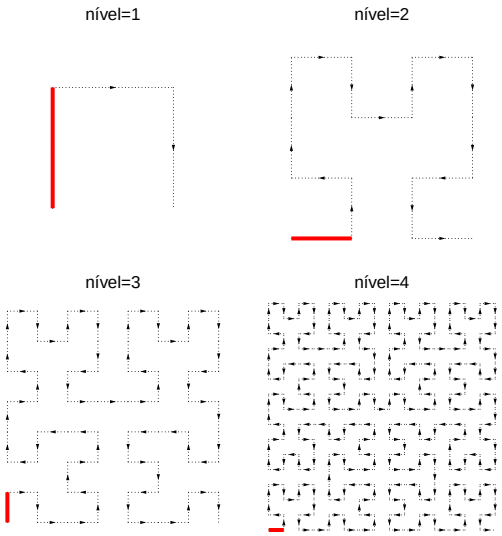

Figura 3.1: Curvas tipo Hilbert para diferentes valores de n. As curvas são construídas dobrando-se recursivamente o segmento de reta $4^n - 1$ vezes.

3.3.6 Comparando a região genômica de produção de lovastatina entre as cepas

As regiões gênicas de todo agrupamento de biossíntese de lovastatina da cepa ATCC 20542 de *A. terreus* e genes ao redor deste foram anteriormente sequenciadas via metodologia Sanger e depositadas no GenBank (códigos de acesso AH007774.2 com 55.428 bases e AF151722 com 11.561 bp) (Kennedy, 1999). Estas foram utilizadas como base de comparação para analisar as sequências correspondente a este agrupamento nos *contigs* construídos pela montagem *de novo* (seção 3.3) das cepas estudadas.

A ferramenta nucmer do pacote MUMmer (Kurtz et al., 2004) com parâmetros "`-maxmatch -c 100`" foi utilizada para alinhar as sequências genômicas no GenBank contra os *contigs* para identificar possível sintenia. Somente blocos sintênicos com porcentagem de identidade maior que 90% e tamanho acima de 500 pares de base foram levados em consideração para análises posteriores.

Para facilitar a análise destes dados, foi desenvolvido pelo autor desta dissertação um visualizador utilizando a linguagem JavaScript e módulo D3.js[3]. A visualização baseia-se no modelo circos proposto por Krzywinski et al. (2009). Os gráficos e atributos, como a porcentagem de identidade entre as sequências comparadas e genes membros em cada montagem podem ser visualizados online (Rocha, 2016).

3.3.7 Reproducibilidade das etapas

As principais etapas desta dissertação que resultaram na construção de *pipelines* como, por exemplo, a busca de variantes e a montagem *de novo* e anotação dos *draft* genomas estão disponíveis online (Rocha, 2016). Estas *pipelines* reprodutíveis foram construídas com o auxilio da ferramenta Snakemake (Köster e Rahmann, 2012). Os códigos referentes às metodologias desenvolvidas nesta dissertação também podem ser acessadas online (Rocha, 2016).

3.3.8 Teste de significância entre múltiplos conjuntos

O estudo das relações entre múltiplos conjuntos oriundos de uma mesma população normalmente é feito através da verificação dos elementos em comum entre os conjuntos (intersecção). Diversas metodologias para medir o grau de similaridade e consequente-

[3] `http://d3js.org`

mente, dissimilaridade, entre dois conjuntos foram propostas, tais como o índice Jaccard e o coeficiente de Sorensen. Enquanto testes estatísticos como o teste exato de Fisher e o teste hipergeométrico podem ser empregados para determinar a significância do número de elementos em comum observados entre conjuntos (interseção), os mesmos não aplicam-se para testes nos quais o número de conjuntos comparados é maior que dois. A representação visual através dos diagramas de Venn facilita a compreensão, exploração e ilustra as relações entre 2 ou mais conjuntos, porém, à medida que o número de conjuntos comparados cresce o número de possíveis intersecções cresce (2^n intersecções para n conjuntos) e a visualização das múltiplas elipses torna-se inviável. Com a finalidade de explorar as relações entre múltiplos conjuntos o pacote SuperExactTest foi proposto (Wang et al., 2015). Simplificadamente, a metodologia consiste em aplicar o teste exato de Fisher com a finalidade de obter o estimador de significância estatística (p-valor) para cada intersecção observada entre as possíveis combinações entre os conjuntos. O método proposto assume que os conjuntos usados no teste são constituídos de amostras aleatórias e independentes de uma população. Portanto, para análises genômicas nas quais a população consiste de milhares de genes e o conjunto de genes testados estão na ordem de centenas a aplicação do SuperExactTest é válida.

Capítulo 4

Resultados

4.1 Quantificação dos níveis de lovastatina

A concentração de lovastatina no sobrenadante das oito cepas sequenciadas foi quantificado em trabalho de colaboração (metodologia adicional de cultivo e quantificação no apêndice A.3). As cepas que mais produziram lovastatina, de acordo com a detecção no sobrenadante, foram a ATCC 20542 com cerca de 1,0 gL^{-1}, BU35 ∼0,6 gL^{-1}, BU33 ∼0,07 gL^{-1}, U26 0,05 gL^{-1} e as cepas BU27, U22, U10, U9 não apresentaram lovastatina detectável no sobrenadante (figura 4.1).

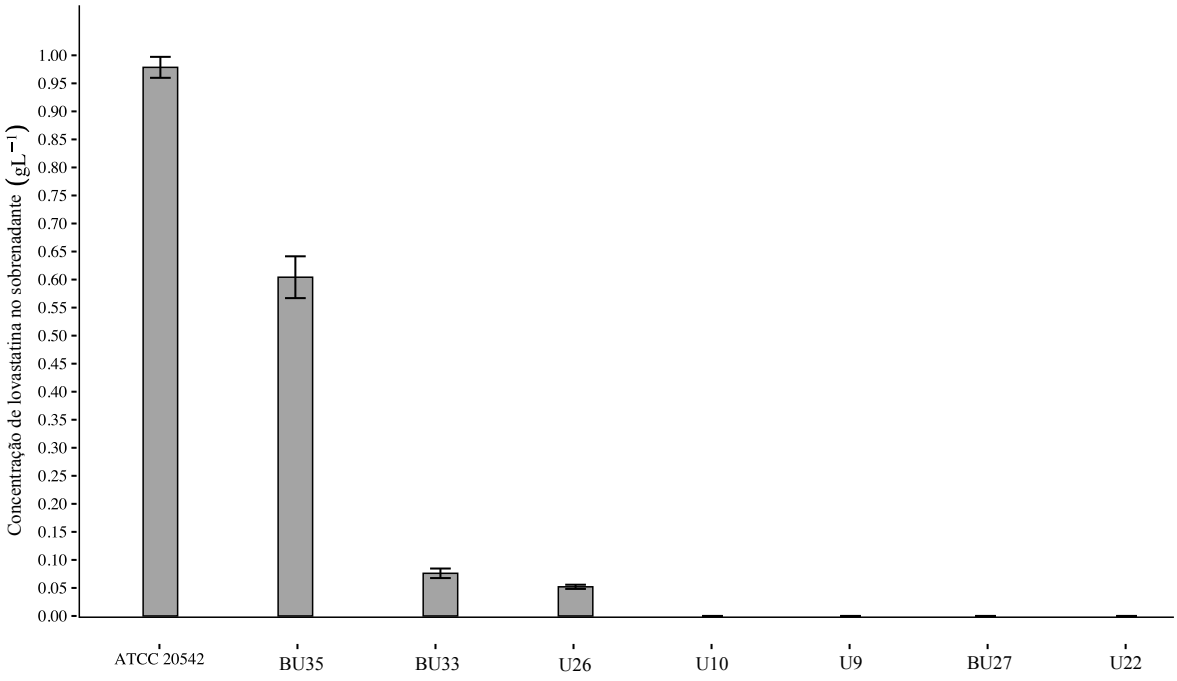

Figura 4.1: Quantificação do metabólito secundário lovastatina no sobrenadante de 8 indivíduos após cultivo das cepas de acordo com metodologia A.3.

Com os dados de quantificação procedeu-se com as análises de sequenciamento no qual buscaram-se as bases moleculares para explicar os diferentes fenótipos observados entre as cepas de *A. terreus*.

4.2 Sequenciamento das cepas de A. terreus

O sequenciamento via plataforma Illumina HiSeq 2500 do DNA genômico da cepa ATCC 20542 resultou em leituras-curtas de tamanho 150 pb. O total de bases sequenciadas, compreendendo as sequências *forward* e *reverse*, totalizaram aproximadamente 2,1 Gb. Após a etapa de controle de qualidade foram filtradas cerca de 0,37 Gb com baixa qualidade, de acordo com parâmetros descritos na metodologia (seção 3.2.1).

Assim como a cepa ATCC 20542, o DNA genômico da cepa U22 também foi sequenciado com plataforma Illumina HiSeq 2500. Este sequenciamento gerou em torno de 1,88 Gb e, devido baixa qualidade do sequenciamento referente às leituras reversas a etapa de clivagem das bases (*quality clipping*) rendeu leituras médias com o mínimo de 105 bases em média (tabela 4.2).

Tabela 4.1 - Métricas avaliadas após a etapa de sequenciamento. Na nomenclatura das amostras as extensões .1 e .2 referem as leituras resultantes do sequenciamento *paired-end*

%Bases ($< Q20$)	%Bases ($\geq Q20$)	Amostra	Qualidade Média (Q)	Tamanho da leitura (bp)	#Bases
3,5	96,5	ATCC.1	36,7	145,00	1.055.497.054
16,7	83,3	ATCC.2	32,6	145,13	1.056.453.229
12,2	87,8	BU35.1	33,3	215,51	736.751.401
14,0	86,0	BU35.2	33,0	216,67	740.702.797
6,7	93,3	BU33.1	35,2	229,33	593.348.148
26,5	73,5	BU33.2	29,4	230,49	596.345.863
6,1	93,9	BU27.1	35,5	217,18	748.001.423
24,4	75,6	BU27.2	30,0	218,98	754.189.159
3,3	96,7	U26.1	36,7	206,35	442.675.951
14,9	85,1	U26.2	32,9	206,98	444.021.451
3,1	96,9	U10.1	36,8	214,03	538.687.515
10,5	89,5	U10.2	34,3	214,32	539.428.853
8,2	91,8	U9.1	34,8	227,66	298.305.385
8,2	91,8	U9.2	35,0	227,65	298.294.318
3,7	96,3	U22.1	36,6	146,61	942.825.412
24,7	75,3	U22.2	30,4	146,86	944.431.401

O restante das cepas foram sequenciadas com Illumina MiSeq v2, produzindo entre 0,6-1.47 Giga bases totais por cepa (tab 4.1). A qualidade média das bases (Phred) manteve-se constante na faixa de 30, ou seja, 0.1% de erro de nomeação.

Tabela 4.2 - Tabela sumarizando os resultados pós filtragem dos dados brutos.

%Bases ($< Q20$)	%Bases ($\geq Q20$)	Amostra	Qualidade Média (Q)	Tamanho da leitura (bp)	#Bases
1,5	98,5	ATCC.1	37,7	138,13	683.871.578
10,1	89,9	ATCC.2	34,5	137,43	680.434.093
8,8	91,2	BU35.1	34,5	180,75	617.380.080
8,3	91,7	BU35.2	34,8	175,01	597.760.175
4,7	95,3	BU33.1	36,0	203,12	524.218.797
14,5	85,5	BU33.2	33,0	152,71	394.115.288
4,3	95,7	BU27.1	36,2	192,50	661.283.375
13,6	86,4	BU27.2	33,3	150,32	516.407.554
2,3	97,7	U26.1	37,2	187,42	401.092.058
9,0	91,0	U26.2	34,8	165,28	353.708.952
2,2	97,8	U10.1	37,3	194,91	489.845.863
6,7	93,3	U10.2	35,6	181,29	455.618.747
6,3	93,7	U9.1	35,5	201,02	263.273.719
5,6	94,4	U9.2	35,9	199,44	261.216.867
3,0	97,0	U22.1	37,1	129,30	830.213.760
17,1	82,9	U22.2	32,6	106,19	681.827.744

Todos os resultados dos sequenciamentos *paired-end* encontram-se sumarizados na tabela 4.1 e, na tabela 4.2 estão os resultados pós correção das leituras brutas. Os experimentos envolveram o sequenciamento de múltiplas cepas na mesma *lane* (*multiplex sequencing*).

As leituras resultantes da etapa de correção dos dados brutos foram usadas como entrada nas próximas etapas que envolvem o mapeamento das leituras-curtas, montagem *de novo* e análises subsequentes.

4.3 Mapeamento das leituras-curtas contra o genoma de referência

A primeira etapa implementada em todos os estudos envolvendo resequenciamento de DNA é o mapeamento das leituras-curtas dos indivíduos contra o genoma de referência da espécie (Alex Buerkle e Gompert, 2013; Sims et al., 2014).

As coberturas dos mapeamentos das cepas contra o genoma de referência de *A. terreus* NIH 2624 variam no intervalo de $15X$ até $41X$. Com isso, a cobertura média do resequenciamento é considerado de baixa a média cobertura (Sims et al., 2014). A extensão de cobertura das leituras-curtas no genoma de referência (*coverage breadth*), isto é, o quanto o genoma de referência foi mapeado pelos dados oriundos da cepa U22 foi de 88,6% apesar de $41X$ na métrica cobertura média (*coverage depth*) de mapeamento (tabela 4.5). Nota-se

que as extensões genômicas amostradas das cepas U9, U10 e U26 cuja cobertura média foi menor que $25X$ são abaixo 90% do genoma de referência *A. terreus* NIH 2624.

Apesar da taxa de cobertura ser $\sim 30X$ para a cepa BU27, identificam-se regiões genômicas com distribuição heterogênea de cobertura, inclusive com cobertura zero, quando visualizamos os mapeamentos contra o genoma de referência (fig. 4.1). As regiões contíguas com cobertura zero não restringem-se à cepa BU27. Destaca-se que a cepa U22 não possui quase nenhuma leitura mapeada ao longo de 40 kb compreendendo o loco com os genes responsáveis pela produção de lovastatina (fig. 4.1).

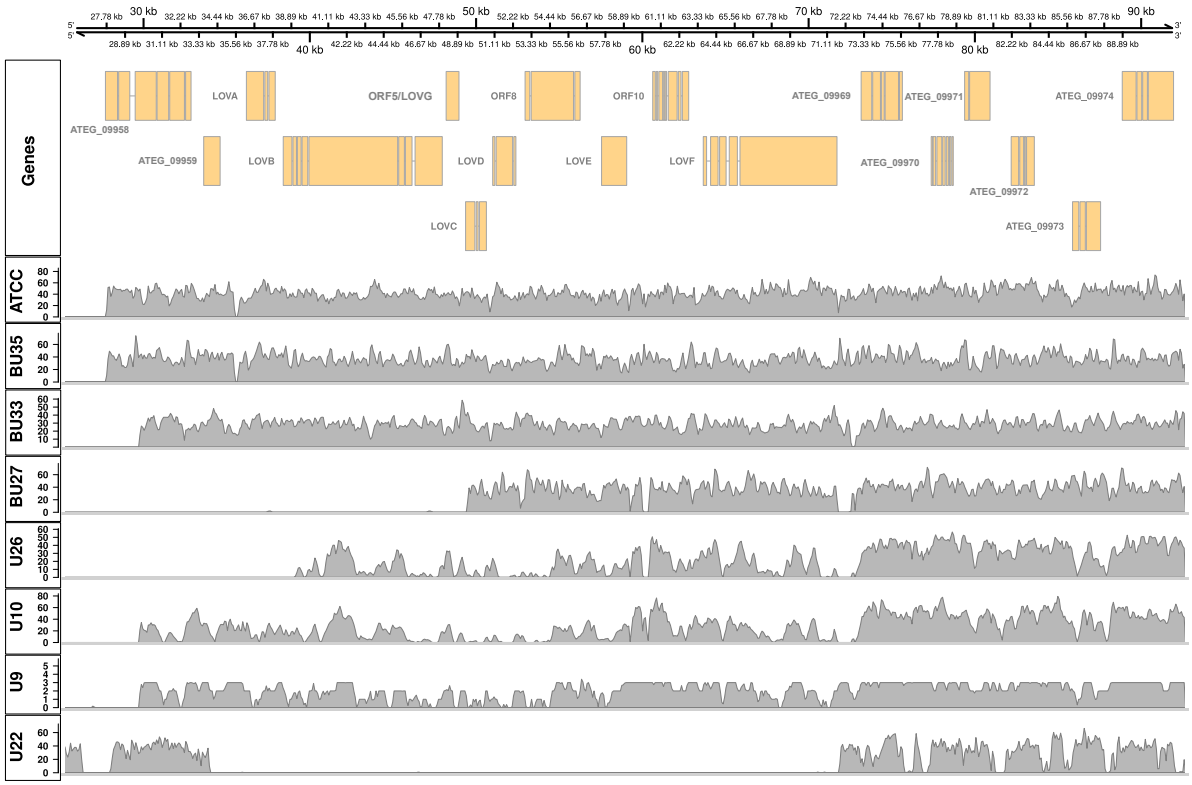

Figura 4.1: Espectro de cobertura do mapeamento das leituras-curtas (*short reads*) das oito cepas de *Aspergillus terreus* contra o genoma de referência NIH 2624. O quadro superior delimita a região genômica entre as posições 22,7 kb e ~93 kb do *supercontig* 1.16 do genoma referência. No painel genes, as anotações dos genes envolvidos na biossíntese de lovastatina (lovA até lovF), e os genes margeando este BCG (ATEG_*).Os painéis restantes indicam a cobertura para cada cepa identificada (ATCC, BU35, BU33, BU27, U26, U10, U9 e U22). Nota-se um padrão uniforme no espectro de cobertura das 3 cepas ATCC, BU35, BU33 em comparação as demais. Embora em algumas posições há queda abrupta na cobertura de uma das cepas, como entre o gene LovF e ATEG_09959 na cepa BU33. Além disso, existem longas extensões contínuas com cobertura zero, como a região compreendida entre a ORF1 e lovC na cepa BU27, o intervalo genômico entre a ORF2 e a ATEG_09959 na cepa U22. Todos os eixos y nos painéis de espectro de cobertura estão na escala absoluta excluindo-se o eixo y da cepa U9 na qual o eixo está na escala $log_e(1 + x)$, onde x é o valor da cobertura em cada posição genômica para a cepa U9.

O perfil contendo a distribuição dos tamanhos das regiões do genoma NIH 2624 com cobertura de mapeamento extremamente baixas ($\leq 2X$) revela a incidência de longas regiões genômicas com mais de 10.000 bases com quase nenhuma leitura mapeada, apesar da prevalência de regiões pequenas e médias com tais características (fig. 4.2).

Com a finalidade de verificar se as regiões genômicas com espectro de cobertura de mapeamento heterogêneo foram causadas por artefatos de sequenciamento ou são variantes estruturais (VEs) entre as cepas desenhou-se primers para algumas regiões do loco de biossíntese de lovastatina.

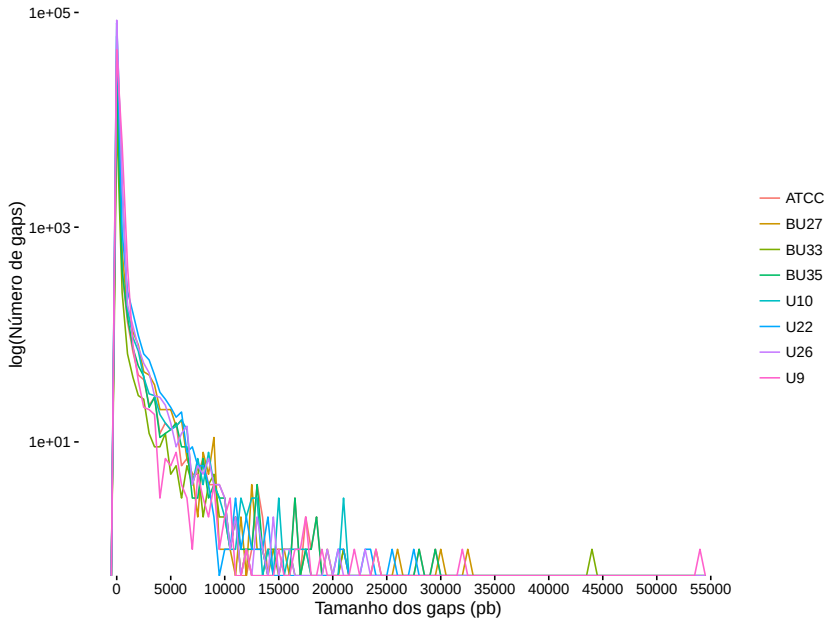

Figura 4.2: Distribuição das regiões genômicas com cobertura de mapeamento $\leq$2x.

4.4 Desenho de primers para o agrupamento de biossíntese de lovastatina

A metodologia proposta para desenhar os primers baseia-se na extração de segmentos de sequência conservados ao longo dos indivíduos e posterior identificação dos pares de primers dentro das regiões que flanqueiam o *loco* alvo usando as informações conjuntas provenientes dos mapeamentos das leituras-curtas dos indivíduos contra o genoma de referência da espécie.

As regiões margeando o *loco* a ser amplificado são as sequências estendidas n pares de base a montante e a jusante, delimitando, respectivamente, as regiões flanqueadoras 5' e 3' (fig. 4.5). Com a sequência genômica das regiões flanqueadoras e o conjunto de leituras

de cada indivíduo mapeadas a ela pode-se formar uma sequência consenso. As bases da sequência consenso final são determinadas selecionando a base com maior probabilidade ao longo da coluna da matriz de alinhamento múltiplo de sequências (MSA). Cada linha da matriz MSA representa uma sequência e cada coluna as letras alinhadas (Lee et al., 2002) (fig. 4.3 A). As letras não limitam-se às bases nitrogenadas podendo abranger inclusive aminoácidos.

Alguns programas como o bcftools, geram a sequência consenso, a partir do mapeamento das leituras-curtas contra uma sequência referência, utilizando informações adicionais além de simplesmente a distribuição empírica das bases contidas nas leituras. Eles possuem modelos probabilísticos que incorporam dados como a qualidade de chamada do nucleotídeo (Phred), frequência alélica, entre outros fatores (Li, 2011). Apesar de cada algoritmo possuir modelos probabilísticos próprios e, desconsiderando-se o mérito de cada modelo, a maioria das metodologias propostas que visam a determinação da sequência consenso possuem uma característica em comum: a impossibilidade de reconstruir a sequência de cada indivíduo a partir da sequência consenso gerada por eles (Lee et al., 2002). Como notado por Lee (2003), essa impossibilidade deve-se ao fato dos alinhadores progressivos, por exemplo clustal e clustalW (Larkin et al., 2007), necessitarem da determinação da sequência linear (1D), ou consenso, de cada par de sequência alinhado para, progressivamente, determinar o alinhamento múltiplo para um conjunto de 2 sequências ou mais. Isto impõe desafios adicionais na determinação da sequência linear já que inserções/deleções levantam a questão sobre o que deverá ser incluído na sequência consenso final do conjunto. Inevitavelmente, a redução na complexidade da representação dos alinhamentos na forma de sequência consenso envolve a perda de informação. Enquanto o MSA contêm toda a informação necessária para a determinação da sequência consenso o mesmo não acontece na operação reversa: obter o MSA a partir da sequência consenso.

4.4.1 Desenvolvendo algoritmos baseados em PO-MSA para desenho de primers

Para abordar essas questões, Lee et al. (2002) desenvolveram uma representação do MSA através de um grafo acíclico direto chamado grafo de ordem parcial (PO-MSA), no qual cada nó representa letras da sequência e arestas ligam letras consecutivas na sequência (fig 4.3 B).

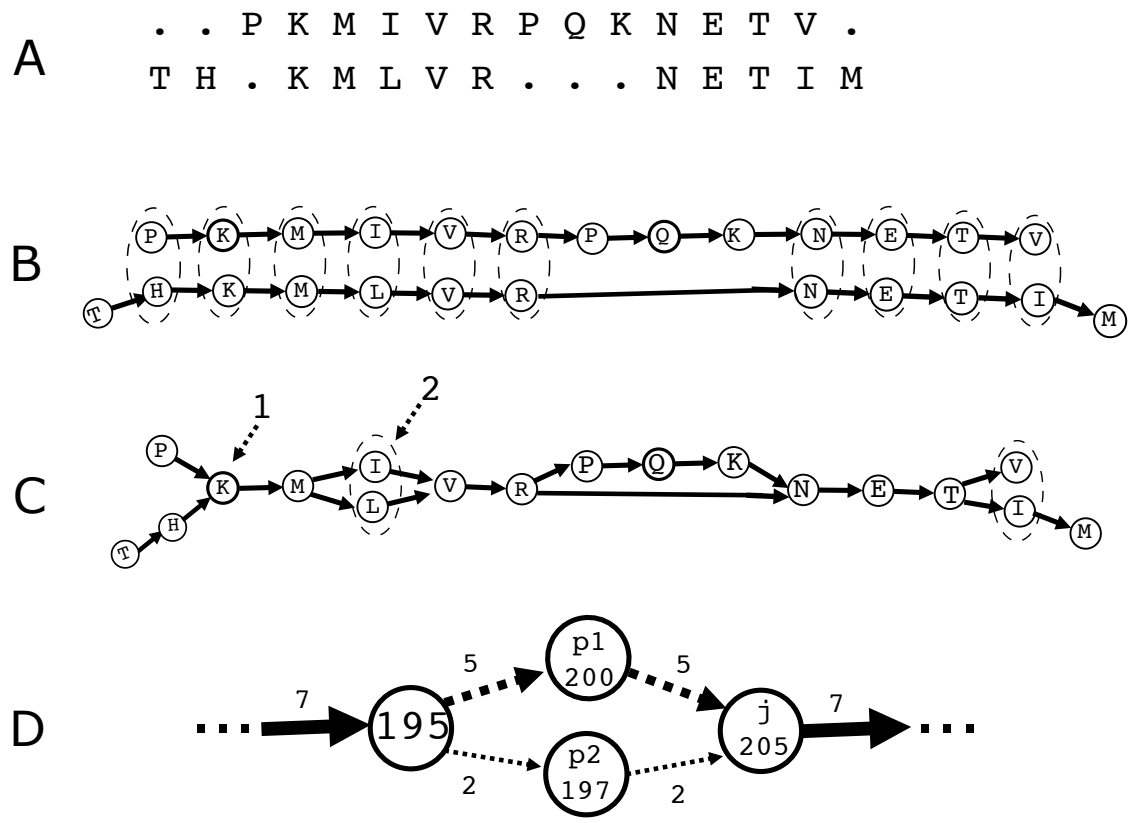

Figura 4.3: A) Esquema representativo da matriz alinhamento de duas sequências de aminoácidos. B) Grafo acíclico direto das duas sequências indicando as relações entre os aminoácidos das sequências. C) Construção do grafo de ordem parcial onde os aminoácidos em comum às duas sequências são colapsados e variantes nas sequências são exibidas como nós independentes. D) Região do grafo no qual existe um polimorfismo entre o nó i até j onde o caminho inferior é suportado por 2 sequências e o superior por 5 sequências. O nó i tem peso acumulado em 195. Como o algoritmo percorre o grafo considerando o peso das arestas e não o peso acumulado, o caminho com maior peso de aresta (superior) é o trilhado. Adaptado de Lee et al. (2002).

Letras alinhadas e idênticas entre sequências são fusionadas em nó único (fig 4.3 C seta 1), enquanto letras alinhadas, mas não idênticas (variantes), são representadas como nós separados e identificados como alinhados no grafo (fig 4.3 C seta 2). Cada letra fusionada vem de uma sequência e a identificação da sequência parental é armazenada no nó fusionado. Um nó pode possuir diversas arestas saindo ou entrando dele representando, respectivamente, a relação de consecutividade e precedência entre duas letras. Um nó com múltiplas arestas é chamado nó de junção e ele forma ramificações no grafo caracterizando sequências com um ou mais nucleotídeos polimórficos (nó entre as setas 1 e 2 da fig 4.3 C). Lee et al. (2002) além de idealizar a representação dos alinhamentos estendeu os algoritmos

de alinhamento de sequências usando programação dinâmica (Needleman e Wunsch, 1970; Smith e Waterman, 1981) para a tarefa de alinhar sequências lineares contra o grafo de ordem parcial.

A estrutura de grafo parcial (PO-MSA) foi utilizada na metologia proposta nesta dissertação para armazenar as informações sobre a composição e estrutura do alinhamento das regiões flanqueadoras entre os indivíduos. Inicia-se com a construção do grafo PO-MSA contendo as regiões flanqueadoras provenientes do genoma de referência e das sequências consenso do mapeamento dos indivíduos contra essa região. Com o grafo PO-MSA aplica-se o algoritmo proposto para identificar regiões altamente conservadas entre indivíduos dentro dos limites das regiões flanqueadoras. O intuito do algoritmo proposto é similar ao desenvolvido por Lee (2003) no qual a pergunta respondida foi: dado uma estrutura de grafo parcial entre sequências, como obter a sequência consenso a partir desta estrutura?

Diferentemente da metodologia de "votação" (maior frequência da letra na posição do alinhamento) para obter a sequência consenso a partir da matriz MSA, o algoritmo proposto por Lee (2003) equivale a percorrer o grafo escolhendo-se o melhor caminho possível, ou seja, maximizando a probabilidade de observação das sequências alinhadas para determinar a(s) sequência(s) consenso(s) do grafo.

Considere um grafo PO-MSA G consistindo dos nós $i, j, \ldots$ e arestas diretas e_{ij} ligando os nós (i, j) com peso w_{ij}. Sem perda de generalização considere G como o grafo da figura 4.4 B construído a partir da matriz de alinhamento mostrada contendo as S_k sequências com $k = A, B, C$ (figura 4.4). Seja a ordem topológica o ordenamento das sequências da esquerda para direita e as arestas e_{ij} a união das arestas e_{ijk} mostradas no grafo.

O peso em cada aresta w_{ij} do grafo é simplesmente a soma dos pesos w_{ijk} das arestas das sequências S_k que atravessam a aresta e_{ij}. Isto é, na figura 4.4 o peso da aresta e_{01} que liga os 2 primeiros nós, `C` e `G`, é a soma do peso da aresta w_{01A} (aresta superior ligando $C \rightarrow G$) + w_{01B} (aresta mediana ligando $C \rightarrow G$) + w_{01C} (aresta inferior ligando $C \rightarrow G$). Portanto o peso é dado por:

$$w_{ij} = \sum_{k}^{\forall S_k : i \rightarrow j} w_{ijk}$$

Na concepção proposta por Lee et al. (2002), a sequência consenso pode ser obtida após percorrer o grafo da esquerda para a direita, isto é, garantindo-se a ordem topológica dos

nós de i a j. Todos os nós no começo do percurso são assinalados com pontuação $s_i = 0$. Para cada nó i com arestas e_{pi} segue-se o caminho de aresta com maior peso de aresta w_{pi} e, então assinala-se a cada nó a pontuação do nó $s_i = s_p + w_{pi}$ (fig 4.3 D). Repete-se a operação recursivamente até visitar todos os nós a partir do primeiro nó (sem arestas de entrada) até, possivelmente, o último nó (sem arestas de saída). O uso do peso das arestas w_{ij} na escolha do caminho, em contrapartida à pontuação acumulada nos nós s_i, segue o princípio da verossimilhança, isto é, o peso seleciona o percurso que maximiza a probabilidade da sequência observada (fig 4.3 D).

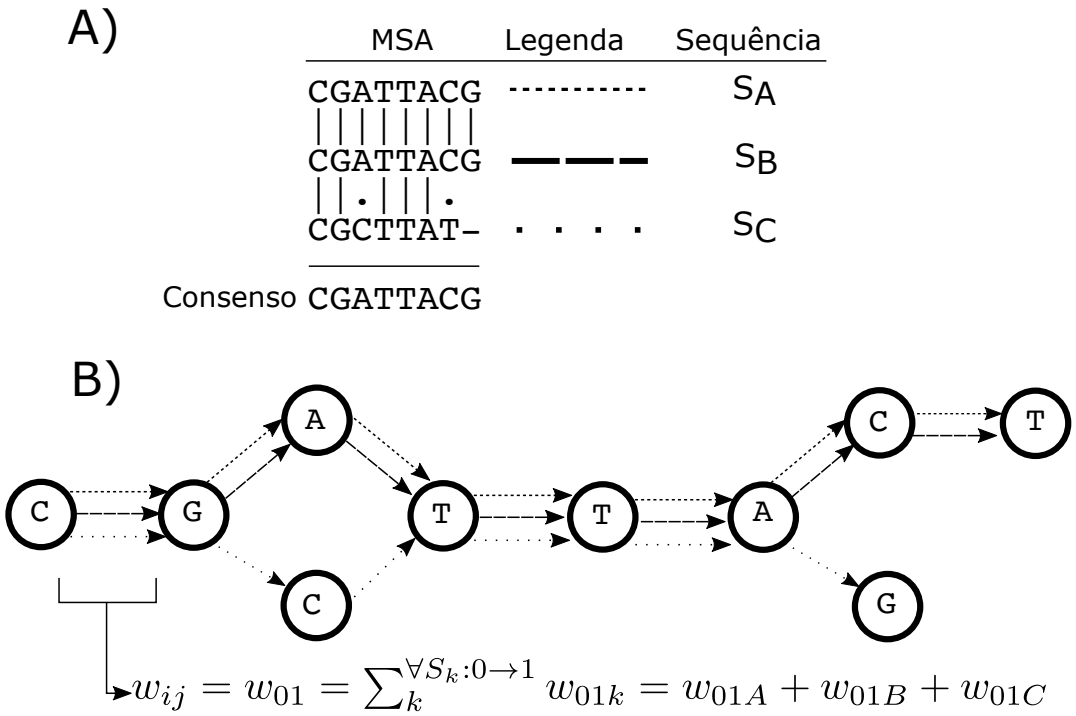

Figura 4.4: Em (A) destaca-se a matriz de alinhamento múltiplo de 3 sequências S_A, S_B e S_C e, em (B), a representação na forma de grafo de ordem parcial (PO-MSA) deste alinhamento. Cada aresta em (B) é suporte para uma das sequências S. O peso da primeira aresta que liga `C` e `G`, isto é, w_{01}, é a soma do peso das três arestas visto que todas as sequências dão suporte para este caminho.

Como a finalidade da metodologia proposta é desenhar pares de primers com dados de resequenciamento, a abordagem de obter a sequência consenso a partir do grafo POA-MSA pode ser adaptada. A diferença fundamental entre o algoritmo proposto neste trabalho e o desenvolvido por Lee (2003) está no tamanho do caminho resultante da travessia do grafo POA-MSA. Lee (2003) impõe que o caminho corresponde a uma travessia completa do grafo, ou seja, começando do nó inicial até o último nó, gerando assim a sequência consenso.

Já o algoritmo proposto neste trabalho visa buscar caminhos "movimentados" sem necessidade de contiguidade entre nó inicial e final. Isto é, gostaríamos a partir de um número mínimo de sequências N, retornar o maior caminho suportado por no mínimo N sequências entre as k sequências alinhadas. Com isso, pode-se evitar posições do alinhamento muito polimórficas e variantes como pequenas inserções e deleções (indels) nas sequências. Por exemplo, na figura 4.5, em (C) o grafo POA-MSA é construído a partir das sequências consenso do mapeamento de 3 indivíduos contra a região flanqueadora 5' (figura 4.5 A). Note que poderíamos obter a sequência consenso final desses indivíduos (quadro B) com as sequências lineares dos indivíduos, entretanto, ao desenharmos um primer para esta região, a partir da sequência final, todas as informações de variantes entre as sequências seriam perdidas e, com isso, reduziríamos a probabilidade dele se anelar. Porém, ao percorrer o grafo POA-MSA (figura 4.5 D) com o algoritmo proposto determinando o mínimo de sequência $N = 2$ e $N = 3$ dentre as $k = 3$ sequências alinhadas, o resultado será o maior subsegmento de sequência suportando N, isto é, contendo 2 ou 3 indivíduos com sequência de segmento idênticas.

A heurística do algoritmo é a seguinte: o grafo com as 3 sequências segue a direção do nó i ao j (arestas direcionadas na figura 4.5 C) onde uma única sequência tem a inserção de adenina (`A`, indivíduo 2) entre os nós $i = 5$ e $i = 6$. Assumindo $w_{ijk} = 1$ para as $k = 3$ sequências, o correto movimento para percorrer o grafo no nó da inserção `A` é selecionar o caminho de i a j com maior peso ($w_{ij} = 2$) ao invés do caminho com a inserção ($w_{ip} = 1$). Se o número mínimo de sequências é 3 (parâmetro $N = 3$) o grafo encontrará um obstáculo entre os nós 5-6 já que $w_{56} < 3$. O algoritmo interrompe o caminho no nó $i = 6$ e armazena o caminho até $i = 6$, na variável C_n. Recursivamente, o algoritmo, busca caminhos maiores que satisfazem a restrição $w_{ij} \geq N$. No exemplo, os maiores caminhos para os parâmetros $N = 2$ e $N = 3$ são mostrados na figura 4.5 C. O algoritmo retorna, quando existe, o maior caminho C_n que satisfaz a condição imposta, as bases e a posição desse segmento no PO-MSA. Aplicando o algoritmo nas sequências flanqueadoras usamos o caminho resultante de cada uma como entrada para o programa Primer3 que buscará o par de primers ideal otimizando os parâmetros químicos e físicos (maiores detalhes checar Untergasser et al. (2012)).

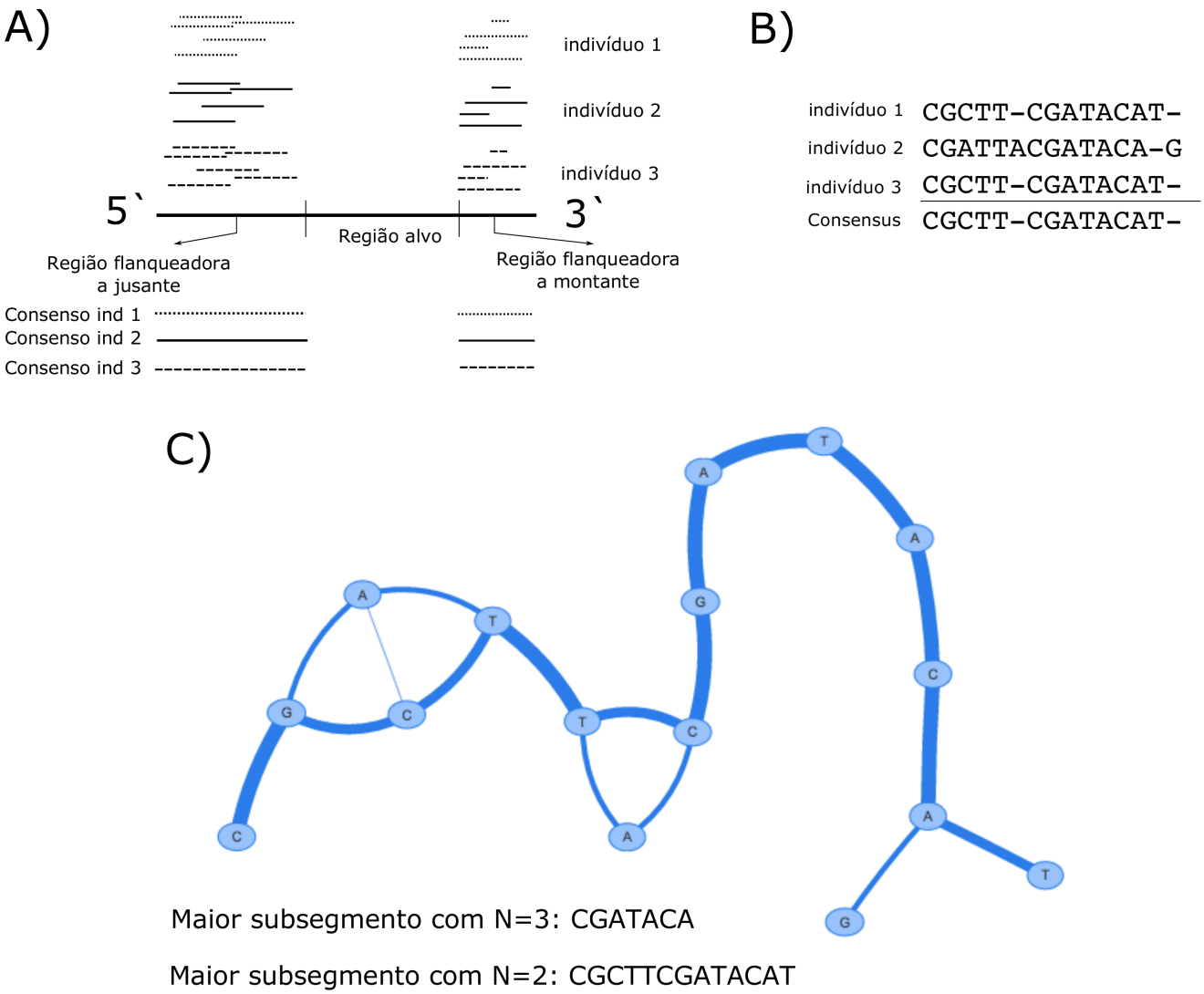

Figura 4.5: Estudos envolvendo resequenciamento de múltiplos indivíduos da mesma espécie, ocasionalmente, são procedidos de etapas na qual mapeiam-se as leituras-curtas contra o genoma de referência da espécie. Após análise das coberturas de sequenciamento identificam-se regiões potenciais, como genes essenciais para vias metabólicas, regiões candidatas a variantes estruturais (VEs), repetições em série de sequências que, na maioria das vezes, necessitam de validação posterior. As validações, na maioria dos casos, são realizadas via ensaio de PCR. Para otimizar o desenho de primers em estudos de resequenciamento determinou-se uma nova metodologia. As etapas compreendem: (A) mapeamento das leituras-curtas de cada indivíduo sequenciado contra o genoma de referência da espécie e determinação da posição da sequência a ser amplificada via PCR no genoma (região alvo), (B) alinhamento múltiplo das sequências flanqueadoras de cada indivíduo determinando a matriz de alinhamento múltiplo, (C) obtenção do grafo POA-MSA contendo as informações referentes à composição nucleotídica e variantes da região flanqueadora de cada indivíduo e travessia do grafo através do algoritmo proposto que resulta na obtenção do maior segmento de sequência conservado na região flanqueadora presente em no mínimo 2 ou 3 indivíduos. O segmento de sequência resultante é a entrada para o programa Primer3 que desenhará o primer neste segmento preservando as restrições termodinâmicas.

4.4.2 Validação dos primers desenhados por PO-MSA

Para validar a metodologia de desenho de primers proposta acima, explorar *loci* genômicos com cobertura irregular na região de biosíntese do metabólito secundário lovastatina foram sintetizados primers genômicos usando a abordagem proposta para amplificação via reação em cadeia da polimerase (PCR).

Após inspeção visual do mapeamento das leituras-curtas de cada uma das oito cepas contra o genoma de referência NIH 2624 (metodologia 3.2.1) procedeu-se com a escolha das regiões para amplificação usando três critérios principais. O primeiro critério levou em consideração a importância do gene para a produção de lovastatina. Com isso, na extensão do gene codificador da PKS lovB foram escolhidos 3 *loci* devido sua importância como enzima chave na produção e seu tamanho. Já para os genes lovC, lovE e lovA foi escolhido 1 *loco* para cada gene.

O segundo critério na determinação das regiões alvo para amplificação foram regiões dentro do agrupamento de lovastatina que apresentaram características de variantes estruturais entre as cepas e o genoma NIH 2624. Por exemplo, a cobertura das leituras na região entre o gene lovF e o gene lovH entre as cepas altera-se de uma cobertura média para zero nesse segmento do genoma sugerindo-se a deleção desse *loco* em algumas cepas.

O terceiro critério foi escolher regiões que abrangem toda a extensão genômica responsável pela produção de lovastatina delimitada pelo agrupamento de biossíntese de lovastatina proposto por Kennedy (1999). Em algumas cepas, como U22 e BU27 observou-se um longo segmento desse agrupamento com cobertura zero. Os primers foram usados para validar se essas regiões com cobertura zero são consequência de artefatos impostos pelo sequenciamento ou, realmente, as cepas perderam esses extensos segmentos genômicos caracterizando grandes variantes estruturais.

Com isso foram escolhidas 12 regiões genômicas dentro do agrupamento de biossíntese de lovastatina proposto por Kennedy (1999) para amplificação via PCR. No genoma NIH 2624, esta região está compreendida entre as bases 20.000 e 92.000 do *supercontig* 1.16. Usando o algoritmo proposto para desenho dos primers obtivemos 12 pares de primers para estas regiões (especificações sumarizadas na tabela 4.3).

Tabela 4.3 - Tabela com as informações dos primers desenhados nesse estudo.

Identificador	Direção	Tm	Tamanho (nt)	GC%	Sequência	Cepas amplificadas
J	forward	59,57	20	55,00	TTAGTCCTCTCGGCGAAGTC	BU33, U10, ATCC, U9, U26, BU35
J	reverse	60,28	20	45,00	GGGGTGAAAAGGGCTTAAAA	BU33, U10, ATCC, U9, U26, BU35
K	forward	59,42	20	45,00	AACCGTGTCAACATCAATGC	BU33, U10, ATCC, U9, U26, BU35
K	reverse	61,15	19	52,63	CGGTTATTGCGAGCCAGAT	BU33, U10, ATCC, U9, U26, BU35
L	forward	59,40	19	57,90	CTTCAGGGACGTGACAAGC	BU33, U10, ATCC, U9, U26, BU35
L	reverse	58,84	18	55,56	CATCATGCCAGCTTCAGG	BU33, U10, ATCC, U9, U26, BU35
M	forward	60,54	20	45,00	TCCTTTGACAGCAGCATGAA	BU33, BU27, U10, ATCC, U9, U26, U22, BU35
M	reverse	58,62	20	55,00	ACTCCAAGAGGCTTCTCGAC	BU33, BU27, U10, ATCC, U9, U26, U22, BU35
O	forward	58,26	21	42,86	GGAAAACGGGCATTTACTAGA	BU33, BU27, U10, ATCC, U9, U26, BU35
O	reverse	59,73	20	50,00	ATCAGAAACGCCACCAGAGT	BU33, BU27, U10, ATCC, U9, U26, BU35
P	forward	59,55	18	50,00	AATAACGCCGCACAAACC	BU33, U10, ATCC, U9, U26, BU35
P	reverse	60,76	20	50,00	CGTGATCTGACACGCACATT	BU33, BU27, U10, ATCC, U9, U26, BU35
Q	forward	59,31	20	45,00	CTTGAAAATGACGGGCTCTT	BU33, U10, ATCC, U9, BU35
Q	reverse	59,93	19	47,37	CAAGCCTCTTGCCAATGAA	BU33, U10, ATCC, U9, BU35
R	forward	58.49	22	40,91	AATGAATCGATCAGCTTAGTGC	BU33, BU27, U10, ATCC, U9, U26, BU35
R	reverse	60,41	20	50,00	ATTTGCGACAGGAGCAACTC	BU33, BU27, U10, ATCC, U9, U26, BU35
S	forward	60,41	19	57,90	CATCGACGTCGGTCTTCAG	BU33, BU27, ATCC, BU35
S	reverse	59,64	22	50,00	AACGGACTCAACGAGATCTACC	BU33,BU27, U10, ATCC, U9, U26, BU35
T	forward	60,25	19	52,63	CCGTTGCCAGAAACATCAG	BU33, BU27, U10, ATCC, U9, U22, BU35
T	reverse	59,72	20	55,00	TAGCCGACGGAGACAGGTAT	BU33, U10, ATCC, U9, U22, BU35
U	forward	58,87	20	50,00	CGGAACTGGATCACAGCTAA	BU33, U26, U22,BU27, BU35, ATCC, U9, U10
U	reverse	59,94	19	57,90	GTGGCGGTAGACCCATCTT	BU33, U26, U22, BU27, BU35, ATCC, U9, U10
X	forward	59,67	19	57,90	CGGAACCGGGACTTCTTAG	BU33, BU27, U10,ATCC, U9, BU35
X	reverse	60,42	20	45,00	TAATAAATCGGCCCACGAGA	BU33, U10, ATCC, U9, BU35

Conforme os resultados dos mapeamentos (figura 4.1) os primers distribuem-se ao longo do agrupamento perfazendo regiões com distribuição de cobertura anômalas como o primer T no qual verificou-se ausência de amplificação da região delimitada por T (retângulos brancos na fig 4.6) nas cepas BU27 e U26, o primer U com amplificação positiva somente nas cepas ATCC e BU35 (retângulos azuis na fig 4.6) e o primer X com ausência de amplificações nas cepas U22, BU27, U26 e U10. O primer Q extende-se ao longo do gene LovA, os identificados pelas letras J, K e L distribuem-se no gene PKS lovB e os primers P, S, R e O foram confeccionados para amostrar os genes auxiliares e de transporte inseridos no agrupamento. Embora próximo ao agrupamento de biossíntese de lovastatina do genoma de referência NIH o primer M abrange o gene cadA envolvido na produção do metabólito primário ácido itacônico. Os resultados de PCR do primer L indicam multiplas amplificações dessa banda (retângulo cinza na fig 4.6) e, nenhuma amplificação foi observada com o primer P.

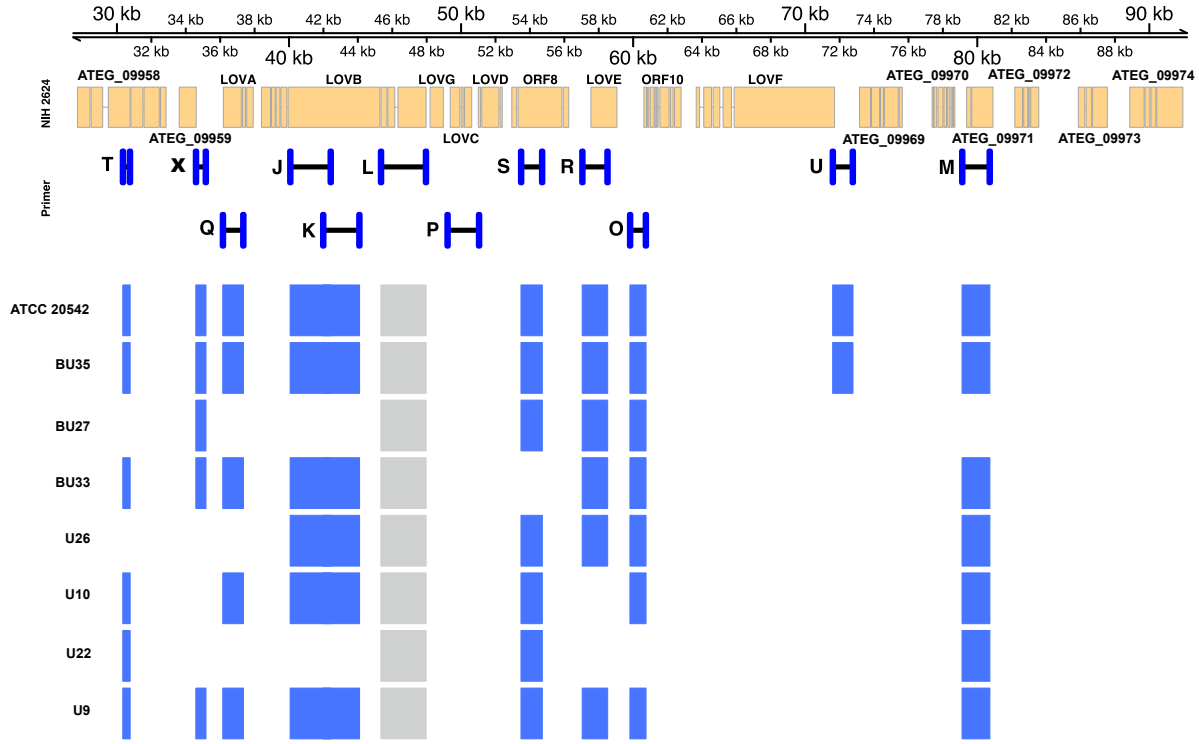

Figura 4.6: Análise dos experimentos de PCR e ancoragem dos primers no genoma de referência NIH 2624. A parte inferior da figura assinala o retângulo com a cor azul quando houve amplificação do fragmento determinado pelo primer acima. Cinza quando múltiplas bandas foram amplificadas e, sem cor, quando nenhuma amplificação foi observada.

Comparou-se os resultados da PCR com as predições das amplificações resultantes da *pipeline* de desenho dos primers proposta na metodologia. O resultado da execução do *pipeline* é a tabela 4.3 na qual é possível verificar as informações sobre os atributos dos primers e na última coluna quais cepas serão supostamente amplificadas por esse primer. Considera-se um primer como verdadeiramente positivo se o primer direto e reverso do conjunto predizer o anelamento para uma cepa. Entretanto, se o conjunto de primer predizer o anelamento para uma cepa e a PCR tiver ausência de *amplicon* detectável consideramos como falso negativo. O falso positivo surge se o programa predizer ausência de anelamento do primer e este mostrar banda no ensaio da PCR. Com os valores da quantidade de primers verdadeiramente positivos (TP), falso positivos (FP), verdadeiramente negativos (TN) e falso negativos (FN) calcularam-se as métricas especificidade ($\frac{TN}{TN+FP}$) e sensibilidade ($\frac{TP}{TP+FN}$). Os resultados das duas métricas foram de aproximadamente 75% (especificidade=0,76 e sensibilidade=0,75).

4.4.3 Análise global de polimorfismo de sequências

Além da ausência de genes essenciais na biossíntese de lovastatina ser o principal indicador do fenótipo não produtor de lovastatina em algumas cepas, os polimorfismos de sequência podem ter papel importante na produção de metabólitos secundários (Cabañes et al., 2015). Portanto, estendendo-se as análises para a identificação de variantes genéticas no genoma inteiro, foram identificados 592.689 sítios contendo os polimorfismos únicos (SNPs – *Single Nucleotide Polymorphism*). O número de genótipos nas cepas estudadas é mostrado na tabela 4.4. A quantidade de genótipos heterozigotos em relação ao genoma de referência NIH 2624 é maior nas cepas U9 e BU33 em comparação às outras cepas, respectivamente, 6.353 sítios e 3.704 sítios. A quantidade de genótipos homozigotos referência é maior na cepa BU27 seguido pelas cepas BU35,ATCC,BU33,U26,U22,U10 e U9. As cepas com maior número de sítios não genotipados estão U10, U22 e U26.

O genoma de referência NIH 2624 (*Ensembl Fungi* versão 29) foi usado como base de dados para anotações funcionais das variantes genéticas e predição dos efeitos pela ferramenta SnpEff (Cingolani et al., 2012). O número de SNPs identificados dentro dos éxons que, podem indicar variantes funcionais, foi de 285.080 variantes. Entre as variantes, 93.118 (32,6%) classificados como SNPs sem sentido (*missense*), 1.169 (0,4%) classificados com perda de sentido (*nonsense*) e 190.793 (66,9%) variantes com efeito silencioso (*silent*).

Tabela 4.4 - Contagem dos polimorfismos únicos de sequência (SNPs). O termo het refere-se às variantes heterozigotas, classificadas em RA quando possui um alelo igual referência e outro alternativo e AA onde os dois alelos são alternativos à referência e divergentes entre si. O termo hom refere-se às variantes homozigotas alternativas (hom AA) e homozigoto referência (hom RR).

Cepa	het AA	het RA	hom AA	hom RR	SNPs ausentes	SNPs (total)
U9	57	6.296	106.961	59.727	114.957	173.041
BU33	34	3.670	184.109	84.858	15.327	272.671
U26	0	0	0	77.852	210.146	77.852
BU35	2	119	34.937	96980	155.960	132.038
BU27	0	0	0	116.229	171.769	116.229
ATCC	2	161	45.751	84900	157.184	130.814
U22	0	0	0	70.391	217.607	70.391
U10	0	0	0	60.426	227.572	60.426

4.4.4 *Análise local de polimorfismo de sequências no agrupamento de biossíntese de lovastatina da ATCC 20542*

A análise global dos polimorfismos, isto é, a identificação e genotipagem das variantes únicas de sequência em todo o genoma revela a nível molecular o grau de variação genética entre as cepas. Entretanto, fazer inferências sobre o agrupamento de biossíntese de lovastatina e o possível impacto das variantes na produção deste metabólito secundário não é o indicado pois não existe dados relacionando a cepa referência com a produção de lovastatina. Uma das maneiras de conectar a funcionalidade das variantes únicas com a biossíntese de lovastatina é identificar os SNPs usando a cepa ATCC 20542. Sendo esta cepa precursora no estudo do composto lovastatina usamos a sequência dela, anotada por Kennedy (1999), como referência para buscar SNPs e pequenas inserções e deleções (INDELS).

Os resultados das anotações funcionais das variantes únicas de sequência identificadas nos 8 genes que compreendem o BCG de lovastatina indicaram a presença de 1.355 sítios polimórficos que encontram-se sumarizados na figura 4.7. Alguns sítios não foram genotipados em algumas cepas devido deleção do sítio ou cobertura de leituras insuficientes, notadamente no caso da cepa U22 em razão da não existência do loco genômico de BCG de lovastatina. Adicionalmente, grande parte dos não genótipos encontrados nas cepas U9, U10 e U20 são consequência da baixa cobertura.

Quase todos os genótipos da cepa ATCC é composto de alelos referência validando a abordagem de filtragem imposta visto que essa cepa possui mesma sequência que a referência usada na chamada das variantes. Não identificou-se nenhum polimorfismo na região codificadora dos genes de biossíntese de lovastatina entre a cepa BU35 e a ATCC (referência). Apesar da constância na quantidade de SNPs categorizados como motivadores de baixo e médio impacto foi possível identificar SNPs de alto impacto funcional: 3 na cepa BU33 e 1 na cepa U10.

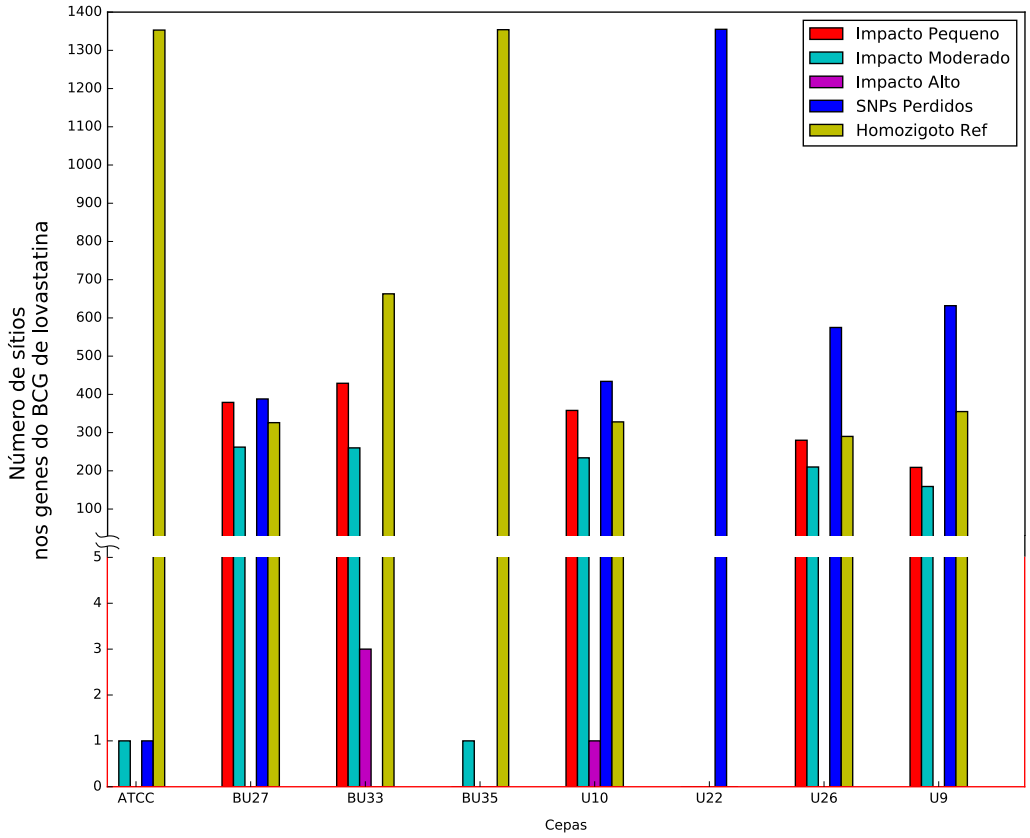

Figura 4.7: Distribuição do número de variantes únicas de sequência com os respectivos impactos funcionais identificadas após mapear as leituras-curtas das cepas contra o loco de biossíntese de lovastatina anotado por Kennedy (1999).

Embora a validação destas variantes seja fundamental, especula-se que a existência de variantes únicas de alto impacto funcional nas regiões codificadoras doo BCG de lovastatina influencie na redução da produção desta molécula na cepa BU33 em comparação à ATCC 20542.

4.5 Análises baseadas nas montagens de novo

Como os resultados de mapeamento das leituras-curtas das cepas indicaram extensas regiões do genoma com cobertura zero caracterizando potenciais variantes estruturais (VEs) entre os indivíduos e o genoma de referência (fig 4.2) procedeu-se uma montagem *de novo* dos genomas utilizando as leituras-curtas, para tentar mitigar os vieses do mapeamento de sequências contra um genoma de referência fixo. A montagem *de novo* seria uma forma de melhor caracterizar as variações estruturais, pois apesar de avanços nas abor-

dagens de detecção de VEs através das informações de mapeamento, a limitação destas metodologias é a baixa sensitividade em regiões repetitivas e a impossibilidade de resolver as sequências quando há eventos de inserção (Tattini et al., 2015).

Os resultados referentes à montagem *de novo* dos genomas, anotação dos genomas e estimativa dos metabólitos secundários das cepas são descritos a seguir. Para melhor compreensão os resultados foram divididos em 4 tópicos. O primeiro tópico define os resultados relacionados a cepa *A. terreus* ATCC 20542, a padrão para a produção de lovastatina. O segundo tópico, intitulado U26, apresenta os resultados da execução, para esta cepa, de diferentes ferramentas de montagem genômica testadas no presente trabalho e finaliza com os resultados das predições gênicas. O terceiro tópico da seção sumariza os resultados para as cepas restantes. O último tópico é focado na análise do agrupamento de biossíntese (BCG) do metabólito secundário lovastatina.

4.5.1 ATCC 20542

O genoma da cepa *A. terreus* ATCC 2624 foi montado com a ferramenta SPAdes (Bankevich et al., 2012). A montagem *de novo* resultou em 173 contigs maiores que 1.000 pb totalizando ∼30,28 Mbp, sendo que o maior contig tem 991.250 pb. A média do conteúdo G+C é 52,16% e o N50 da montagem é 359.971 pb. Quando comparada com o genoma de referência, a montagem da ATCC 20542 possui ∼1 Mb distribuídos em 49 contigs que não foram totalmente ou parcialmente alinhados contra o genoma NIH 2624. O número de contigs necessários para cobrir 75% das bases do genoma de referência (métrica LG75) foi 49. A fração do genoma de NIH 2624 coberto pela cepa ATCC 20540 equivale a 88,81%. A porcentagem de marcadores, sequências de proteínas e elementos de DNA evolutivamente conservados ao longo do genoma de 246 fungos, foi usada como métrica para quantificar a completude das montagens *de novo* (Cisse e Stajich, 2016). A porcentagem de marcadores encontrados na montagem da ATCC 20542 foi de 97,6%.

Foram preditos 11.492 genes codificadores de proteína com a ferramenta Coding Quarry (Testa et al., 2015). Isso equivale a ∼1000 genes preditos a mais no genoma da cepa ATCC 20542 em comparação ao genoma de referência NIH 2624. A taxa média de 2,63 éxons por transcrito também destoou da taxa 3,14 encontrada na anotação da NIH 2624.

4.5.2 U26: Avaliando a montagem genômica

Com o propósito de avaliar a influência de diferentes algoritmos na qualidade das montagens *de novo*, adotou-se os dados de sequenciamento da cepa *A. terreus* U26 como cepa teste. Esta cepa foi escolhida pela alta taxa de fragmentação da montagem pelo software SPAdes e, consequentemente, baixo valor de N50 obtido (tab 4.5). Para isso o conjunto de leituras-brutas da cepa U26 foi sujeito à montagem *de novo* pelos softwares AByss, SOAPdeNovo e IDBA, além do SPAdes.

Usamos a ferramenta QUAST (Gurevich et al., 2013) para cálculo das métricas relacionadas à montagem e comparação da montagem *de novo* com genoma de referência da espécie (NIH 2624)

Como não existiram diferenças significativas entre as métricas que atestam a qualidade obtidas pela montagem via SPAdes em comparação às demais, e corroborado por estudos que atestam a alta qualidade desse software, decidiu-se manter um padrão nas montagens. O padrão foi a montagem *de novo* com o SPAdes de acordo com o citado na metodologia (seção 3.3).

4.5.3 BU35, BU33, BU27, U9, U10 e U22

O grau de fragmentação do genoma montado *de novo* variou entre as cepas. As cepas BU35 e BU27 apresentaram número de contigs gerados nas montagens comparável à montagem *de novo* da cepa ATCC 20542. Enquanto o número de contigs das montagens referentes às cepas U9, U10 e U26 atingiram a ordem dos milhares (>9.000 contigs). Importante observar a baixa porcentagem, ∼76%, do genoma de referência NIH 2624 coberto pela montagem produzida das cepas U9, U10 e U26, indicando baixa cobertura de sequenciamento nessas amostras. A porcentagem de marcadores encontrados, que estima a completude da montagem, para essas três cepas também foi baixo. A porcentagem de marcadores encontrados ficou entre 84,8-90,4%, em contrapartida aos mais de 95% encontrados nos genomas *de novo* das demais cepas. Apesar do tamanho das montagens *de novo* das cepas U9, U10 e U26 serem consideravelmente menores do que os 30 Mb estimados para o tamanho do genoma da espécie *Aspergillus terreus*, a quantidade de genes preditos apresenta pouco desvio entre as cepas. Entretanto, quando comparamos o desvio considerando as predições de genes codificadores de proteína com no mínimo 100 aminoácidos a

Tabela 4.5 - Sumário das métricas avaliadas na montagem *de novo*, predição dos genes e mapeamento.

Cepa	ATCC 20542	BU35	BU33	BU27	U26	U10	U9	U22
Métricas gerais								
Tamanho da montagem (pb)	∼30,28 Mb	∼30,29 Mb	∼30,13 Mb	∼30,18 Mb	∼26,30 Mb	∼26,55 Mb	∼25,14 Mb	∼30,75 Mb
Quantidade de contigs (>0 pb)	173	228	443	252	9528	9064	9833	343
Tamanho do maior contig (pb)	991250	1311690	1397437	1322790	51119	74218	54426	890692
FGMP % completude	97,6%	97,3%	97,5%	98,0%	88,8%	90,4%	84,8%	97,6%
Cobertura (%) contra genoma ref NIH 2624	0,90	0,90	0,94	0,87	0,75	0,76	0,78	0,83
GC%	52,16	52,16	52,46	52,05	50,61	51	50,09	52,27
Genes codificadores de proteínas	11492	11468	11601	11474	12232	12328	11215	10988
Genes codificadores de proteínas >100 aminoácidos	11492	10580	10651	10499	8573	8837	7685	10104
Sequências codificadoras de proteínas >100 aminoácidos								
Mediana do tamanho dos genes	1238	1239	1242	1240	813	832	738	1224
Média do tamanho dos genes	1487	∼1488	∼1486	∼1504	∼1094	∼1122	∼1052	∼1461
Média de éxons por gene	2,63	2,63	2,62	2,64	2,38	2,45	2,32	2,60
Métricas gerais mapeamento								
Cobertura média do mapeamento (Depth X)	39,73	36,10	28,78	33,95	20,80	25,74	15,57	41,14
Extensão do mapeamento (Breadth)	0,936	0,931	0,962	0,917	0,862	0,866	0,851	0,886

quantidade de genes preditos nas cepas U9, U10 e U26 reduzem drasticamente.

4.5.4 O agrupamento de biossíntese de lovastatina

A visualização do alinhamento dos contigs gerados das oito cepas contra a sequência anotada do agrupamento de biossíntese de lovastatina, de acordo com a metodologia 3.3.6, é mostrada nas figuras 4.8 e 4.9. O agrupamento é composto por duas anotações totalizando 18 genes anotados, entretanto, fragmentada em 3 contigs para melhor visualização. A primeira anotação (código GenBank AF151722) refere-se à sequência genômica de 11.561 pares de base constituindo o gene codificador da PKS lovB, a segunda anotação depositada (código AH007774) tem 55.428 pb e equivale aos 17 genes flanqueando o gene lovB identificados e sequenciados por Kennedy (1999). Nota-se que estas sequências de DNA depositadas pertencem ao genoma da cepa ATCC 20542 que foi isolada pelo mesmo grupo Kennedy (1999). Os genes essenciais para a biossíntese de lovastatina são conhecidos (Guo e Wang, 2014). Eles compreendem genes codificando 2 PKS redutoras, lovB (lovastatina nonacetídeo sintase) e lovF (lovastatina dicetídeo sintase), 1 fator de transcrição (lovE) e 4 genes auxiliares (lovC, lovG, lovA e lovD). Dentre as funções dos genes auxiliares está, por exemplo, a complementação da ação das PKS disponibilizando domínios enoil-redutase e liberação de substrato através do domínio tioesterase (Yin et al., 2016).

Todos os 7 genes essenciais para produção de lovastatina mais 11 genes restantes foram detectados em um único contig nas montagens das cepas ATCC 20542 e BU35 e mostraram identidade com as sequências depositadas por Kennedy (1999). As identidades em nível nucleotídico são, respectivamente, 99,9% e 99,98% (fig. 4.8-A e B). A conservação entre nucleotídeos e arquitetura do agrupamento entre os clones ATCC 20542 comparando-se a montagem *de novo* por leituras-curtas e a sequência obtida via sequenciamento Sanger Kennedy (1999) foi praticamente total. Portanto, a cepa ATCC 20542 pode ser vista como um controle positivo para o protocolo de montagem *de novo*.

Surpreendentemente, os limites gênicos do contig montado com NGS da ATCC 20542 são os iguais aos obtidos anteriormente por Kennedy (1999), ou seja os contigs montados por ambas as estratégias possuem tamanho praticamente igual. Kennedy (1999) utilizou o *probe fishing* para isolar as regiões genômicas flanqueando o gene lovB e subsequente sequenciamento destas via Sanger. Sugere-se que os mesmos limites genômicos obtidos pelas duas metodologias são inerentes ao genoma da ATCC 20542. Na porção a jusante do loco

de biossíntese de lovastatina, o obstáculo que dificultou a montagem parece ser a região telomérica, sendo que as repetições características das pontas cromossomais dificultam a montagem *de novo*. No outro extremo, a montante do contig, a impossibilidade da montagem e da extensão de sequência via metodologia proposta por Kennedy (1999) parecem relacionar-se com assinaturas de sequências repetitivas que podem causar fragmentação de contigs ou dificultar anelamento de primers requerido pela metodologia usada por Kennedy (1999). A assinatura de sequência de DNA repetitiva neste limite foi confirmada quando comparou-se os contigs contendo o loco de biossíntese de lovastatina montado das cepas ATCC 20542 e BU35. A extensão do primeiro foi de 76.397 pb (figura 4.8-A) e da cepa BU35 (figura 4.8-B) estende-se por 303.225 pb. Como o contig para ATCC 20542 está contido nos 303.225 pb do contig obtido da cepa BU35 [1] foi possível verificar os limites genômicos e, corroborando a hipótese de que sequências repetitivas estão funcionando como obstáculos à montagem, foi encontrada a assinatura de repetição simples `(CCTG)n` alongando-se por 140 bases neste limite.

As análises baseadas em homologia não identificaram nenhum contig na cepa BU27 com similaridade aos genes codificadores das putativas enzimas ORF1, ORF2 e das enzimas lovA e lovB. Ademais, os genes restantes, incluindo os essências à biossíntese de lovastatina, estão distribuídos em dois contigs (figura 4.8-C). Identificou-se na cepa BU33, ao longo de dois contigs, genes homólogos a todos os genes anotados por Kennedy (1999), com exceção do putativo gene ORF12 como nota-se pelo vão entre os arcos verdes na figura 4.8-D. Ressalta-se que o nível de identidade nucleotídica é de, aproximadamente, 95%.

Diversos pequenos contigs, isto é, resultantes de montagens fragmentadas mostraram homologia com as anotações propostas por Kennedy (1999) analisando-se as cepas U26, U10 e U9 (figura 4.9-A,B,D). Somente as regiões putativas da ORF1, ORF2, ORF14, ORF15 e ORF16 mostraram homologia com algum contig na cepa U22 (figura 4.9-C).

4.6 Alinhamento Múltiplo dos Genomas

Computamos o alinhamento múltiplo dos genomas com a ferramenta Mugsy (Angiuoli e Salzberg, 2011) usando o conjunto de contigs e scaffolds obtidos das montagens *de novo* das 8 cepas de *A. terreus* e os 26 supercontigs do genoma de referência NIH 2624. Os arquivos

[1] Repare na figura 4.8-B que somente os 80.000 pb do contig da cepa BU35 são mostradas a fim de facilitar visualização.

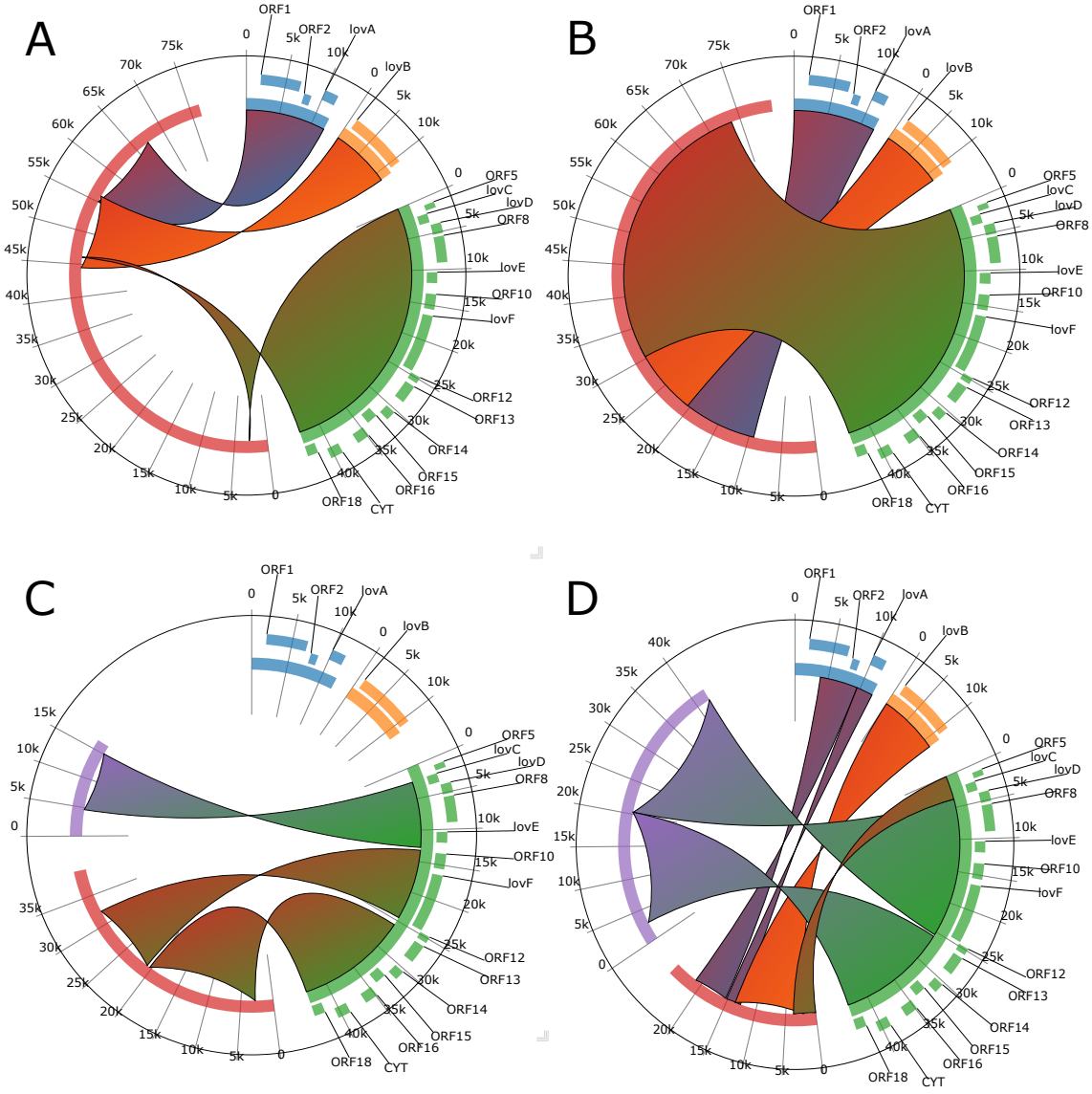

Figura 4.8: Visualização com layout circos dos contigs das cepas ATCC 20542 (`A`), BU35 (`B`), BU27 (`C`) e BU33 (`D`) homólogos às anotações da sequência de DNA responsável pela produção de lovastatina determinadas por Kennedy (1999)

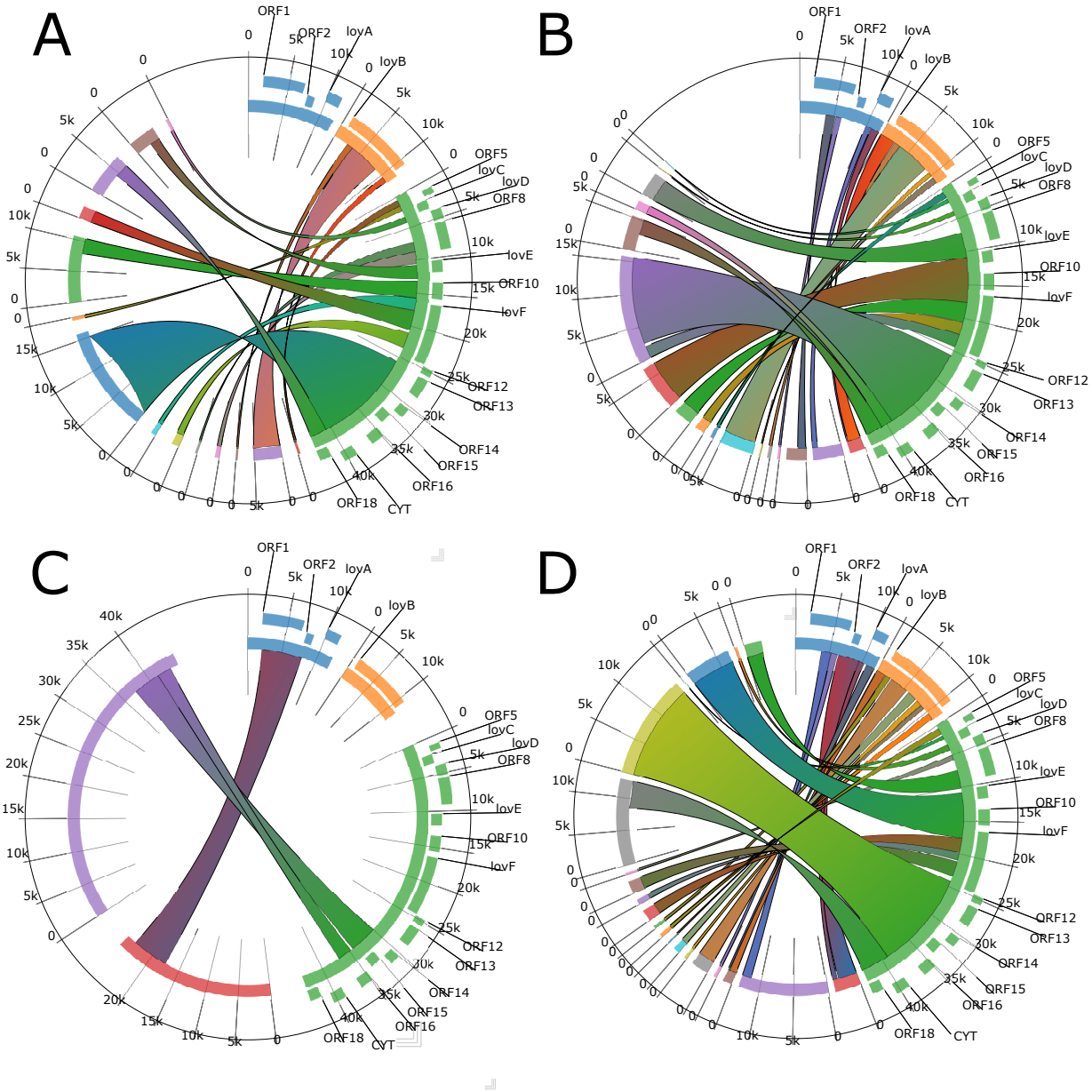

Figura 4.9: Visualização com layout circos dos contigs das cepas U9 (A), U10 (B), U22 (C) e U26 (D) homólogos às anotações da sequência de DNA responsável pela produção de lovastatina determinadas por Kennedy (1999)

contendo os alinhamentos múltiplos, MAF, foram processados com MafTools para obter as estatísticas descritivas dos blocos sintênicos. A porcentagem de conservação genômica entre as montagens de todas as cepas e o genoma NIH 2624 é alta, 86%, com blocos sintênicos em 25,22 Mb das, aproximadamente, 29,33 Mb do tamanho total do genoma de referência (linha 8, tabela 4.6). Apesar disso, o tamanho dos blocos sintênicos é, em média, menor que 2.000 pares de base. No outro extremo, aproximadamente 90 kb são únicos a alguma cepa (linha 1 , tabela 4.6). A tabela 4.6 mostra os resultados para a comparação dos blocos sintênicos de N números de cepas, com N variando de 1 até 8 (coluna 1). Por exemplo, na linha 1 a tabela 4.6 mostra os blocos sintênicos onde há conservação entre o genoma e uma ($N = 1$) das 8 cepas analisadas. Na linha 5, as regiões conservadas entre 5 ($N = 5$) cepas e o genoma de referência e assim por diante.

Tabela 4.6 - A coluna denominada "Conservação" representa a quantidade de cepas pertencentes ao bloco sintênico excluindo-se a cepa referência. Por exemplo, na linha 8, todas as cepas pertencem ao bloco sintênico considerado. As colunas com estatísticas descritivas são a média do tamanho dos bloco sintênicos, o desvio padrão entre os blocos e o maior bloco em pares de base (pb). A cobertura analisada em Mb e porcentagem são referentes ao genoma de referência NIH 2624.

Conservação (#cepas)	Tamanho médio dos blocos (pb)	Desvio padrão (pb)	Tamanho máx do bloco (pb)	Cobertura (Mb)	Cobertura (%)
1	604.66	993.65	5719	0.09	0.003
2	811.83	2271.28	29722	0.19	0.006
3	749.06	1710.73	17896	0.20	0.007
4	698.02	1142.80	11232	0.29	0.010
5	728.30	1259.27	12942	0.25	0.009
6	650.14	857.88	15395	0.65	0.022
7	651.22	783.45	11440	1.39	0.047
8	1892.26	1932.04	25389	25.22	0.860
Média	848.19	426.86	29722	28.26	0.964

A figura 4.10 mostra as informações sobre a conservação genômica entre as cepas e o genoma de referência NIH 2624 na forma de curva Hilbert (métodos seção 3.3.5). Esta representação permite que tenhamos uma visão consolidada das regiões conservadas entre as montagens *de novo* e o genoma de referência. Na figura 4.10 encontram-se os 26 "cromossomos" da referência NIH 2624, denominados supercontigs, porém somente até o supercontig 1.17 garante-se resolução considerável com a curva hilbert usada. O principal motivo é a redução exacerbada no tamanho dos supercontigs de 1.18 até 1.26 (porção inferior direita da figura 4.10). Repare que as regiões ditas órfãs, ou seja, que são únicas à sequência genômica da referência - na figura os caminhos em azul - são, na maioria das vezes, grandes blocos contíguos de sequência. Além disso, nos limites cromossomais parece

haver maior divergência de sequência. Isto é demonstrado na figura 4.10 pelo espectro de cores dos círculos pertencentes ao segmento de linha que interceptam os blocos. Por exemplo, percorrendo-se o caminho do supercontig 1.1 ao 1.2 os círculos começam a clarear a medida que a fronteira desses "cromossomos" se aproxima e, após a passagem da fronteira os círculos tendem a escurecer (conservação aumenta).

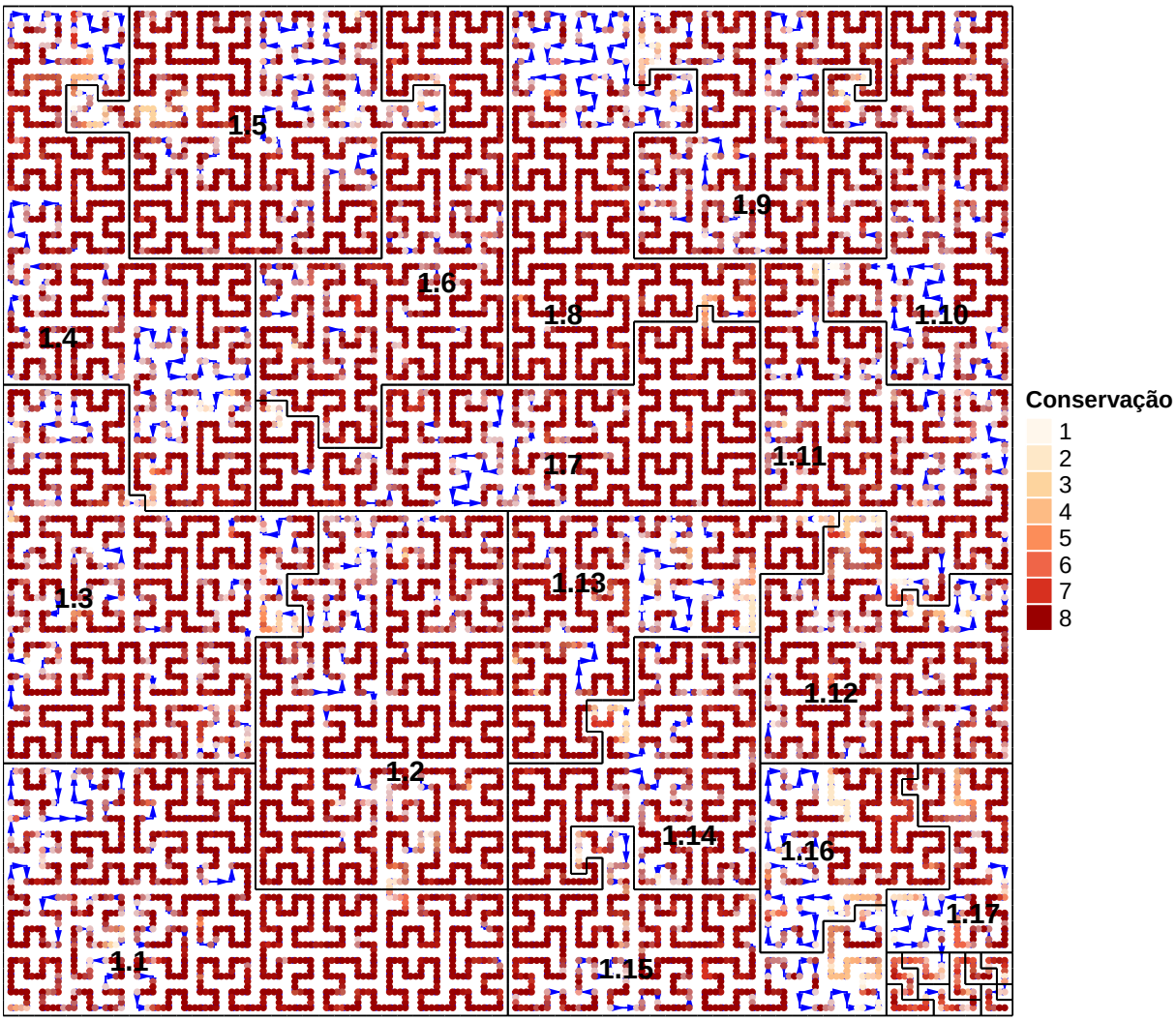

Figura 4.10: Representação das regiões conservadas no genoma de referência NIH para as oito cepas utilizando curvas tipo Hilbert. Cada ponto representa, aproximadamente, 7.162 pares de base do genoma. Os pontos vermelhos mais escuros (legenda item 8) na curva são regiões de alta conservação entre as 8 cepas (ATCC, BU35, BU33, BU27, U26, U9, U10, U22) e o genoma de referência NIH. A legenda indica as cores para cada ponto representativas do grau de conservação. Ou seja, a cor mais clara (item 1) indica a região é conservada entre uma das cepas presentes no estudo e o genoma de referência. O item 2, entre duas cepas presentes no estudo e a referência e assim por diante. Quando uma região do genoma de referência é órfã, ou seja, não possui conservação com nenhuma das 8 cepas o caminho é mostrado em azul. Os 26 cromossomos são separados pelas bordas e, as devidas identificações, com exceção dos 9 últimos, estão dentro da área delimitada de cada cromossomo.

4.7 Predição de agrupamentos de genes biossintéticos

Como um dos objetivos do trabalho é o entendimento das bases genéticas para os diferentes fenótipos de produção de lovastatina pelas cepas, procedeu-se uma anotação especializada para busca de agrupamentos gênicos envolvidos na biossíntese de metabólitos secundários (BCGs) de maneira ampla para todos os genomas montados *de novo*. Para tanto, empregou-se uma metodologia padrão e criou-se uma nova abordagem algorítmica, que são descritas a seguir.

4.7.1 Predição de BCGs com ferramentas tradicionais

Os resultados para predição de metabólitos secundários usando o programa antiSMASH (Weber et al., 2015) na montagem *de novo* da ATCC 20542 apontou 14 agrupamentos de biossíntese de metabólitos secundários (BCGs) classificados como PKS, 18 NRPS, 2 híbridos PKS-NRPS, 6 terpenos, 3 alcaloides indólicos e 13 outros agrupamentos classificados com a categoria outros (figura 4.11). Já o uso do antiSMASH contra as montagens das cepas U9, U10 e U26 indica a quase ausência na predição de possíveis BCGs (fig. 4.11). Na cepa U26, por exemplo, somente um agrupamento de biossíntese de metabólito secundário (BCG) pertencente à classe NRPS foi predito.

Sugere-se que estes resultados são consequência do reduzido tamanho dos contigs obtidos nas montagens destas cepas visto que a ferramenta antiSMASH necessita que o loco do BCG predito esteja contido num único contig/scaffold. Todavia, as predições genômicas de BCGs para a cepa U22, por exemplo, indicam a presença de 13 PKS, 15 NRPS, 4 híbridos PKS-NRPS, 2 sideróforos, 8 terpenos e 14 aglomerados gênicos classificados como outros. Totalizando 53 putativos BCGs encontrados. Fora as predições das cepas U9, U10 e U26 (montagens fragmentadas) as predições das 5 outras cepas estudadas - ATCC 20542, BU35, BU33, BU27, U22 - variam entre 58 BCGs para a cepa ATCC 20542 e 63 para a cepa BU35 (figura 4.11). A lista detalhando os BCGs preditos e que exibiram homologia com BCGs conhecidos encontra-se na tabela do apêndice B.2.

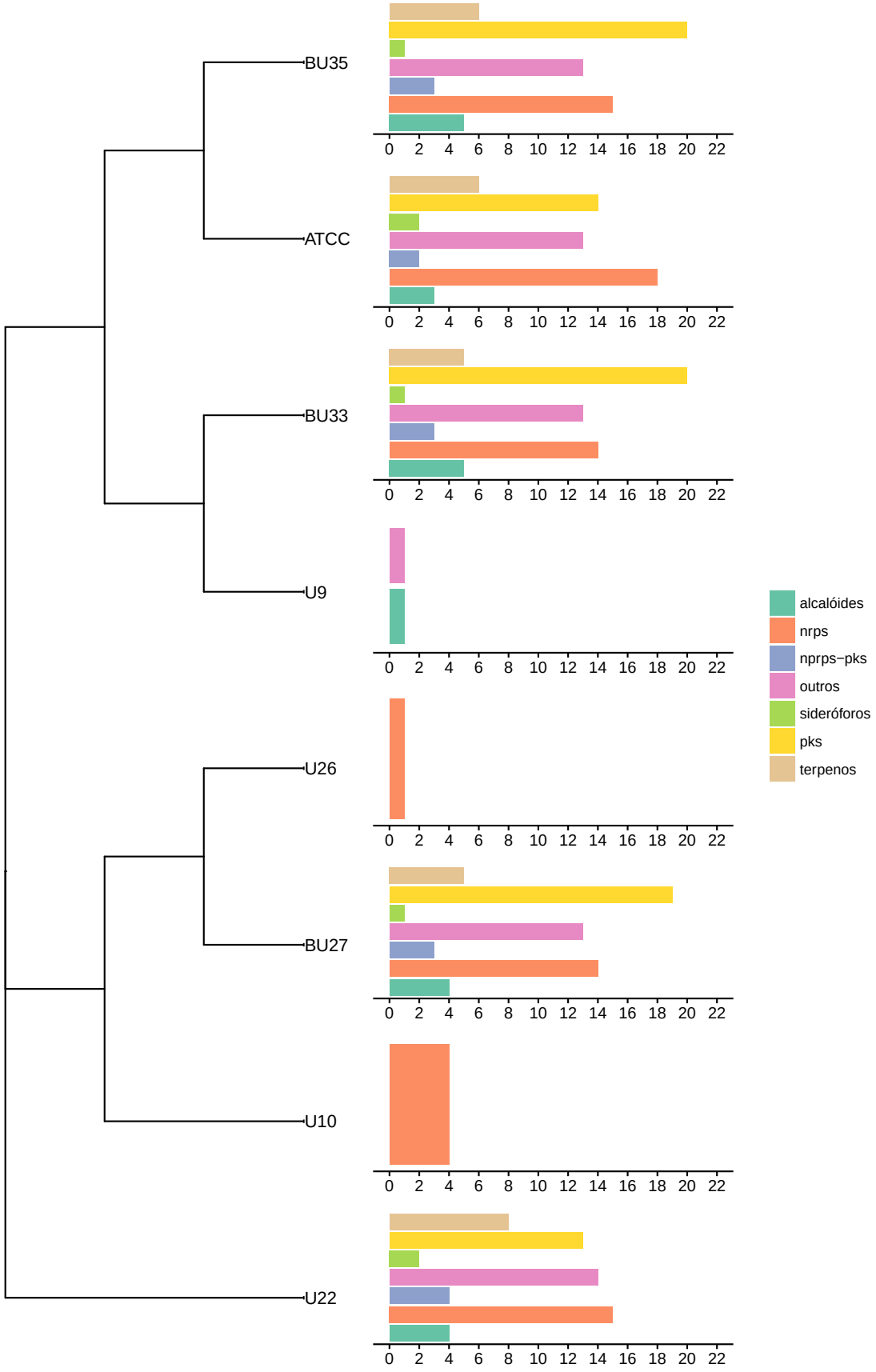

Figura 4.11: Predição dos metabólitos secundários usando antiSMASH nos genomas montados *de novo* das oito cepas. A ferramenta prediz as principais classes de SM e, na figura, a classe "outros" é composta pelos agrupamentos detectados pelo módulo *ClusterFinder* definidos por Cimermancic et al. (2014).

4.7.2 *Metodologia para detecção de BCGs usando estruturas de grafo de Bruijn multicolorido*

Nesta seção pretende-se apresentar uma abordagem alternativa para identificação de BCGs baseada em ressequenciamento e genômica comparativa. A abordagem inicia com a representação do genoma de referência e das amostras sequenciadas na forma de estruturas de dados denominada grafo de Bruijn colorido (Iqbal et al., 2012). O resultado é um grafo concatenado (*joint assembly*) onde cada amostra é indicada por uma cor. Na construção do grafo podem ser usadas tanto o conjunto de leituras-curtas quanto sequências montadas como o genoma de referência ou montagens *de novo*. A figura 4.12-A, retrata duas amostras e o genoma de referência na estrutura de dados grafo de Bruijn colorido: o genoma referência (azul), uma amostra com boa cobertura (preto) neste segmento do genoma e a amostra, em verde, com baixa cobertura e consequente baixa contiguidade. Com a montagem do grafo de Bruijn colorido pode-se seguir o caminho da cor de referência e, buscar caminhos divergentes entre a cor referência e a(s) amostra(s). Na figura 4.12-A observa-se uma deleção de sequência de tamanho maior que 10 kb na amostra em preto caracterizando uma grande variante estrutural (VE) em relação ao genoma em azul no ponto de quebra (*breakpoint*).

Com a estrutura de grafo de Bruijn colorido podemos, portanto, percorrer o genoma de referência (a cor azul) e identificar variantes estruturais (VEs) relacionadas a deleções de sequência nas amostras em relação ao genoma de referência. Na figura 4.12, tais deleções são os caminhos do genoma (em azul) que não possuem correspondente às amostras (preto ou verde) explicitados em vermelho.

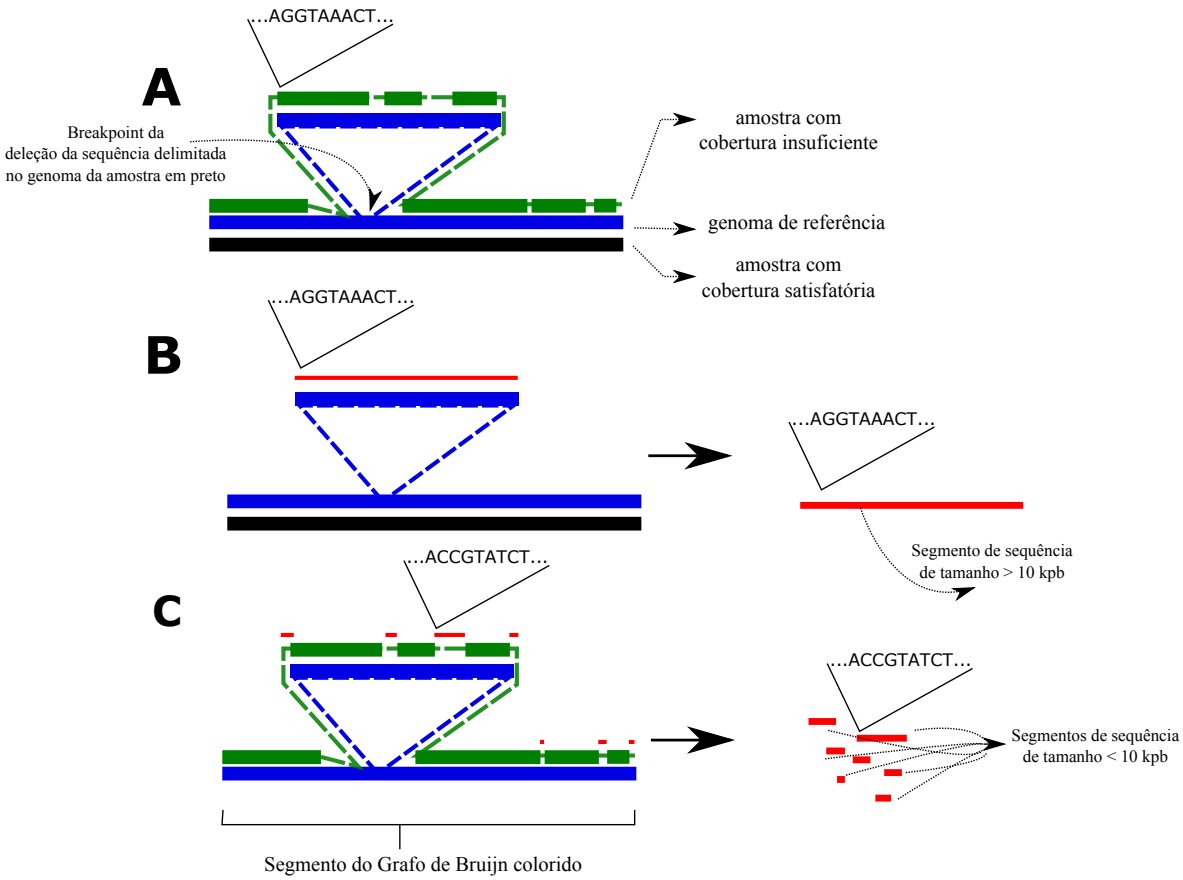

Figura 4.12: Segmentos de sequências de DNA genômico de 3 indivíduos da mesma espécie é mostrado na estrutura de dados grafo de Bruijn colorido. Um dos indivíduos é aquele no qual o genoma de referência foi montado, mostrado em azul na figura. Os outros são oriundos de resequenciamento no qual utilizou-se as leituras resultantes para a construção do grafo e exibem variantes estruturais e coberturas de sequenciamento diferentes. O indivíduo em verde, apesar da baixa cobertura de sequenciamento (falta de contiguidade) tem a mesma arquitetura que o genoma de referência. O outro indivíduo (em preto) tem cobertura satisfatória mas exibe um evento de deleção em relação à referência e ao outro indivíduo. O objetivo da metodologia proposta é identificar estas variantes estruturais de tamanho igual ou superior a 10 kb. Note que a metodologia pode ser aplicada em dados de resequenciamento com baixa cobertura visto que ancoramos as leituras nos genomas dos outros indivíduos e grandes extensões genômicas com cobertura zero devido viés de sequenciamento tem probabilidade mínima de ocorrer.

Na figura 4.12-B, o grafo da amostra, representado pela cor preta, é contíguo indicando que houve amostragem das regiões genômicas, com exceção da deleção de sequência caracterizando a variante estrutural. Percorrendo o grafo em azul, o caminho divergente é mostrado em vermelho. Além do mais, na figura 4.12-C, percorrendo o caminho divergente entre o genoma e a amostra verde obtêm diversas regiões únicas à referência (pequenos segmentos em vermelho). Isto ilustra a situação na qual a amostra não foi sequenciada com cobertura suficiente e portanto existem regiões não amostradas que não são reais variantes

genéticas. A metodologia proposta tenta superar este fato filtrando caminhos divergentes menores que 10 kb.

A aplicação da metodologia proposta na detecção de agrupamentos de biossíntese (BCGs) de metabólitos secundários é baseada em quatro características observadas dessas regiões. A primeira, refere-se à co-localização dos genes essenciais à produção de um metabólito numa única região, isto é, formando um *cluster* com tamanhos médios maiores que 10.000 bases (Bentley, 1999; Sanchez et al., 2012; Hoffmeister e Keller, 2007; Keller et al., 2005a). A segunda refere-se ao fato dos BCGs, supostamente, distribuírem-se em regiões não sintênicas quando diversos genomas de uma espécie são analisados (Andersen et al., 2011). As outras duas suposições envolvem os atributos genéticos dos BCGs. O perfil de produção de metabólitos secundários é cepa específico (Nierman et al., 2005; Andersen et al., 2013) e, em diversas ocasiões, observou-se que a perda ou ganho de BCGs ocorre via único evento de transferência horizontal ou lateral de toda a extensão da sequência (Schönknecht et al., 2014; Fitzpatrick, 2012; Marcet-Houben e Gabaldón, 2010).

Em vista destas características, especula-se que grandes variantes estruturais (VEs) de inserção e deleção no genoma podem ser putativos locos de BCG. A metologia proposta supõe que os caminhos divergentes nos grafos de Bruijn entre dados de indivíduos da mesma espécie podem constituir putativos locos de BCG e, os *breakpoints* delimitam as verdadeiras margens desses agrupamentos. Quando a cobertura de sequenciamento de uma amostra não é boa, ainda assim, com a metodologia é possível verificar quais putativos locos de BCG a amostra conterá em relação às anotações da referência ancorada no grafo de Bruijn.

A vantagem desta metodologia é que se podem utilizar diretamente dados de resequenciamento, em contraste a montagens genômica de alta qualidade, como requerido no caso da detecção com programas baseados em motivos de sequência como antiSMASH (Weber et al., 2015) e SMURF (Khaldi et al., 2010). O esquema na figura 4.13 mostra as diferenças entre as etapas da metodologia de Bruijn colorido e a tradicional ferramenta de predição de MS antiSMASH. Outra vantagem da metodologia, como exemplificaremos posteriormente, é o quase correto delineamento das margens dos BCGs pela metodologia já que programas baseados em homologia tendem a superestimar a quantidade de genes membros dos BCGs (Inglis et al., 2013).

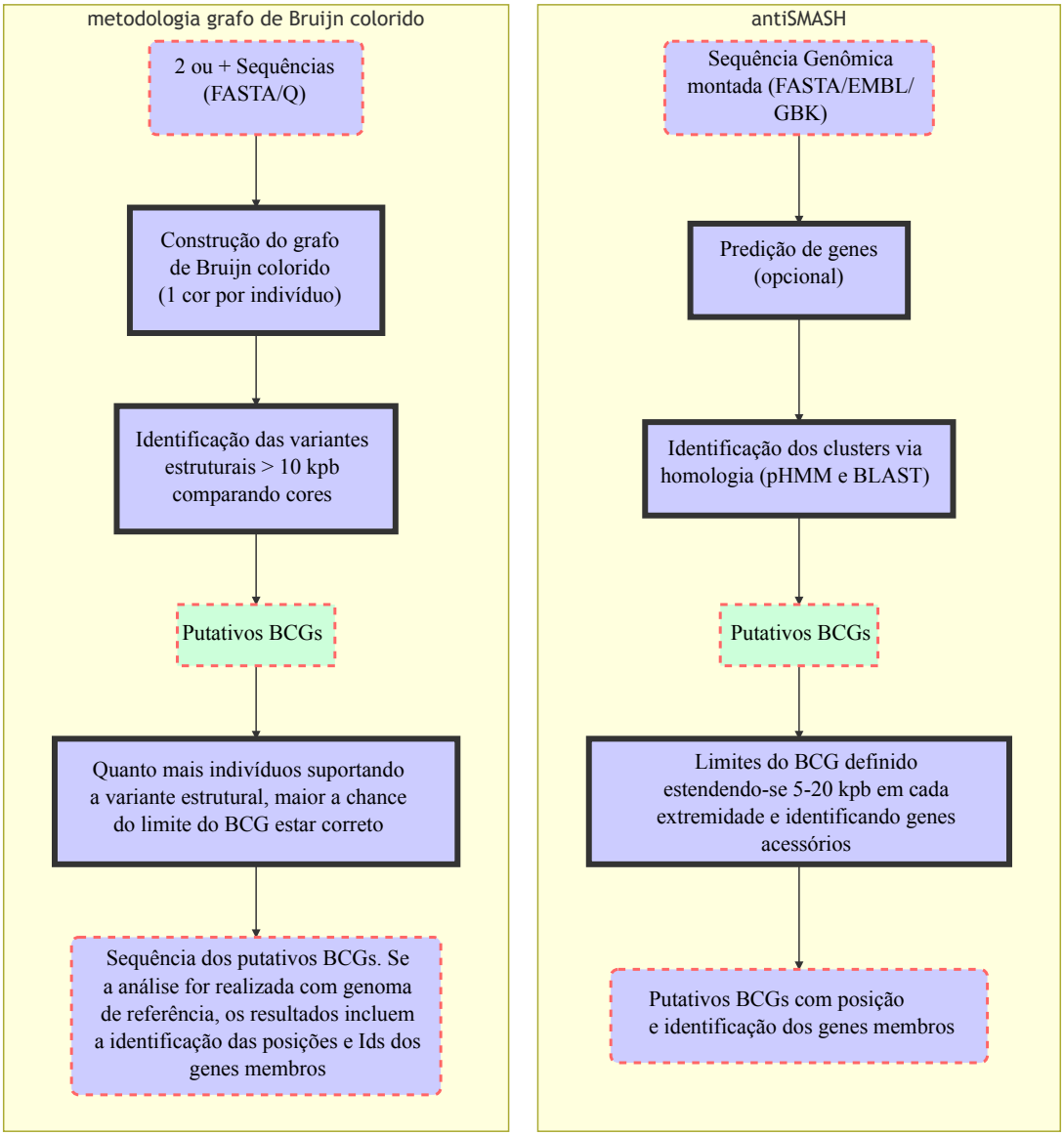

Figura 4.13: Comparando a metodologia de Bruijn colorida contra a ferramenta tradicional para predição de agrupamentos gênicos de biossíntese de metabólitos secundários antiSMASH.

Dado o variável grau de cobertura de sequenciamento das amostras usadas nesta dissertação, bem como da alta fragmentação das montagens para algumas cepas, a metodologia proposta (de Bruijn colorido) foi aplicada a fim de identificar BCGs posicionados no genoma de referência de *A. terreus*.

4.7.3 *Identificação de loci referência putativamente envolvidos em vias de biossíntese de MSs em Aspergillus terreus*

A alta fragmentação dos contigs obtidos nas montagens *de novo* das cepas U9, U10 e U26 impossibilita a busca por homologia de regiões de biossíntese de metabólitos secundários com ferramentas tradicionais como antiSMASH (Weber et al., 2015) e SMURF (Khaldi et al., 2010), motivando o emprego da metodologia de detecção de BCG usando grafos de Bruijn coloridos.

O resultado da metodologia, após percorrer o grafo colorido para cada cepa e o genoma de referência NIH 2624, é uma lista de sequências em fasta com os putativos loci de produção de MSs. No caso deste trabalho, usamos o grafo do genoma de referência como ancora para as amostras. Com isso, não é possível determinar os BCGs da amostra analisada mas identificar quais BCGs estão ausentes em relação à referência. Além disso, avaliamos se a metodologia refina e identifica putativos loci de BCG na referência. Note que não é necessário existir um genoma de referência para aplicar a metodologia, entretanto, deve-se escolher uma amostra para determinar-se o caminho divergente.

Portanto, a metodologia resulta em putativos loci BCG não contidos na cepa (amostra) analisada, mas pertencente na referência. Isto é, determina os loci putativamente deletados na amostra. Com isso, podemos ter uma ideia do perfil alternativo de produção de metabólitos secundários da cepa em relação à referência. Ademais, a metodologia pode, teoricamente, predizer novas regiões genômicas relacionadas à produção de metabólitos pois, não baseia-se no uso de motivo de sequência (Weber, 2014a). Além disso, avaliamos a capacidade da metodologia em identificar os limites dos BCGs, já que ferramentas baseadas em homologia são falhas nesse ponto (Inglis et al., 2013).

A escolha de identificar o caminho divergente seguindo a sequência da referência NIH 2624 é sustentada pela validação da metodologia. Utilizamos a ferramenta antiSMASH (versão 3.0) aplicada contra os putativos BCGs identificados pela metodologia proposta para checar se existem anotações a respeito do loci detectado.

A ferramenta antiSMASH reportou 4 putativas BCGs nas 24 identificadas pela aplicação do grafo colorido na cepa ATCC 20542. Interessante notar que dois agrupamentos dos 4 detectados pela antiSMASH foram identificados como agrupamento de biossíntese dos metabólitos terretonina e epi-aszonaleninas.

O grafo colorido da cepa BU33 gerou 18 putativos agrupamentos gênicos de biossíntese (BCGs), sendo que antiSMASH detectou um deles sem determinar o metabólito produzido. Trinta e quatro BCGs foram identificados na cepa BU27 pela metodologia, sendo oito dessas preditas pela antiSMASH e quatro delas envolvidas na produção de lovastatina, acetilaranotina, terretonina e epi-aszonaleninas. Das 31 BCGs encontradas usando a cepa BU27, 5 foram preditas pela antiSMASH e identificadas como responsáveis pela produção de acetilaranotina, terretonina e epi-aszonaleninas. Usando os dados da cepa U9, vinte regiões referência foram preditas com grafo colorido e dessas somente o agrupamento de terretonina foi identificado pela antiSMASH. Com as informações da cepa U26, trinta e três putativas regiões envolvidas na produção de MS foram identificadas e, 8 destas preditas pela antiSMASH com algumas envolvidas na produção de lovastatina, acetilaranotina, terretonina e epi-aszonaleninas.

A maior quantidade de loci referência envolvidos na produção de MS foi observado na aplicação da metodologia com grafo colorido usando os dados da cepa U22. Identificou-se 70 putativas regiões de BCGs. Já a ferramenta antiSMASH foi capaz de predizer 12 dessas regiões como putativas BCGs e identificou envolvimento na produção dos metabólitos secundários lovastatina e acetilaranotina.

Como é impossível determinar a sensitividade e especificidade da metodologia em detectar BCGs inéditos, a metodologia proposta foi avaliada comparando-se as suas predições contra as predições de BCGs conhecidas detectadas pelas ferramentas antiSMASH e SMURF aplicadas contra o genoma de referência NIH 2624. Estipula-se BCG conhecida, neste trabalho, como aquelas que foram caracterizadas experimentalmente e, seus limites gênicos na referência *A. terreus* NIH 2624 estão estabelecidos na literatura (Guo e Wang, 2014; Yin et al., 2016). Sendo assim, apenas os agrupamentos gênicos de biossíntese (BCGs) dos metabólitos terretonina, acetilaranotina e lovastatina detectados pela metodologia que emprega grafo de Bruijn colorido foram usados para avaliar a metodologia.

A predição dos genes-chave na biossíntese de metabólitos secundários na cepa *Aspergillus terreus* NIH 2624 usando *Secondary Metabolite Unique Regions Finder* (SMURF) identificou 28 genes policetídeos sintases (PKS), 20 genes peptídeos sintases não-ribossomais (NRPS), 2 tipo-PKS, 1 híbrido PKS/NRPS, 14 genes tipo-NRPS e 10 genes triptofano dimetilalil sintase (DMAT). De acordo com Khaldi et al. (2010), existem 15 genes tipo-NRPS preditos, 28 PKS, 22 NRPS, 2 tipo-PKS e 1 híbrido PKS/NRPS. A versão anterior

do software não predizia genes da classe alcalóide cuja enzima chave é DMAT e, além disso, tanto a versão 2.0 quanto a atual (v 3.0) não predizem MS da classe dos terpenos (Weber, 2014b).

4.7.4 Identificando o agrupamento gênico de biossíntese de acetilaranotina

Um desses agrupamentos detectados por (Khaldi et al., 2010), *cluster 47*, foi posteriormente identificado como responsável pela produção do metabólito acetilaranotina (Guo e Wang, 2014). A enzima-chave dessa via é codificada pelo gene ATEG_03470[2]. A ferramenta SMURF identificou 9 genes a montante do gene-chave e 8 genes a jusante (totalizando 18 genes membros para esta BCG).

Aplicando a metodologia com grafo de Bruijn colorido nas cepas BU27, U10, U22 e U26 foram preditos 4 genes a montante e 6 a jusante do gene-chave ATEG_03470, totalizando 11 genes putativos responsáveis pela biossíntese deste metabólito secundário. Os genes detectados pela metodologia de Bruijn condizem com os 9 genes envolvidos na produção de acetilaranotina determinados experimentalmente via deleção gênica realizada por Guo et al. (2013). Além disso, a metodologia de Bruijn detectou 2 genes a mais, o ATEG_03467 que codifica putativo transportador e ATEG_03476 que não parece estar envolvido na produção de acetilaranotina (Guo et al., 2013).

A predição usando a ferramenta antiSMASH (Weber et al., 2015) contra o genoma de referência *A. terreus* NIH 2624 identificou 9 putativas BCGs da classe NRPS, 21 PKS, 2 híbridos PKS/NRPS, 1 sideróforo e 12 putativas BCGs classificadas como outras. Dentre as BCGs preditas estão o putativo agrupamento de biossíntese de terretonina (fig. 4.14) e o de lovastatina (fig. 4.15).

4.7.5 Identificando o agrupamento gênico de biossíntese de terretonina

Após a aplicação da metodologia de Bruijn colorida, identifica-se nas cepas ATCC 20542, BU35, BU27, U9, U10 e U26 a deleção da região responsável pela produção de terretonina e, com isso, delimita-se o loco de biossíntese deste metabólito (figura 4.14-C). A predição via antiSMASH detectou 20 genes totais como putativos membros da BCG de terretonina quando aplicado contra o genoma de referência NIH 2624 (figura 4.14-B).

[2] Esta é a nomenclatura oficial do Ensembl para as anotações gênicas do genoma de referência NIH 2624.

Já os 10 genes identificados pela metodologia de grafos de Bruijn condizem com os genes experimentalmente validados por Yin et al. (2016) (figure 4.14-A).

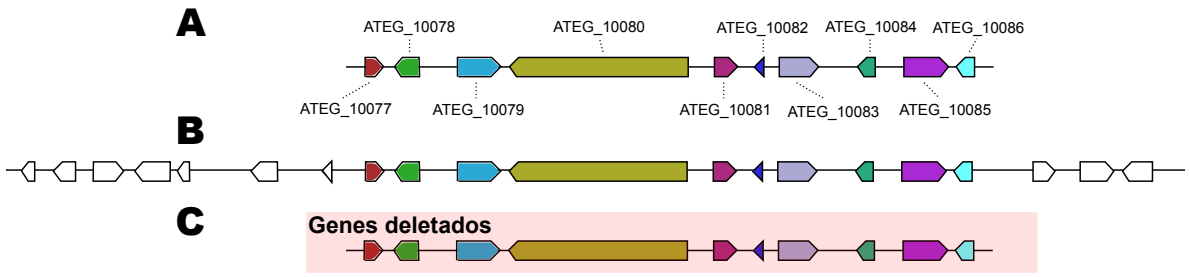

Figura 4.14: Comparação do agrupamento gênico de biossíntese (BCG) do metabólito secundário terretonina identificado com diferentes metodologias. A BCG de terretonina foi experimentalmente validada e contêm 10 genes membros (**A**) anotados com a identificação dos genes de acordo com as anotações do NIH 2624 (MIBiG[3]: BGC0000682). Em **B**, a predição *in silico* desta BCG contra a sequência do genoma de referência NIH 2624 usando a ferramenta antiSMASH. O resultado da aplicação da metodologia usando grafo de Bruijn colorido identificou os mesmos genes membros e limites gênicos que foram estipulados experimentalmente (**C**). Toda a região deste BCG foi deletada nas cepas ATCC 20542, BU35, BU27, U9, U10 e U26 caracterizando uma grande variante estrutural.

4.7.6 Identificando o agrupamento gênico de biossíntese de lovastatina

A predição do BCG de lovastatina no genoma de referência *in silico* usando antiSMASH prediz 14 genes neste agrupamento (fig 4.15-B). Entretanto, é conhecido que o agrupamento de biossíntese de lovastatina compreende 9 genes no total ((Yin et al., 2016); figura 4.15-A), portanto, como mostrado nos casos anteriores a ferramenta antiSMASH superestima a quantidade de genes membros.

Os resultados da aplicação da metodologia grafos de Bruijn coloridos (figura 4.15-C) nos dados brutos das cepas e o genoma de referência como base apontam 3 genes deletados na cepa U26, 5 genes deletados na cepa BU27 na região do BCG de lovastatina, e todos os 9 genes deletados nesta região na cepa U22.

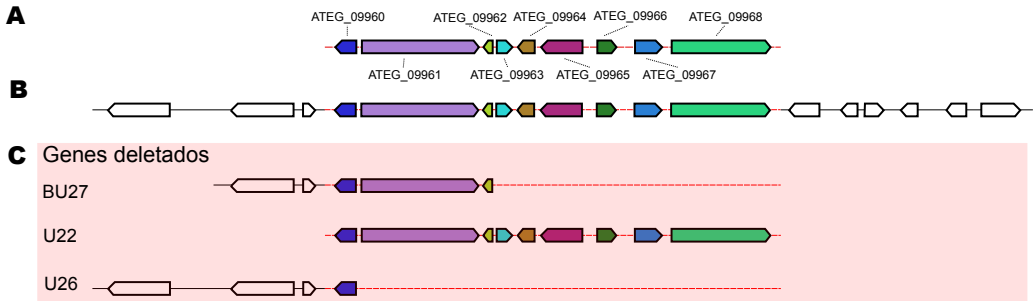

Figura 4.15: Representação dos genes pertencentes ao BCG de lovastatina identificados ilustrando o polimorfismo nesta região entre os diferentes indivíduos. Em **A**, os genes identificados na literatura e, em sua maioria, validados experimentalmente pertencentes ao BCG de lovastatina. Em **B**, a predição *in silico* deste agrupamento no genoma de referência NIH 2624 usando antiSMASH. Em **C**, os resultados da aplicação da metodologia de Bruijn colorido com os dados brutos das cepas. A aplicação detectou a deleção dos genes indicados em três cepas. Representando a incidência de variantes estruturais nesta região. A cepa U22 não possui todos os genes membros do BCG de lovastatina com os limites concordando com os detectados experimentalmente.

4.7.7 Consenso dos pan-genes da cepa NIH 2624

Nas análises anteriores comparando-se os genomas das cepas, observou-se uma alta incidência de variações estruturais, com blocos de deleção englobando diversos genes nas regiões de BCSs. Mas qual seria a variação na composição dos genes ao longo de todo o genoma destas cepas? Notando que pode ocorrer uma grande variação no número de genes entre cepas de microrganismos relacionados, foi introduzido recentemente o conceito de "pan-genoma", que descreve uma espécie como contendo um grupo de genes presentes em todas as cepas, o "genoma cerne", e um menor grupo de genes que ocorre em apenas algumas cepas, o "genoma dispensável" Medini et al. (2005).

Análises deste tipo podem ser feitas usando o programa BLAST buscando os melhores *hits* recíprocos entre os genomas, Mas esta abordagem sofre de escolhas arbitrárias de parâmetros para selecionar os melhores alinhamentos baseado na porcentagem de identidade e extensão da região alinhada Vernikos et al. (2015).

Mais uma vez empregou-se a metodologia de grafo de Bruijn colorido, desta vez para realizar análises pan-genômicas. Nesta abordagem, para buscar genes ausentes em uma cepa constrói-se um grafo de Bruijn colorido com 2 indivíduos (par-a-par) entre a cepa e o genoma de referência. A construção de grafos foi feita entre a referência e cada uma das

8 cepas. Portanto, existem 8 conjuntos compreendendo os genes deletados em cada uma das cepas em relação ao genoma de referência NIH 2624.

Assumindo que os genes foram amostrados de forma independente na população constituída pelos 10.402 genes anotados no genoma de referência, a intersecção entre os 8 conjuntos (ATCC 20542, BU35, BU33, BU27, U9, U10, U22, U26) é estatisticamente significante (p-valor ajustado pelo método de Bonferroni = 1.04×10^{-203}) e contêm 17 genes. Os elementos dessa intersecção são os genes únicos à cepa de referência NIH 2624, ou seja, não possuem genes similares em nenhuma outra cepa analisada. A tabela 4.7 fornece o identificador de cada um desses genes assim como a putativa função, quando determinada pela anotação GO (*Gene Ontology*).

Tabela 4.7 - Genes únicos ao genoma de referência *A. terreus* NIH 2624 em comparação às oito cepas usadas no estudo.

Cromossomo	Início	Fim	Fita	Gene ID	Descrição do Gene Ontology (GO)	Identificador GO
1.3	1793866	1795249	-1	ATEG_02535		
1.3	1795680	1796525	-1	ATEG_02536	Atividade de hidrolase	GO:0016787
1.3	1796788	1798757	1	ATEG_02537	Componente integral de membrana	GO:0016021
1.3	1796788	1798757	1	ATEG_02537	Transporte transmembrana	GO:0055085
1.3	1799430	1800371	-1	ATEG_02538		
1.3	1801766	1803515	1	ATEG_02539		
1.15	1257559	1258113	-1	ATEG_09934		
1.15	1260484	1264114	1	ATEG_09935		
1.15	1264370	1265135	-1	ATEG_09936		
1.15	1266021	1266341	1	ATEG_09937		
1.15	1267143	1268525	-1	ATEG_09938		
1.15	1269737	1270606	1	ATEG_09939	Atividade de hidrolase	GO:0016787
1.15	1271348	1272478	1	ATEG_09940	Atividade oxiredutora	GO:0016491
1.15	1273376	1274330	-1	ATEG_09941		
1.16	577231	581099	-1	ATEG_10158	Atividade hidrolítica, atua em lig. glicosídicas	GO:0016798
1.16	581837	583705	1	ATEG_10159		
1.16	584501	585367	1	ATEG_10160		
1.16	585567	586528	-1	ATEG_10161		

Contratando-se cada amostra de sequenciamento com a cepa NIH 2624 usando a metodologia de grafos observou-se um número significativo de putativos genes que foram deletados nas cepas. O número de genes ausentes nas cepas varia de 101 (cepa BU33) até 351 (cepa U22), o que atesta a noção de "genoma dispensável" no contexto de *Aspergillus terreus*.

Por outro lado, é desejável avaliar o quanto que a ausência de gene se replica em todas as cepas, isto é, dado que um gene é deletado em uma cepa, em quantas outras cepas também seria ausente. Para uma medida global deste comportamento, introduz-se o conceito de

grau de consistência. Em termos gerais, quando a quantidade de genes ausentes em uma cepa também estão ausentes nas demais obtêm-se um maior grau de consistência para a cepa. Isto é, a métrica avalia se os genes ausentes também estão ausentes nas outras cepas.

Formalizando, o grau de consistência r_{ij} é o ranque do conjunto i quando comparado ao conjunto j, onde $j \neq i$ e $j, i = 1, 2, \ldots, 7, 8$. Seguindo a métrica para ranquear genes regulatórios proposta por Wang et al. (2015) calcula-se a pontuação da consistência cumulativa para cada conjunto i como $s_i = \prod \frac{n - r_{ij}}{n}$ para $n = 8$.

A cepa U10 tem maior pontuação de consistência (∼0,15) seguido das cepas BU35, ATCC, BU27, U26, U9, U22 e BU33. Os resultados encontram-se sumarizados na figura 4.16, onde o diâmetro dos círculos é proporcional à quantidade de genes deletados em cada cepa e quanto mais vermelho maior o grau de consistência da cepa. Destaca-se que o maior diâmetro é o referente à cepa U22 representando maior quantidade de genes ausentes em comparação à referência, e a cepa BU33 tem o menor diâmetro, ou seja, esta possui a menor discrepância em termos de composição de genes com a NIH 2624. Além disso, a espessura das arestas indicam o nível de intersecção entre os elementos dos conjuntos ligados (relação entre as cepas). A maior espessura da aresta entre o conjunto da cepa ATCC 20542 (ATCC) e BU35 indica que os genes ausentes nas duas cepas são quase os mesmos.

Em uma visão geral, os genes ausentes não parecem ser únicos à uma cepa. Isto é, não parecem eventos pontuais onde há deleção em uma única cepa. Os genes ausentes ocorrem em mais de um indivíduo especulando-se a coevolução de diversas linhagens carregando determinados genes. Adicionalmente, fica claro que somente um genoma de referência não é o suficiente para representar todos os indivíduos de uma espécie.

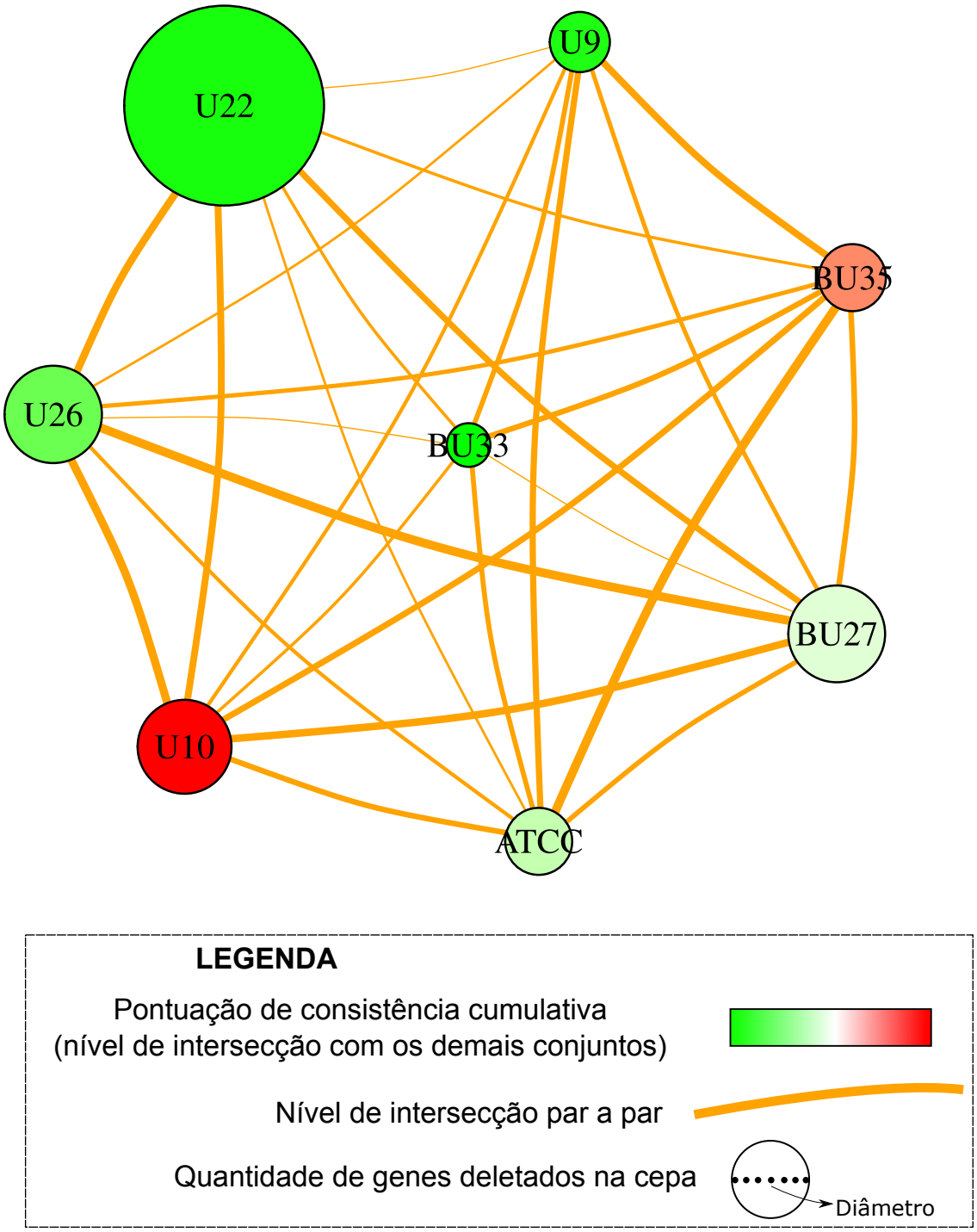

Figura 4.16: Resultados dos genes ausentes em cada cepa determinados pela aplicação da metodologia de grafo de Bruijn colorido. O diâmetro dos círculos é proporcional à quantidade de genes deletados nesta cepa, em relação ao genoma de referência NIH 2624, e do espectro verde para vermelho o grau crescente de consistência. Já a espessura das arestas indicam o nível de intersecção entre os elementos dos conjuntos ligados.

Capítulo 5

Discussão

Nesta dissertação foram analisados dados de resequenciamento genômico de oito cepas de *Aspergillus terreus* através da utilização de plataformas de sequenciamento NGS de leituras-curtas. Esta espécie produz diversos metabólitos secundários de grande importância biotecnológica, dentre os quais a lovastatina, importante composto utilizado como fármaco para redução dos níveis de colesterol.

Entre as cepas sequenciadas, encontra-se a cepa ATCC 20542, super produtora de lovastatina e a utilizada comercialmente para sua produção em larga escala (Boruta e Bizukojc, 2015). Já as outras cepas foram isoladas no Brasil e apresentaram diferentes níveis de produção deste composto, desde a ausência total até níveis intermediários ao da ATCC 20542 (fig 4.1).

Inicialmente aventou-se algumas hipóteses para explicação, em nível molecular, das razões desta diferença fenotípica, dentre as quais podemos citar:

I) Cinética

- Polimorfismos únicos de sequência de DNA (SNPs) poderiam acarretar alterações de resíduos de aminoácidos importantes para a cinética enzimática de enzimas da via biossintética, diminuindo ou abolindo a produção do composto final;

II) Transcricional

- Alterações em regiões *cis*-regulatórias dos genes (ex: promotores) poderiam resultar em diferentes taxas transcricionais dos genes;
- Alterações de elementos regulatórios em *trans*, como fatores de transcrição ou micro RNAs;

III) Estrutura genômica

- Variações estruturais no BCG de lovastatina, como deleções de genes ou inserções de elementos transponíveis;

Para auxiliar na identificação do(s) mecanismo(s) dentre as possibilidades acima foi utilizado o resequenciamento genômico como uma abordagem para correlacionar o fenótipo de produção de lovastatina das cepas com base nas sequências reconstruídas dos genes do agrupamento de biossíntese (BCG) deste composto.

5.1 O resequenciamento de cepas de A. terreus

A espécie *A. terreus* possui um genoma de referência de uma cepa isolada de um caso clínico (NIH 2624), para a qual não foram feitos estudos específicos de medição dos níveis de lovastatina. A sequência deste genoma foi utilizada como referência para o mapeamento de leituras-curtas geradas pelo sequenciamento com a tecnologia Illumina das oito cepas estudadas.

Reparou-se pela distribuição do tamanho das regiões com baixa cobertura ($\leq 2X$), que o perfil de cobertura caracterizado pela queda abrupta da quantidade de leituras mapeadas em longos trechos, como a verificada no perfil do BCG de lovastatina (fig 4.1), e repetindo-se ao longo do genoma (fig 4.2). Portanto, apesar da abundância de pequenas inserções/deleções de sequência entre as cepas sequenciadas e o genoma de referência *A. terreus* NIH 2624 (fig 4.2), a variabilidade de grandes blocos com tamanho $\geq$ 10.000 pb com cobertura nula ou baixa, intitulados ilhas de baixa cobertura, no genoma é notável.

Entre as causas especuladas para explicar a baixa mapeabilidade em algumas regiões estão: deleções no genoma da amostra sequenciada ou inserções na referência, sequências de baixa complexidade, sequências repetitivas ou erros no sequenciamento (Sims et al., 2014). Interessante ressaltar que o protocolo de sequenciamento utilizado fragmenta o genoma aleatoriamente (van Dijk et al., 2014), logo, extensas regiões com cobertura zero são, possivelmente, ligadas a variações genômicas reais e não a erros no processo de sequenciamento. Entretanto, pequenas regiões genômicas com cobertura zero distribuídas aleatoriamente ao longo do genoma podem indicar imprecisão na etapa de sequenciamento, como no caso das cepas U9, U10 e U26.

Portanto a existência das ilhas de baixa cobertura não exclui possíveis erros no processo de sequenciamento. Sendo assim, as regiões dentro do *loco* de produção de lovastatina com cobertura de mapeamento $\leq 2X$ nas cepas U9, U10 e U26 (fig 4.1) são ocasionadas por variantes genéticas reais ou artefatos do sequenciamento? Ressalta-se que além das lacunas de baixa cobertura observados nestas cepas, as métricas gerais de mapeamento, como a cobertura e extensão do mapeamento, para elas são abaixo da média das outras cepas (tabela 4.5). Portanto, executou-se abordagens experimentais envolvendo amplificação via cadeia de polimerase (PCR) para validar as regiões dúbias quanto à cobertura do mapeamento das amostras contra o agrupamento de biossíntese de lovastatina da NIH 2624.

5.1.1 Metodologia para detecção de anomalias de mapeamento

Afim de confirmar a natureza das regiões de cobertura anômala, foi desenvolvida uma metodologia para descoberta destas a partir de análise comparativa de múltiplos dados de mapeamento genômico. Após a identificação das regiões de cobertura anômala são desenhados primers de PCR que as flanqueiam.

Usualmente, os primers são desenhados com base na sequência do genoma de referência. Escolhem-se regiões a montante e a jusante da região alvo e, programas de desenho de primers retornam os melhores candidatos a pares de primers através da otimização de parâmetros que potencialmente afetam a metodologia de PCR como a temperatura de *melting* (T_m), existência de grampos, autocomplementariedade da sequência, conteúdo GC, tamanho do primer, estabilidade 3', possibilidade de dímeros (You et al., 2008). Embora existam programas conhecidos para desenho de primers, como o Primer3 (Untergasser et al., 2012), que incorpora os parâmetros citados para otimização além de modelos termodinâmicos, estes limitam-se ao uso de uma sequência referência para predição dos primers. Essa estratégia, com referência única, é adequada para estudos no qual o objetivo é genotipar, via PCR, alelos variantes já estabelecidos ou validar com PCR em tempo real a diferença na expressão de transcritos conhecidos.

Entretanto, quando dispõe-se de dados de sequenciamento de diversos indivíduos de uma espécie ou de espécies correlatas, as informações sobre as variações genéticas entre eles devem ser consideradas. Ao desenhar primers levando em consideração apenas uma sequência (ex: genoma de referência) a análise torna-se enviesada e sujeita a falhas.

Por exemplo, uma situação de pesquisa comum é o desenho primers a partir de dados de resequenciamento para validar variantes estruturais (VE) devido à inserção de uma sequência em um conjunto de indivíduos em comparação a outros. As etapas comumente seguidas são: mapeamento das leituras brutas contra a sequência referência, determinação das posições no genoma de referência de regiões putativamente importantes via análise visual (Feuk et al., 2006; Spies et al., 2015) ou softwares de predição de VEs (Chen et al., 2009; Rausch et al., 2012; Zhang et al., 2012) e desenho dos primers flanqueando essas regiões.

Os primers são, portanto, desenhados levando-se em conta unicamente a sequência nucleotídica do genoma de referência. Caso existam variações genéticas como polimorfismos únicos de sequência (SNPs), pequenas inserções ou deleções (INDELs), repetições de sequências simples (SSR) ou variações estruturais na região de anelamento do par de primers no genoma dos indivíduos resequenciados não ocorrerá amplificação da região alvo ocasionando um potencial falso negativo. A complexidade aumenta a medida que acrescentam-se indivíduos a genotipar e, com isso, cresce a possibilidade de divergência nucleotídica na região de anelamento devido amostragem de variantes com baixa frequência populacional ou variantes indivíduo específicas inviabilizando a amplificação da região alvo.

Foi proposta uma metodologia (seção 4.4) para desenho de pares de primers que maximiza o uso das informações de sequência contidas no genoma de referência e nas amostras resequenciadas. A metodologia foi aplicada na determinação de alguns conjuntos de primers de regiões selecionadas que apresentavam cobertura anômala entre as cepas (tabela 4.3). As métricas especificidade e sensibilidade foram calculadas comparando-se as predições das amplificações *in silico* e os resultados das análises experimentais de PCR resultando em, aproximadamente, 75% para ambas.

5.1.2 Mapeamento no BCG de lovastatina

As análises visuais das leituras mapeadas na posição do agrupamento gênico de biossíntese de lovastatina no genoma NIH 2624 mostraram espectros de cobertura extremamente desiguais (fig 4.1). Com exceção dos perfis de cobertura das cepas ATCC 20542, BU33 e BU35 que são relativamente constantes, o espectro das demais cepas são extremamente variáveis. No caso das cepas BU27 e U22 existe ausência de cobertura em uma extensa região genômica. Já as cepas U9, U10 e U26 apresentaram amplitude do espectro

de mapeamento muito variável, com pequenas e grandes porções da região analisada com cobertura zero ou, próximas a zero.

Em função deste fato, focalizou-se neste agrupamento de genes para aplicar a metodologia de identificação de regiões de baixa cobertura, otimizando as informações contidas nas cepas para aperfeiçoar o desenho de primers (seção 4.4).

Foram identificadas doze regiões alvo no BCG de lovastatina. Como especulado após análise visual dos espectros de cobertura dos mapeamentos na região genômica de biossíntese de lovastatina na NIH 2624 (fig 4.1) os primers não anelaram nos extensos *loci* com cobertura zero, confirmando uma grande deleção de sequência de DNA entre os genes ORF1 e lovD na cepa BU27, ORF1 até lovH na cepa U22 e ORF1 até lovB na cepa U26 (fig 4.6). Desconsiderou-se nas análises locais a montante do gene ORF1 (fig 4.1) pois, como são regiões teloméricas, o mapeamento das leituras-curtas é enviesado (Trapnell e Salzberg, 2009; Trapnell et al., 2012).

A cepa U22 sofreu maior perda gênica do agrupamento de biossíntese de lovastatina em relação à referência. O tamanho da sequência de DNA perdida é de, aproximadamente, 65 kb e compreende os *loci* do gene ORF2 até ORF14 (fig 4.1). Apesar da amplificação do primer S na cepa U22 especula-se que o loco ORF8 é ausente nesta cepa e que, o motivo da amplificação desta banda é a existência de um gene com alta similaridade à ORF8 fora da BCG de lovastatina. A ORF8 está potencialmente ligada à resistência endógena contra lovastatina pois, essa sequência codifica putativamente uma HMG-CoA redutase (Martín et al., 2014). Com isso, a amplificação observada na figura 4.6 para o primer S desenhado dentro dos limites da ORF8 parece ser devido ao gene homólogo posicionado externamente ao agrupamento gênico de biossíntese de lovastatina.

Note que as cepas que apresentaram os menores valores de cobertura de sequência (tabela 4.5) são as cepas U9, U26 e U10. A amplitude dos espectros de mapeamento dessas cepas são os mais variáveis e, as amplificações de *loci* indicam que a baixa amostragem das leituras-curtas na etapa de sequenciamento é a causa dessas oscilações na cobertura de mapeamento.

Além de verificar a inconsistência nas etapas de sequenciamento das cepas U9, U10 e U26, os experimentos de PCR corroboram para indicar alguns fenótipos. Assim como na cepa U22, a cepa BU27 perdeu o gene codificador da enzima-chave nonacetídeo sintase (lovB) resultando em fenótipo não produtor de lovastatina. A perda do gene codificador

da enzima auxiliar lovA parece impactar na produção de lovastatina na cepa U26. O gene lovA codifica a enzima auxiliar P450 cuja função é oxidar duas vezes dihidromonacolina L para formar monacolina J (Yin et al., 2016).

A variabilidade intraespecífica na quantidade de genes dentro do agrupamentos de biossíntese de metabólitos secundários foi notada anteriormente em *A. flavus* através do resequenciamento de cepas desta espécie (Gibbons et al., 2012). Os resultados obtidos pela metodologia descrita no presente trabalho permitiram identificar que o contexto genômico das regiões de baixa cobertura foi causado por grandes deleções, ou seja, variações estruturais nos genomas das cepas, as quais podem ser diretamente correlacionadas com a produção da lovastatina. Mesmo com essa pequena amostragem de genomas a plasticidade genômica pode ser apreciada.

5.1.3 Montagem genômica de novo das cepas

Para suprir as limitações impostas pela abordagem de mapeamento, como por exemplo, a detecção e caracterização de variantes estruturais (Tattini et al., 2015) como as notadas no BCG de lovastatina ou nas extensas regiões genômicas que exibiram cobertura zero, optou-se por proceder a montagem *de novo* dos genomas das oito cepas. Nesta estratégia as leituras de sequenciamento de cada cepa são utilizadas para obter-se um rascunho do respectivo genoma, ao invés de se basear totalmente no genoma de referência, como é o caso da estratégia de mapeamento.

Aproximadamente 86% do genoma de referência *A. terreus* NIH 2624 é conservado entre as oito cepas analisadas neste estudo (tabela 4.6). Mas acreditamos que esse valor esteja subestimado pelas montagens fragmentadas e, possivelmente, pela distância genética entre NIH 2624 e a cepa U22. Gibbons e Rokas (2013) relataram estabilidade na estrutura de 20 genomas de *Aspergillus* comparados, todos são constituídos por 8 cromossomos e apresentam tamanho total entre 30 e 40 Mpb. O tamanho do genoma de *A. terreus* NIH 2624 é de, aproximadamente, 29,3 Mpb e a montagem dele resultou em 26 supercontigs (Guo e Wang, 2014). Entre as montagens *de novo* realizadas neste estudo, o tamanho médio estimado da maioria dos genomas montados, ∼30,3 Mb, concorda com o tamanho estimado para a espécie.

Entretanto, as montagens das cepas U9, U26 e U10 têm tamanho médio estimado (∼26,6 Mpb) bem inferior ao restante das montagens corroborando, mais uma vez, que a

baixa cobertura de sequenciamento pode ter afetado negativamente a montagem *de novo* e o mapeamento genômico. As montagens fragmentadas como das cepas acima, além de reduzirem o tamanho do genoma montado, também impactam no tamanho dos blocos sintênicos.

Apesar da subestimação do grau de conservação entre os genomas, pode-se inferir pela figura 4.10 que a conservação no cerne do genoma é evidente e que a variabilidade encontra-se concentrada em regiões pontuais do genoma. Acredita-se que os genes contidos nestas regiões pontuais de variabilidade (ilhas genômicas), ou seja, os genes cepa-específicos contribuem pelos diferentes perfis fenotípicos das cepas pois, constatou-se que ilhas genômicas são enriquecidas de genes cepa-específicos envolvidos no transporte e catabolismo de carboidrato, detoxificação e metabolismo secundário em espécies de *Aspergillus* (Fedorova et al., 2008; Andersen et al., 2011).

5.1.3.1 Qualidade das montagens de novo

Observou-se também que as montagens apresentaram métricas estatísticas bastante dissimilares, assim como na variabilidade observada nas métricas de mapeamento (tabela 4.5).

A montagem *de novo* da cepa ATCC 20542 apresentou boas métricas considerando-se a taxa média ($40X$) de cobertura de sequenciamento. As métricas N_{50} do tamanho de contigs (359.971 pb), quantidade de contigs (173) e tamanho do maior contig (991.250 pb) foram comparáveis à publicações de *draft* genomas do gênero *Aspergillus* usando NGS com cobertura de sequenciamento $> 75X$ (Pi et al., 2015; Singh et al., 2016; Gong et al., 2016). Embora existam estudos recentes sobre o perfil de produção de metabólitos secundários da cepa ATCC 20542 (Boruta e Bizukojc, 2015), nenhum estudo objetivou montar o genoma dessa cepa e conectar os metabólitos às sequências genômicas da ATCC 20542. Os resultados obtidos neste trabalho apontam para esta direção.

As montagens das cepas restantes variam em qualidade de acordo, principalmente, com a cobertura de sequenciamento. Esse fato é constatado na alta fragmentação das montagens *de novo* do genoma das cepas U9, U10 e U26. Pode-se notar nas visualizações comparativas das montagens (gráficos do tipo *circos*) que diversos contigs com tamanho menor do que 1 kb alinham-se contra o agrupamento de biossíntese de lovastatina anotado por Kennedy (1999) (cepas U9, U26 e U10 da fig 4.9), principalmente entre os genes lovA e lovF, onde a contiguidade é bem inferior à obtida nas regiões genômicas entre a lovH e ORF18. Sabe-se

que os genes membros responsáveis pela biossíntese de lovastatina compreendem os genes de lovA a lovF (Yin et al., 2016). Os genes ORF14 (mttA), ORF15 (cadA) e ORF16 (mfsA) são genes que codificam proteínas envolvidas no metabolismo primário do fungo, mais especificamente, participam da via de produção do ácido itacônico (Huang et al., 2014), os genes flanqueando o BCG de lovastatina ORF1, ORF2 e lovH não têm função descrita na literatura.

A porcentagem de indicadores que atestam a completude das montagens *de novo* de genomas fúngicos FGMP (Cisse e Stajich, 2016) para as três cepas com montagem fragmentada (U9, U10 e U26) foi igual ou abaixo 90%, enquanto a taxa de marcadores conservados encontrados nas montagens das outras cepas foi acima de 97% (tab 4.5). Novamente, deve ser ressaltado que para estas cepas a cobertura de sequenciamento obtido pode ser uma explicação da não amostragem de porções dos genomas.

No caso da cepa U22, o fenótipo não produtor de lovastatina é corroborado pela ausência dos 9 genes essenciais à produção deste metabólito. Três fatos sustentam a ausência dessa BCG: (1) ausência de mapeamento das leituras da cepa U22 contra o BCG de lovastatina de NIH 2624 (fig 4.1), (2) inexistência de contigs homólogos (fig 4.9) da montagem *de novo* contra as anotações de BCG da ATCC 20542 (Kennedy, 1999) e (3) não amplificação das regiões alvo dentro desse agrupamento (fig 4.6).

Uma possível explicação para a grande variabilidade do agrupamento de biossíntese de lovastatina em indivíduos *A. terreus* é sua localização no genoma. Este agrupamento localiza-se na região subtelomérica do *supercontig* 1.16 no genoma da referência NIH 2624. Uma característica marcante das sequências de DNA subteloméricas é a presença de elementos transponíveis (Palmer e Keller, 2010) associados a eventos de transferência gênica em fungos (Fitzpatrick, 2012).

5.1.4 Exploração de agrupamentos de metabólitos secundários

Agrupamentos de metabólitos secundários são caravanas viajando entre genomas. A sentença anterior faz parte do título do artigo de Wisecaver e Rokas (2015) que aglutinou diversos fatos que sustentam as hipóteses sobre o transferência gênica horizontal (HGT – *Horizontal Gene Transfer*) como agente de inovação dos genomas fúngicos. As inovações genômicas, em muitas ocasiões, impactam diretamente o perfil metabólico do fungo incorrendo, por exemplo, no aumento do repertório de enzimas e metabólitos secretados que

podem garantir ao microrganismo a capacidade de utilizar formas complexas de nutrientes ou vantagens para suportar pressões ecológicas inerentes a competições com outras espécies.

Entretanto, para um fungo adquirir a capacidade de formar um novo metabólito secundário, uma série de genes da via de biossíntese do metabólito deve ser transferida da espécie/cepa que produzia o metabólito (doadora) para a recipiente. A colocalização dos genes envolvidos na biossíntese de metabólitos secundários em blocos genômicos, definindo os agrupamentos de biossíntese (BCGs), é a característica marcante que possibilita a transferência de toda a via de produção do metabólito durante um único evento de transferência gênica (Slot e Rokas, 2011).

Ainda que não exista consenso sobre os mecanismos causais, não se pode duvidar da importância dos genes cepa-específicos para o perfil de enzimas e metabólitos observados em cada indivíduo ou espécie de fungos filamentosos (Balajee et al., 2007; Nierman et al., 2005). A comparação entre *A. oryzae* e *A. flavus* estimou, aproximadamente, 99,5% de identidade em nível nucleotídico entre os genomas das duas espécies embora a primeira espécie é usada na produção de bebidas orientais e possui certificado GRAS (*Generally Recognized as Safe*), ou seja, totalmente seguro para homens e para os animais, e a segunda espécie *A. flavus* é produtora da aflatoxina que induz câncer de fígado (aflatoxina) (Gibbons et al., 2012).

Diante do contexto de que a maior parte da variabilidade entre os fenótipos observados são causadas pelos metabólitos secundários produzidos por cada espécie (Gibbons e Rokas, 2013), procedeu-se à predição de agrupamentos gênicos de biossíntese (BCG) de metabólitos secundários nas montagens *de novo* utilizando a ferramenta antiSMASH (Weber et al., 2015).

Os resultados indicaram uma variação tanto na quantidade de BCGs totais (fig 4.11) quanto no perfil de metabólitos secundários em cada cepa (tabela B.2). Um exemplo foi a identificação de uma putativa BCG que poderia ser responsável pela produção de terretonina em *A. terreus* NIH 2624, que é uma micotoxina pertencente à complexa classe compartilhada policetídeos-terpenóides (meroterpenóides) (Guo et al., 2012). Experimentalmente, o agrupamento de biossíntese de terretonina foi definido como sendo constituído por 10 genes (Guo et al., 2012), sendo os dois principais o gene ATEG_10080 codificador da PKS e ATEG_10077 da enzima terpeno ciclase (fig 4.14). A predição *in silico* com a

ferramenta antiSMASH do putativo BCG de terretonina em NIH 2624 reportou um agrupamento contendo 20 genes (fig 4.14). Visto que a quantidade de genes membros é superior aos 10 genes experimentalmente caracterizados pressupõe-se que o erro em estabelecer os limites do agrupamento de terretonina pelo algoritmo da ferramenta antiSMASH (Weber et al., 2015) seja devido a proximidade de putativos genes com funções auxiliares, como monoxigenases, hidrolases, oxirredutases. Reciprocamente, Inglis et al. (2013) reportaram disparidades entre as predições dos limites dos BCGs identificados pelos tradicionais programas antiSMASH e SMURF em comparação aos BCGs experimentalmente delimitados. Apesar dos erros nos limites, as informações *in silico* das predições de BCGs são importantes para nortear a ligação dos genes essenciais aos metabólitos secundários produzidos (Sanchez et al., 2012; Cacho et al., 2015).

Nesse contexto, uma nova metodologia para identificar agrupamentos de biossíntese (BCGs) de metabólitos secundários foi proposta nesta dissertação. O algoritmo baseia-se em três premissas: (1) os genes membros dos BCGs encontram-se posicionados adjacentemente em sequências de DNA maiores que 10 kb; (2) os perfis de produção de metabólitos secundários variam entre indivíduos da mesma espécie sendo as sequências de DNA de BCGs contribuem diretamente pela variação genética intraespecífica; e (3) as BCGs são transferidas, inseridas ou deletadas entre genomas respeitando-se seus limites gênicos, em sua maioria.

O algoritmo baseia-se, precipuamente, na estrutura de grafos de Bruijn coloridos proposta por Iqbal et al. (2012). Em contraste aos programas antiSMASH e SMURF, baseados em homologia, que usam motivos de sequências conhecidas como principal forma de identificar BCGs, a metodologia proposta não utiliza as informações prévias de motivos conservados. Uma das desvantagens das predições baseadas em homologia é a necessidade de um genoma com boa qualidade de montagem, isto é, com boa contiguidade. Pois, se os genes responsáveis pela produção de um metabólito secundário estão distribuídos em vários contigs, devido fragmentação na montagem, os programas baseados em homologia não são capazes de identificar a BCG ou a identificarão parcialmente. Este foi o caso observado na predição da lovastatina nas montagens *de novo* da cepa BU33 (tabela B.2) onde a montagem *de novo* construiu o BCG de lovastatina em 2 contigs (fig 4.8). Adicionalmente, a utilização de motivos conservados impossibilita a identificação de novas BCGs que não tenham motivos descritos em bancos de dados (Umemura et al., 2015).

A metodologia propõe a identificação de BCGs usando informações da comparação do genoma de diversos indivíduos da mesma espécie ou espécies correlatas visto que o perfil de produção de metabólitos secundários é cepa específico (Andersen et al., 2011; Gibbons et al., 2012; Inglis et al., 2013). Com isso, não existe necessidade de um genoma de referência, bastando a informação de sequenciamento genômico de dois ou mais indivíduos, embora neste trabalho a metodologia foi aplicada usando como âncora o genoma de referência *A. terreus* NIH 2624.

A metodologia é fundamentada na construção da estrutura de dados grafo de Bruijn conjunto, utilizando as informações das cepas sequenciadas, onde cada indivíduo é discernido no grafo por uma "cor" determinando o grafo de Bruijn colorido (fig 4.12). Percorrendo-se cores únicas do genoma de referência ou da cepa base obtemos sequencias candidatas a variantes estruturais que na concepção de fungos filamentosos são propensas a constituir BCGs (Fedorova et al., 2008; Andersen et al., 2011). Esta é essencialmente uma estratégia de genômica comparativa e quanto maior o número de genomas (ou dados de sequenciamento), maior o número e grau de confiança nos BCGs preditos.

Por exemplo, pelas predições da metodologia proposta, as cepas ATCC 20542, BU27, BU35, U9, U10 e U26 não possuem a sequência referente ao agrupamento gênico de biossíntese de terretonina. Embora experimentos que visem confirmar a ausência de produção de terretonina sejam requeridos para validar as predições, Boruta e Bizukojc (2015) não identificaram este metabólito secundário na determinação experimental dos MS produzidos pela cepa ATCC 20542, corroborando com as nossas predições *in silico*. Além disso, a delimitação dos genes membros desta BCG está de acordo com os estabelecidos experimentalmente para esse agrupamento gênico (Guo et al., 2012).

Entretanto, a metodologia proposta não prediz BCGs de metabólitos secundários compartilhados entre a cepa analisada e a cepa base - no caso desta dissertação a cepa base é o genoma de referência NIH 2624 - isto é, ela só prediz BCGs que putativamente representam deleções em alguma das cepas. Por isso, para gerar um putativo perfil de BCGs em montagens com boa qualidade a utilização de ferramentas como antiSMASH e SMURF são fundamentais. As vantagens na utilização da metodologia proposta nesta dissertação é a correta estipulação dos limites dos putativos agrupamentos de biossíntese detectados, a possibilidade de predizer os fenótipos não produtores de indivíduos sequenciados com baixa cobertura comparando-os com um indivíduo bem anotado e com genoma de referência de

qualidade. Por exemplo, as montagens *de novo* das cepas U9, U10 e U26 foram extremamente fragmentadas (tab 4.5) entretanto a metodologia proposta reportou os putativos *loci* ausentes em relação ao genoma de referência NIH 2624 e, com isso, estima-se que a cepa U26 não possua a porção de DNA 5' do BCG de lovastatina e, consequentemente o gene lovA, bem como os BCGs dos conhecidos metabólitos acetilaranotina, epi-aszonalenina e terretonina. Ou seja, sugere-se que o fenótipo da cepa U26 é não produtor para os metabólitos acetilaranotina, epi-aszonalenina e terretonina, pois os genes que codificam as enzimas-chave para estes MS estão ausentes. Além disso, acredita-se que a ausência da lovA impacta na produção de lovastatina. Isso aumenta a quantidade de informações extraídas desta cepa visto que o uso de programas tradicionais de predição baseados em homologia, antiSMASH e SMURF, detectaram somente a possibilidade de biossíntese de um metabólito.

Os limites do agrupamento gênico de biossíntese de lovastatina não foram estipulados corretamente como no caso do metabólito terretonina. Portanto, o nível de confiança na predição dos limites dos BCGs reportados pela metodologia proposta aumenta com a quantidade de indivíduos da espécie comparados dado que na cepa U22 o limite do BCG de lovastatina foi corretamente estipulado. Este fato é generalizado na figura 4.16 onde comparou-se o nível de intersecção par-a-par dos genes deletados reportados pela metodologia proposta. Os genes ausentes na cepa U10 parecem ser os mesmos ausentes em ao menos um outro indivíduo. A mesma situação repete-se para as cepas ATCC 20542, BU27 e BU35. Já a cepa U22 parece possuir mais deleções gênicas únicas a todas as cepas analisadas corroborado pela maior distância filogenética proposta (fig 4.11). Por outro lado, a cepa BU33 tem a menor quantidade de genes deletados em relação ao genoma de referência.

A quantidade de BCGs detectadas tanto pela metodologia usando grafo de Bruijn colorido quanto pelo antiSMASH ou SMURF é reduzida. Apesar disso, os exemplos dos quatro metabólitos (lovastatina, acetilaranotina, terretonina e epi-aszonaleninas) identificados pela metodologia fundamentam dois fatos. O primeiro, estabelece que os loci de BCGs são "ilhas" e, quando submetidas a eventos moleculares não determinados, são deletadas ou inseridas em blocos carreando consigo todos os genes necessários para a biossíntese do metabólito secundário (Wisecaver e Rokas, 2015).

Em suma, nesta dissertação aplicou-se pela primeira vez a metodologia de grafo de

Bruijn colorido para resolver questões pertinentes à metabólitos secundários. Essa abordagem pode ser usada para priorizar a caracterização experimental de BCGs, de classe conhecida ou desconhecida, facilitando o trabalho e tempo consumido no laboratório. Por exemplo, a detecção precisa dos limites do agrupamento gênico pode ser utilizada para a construção de um cassete de expressão heterólogo e ligar o putativo BCG com o metabólito secundário produzido (Unkles et al., 2014).

Esta metodologia aliada à utilização de dados de resequenciamento de diversos indivíduos da mesma espécie de *Aspergillus* que foram submetidos a distintas pressões evolutivas pode auxiliar na descoberta de agrupamentos gênicos de biossíntese de metabólitos secundários crípticos (Fischer et al., 2016), ajudar na determinação dos genes membros do agrupamento facilitando e reduzindo as laboriosas metodologias experimentais de validação de BCG (Inglis et al., 2013) e, possivelmente, complementar para a estipulação de um pangenoma para as espécies de fungos filamentosos visto que a variabilidade intraespecífica é notável.

Capítulo 6

Conclusão

Nesta dissertação demonstrou-se a utilidade da genômica comparativa em responder questões biológicas relacionadas aos metabólitos secundários de fungos filamentosos. Combinando o resequenciamento de DNA de oito cepas de *Aspergillus terreus* de diferentes nichos com fenótipos contrastantes para biossíntese de lovastatina foi possível caracterizar fatores genéticos determinantes na produção deste metabólito secundário. Em particular, observou-se que Variantes estruturais alteram a arquitetura gênica do agrupamento de biossíntese de lovastatina nas cepas. Estas variantes genéticas se manifestaram, em diversas cepas, como deleções de genes essenciais para a biossíntese deste metabólito, sendo validadas experimentalmente.

Apesar da insuficiência na cobertura de sequenciamento para algumas cepas observou-se grandes extensões conservadas entre os genomas das cepas caracterizando um "núcleo" em comum e diversos blocos de sequências esparsos pertencentes a uma ou mais cepas. A composição gênica destes blocos esparsos revelou sequências candidatas a influenciar diretamente no fenótipo das cepas. Como a singularidade do gênero *Aspergillus* é a produção de metabólitos secundários (Gibbons e Rokas, 2013) e, estes são os agentes da diferença fenotípica constatada entre indivíduos da mesma espécie (Andersen et al., 2011; Gibbons et al., 2012; Wisecaver e Rokas, 2015), acredita-se que diversos blocos esparsos contenham locos com putativos agrupamentos gênicos de biossíntese de metabólitos secundários.

Estas conjecturas foram as bases para o desenvolvimento da metodologia de detecção de agrupamentos genes de biossíntese de metabólitos secundários (BCGs) usando grafos de Bruijn coloridos. Essa é a primeira vez que esta estrutura de dados é aplicada na identificação de BCGs e mostrou-se eficiente em identificar e refinar os limites gênicos dos agrupamentos e determinar a ausência de genes de produção de um metabólito em relação

a uma cepa referência. Entre as vantagens da metodologia proposta estão a viabilidade em descobrir BCGs sem enzimas-chave características, pois baseia-se apenas em genômica comparativa, em contrapartida às buscas por motivos de sequências usadas pelas tradicionais ferramentas de busca de BCGs. Outro benefício é a correta predição dos limites gênicos dos BCGs a medida que mais cepas forem comparadas facilitando as laboriosas etapas de validação experimental. Além disso a metodologia não requer montagem *de novo* prévia com boa qualidade, como nas tradicionais metodologias de busca, para detectar os agrupamentos, funcionando apenas com dados de resequenciamento de DNA. No entanto, a dependência no número de cepas com perfis fenotípicos divergente é a condição que a metodologia necessita.

O fato particular encontrado sobre a individualidade no perfil de produção de metabólitos secundários em cada cepa faz questionar se existem possíveis mecanismos moleculares governando esta variabilidade intraespecífica. Especula-se sobre possíveis transferências laterais e horizontais envolvidas nestes eventos (Slot e Rokas, 2011; Fitzpatrick, 2012; Schönknecht et al., 2014) porém a observância, a partir das análises comparativas, que os contigs contendo BCGs montados nas cepas com cobertura média a alta têm os mesmos limites de sequência, levou ao questionamento sobre contextos de sequências favorecendo a inserção ou deleção das BCGs. Presume-se que estes contextos sejam sequências repetitivas, como elementos transponíveis que, assim como a cobertura de sequência, induzem a fragmentação das montagens.

Este fato aliado à observância que o aumento na cobertura de sequenciamento parece não influenciar significativamente na melhora das montagens hipotetiza-se que o sequenciamento com leituras longas contribuirá para responder estas questões. O sequenciamento com leituras longas, seja via as tecnologias das empresas Pacific Biosciences (PacBio) ou Oxford Nanopore (MinION), é perspectiva futura para este trabalho e auxiliará na finalização do *draft* genoma da cepa ATCC 20542 devido sua importância biotecnológica(Boruta e Bizukojc, 2015).

Referências Bibliográficas

Abbas M. M., Malluhi Q. M., Balakrishnan P., Assessment of de Novoassemblers for Draft Genomes: A Case Study with Fungal Genomes, BMC Genomics, 2014, vol. 15, p. 1

Alberts A. W., Chen J., Kuron G., Hunt V., Huff J., Hoffman C., Rothrock J., Lopez M., Joshua H., Harris E., Patchett A., Monaghan R., Currie S., Springer J., Mevinolin: A Highly Potent Competitive Inhibitor of Hydroxymethylglutaryl-Coenzyme A Reductase and a Cholesterol-Lowering Agent, Proceedings of the National Academy of Sciences, 1980, vol. 77, p. 3957

Alex Buerkle C., Gompert Z., Population Genomics Based on Low Coverage Sequencing: How Low Should We Go?, Molecular Ecology, 2013, vol. 22, p. 3028

Alkan C., Sajjadian S., Eichler E. E., Limitations of next-Generation Genome Sequence Assembly, Nature Methods, 2011, vol. 8, p. 61

Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J., Basic Local Alignment Search Tool, Journal of Molecular Biology, 1990, vol. 215, p. 403

Ames B. D., Nguyen C., Bruegger J., Smith P., Xu W., Ma S., Wong E., Wong S., Xie X., Li J. W.-H., Vederas J. C., Tang Y., Tsai S.-C., Crystal Structure and Biochemical Studies of the Trans-Acting Polyketide Enoyl Reductase LovC from Lovastatin Biosynthesis, Proceedings of the National Academy of Sciences, 2012, vol. 109, p. 11144

Anand S., Prasad M. V. R., Yadav G., Kumar N., Shehara J., Ansari M. Z., Mohanty D., SBSPKS: Structure Based Sequence Analysis of Polyketide Synthases, Nucleic Acids Research, 2010, vol. 38, p. W487

Anders S., Visualization of Genomic Data with the Hilbert Curve, Bioinformatics, 2009, vol. 25, p. 1231

Andersen M. R., Nielsen J. B., Klitgaard A., Petersen L. M., Zachariasen M., Hansen T. J., Blicher L. H., Gotfredsen C. H., Mortensen U. H., Accurate Prediction of Secondary Metabolite Gene Clusters in Filamentous Fungi, Proceedings of the National Academy of Sciences of the United States of America, 2013, vol. 110, p. E99

Andersen M. R., Salazar M. P., Schaap P. J., van de Vondervoort P. J. I., Culley D., Thykaer J., Frisvad J. C., Nielsen K. F., Baker S. E., Comparative Genomics of Citric-Acid-Producing Aspergillus Niger ATCC 1015 versus Enzyme-Producing CBS 513.88, Genome Research, 2011, vol. 21, p. 885

Angiuoli S. V., Salzberg S. L., Mugsy: Fast Multiple Alignment of Closely Related Whole Genomes, Bioinformatics, 2011, vol. 27, p. 334

Ansari M. Z., Yadav G., Gokhale R. S., Mohanty D., NRPS-PKS: A Knowledge-Based Resource for Analysis of NRPS/PKS Megasynthases, Nucleic Acids Research, 2004, vol. 32, p. W405

Askenazi M., Driggers E. M., Holtzman D. A., Norman T. C., Iverson S., Zimmer D. P., Boers M.-E., Blomquist P. R., Martinez E. J., Madden K. T., Integrating Transcriptional and Metabolite Profiles to Direct the Engineering of Lovastatin-Producing Fungal Strains, Nature Biotechnology, 2003, vol. 21, p. 150

Balajee S. A., Tay S. T., Lasker B. A., Hurst S. F., Rooney A. P., Characterization of a Novel Gene for Strain Typing Reveals Substructuring of Aspergillus Fumigatus across North America, Eukaryotic Cell, 2007, vol. 6, p. 1392

Bankevich A., Nurk S., Antipov D., Gurevich A. A., Dvorkin M., Kulikov A. S., Lesin V. M., Nikolenko S. I., Pham S., Pevzner P. A., SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, Journal of Computational Biology, 2012, vol. 19, p. 455

Barriuso J., Nguyen D. T., Li J. W.-H., Roberts J. N., MacNevin G., Chaytor J. L., Marcus S. L., Vederas J. C., Ro D.-K., Double Oxidation of the Cyclic Nonaketide

Dihydromonacolin L to Monacolin J by a Single Cytochrome P450 Monooxygenase, LovA, Journal of the American Chemical Society, 2011, vol. 133, p. 8078

Bennett J. W., Bentley R., , 1989 in Neidleman S. L., ed., , Vol. 34, Advances in Applied Microbiology. Academic Press pp 1–28

Bennett J. W., Bentley R., Pride and Prejudice: The Story of Ergot, Perspectives in Biology and Medicine, 1999, vol. 42, p. 333

Bentley R., Secondary Metabolite Biosynthesis: The First Century, Critical Reviews in Biotechnology, 1999, vol. 19, p. 1

Bergmann S., Schümann J., Scherlach K., Lange C., Brakhage A. A., Hertweck C., Genomics-Driven Discovery of PKS-NRPS Hybrid Metabolites from Aspergillus Nidulans, Nature Chemical Biology, 2007, vol. 3, p. 213

Birney E., Assemblies: The Good, the Bad, the Ugly, Nature Methods, 2011, vol. 8, p. 59

Bok J. W., Keller N. P., LaeA, a Regulator of Secondary Metabolism in Aspergillus Spp, Eukaryotic Cell, 2004, vol. 3, p. 527

Bolger A. M., Lohse M., Usadel B., Trimmomatic: A Flexible Trimmer for Illumina Sequence Data, Bioinformatics, 2014, p. btu170

Boruta T., Bizukojc M., Induction of Secondary Metabolism of Aspergillus Terreus ATCC 20542 in the Batch Bioreactor Cultures, Applied Microbiology and Biotechnology, 2015, vol. 100, p. 3009

Brakhage A. A., Regulation of Fungal Secondary Metabolism, Nature Reviews Microbiology, 2013, vol. 11, p. 21

Cabañes F. J., Sanseverino W., Castellá G., Bragulat M. R., Cigliano R. A., Sánchez A., Rapid Genome Resequencing of an Atoxigenic Strain of Aspergillus Carbonarius, Scientific Reports, 2015, vol. 5, p. 9086

Cacho R. A., Tang Y., Chooi Y.-H., Next-Generation Sequencing Approach for Connecting Secondary Metabolites to Biosynthetic Gene Clusters in Fungi, Frontiers in Microbiology, 2015, vol. 5

Cerqueira G. C., Arnaud M. B., Inglis D. O., Skrzypek M. S., Binkley G., Simison M., Miyasato S. R., Binkley J., Orvis J., Wortman J. R., The \emphAspergillus Genome Database: Multispecies Curation and Incorporation of RNA-Seq Data to Improve Structural Gene Annotations, Nucleic Acids Research, 2014, vol. 42, p. D705

Chaisson M. J. P., Huddleston J., Dennis M. Y., Sudmant P. H., Malig M., Hormozdiari F., Antonacci F., Surti U., Sandstrom R., Eichler E. E., Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing, Nature, 2014, vol. 517, p. 608

Chaisson M. J. P., Wilson R. K., Eichler E. E., Genetic Variation and the de Novo Assembly of Human Genomes, Nature Reviews Genetics, 2015, vol. 16, p. 627

Chang K., Creighton C. J., Davis C., Donehower L., Drummond J., Wheeler D., Ally A., Balasundaram M., Birol I., Stuart J. M., The Cancer Genome Atlas Pan-Cancer Analysis Project, Nature Genetics, 2013, vol. 45, p. 1113

Chen K., Wallis J. W., McLellan M. D., Larson D. E., Kalicki J. M., Pohl C. S., McGrath S. D., Wendl M. C., Zhang Q., Locke D. P., Mardis E. R., BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation, Nature Methods, 2009, vol. 6, p. 677

Chen Y.-C., Liu T., Yu C.-H., Chiang T.-Y., Hwang C.-C., Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly, PLoS ONE, 2013, vol. 8, p. e62856

Cimermancic P., Medema M. H., Claesen J., Kurita K., Wieland˜ Brown L. C., Mavrommatis K., Pati A., Godfrey P. A., Fischbach M. A., Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters, Cell, 2014, vol. 158, p. 412

Cingolani P., Platts A., Wang L. L., Coon M., Nguyen T., Wang L., Land S. J., Lu X., Ruden D. M., A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; Iso-2; Iso-3, Fly, 2012, vol. 6, p. 80

Cisse O. H., Stajich J. E., , 2016 Technical Report biorxiv;049619v1 FGMP: Assessing Fungal Genome Completeness and Gene Content

Clarke J., Wu H.-C., Jayasinghe L., Patel A., Reid S., Bayley H., Continuous Base Identification for Single-Molecule Nanopore DNA Sequencing, Nature Nanotechnology, 2009, vol. 4, p. 265

Compeau P. E. C., Pevzner P. A., Tesler G., How to Apply de Bruijn Graphs to Genome Assembly, Nature Biotechnology, 2011, vol. 29, p. 987

Demain A. L., , 2014 in Martín J.-F., García-Estrada C., Zeilinger S., eds, Fungal Biology, Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites. Springer New York pp 1–15

Dietrich D., Vederas J. C., , 2014 in Martín J.-F., García-Estrada C., Zeilinger S., eds, Fungal Biology, Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites. Springer New York pp 263–287

Durbin R. M., Altshuler D. L., Durbin R. M., Lander E. S., Peterson J. L., Schafer A. J., Abecasis G. R., Altshuler D. L., Auton A., Brooks L. D., Durbin R. M., Gibbs R. A., Hurles M. E., McVean G. A., A Map of Human Genome Variation from Population-Scale Sequencing, Nature, 2010, vol. 467, p. 1061

Dutheil J. Y., Gaillard S., Stukenbrock E. H., MafFilter: A Highly Flexible and Extensible Multiple Genome Alignment Files Processor, BMC Genomics, 2014, vol. 15, p. 53

Eddy S. R., Profile Hidden Markov Models, Bioinformatics, 1998, vol. 14, p. 755

Eid J., Fehr A., Gray J., Luong K., Lyle J., Otto G., Peluso P., Rank D., Baybayan P., Bettman B., Bibillo A., Bjornson K., Chaudhuri B., Christians F., Turner S., Real-Time DNA Sequencing from Single Polymerase Molecules, Science, 2009, vol. 323, p. 133

Endo A., Monacolin K, a New Hypocholesterolemic Agent Produced by a Monascus Species., The Journal of Antibiotics, 1979, vol. 32, p. 852

Endo A., A Historical Perspective on the Discovery of Statins, Proceedings of the Japan Academy, Series B, 2010, vol. 86, p. 484

Endo A., Kuroda M., Tsujita Y., ML-236A, ML-236B, and ML-236C, New Inhibitors of Cholesterogensis Produced by Penicillium Citrinum., The Journal of Antibiotics, 1976, vol. 29, p. 1346

Endo A., Negishi Y., Iwashita T., Mizukawa K., Hirama M., Biosynthesis of ML-236B (Compactin) and Monacolin K., The Journal of Antibiotics, 1985, vol. 38, p. 444

Fedorova N. D., Khaldi N., Joardar V. S., Maiti R., Amedeo P., Anderson M. J., Crabtree J., Silva J. C., Badger J. H., Nierman W. C., Genomic Islands in the Pathogenic Filamentous Fungus Aspergillus Fumigatus, PLOS Genet, 11-Apr-2008, vol. 4, p. e1000046

Feuk L., Carson A. R., Scherer S. W., Structural Variation in the Human Genome, Nature Reviews Genetics, 2006, vol. 7, p. 85

Fischer J., Schroeckh V., Brakhage A. A., , 2016 in Schmoll M., Dattenböck C., eds, , Gene Expression Systems in Fungi: Advancements and Applications. Springer International Publishing Cham pp 253–273

Fitzpatrick D. A., Horizontal Gene Transfer in Fungi, FEMS Microbiology Letters, 2012, vol. 329, p. 1

Fujii I., Watanabe A., Sankawa U., Ebizuka Y., Identification of Claisen Cyclase Domain in Fungal Polyketide Synthase WA, a Naphthopyrone Synthase of Aspergillus Nidulans, Chemistry & Biology, 2001, vol. 8, p. 189

Galagan J. E., Genomics of the Fungal Kingdom: Insights into Eukaryotic Biology, Genome Research, 2005, vol. 15, p. 1620

Galagan J. E., Calvo S. E., Cuomo C., Ma L.-J., Wortman J. R., Batzoglou S., Lee S.-I., Baştürkmen M., Spevak C. C., Birren B. W., Sequencing of Aspergillus Nidulans and Comparative Analysis with A. Fumigatus and A. Oryzae, Nature, 2005, vol. 438, p. 1105

Geiser D., Klich M., Frisvad J., Peterson S., Varga J., Samson R., The Current Status of Species Recognition and Identification in Aspergillus, Studies in Mycology, 2007, vol. 59, p. 1

Gibbons J. G., Beauvais A., Beau R., McGary K. L., Latge J.-P., Rokas A., Global Transcriptome Changes Underlying Colony Growth in the Opportunistic Human Pathogen Aspergillus Fumigatus, Eukaryotic Cell, 2012, vol. 11, p. 68

Gibbons J. G., Rokas A., The Function and Evolution of the Aspergillus Genome, Trends in microbiology, 2013, vol. 21, p. 14

Gibbons J. G., Salichos L., Slot J. C., Rinker D. C., McGary K. L., King J. G., Klich M. A., Tabb D. L., McDonald W. H., Rokas A., The Evolutionary Imprint of Domestication on Genome Variation and Function of the Filamentous Fungus Aspergillus Oryzae, Current Biology, 2012, vol. 22, p. 1403

Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., Galibert F., Hoheisel J. D., Jacq C., Johnston M., Louis E. J., Mewes H. W., Murakami Y., Philippsen P., Tettelin H., Oliver S. G., Life with 6000 Genes, Science, 1996, vol. 274, p. 546

Gong W., Cheng Z., Zhang H., Liu L., Gao P., Wang L., Draft Genome Sequence of \emphAspergillus\emph Niger Strain An76, Genome Announcements, 2016, vol. 4, p. e01700

Goodwin S., McPherson J. D., McCombie W. R., Coming of Age: Ten Years of next-Generation Sequencing Technologies, Nature Reviews Genetics, 2016, vol. 17, p. 333

Greenspan M. D., Yudkovitz J. B., Mevinolinic Acid Biosynthesis by Aspergillus Terreus and Its Relationship to Fatty Acid Biosynthesis, Journal of Bacteriology, 1985a, vol. 162, p. 704

Greenspan M. D., Yudkovitz J. B., Mevinolinic Acid Biosynthesis by Aspergillus Terreus and Its Relationship to Fatty Acid Biosynthesis., Journal of Bacteriology, 1985b, vol. 162, p. 704

Grigoriev I. V., , 2013 in Martin F., ed., , The Ecological Genomics of Fungi. John Wiley & Sons, Inc Hoboken, NJ pp 1–20

Grigoriev I. V., Nikitin R., Haridas S., Kuo A., Ohm R., Otillar R., Riley R., Salamov A., Zhao X., Korzeniewski F., Shabalov I., MycoCosm Portal: Gearing up for 1000 Fungal Genomes, Nucleic Acids Research, 2014, vol. 42, p. D699

Guenzi E., Galli G., Grgurina I., Gross D. C., Grandi G., Characterization of the Syringomycin Synthetase Gene Cluster A LINK BETWEEN PROKARYOTIC AND EUKARYOTIC PEPTIDE SYNTHETASES, Journal of Biological Chemistry, 1998, vol. 273, p. 32857

Guo C.-J., Knox B. P., Chiang Y.-M., Lo H.-C., Sanchez J. F., Lee K.-H., Oakley B. R., Bruno K. S., Wang C. C. C., Molecular Genetic Characterization of a Cluster in \emphA\emph. Terreus for Biosynthesis of the Meroterpenoid Terretonin, Organic Letters, 2012, vol. 14, p. 5684

Guo C.-J., Knox B. P., Sanchez J. F., Chiang Y.-M., Bruno K. S., Wang C. C. C., Application of an Efficient Gene Targeting System Linking Secondary Metabolites to their Biosynthetic Genes in \emphAspergillus\emph Terreus, Organic Letters, 2013, vol. 15, p. 3562

Guo C.-J., Wang C. C. C., Recent Advances in Genome Mining of Secondary Metabolites in Aspergillus Terreus, Frontiers in Microbiology, 2014, vol. 5

Guo C.-J., Yeh H.-H., Chiang Y.-M., Sanchez J. F., Chang S.-L., Bruno K. S., Wang C. C. C., Biosynthetic Pathway for the Epipolythiodioxopiperazine Acetylaranotin in Aspergillus Terreus Revealed by Genome-Based Deletion Analysis, Journal of the American Chemical Society, 2013, vol. 135, p. 7205

Gurevich A., Saveliev V., Vyahhi N., Tesler G., QUAST: Quality Assessment Tool for Genome Assemblies, Bioinformatics, 2013, vol. 29, p. 1072

Hendrickson L., Ray Davis C., Roach C., Kim Nguyen D., Aldrich T., McAda P. C., Reeves C. D., Lovastatin Biosynthesis in Aspergillus Terreus: Characterization of Blocked Mutants, Enzyme Activities and a Multifunctional Polyketide Synthase Gene, Chemistry & Biology, 1999, vol. 6, p. 429

Hertweck C., Hidden Biosynthetic Treasures Brought to Light, Nature Chemical Biology, 2009, vol. 5, p. 450

Hoffmeister D., Keller N. P., Natural Products of Filamentous Fungi: Enzymes, Genes, and their Regulation, Natural Product Reports, 2007, vol. 24, p. 393

Holcomb C. L., Höglund B., Anderson M. W., Blake L. A., Böhme I., Egholm M., Ferriola D., Gabriel C., Gelber S. E., Goodridge D., Hawbecker S., Erlich H. A., A Multi-Site Study Using High-Resolution HLA Genotyping by next Generation Sequencing, Tissue Antigens, 2011, vol. 77, p. 206

Holtgrewe M., Kuchenbecker L., Reinert K., Methods for the Detection and Assembly of Novel Sequence in High-Throughput Sequencing Data, Bioinformatics, 2015, p. btv051

Huang S., Holt J., Kao C.-Y., McMillan L., Wang W., A Novel Multi-Alignment Pipeline for High-Throughput Sequencing Data, Database, 2014, vol. 2014, p. bau057

Hutchison C. A., DNA Sequencing: Bench to Bedside and beyond, Nucleic Acids Research, 2007, vol. 35, p. 6227

Inglis D. O., Binkley J., Skrzypek M. S., Arnaud M. B., Cerqueira G. C., Shah P., Wymore F., Wortman J. R., Sherlock G., Comprehensive Annotation of Secondary Metabolite Biosynthetic Genes and Gene Clusters of Aspergillus Nidulans, A. Fumigatus, A. Niger and A. Oryzae, BMC Microbiology, 2013, vol. 13, p. 91

Iqbal Z., Caccamo M., Turner I., Flicek P., McVean G., De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs, Nature Genetics, 2012, vol. 44, p. 226

Jones T., Federspiel N. A., Chibana H., Dungan J., Kalman S., Magee B. B., Newport G., Thorstenson Y. R., Agabian N., Magee P. T., Davis R. W., Scherer S., The Diploid Genome Sequence of Candida Albicans, Proceedings of the National Academy of Sciences, 2004, vol. 101, p. 7329

Keller N. P., Turner G., Bennett J. W., Fungal Secondary Metabolism —from Biochemistry to Genomics, Nature Reviews Microbiology, 2005a, vol. 3, p. 937

Keller N. P., Turner G., Bennett J. W., Fungal Secondary Metabolism —from Biochemistry to Genomics, Nature Reviews Microbiology, 2005b, vol. 3, p. 937

Kennedy J., Modulation of Polyketide Synthase Activity by Accessory Proteins During Lovastatin Biosynthesis, Science, 1999, vol. 284, p. 1368

Kersey P. J., Allen J. E., Armean I., Boddu S., Bolt B. J., Carvalho-Silva D., Christensen M., Davis P., Falin L. J., Grabmueller C., Humphrey J., Kerhornou A., Khobova J., Aranganathan N. K., Maslen G., Staines D. M., Ensembl Genomes 2016: More Genomes, More Complexity, Nucleic Acids Research, 2016, vol. 44, p. D574

Khaldi N., Seifuddin F. T., Turner G., Haft D., Nierman W. C., Wolfe K. H., Fedorova N. D., SMURF: Genomic Mapping of Fungal Secondary Metabolite Clusters, Fungal Genetics and Biology, 2010, vol. 47, p. 736

Kircher M., Kelso J., High-Throughput DNA Sequencing - Concepts and Limitations, BioEssays, 2010, vol. 32, p. 524

Koboldt D. C., Larson D. E., Sullivan L. S., Bowne S. J., Steinberg K. M., Daiger S. P., Exome-Based Mapping and Variant Prioritization for Inherited Mendelian Disorders, The American Journal of Human Genetics, 2014, vol. 94, p. 373

Köster J., Rahmann S., Snakemake\textemdasha Scalable Bioinformatics Workflow Engine, Bioinformatics, 2012, vol. 28, p. 2520

Kozarewa I., Ning Z., Quail M. A., Sanders M. J., Berriman M., Turner D. J., Amplification-Free Illumina Sequencing-Library Preparation Facilitates Improved Mapping and Assembly of (G+C)-Biased Genomes, Nature Methods, 2009, vol. 6, p. 291

Krzywinski M., Schein J., Birol I., Connors J., Gascoyne R., Horsman D., Jones S. J., Marra M. A., Circos: An Information Aesthetic for Comparative Genomics, Genome Research, 2009, vol. 19, p. 1639

Kupfer D. M., Drabenstot S. D., Buchanan K. L., Lai H., Zhu H., Dyer D. W., Roe B. A., Murphy J. W., Introns and Splicing Elements of Five Diverse Fungi, Eukaryotic Cell, 2004, vol. 3, p. 1088

Kurtz S., Phillippy A., Delcher A. L., Smoot M., Shumway M., Antonescu C., Salzberg S. L., Versatile and Open Software for Comparing Large Genomes, Genome Biology, 2004, vol. 5, p. R12

Lander E. S., Initial Impact of the Sequencing of the Human Genome, Nature, 2011, vol. 470, p. 187

Lander E. S., Waterman M. S., Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis, Genomics, 1988, vol. 2, p. 231

Langvad F., A Rapid and Efficient Method for Growth Measurement of Filamentous Fungi, Journal of Microbiological Methods, 1999, vol. 37, p. 97

Larkin M., Blackshields G., Brown N., Chenna R., McGettigan P., McWilliam H., Valentin F., Lopez R., Thompson J., Gibson T., Higgins D., Clustal W and Clustal X Version 2.0, Bioinformatics, 2007, vol. 23, p. 2947

Lee C., Generating Consensus Sequences from Partial Order Multiple Sequence Alignment Graphs, Bioinformatics, 2003, vol. 19, p. 999

Lee C., Grasso C., Sharlow M. F., Multiple Sequence Alignment Using Partial Order Graphs, Bioinformatics, 2002, vol. 18, p. 452

Li H., A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data, Bioinformatics (Oxford, England), 2011, vol. 27, p. 2987

Li H., Durbin R., Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform, Bioinformatics (Oxford, England), 2009, vol. 25, p. 1754

Li M. H., Ung P. M., Zajkowski J., Garneau-Tsodikova S., Sherman D. H., Automated Genome Mining for Natural Products, BMC Bioinformatics, 2009, vol. 10, p. 185

Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L., Law M., Comparison of Next-Generation Sequencing Systems, Journal of Biomedicine and Biotechnology, 2012, vol. 2012, p. 1

Loftus B. J., The Genome of the Basidiomycetous Yeast and Human Pathogen Cryptococcus Neoformans, Science, 2005, vol. 307, p. 1321

Luo R., Liu B., Xie Y., Li Z., Li Y., Yang H., Wang J., Lam T.-W., Wang J., SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler, GigaScience, 2012, vol. 1

Ma S. M., Li J. W.-H., Choi J. W., Zhou H., Lee K. K. M., Moorthie V. A., Xie X., Kealey J. T., Silva N. A. D., Vederas J. C., Tang Y., Complete Reconstitution of a Highly Reducing Iterative Polyketide Synthase, Science, 2009, vol. 326, p. 589

Mabey Gilsenan J., Cooley J., Bowyer P., CADRE: The Central Aspergillus Data REpository 2012, Nucleic Acids Research, 2012, vol. 40, p. D660

McGuire A. M., Pearson M. D., Neafsey D. E., Galagan J. E., Cross-Kingdom Patterns of Alternative Splicing and Splice Recognition, Genome Biology, 2008, vol. 9, p. R50

Machida M., Yamada O., Gomi K., Genomics of Aspergillus Oryzae: Learning from the History of Koji Mold and Exploration of Its Future, DNA Research, 2008, vol. 15, p. 173

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M. A., The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data, Genome Research, 2010, vol. 20, p. 1297

McVean G. A., Altshuler (Co-Chair) D. M., Durbin (Co-Chair) R. M., Abecasis G. R., Bentley D. R., Chakravarti A., Clark A. G., McVean G. A., An Integrated Map of Genetic Variation from 1,092 Human Genomes, Nature, 2012, vol. 491, p. 56

Maiya S., Grundmann A., Li S.-M., Turner G., The Fumitremorgin Gene Cluster of Aspergillus Fumigatus: Identification of a Gene Encoding Brevianamide F Synthetase, ChemBioChem, 2006, vol. 7, p. 1062

Marcet-Houben M., Gabaldón T., Acquisition of Prokaryotic Genes by Fungal Genomes, Trends in Genetics, 2010, vol. 26, p. 5

Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J., Braverman M. S., Chen Y.-J., Chen Z., Rothberg J. M., Genome Sequencing in Microfabricated High-Density Picolitre Reactors, Nature, 2005, vol. 437, p. 376

Martín J.-F., García-Estrada C., Zeilinger S., eds, Biosynthesis and Molecular Genetics of Fungal Secondary Metabolites. Fungal Biology, Springer New York New York, NY, 2014

Martín J. F., Gil J. A., Cloning and Expression of Antibiotic Production Genes, Nature Biotechnology, 1984, vol. 2, p. 63

Medema M. H., Blin K., Cimermancic P., de Jager V., Zakrzewski P., Fischbach M. A., Weber T., Takano E., Breitling R., antiSMASH: Rapid Identification, Annotation and Analysis of Secondary Metabolite Biosynthesis Gene Clusters in Bacterial and Fungal Genome Sequences, Nucleic Acids Research, 2011, vol. 39, p. W339

Medema M. H., Kottmann R., Yilmaz P., Cummings M., Biggins J. B., Blin K., de Bruijn I., Chooi Y. H., Claesen J., Glöckner F. O., Minimum Information about a Biosynthetic Gene Cluster, Nature Chemical Biology, 2015, vol. 11, p. 625

Medini D., Donati C., Tettelin H., Masignani V., Rappuoli R., The microbial pan-genome, Current Opinion in Genetics & Development, 2005, vol. 15, p. 589

Metzker M. L., Sequencing Technologies —the next Generation, Nature Reviews Genetics, 2010, vol. 11, p. 31

Michiels A., Van den Ende W., Tucker M., Van Riet L., Van Laere A., Extraction of High-Quality Genomic DNA from Latex-Containing Plants, Analytical Biochemistry, 2003, vol. 315, p. 85

Miller J. R., Koren S., Sutton G., Assembly Algorithms for next-Generation Sequencing Data, Genomics, 2010, vol. 95, p. 315

Moore G. G., Mack B. M., Beltz S. B., Draft Genome Sequences of Two Closely Related Aflatoxigenic \emphAspergillus Species Obtained from the Ivory Coast, Genome Biology and Evolution, 2016, vol. 8, p. 729

Mulder K. C., Mulinari F., Franco O. L., Soares M. S., Magalhães B. S., Parachin N. S., Lovastatin Production: From Molecular Basis to Industrial Process Optimization, Biotechnology Advances, 2015, vol. 33, p. 648

Mulinari F., Produção de lovastatina em *Aspergillus terreus*: bioprospecção e genômica da biodiversidade brasileira, Universidade Católica de Brasília, 2016, Tese de Doutorado

Myers E. W., Sutton G. G., Delcher A. L., Dew I. M., Fasulo D. P., Flanigan M. J., Kravitz S. A., Mobarry C. M., Reinert K. H., Remington K. A., Anson E. L., Venter J. C., A Whole-Genome Assembly of Drosophila, Science (New York, N.Y.), 2000, vol. 287, p. 2196

Needleman S. B., Wunsch C. D., A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, Journal of Molecular Biology, 1970, vol. 48, p. 443

Nierman W. C., Pain A., Anderson M. J., Wortman J. R., Kim H. S., Arroyo J., Berriman M., Abe K., Archer D. B., Denning D. W., Genomic Sequence of the Pathogenic and Allergenic Filamentous Fungus Aspergillus Fumigatus, Nature, 2005, vol. 438, p. 1151

Pain A., Woodward J., Quail M. A., Anderson M. J., Clark R., Collins M., Fosker N., Fraser A., Harris D., Warren T., Denning D. W., Barrell B., Hall N., Insight into the Genome of Aspergillus Fumigatus: Analysis of a 922kb Region Encompassing the Nitrate Assimilation Gene Cluster, Fungal Genetics and Biology, 2004, vol. 41, p. 443

Palmer J. M., Keller N. P., Secondary Metabolism in Fungi: Does Chromosomal Location Matter?, Current Opinion in Microbiology, 2010, vol. 13, p. 431

Parra G., Bradnam K., Korf I., CEGMA: A Pipeline to Accurately Annotate Core Genes in Eukaryotic Genomes, Bioinformatics, 2007, vol. 23, p. 1061

Payne G. A., Brown M. P., Genetics and Physiology of Aflatoxin Biosynthesis, Annual Review of Phytopathology, 1998, vol. 36, p. 329

Pecyna M., Bizukojc M., Lovastatin Biosynthesis by Aspergillus Terreus with the Simultaneous Use of Lactose and Glycerol in a Discontinuous Fed-Batch Culture, Journal of Biotechnology, 2011, vol. 151, p. 77

Pel H. J., de Winde J. H., Archer D. B., Dyer P. S., Hofmann G., Schaap P. J., Turner G., de Vries R. P., Albang R., Albermann K., Stam H., Genome Sequencing and Analysis of the Versatile Cell Factory Aspergillus Niger CBS 513.88, Nature Biotechnology, 2007, vol. 25, p. 221

Peng Y., Leung H. C. M., Yiu S. M., Chin F. Y. L., , 2010 in Hutchison D., Kanade T., Kittler J., Kleinberg J. M., Mattern F., Mitchell J. C., Steffen B., Sudan M., Terzopoulos D., Tygar D., Vardi M. Y., Weikum G., Berger B., eds, , Vol. 6044, Research in Computational Molecular Biology. Springer Berlin Heidelberg Berlin, Heidelberg pp 426–440

Peterson S. W., Phylogenetic Analysis of Aspergillus Species Using DNA Sequences from Four Loci, Mycologia, 2008, vol. 100, p. 205

Pevzner P. A., Tang H., Waterman M. S., An Eulerian Path Approach to DNA Fragment Assembly, Proceedings of the National Academy of Sciences, 2001, vol. 98, p. 9748

Pi B., Yu D., Dai F., Song X., Zhu C., Li H., Yu Y., A Genomics Based Discovery of Secondary Metabolite Biosynthetic Gene Clusters in Aspergillus Ustus, PLOS ONE, 2015, vol. 10, p. e0116089

Qingdao, 2015 Tolerance mechanisms towards high-temperature and acid stress conditions during itaconic acid biosynthesis in Aspergillus terreus

Rangaswamy V., Jiralerspong S., Parry R., Bender C. L., Biosynthesis of the Pseudomonas Polyketide Coronafacic Acid Requires Monofunctional and Multifunctional Polyketide Synthase Proteins, Proceedings of the National Academy of Sciences, 1998, vol. 95, p. 15469

Raper K. B., Fennel D. I., The genus Aspergillus.. Baltimore, Williams AND Wilkins Co., 1965

Rausch C., Specificity Prediction of Adenylation Domains in Nonribosomal Peptide Synthetases (NRPS) Using Transductive Support Vector Machines (TSVMs), Nucleic Acids Research, 2005, vol. 33, p. 5799

Rausch T., Zichner T., Schlattl A., Stutz A. M., Benes V., Korbel J. O., DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis, Bioinformatics, 2012, vol. 28, p. i333

Reuter J. A., Spacek D. V., Snyder M. P., High-Throughput Sequencing Technologies, Molecular Cell, 2015, vol. 58, p. 586

Rocha R. T., Arquivos da dissertação, https://doi.org/10.5281/zenodo.200423, 2016

Rokas A., Payne G., Fedorova N., Baker S., Machida M., Yu J., Georgianna D. R., Dean R. A., Bhatnagar D., Cleveland T., Wortman J., Maiti R., Nierman W., What Can Comparative Genomics Tell Us about Species Concepts in the Genus Aspergillus?, Studies in Mycology, 2007, vol. 59, p. 11

Ross M. G., Russ C., Costello M., Hollinger A., Lennon N. J., Hegarty R., Nusbaum C., Jaffe D. B., Characterizing and Measuring Bias in Sequence Data, Genome Biology, 2013, vol. 14, p. R51

Rottig M., Medema M. H., Blin K., Weber T., Rausch C., Kohlbacher O., NRPSpredictor2–a Web Server for Predicting NRPS Adenylation Domain Specificity, Nucleic Acids Research, 2011, vol. 39, p. W362

Röttig M., Rausch C., Kohlbacher O., Combining Structure and Sequence Information Allows Automated Prediction of Substrate Specificities within Enzyme Families, PLoS Computational Biology, 2010, vol. 6, p. e1000636

Rypien K. L., Andras J. P., Harvell C. D., Globally Panmictic Population Structure in the Opportunistic Fungal Pathogen \emphAspergillus\emph Sydowii, Molecular Ecology, 2008, vol. 17, p. 4068

Sanchez J. F., Somoza A. D., Keller N. P., Wang C. C. C., Advances in Aspergillus Secondary Metabolite Research in the Post-Genomic Era, Natural Product Reports, 2012, vol. 29, p. 351

Sanger F., Nicklen S., Coulson A. R., DNA Sequencing with Chain-Terminating Inhibitors, Proceedings of the National Academy of Sciences of the United States of America, 1977, vol. 74, p. 5463

Schönknecht G., Weber A. P. M., Lercher M. J., Horizontal Gene Acquisitions by Eukaryotes as Drivers of Adaptive Evolution: Insights & Perspectives, BioEssays, 2014, vol. 36, p. 9

Sharma K. K., Fungal Genome Sequencing: Basic Biology to Biotechnology, Critical Reviews in Biotechnology, 2015, pp 1–17

Simpson J. T., Wong K., Jackman S. D., Schein J. E., Jones S. J., Birol I., ABySS: A Parallel Assembler for Short Read Sequence Data, Genome Research, 2009, vol. 19, p. 1117

Sims D., Sudbery I., Ilott N. E., Heger A., Ponting C. P., Sequencing Depth and Coverage: Key Considerations in Genomic Analyses, Nature Reviews Genetics, 2014, vol. 15, p. 121

Singh N. K., Blachowicz A., Checinska A., Wang C., Venkateswaran K., Draft Genome Sequences of Two \emphAspergillus\emph Fumigatus Strains, Isolated from the International Space Station, Genome Announcements, 2016, vol. 4, p. e00553

Slot J. C., Rokas A., Horizontal Transfer of a Large and Highly Toxic Secondary Metabolic Gene Cluster between Fungi, Current Biology, 2011, vol. 21, p. 134

Smith T., Waterman M., Identification of Common Molecular Subsequences, Journal of Molecular Biology, 1981, vol. 147, p. 195

Sorensen J. L., Auclair K., Kennedy J., Hutchinson C. R., Vederas J. C., Transformations of Cyclic Nonaketides by Aspergillus Terreus Mutants Blocked for Lovastatin Biosynthesis at the lovA and lovC Genes, Organic & Biomolecular Chemistry, 2003, vol. 1, p. 50

Spies N., Zook J. M., Salit M., Sidow A., Svviz: A Read Viewer for Validating Structural Variants, Bioinformatics, 2015, p. btv478

Stanke M., Steinkamp R., Waack S., Morgenstern B., AUGUSTUS: A Web Server for Gene Finding in Eukaryotes, Nucleic Acids Research, 2004, vol. 32, p. W309

Takeda I., Umemura M., Koike H., Asai K., Machida M., Motif-Independent Prediction of a Secondary Metabolism Gene Cluster Using Comparative Genomics: Application to Sequenced Genomes of Aspergillus and Ten Other Filamentous Fungal Species, DNA Research, 2014, p. dsu010

Tattini L., D'Aurizio R., Magi A., Detection of Genomic Structural Variants from Next-Generation Sequencing Data, Frontiers in Bioengineering and Biotechnology, 2015, vol. 3

terreus Broad A., , 2006 About the Aspergillus terreus genome NIH 2624

Testa A. C., Hane J. K., Ellwood S. R., Oliver R. P., CodingQuarry: Highly Accurate Hidden Markov Model Gene Prediction in Fungal Genomes Using RNA-Seq Transcripts, BMC Genomics, 2015, vol. 16, p. 170

Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D. R., Pimentel H., Salzberg S. L., Rinn J. L., Pachter L., Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks, Nature Protocols, 2012, vol. 7, p. 562

Trapnell C., Salzberg S. L., How to Map Billions of Short Reads onto Genomes, Nature biotechnology, 2009, vol. 27, p. 455

Treangen T. J., Salzberg S. L., Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions, Nature Reviews Genetics, 2011

Treiber L. R., Reamer R. A., Rooney C. S., Ramjit H. G., Origin of Monacolin L from Aspergillus Terreus Cultures., The Journal of Antibiotics, 1989, vol. 42, p. 30

Umemura M., Koike H., Machida M., Motif-Independent de Novo Detection of Secondary Metabolite Gene Clusters-toward Identification from Filamentous Fungi, Frontiers in Microbiology, 2015, vol. 6, p. 371

Umemura M., Koike H., Nagano N., Ishii T., Kawano J., Yamane N., Kozone I., Horimoto K., Shin-ya K., Asai K., Machida M., MIDDAS-M: Motif-Independent De Novo Detection of Secondary Metabolite Gene Clusters through the Integration of Genome Sequencing and Transcriptome Data, PLOS ONE, 31-Dec-2013, vol. 8, p. e84028

Unkles S. E., Valiante V., Mattern D. J., Brakhage A. A., Synthetic Biology Tools for Bioprospecting of Natural Products in Eukaryotes, Chemistry & Biology, 2014, vol. 21, p. 502

Untergasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B. C., Remm M., Rozen S. G., Primer3–New Capabilities and Interfaces, Nucleic Acids Research, 2012, vol. 40, p. e115

Utturkar S. M., Klingeman D. M., Land M. L., Schadt C. W., Doktycz M. J., Pelletier D. A., Brown S. D., Evaluation and Validation of de Novo and Hybrid Assembly Techniques to Derive High-Quality Genome Sequences, Bioinformatics, 2014, vol. 30, p. 2709

Van der Auwera G. A., Carneiro M. O., Hartl C., Poplin R., del Angel G., Levy-Moonshine A., Jordan T., Shakir K., DePristo M. A., , 2013 in Bateman A., Pearson W. R., Stein L. D., Stormo G. D., Yates J. R., eds, , Current Protocols in Bioinformatics. John Wiley & Sons, Inc. Hoboken, NJ, USA pp 11.10.1–11.10.33

van Dijk E. L., Auger H., Jaszczyszyn Y., Thermes C., Ten Years of next-Generation Sequencing Technology, Trends in Genetics, 2014, vol. 30, p. 418

Vernikos G., Medini D., Riley D. R., Tettelin H., Ten years of pan-genome analyses, Current Opinion in Microbiology, 2015, vol. 23, p. 148

Wang M., Zhao Y., Zhang B., Efficient Test and Visualization of Multi-Set Intersections, Scientific Reports, 2015, vol. 5, p. 16923

Wang Y., Yang Q., Wang Z., The Evolution of Nanopore Sequencing, Frontiers in Genetics, 2015, vol. 5

Weber T., In Silico Tools for the Analysis of Antibiotic Biosynthetic Pathways, International Journal of Medical Microbiology, 2014a, vol. 304, p. 230

Weber T., In Silico Tools for the Analysis of Antibiotic Biosynthetic Pathways, International Journal of Medical Microbiology, 2014b, vol. 304, p. 230

Weber T., Blin K., Duddela S., Krug D., Kim H. U., Bruccoleri R., Lee S. Y., Fischbach M. A., Müller R., Wohlleben W., Medema M. H., antiSMASH 3.0\textemdasha Comprehensive Resource for the Genome Mining of Biosynthetic Gene Clusters, Nucleic Acids Research, 2015, vol. 43, p. W237

Wisecaver J. H., Rokas A., Fungal Metabolic Gene Clusters\textemdashcaravans Traveling across Genomes and Environments, Microbial Physiology and Metabolism, 2015, vol. 6, p. 161

Xie X., Watanabe K., Wojcicki W. A., Wang C. C. C., Tang Y., Biosynthesis of Lovastatin Analogs with a Broadly Specific Acyltransferase, Chemistry & Biology, 2006, vol. 13, p. 1161

Xu W., Chooi Y.-H., Choi J. W., Li S., Vederas J. C., Da Silva N. A., Tang Y., LovG: The Thioesterase Required for Dihydromonacolin L Release and Lovastatin Nonaketide Synthase Turnover in Lovastatin Biosynthesis, Angewandte Chemie (International Ed. in English), 2013, vol. 52, p. 6472

Yadav G., Gokhale R. S., Mohanty D., SEARCHPKS: A Program for Detection and Analysis of Polyketide Synthase Domains, Nucleic Acids Research, 2003, vol. 31, p. 3654

Yin Y., Cai M., Zhou X., Li Z., Zhang Y., Polyketides in Aspergillus Terreus: Biosynthesis Pathway Discovery and Application, Applied Microbiology and Biotechnology, 2016, pp 1–12

You F. M., Huo N., Gu Y. Q., Luo M.-c., Ma Y., Hane D., Lazo G. R., Dvorak J., Anderson O. D., BatchPrimer3: A High Throughput Web Application for PCR and Sequencing Primer Design, BMC Bioinformatics, 2008, vol. 9, p. 253

Yu J., Fedorova N. D., Montalbano B. G., Bhatnagar D., Cleveland T. E., Bennett J. W., Nierman W. C., Tight Control of Mycotoxin Biosynthesis Gene Expression in Aspergillus Flavus by Temperature as Revealed by RNA-Seq: Aflatoxin Biosynthesis Pathway Gene Regulation, FEMS Microbiology Letters, 2011, vol. 322, p. 145

Zamosky L., , 2012 Special to the Los Angeles Times

Zhang J., Wang J., Wu Y., An Improved Approach for Accurate and Efficient Calling of Structural Variations with Low-Coverage Sequence Data, BMC bioinformatics, 2012, vol. 13, p. S6

Apêndice

Apêndice A

Metodologias Adicionais

As etapas de cultivo, extração de DNA genômico e quantificação de lovastatina no sobrenadante foram planejadas e executadas por Mulinari (2016) como parte da sua tese. Como o projeto é em colaboração com a presente dissertação, e complementam-se em diversas ocasiões, as metodologias experimentais utilizadas pela colaboradora foram descritas neste apêndice.

A.1 Cultivo das cepas de A. terreus

As oito cepas foram cultivadas, em meio sólido, em placas com 20 mL de meio BDA (Batata 100mg/L, Dextrose 10g/L e Agar 15g/L), por 7 dias à 28° C, para crescimento. Após o tempo de crescimento as placas foram mantidas a 4°C.

Para a realização do inóculo, empregado na etapa de extração de DNA genômico e quantificação de lovastatina, foram utilizados 200 mL do meio líquido de cultivo descrito por Pecyna e Bizukojc (2011) e 0,05 g de célula. O crescimento de fungos filamentosos em meio líquido foi mensurado através do peso seco Langvad (1999).

Com o volume determinado correspondente à concentração celular requerida, este foi adicionado a 200 mL de meio líquido, e o inóculo realizado à temperatura de 28° C sob agitação de 180 rpm por aproximadamente 120 horas, segundo protocolo descrito por Pecyna e Bizukojc (2011) para quantificação de lovastatina. Todos os cultivos foram realizados em triplicadas para todas as cepas.

A.2 Extração de DNA Genômico

A extração de DNA foi realizada conforme descrito por Michiels et al. (2003). Primeiramente as amostras foram cultivadas em meio líquido por 48 horas a 180 rpm e 28 °C. O cultivo celular foi centrifugado e o pellet macerado em almofariz de porcelana contendo nitrogênio líquido. Aproximadamente 2g do macerado foi transferido para um falcon de 50 mL e adicionado 15 mL do Tampão de Extração (Tris 100 mM, pH 8,0; NaCl 1,4 M; EDTA 20 mM, pH 8,0; β- mercaptoetanol 0,2%; 2% de Polivinilpirimidina e 2% de CTAB) pré aquecido e mantido sob incubação por 60 minutos a 60° C.

Após a incubação, foram adicionados 15 mL de clorofórmio: álcool isoamílico (24/1), agitados fortemente e centrifugados à 3000 g a 20° C durante 5 minutos. A fase superior foi transferida para um tubo limpo e a lavagem com 15 mL de clorofórmio: álcool isoamílico (24/1) repetida por mais duas vezes. A última coleta de fase aquosa foi homogeneizada por por inversão, e foram adicionados 2/3 do volume de Isopropanol mantidos a 25° C ao longo da noite, para a precipitação ácidos nucleicos. Em seguida as amostras foram centrifugadas a 4500 xg por 15 minutos a 20° C o sobrenadante descartado. O precipitado foi lavado com 15 mL da solução de lavagem (Acetato de Amônio 10 mM e Etanol 70%) e incubado por 15 minutos a temperatura ambiente. Após esse tempo a amostra foi centrifugada a 3000 g a 20° C. A lavagem foi repetida por mais uma vez.

O precipitado resultante da lavagem foi resuspendido em 1 mL de Tampão TE (10 mM Tris, pH 8,0 e EDTA 1 mM) e incubado a 37° C com RNAse por 30 minutos. À amostra tratada com RNAse foi adicionado 1 mL de Fenol e homogeneizada até formar uma emulsão, a qual, foi centrifugada por 5 minutos a 3000 g a 20° C e o sobrenadante transferido para um tubo estéril. A extração foi repetida com fenol: clorofórmio: álcool isoamílico (25/24/1) e com clorofórmio: álcool isoamílico (24/1).

À fase aquosa coletada foi adicionado Acetato de Amônia 7,5 M, pH 7,7; para uma concentração final de 2,5M e dois volumes de Etanol gelado. A amostra foi homogeneizada e incubada no gelo por 5 minutos e em seguida centrifugada a 4500 xg por 15 minutos a 4° C. Ao final, as amostras foram lavas com Etanol 70%.

Após a secagem das amostras, estas foram ressuspendidas em 60 μL de água MilliQ autoclavada e quantificadas no NanoDrop.

A.3 Extração e Quantificação de lovastatina no sobrenadante

A curva de calibração foi realizada com as concentrações de 0,05 g.L^{-1}; 0,075 g.L^{-1}; 0,125 g.L^{-1}; 0,250 g.L^{-1}; 0,5 g.L^{-1}; 0, 75 g.L^{-1}; 1,0 g.L^{-1} e 1,5 g.L^{-1}, utilizando-se o padrão Mevinolin M2147 (Sigma®) solubilizado em acetonitrila 95%. Visto que a lovastatina pode ser encontrada tanto na forma ácida quando na forma lactona, o padrão, que encontra-se na forma lactona, foi acidificado com 50% de NaOH 0,1 M (v/v) incubados por 2 horas a 50°C.

A detecção e quantificação de lovastatina foi realizada por cromatografia líquida de fase reversa utilizando o sistema de HPLC (High-performance Liquid Chromatography) (Shimadzu, Kyoto, Japão) e a coluna Waters Symmetry C18 (4.6mm × 250mm × 5μm). Os eluentes utilizados foram água/0.1% ácido trifluoroacético (eluente A) e acetonitrila/0.1% ácido trifluoroacético (eluente B). O método foi adaptado segundo descrito por Li e colaboradores (2004), onde a amostra foi eluída durante 5 minutos a 35% do eluente B, indo de 35% a 95% do eluente B durante 12,5 minutos e voltando para 35% de eluente B nos últimos 3,5 minutos. O sistema foi mantido a 25° C e lovastatina será detectada usando o sistema PDA (photodiode array assay Shimadzu, Kyoto, Japão) com o comprimento de onda a 238nm. A detecção e quantificação foi realizada nas 8 cepas.

Após o estabelecimento do método de detecção e quantificação, a 1 mL dos sobrenadantes de cada cepa foram acrescentados 500 μL de acetato de etila para extração de lovastatina. Após 5 minutos de centrifugação a 1000 xg, o sobrenadante foi coletado e a extração foi novamente realizada, totalizando um volume de final de 1 mL. As amostras foram secas em speed vacuum e ressuspendidas em 200 μL de acetonitrila para análise em HPLC.

Os picos para lovastatina na forma ácida e forma lactona, foram identificados com aproximadamente 9,5 minutos e 11,5 minutos de tempo de retenção, respectivamente. Para a quantificação das amostras, foi utilizado o software LabSolutions, e todas as amostras foram quantificadas em triplicatas biológicas e técnicas.

Apêndice B

Tabelas adicionais

Tabela B.1 - Supercontigs da montagem de referência NIH 2624.

Cromossomo	Tamanho (pb)
1.1	2751824
1.2	2563198
1.3	2486407
1.4	2214963
1.5	1970408
1.6	1921322
1.7	1912493
1.8	1704888
1.9	1632022
1.10	1587754
1.11	1534813
1.12	1435188
1.13	1423202
1.14	1353958
1.15	1314789
1.16	778656
1.17	525030
1.18	32905
1.19	32146
1.20	29506
1.21	29485
1.22	28116
1.23	21829
1.24	21057
1.25	13141
1.26	12095
	29331195

Tabela B.2 - Lista de metabólitos secundários (SMs) conhecidos preditos pela ferramenta antiSMASH (Weber et al., 2015). O identificador associa o SM com o banco de dados de anotações de agrupamentos de biossíntese de SM, MIBiG (Medema et al., 2015). A tabela mostra a predição para seis cepas sequenciadas no estudo e o genoma de referência NIH 2624. As cepa U9, U26 e U10 não foram mostradas na tabela pois não houve predição de BCGs conhecidas nestas cepas. O valor em cada célula significa a porcentagem de genes compartilhados entre a BCG conhecida e anotada no MIBiG contra a encontrada na montagem *de novo* da cepa considera na coluna.

Id MiBiG	Classe Predita	Metabólito	BU35	BU33	BU27	U22	ATCC	NIH	Gene-chave
BGC0000292	NRPS	Acetylaranotin	80%	90%	X	X	100%	X	ATEG_03470 (ataP)
BGC0000293	NRPS	Acetylaszonalenin	X	66%	X	X	66%	100%	
BGC0000022	PKS-NRPS	Asperfuranone	72%	72%	72%	72%	81%	72%	ATEG_07659 (ateafoG), ATEG_07661 (ateafoE)
BGC0000027	PKS	Azaphilone	25%	29%	25%	12%	29%	16%	
BGC0001239	Other	Biotin	66%	0%	66%	0%	X	X	
BGC0000045	PKS	Dehydrocurvularin	X	X	25%	50%	X	X	
BGC0000046	PKS	Depudecin	X	50%	0%	0%	X	X	
BGC0000348	NRPS-INDOLE	Ergovaline	X	X	X	X	33%	X	
BGC0000355	NRPS	Fumiquinazolines	X	X	X	X	40%	X	
BGC0000070	PKS-NRPS	Griseofulvin	X	X	X	23%	X	X	
BGC0000372	INDOLE-NRPS	Hexadehydro-astechrome	50%	62%	62%	37%	87%	75%	
BGC0001122	NRPS	Isoflavupucine	50%	43%	52%	0%	100%	93%	ATEG_00325
BGC0000089	PKS	Lovastatin	71%	50%	28%	0%	70%	85%	ATEG_09961 (lovB), ATEG_09968 (lovF)
BGC0000098	PKS	Monacolin_K	0%	44%	0%	62%	88%	X	ATEG_09961 (lovB), ATEG_09969 (lovF)
BGC0001278	TERPENE	Nivalenol	8%	0%	0%	9%	X	X	
BGC0001084	NRPS	Notoamide	X	X	X	10%	16%	X	
BGC0000121	PKS	Pestheic acid	X	40%	40%	10%	X	40%	
BGC0000438	PKS	Syringopeptin	X	X	X	X	66%	X	
BGC0000160	PKS	Terreic Acid	88%	100%	88%	88%	100%	72%	ATEG_06275 (atX)
BGC0000161	PKS	Terrein	45%	36%	36%	18%	63%	72%	ATEG_00145 (terA)
BGC0000442	NRPS	Terrequinone	X	X	X	X	60%	X	ATEG_00700
BGC0000682	TERPENE	Terretonin	X	70%	0%	0%	X	90%	ATEG_10080 (trt4)
BGC0001187	NRPS	Xenolozoyenone	X	X	X	X	100%	X	