Universidade Estadual de Campinas
Instituto de Computação

Lucas Miguel de Carvalho

Avaliação de montadores *de novo* de RNA-Seq para análise de expressão diferencial de transcritos

CAMPINAS
2015

Lucas Miguel de Carvalho

Avaliação de montadores *de novo* de RNA-Seq para análise de expressão diferencial de transcritos

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Zanoni Dias
Coorientador: Dr. Felipe Rodrigues da Silva

Este exemplar corresponde à versão final da Dissertação defendida por Lucas Miguel de Carvalho e orientada pelo Prof. Dr. Zanoni Dias.

CAMPINAS
2015

Agência(s) de fomento e nº(s) de processo(s): CNPq, 134480/9-2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

C253a

Carvalho, Lucas Miguel de, 1991-
Avaliação de montadores de novo de RNA-Seq para análise de expressão diferencial de transcritos / Lucas Miguel de Carvalho. – Campinas, SP : [s.n.], 2015.

Orientador: Zanoni Dias.
Coorientador: Felipe Rodrigues da Silva.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. RNA-seq. 2. Bioinformática. 3. Transcriptoma. 4. Genética - Expressão. I. Dias, Zanoni,1975-. II. Silva, Felipe Rodrigues da. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Evaluation of de novo RNA-Seq assemblers in a differentially expressed transcripts analysis
Palavras-chave em inglês:
Bioinformatics
RNA-seq
Transcriptome
Gene expression
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Zanoni Dias [Orientador]
Francisco Pereira Lobo
Benilton de Sá Carvalho
Data de defesa: 10-11-2015
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Lucas Miguel de Carvalho

Avaliação de montadores *de novo* de RNA-Seq para análise de expressão diferencial de transcritos

Banca Examinadora:

- Dr. Zanoni Dias
  Instituto de Computação - Unicamp (Orientador)
- Dr. Francisco Pereira Lobo
  Embrapa Informática Agropecuária
- Dr. Benilton de Sá Carvalho
  Instituto de Matemática e Computação Científica - Unicamp

A ata da defesa com as respectivas assinaturas dos membros da banca encontra-se no processo de vida acadêmica do aluno.

Campinas, 10 de novembro de 2015

Agradecimentos

Eu gostaria de agradecer primeiramente a Deus por me dar forças para driblar todas as dificuldades encontradas ao longo do Mestrado.

Eu gostaria de agradecer as pessoas que me apoiaram e não me deixaram desistir, como meu pai Antônio Miguel, minha mãe Roseli, meu irmão Nicolas, minha namorada Jaqueline Geraldis, e meus amigos, em especial, Lucas Batista, Jacqueline Midlej, Eduardo Amorim, Giovanni Marques, Edmar Santos, Eijy Nagai e Jorge Hongo.

Eu gostaria de agradecer aos meus orientadores Felipe Rodrigues da Silva e Zanoni Dias pela disposição e atenção durante o mestrado.

Eu gostaria de agradecer a todos os pesquisadores da Embrapa que me ajudaram nas horas difćeis, inclusive a própria Embrapa por ceder um ótimo espaço de pesquisa.

Obrigado ao Instituto de Computação pela infraestrutura cedida, e ao CNPq pela ajuda de fomento através da bolsa de mestrado.

Que esta dissertação possa ser usufruída por pesquisadores que um dia possam mudar a história do mundo, sempre lembrando que a única coisa que ninguém consegue tirar de você é o seu conhecimento, logo, sempre o expanda e o compartilhe.

De fato, cada obstáculo em vida pode ser superado por muita determinação, perseverança e empenho, qualidades significativas em um ser humano. Nunca desista, sempre acredite, e lembre-se: se você quiser acertar a montanha, você deve mirar no sol.

Resumo

RNA-Seq é uma tecnologia desenvolvida a partir de dados de sequenciamento de nova geração (NGS) para estudos de transcriptomas. Um pesquisador pode reconstruir isoformas a partir de dados de RNA-Seq sem utilizar um genoma de referência (montagem *de novo*). Uma das diversas análises possíveis utilizando dados de RNA-Seq é encontrar genes ou transcritos diferencialmente expressos. O objetivo deste trabalho é avaliar metodologias de análises em larga escala aplicadas na área da transcriptômica para encontrar transcritos diferencialmente expressos, propondo um critério de classificação que maximize a chance da escolha de algum transcrito montado por um montador *de novo* ser diferencialmente expresso. Essas classificações podem auxiliar a eliminar transcritos falsos positivos a serem analisados posteriormente em bancada por metodologias, como *Real Time PCR (Real Time Polimerase Chain Reaction)*. Dados reais foram testados para validar as montagens *de novo* na procura de transcritos verdadeiramente diferencialmente expressos e resultados mostram que na alteração do volume de dados, a quantidade de verdadeiros positivos (transcritos verdadeiramente diferencialmente expressos) se altera. Concluímos que o melhor montador *de novo* testeado neste estudo é o Trinity.

Abstract

RNA-Seq is a next-generation sequencing data (NGS) technology developed for transcriptome studies. For an organism, a researcher can perform isoform reconstructions from RNA-Seq data without the reference genome (*de novo* assembly). One of the several possible analyses using RNA-Seq data is finding differentially expressed genes or transcripts. This study evaluates analytic methods used in large-scale transcriptome studies for finding differentially expressed transcripts, proposing a data classification criterium that maximizes the chance of choosing a differentially expressed transcript in a *de novo* assembly. This criterium helps eliminate false positives that hinder posterior methods, such as *Real-Time PCR (Polymerase Chain Reaction Real Time)*. Real data were tested to evaluate *de novo* assemblies, searching for differentially expressed transcripts, and the results show that the amount of true positives (truly differentially expressed transcripts) varies with the data volume, favoring libraries with more data. We concluded that the best *de novo* assembler is Trinity.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Até a primeira metade do século XX acreditava-se que moléculas biológicas, como carboidratos, proteínas ou lipídeos, poderiam ser responsáveis pela transmissão da herdabilidade gênica, porém, vários experimentos levaram os cientistas a concluírem que o material genético estava contido em moléculas de DNA [1].

Em 1953, James Watson e Francis Crick descobriram a estrutura dessas moléculas, apresentando um modelo de estrutura revolucionária, de dupla hélice, tomando como base experimentos de difração de raio X, feitos por Rosalind Franklin. Essa estrutura foi definida como uma molécula simples compostas por blocos estruturais, posteriormente denominadas nucleotídeos. Essas descobertas permitiram a compreensão da ação gênica e da hereditariedade a nível molecular.

Paralelamente a esses experimentos, sabia-se da existência de outra molécula de ácido nucléico com propriedades distintas às do DNA, o ácido ribonucléico (RNA), com diferenças estruturais e em composição. Alguns cientistas da época, ao observarem algumas características dessa nova molécula, criaram uma hipótese de que ela estaria relacionada a transmissão de informações entre DNA e proteínas, atuando como intermediário. Essa hipótese foi confirmada em 1957 por Elliot Volkin e Lawrence Astrachan [24].

Todo esse conhecimento gerado com as descobertas das estruturas do DNA e RNA possibilitou o início do sequenciamento dessas moléculas, que desde então, vem sendo aprimoradas. Duas vertentes podem ser observadas com essa técnica, o sequenciamento do genoma, conjunto dos genes de um organismo, com o intuito de identificação da funcionalidade de genes em diversos organismos (genômica), e sequenciamento dos transcritos para identificar a expressão gênica (transcriptômica), porém, com o uso de técnicas de conversão de RNA em DNA complementar (cDNA). Além da funcionalidade dos genes, houve a necessidade de analisar como cada gene agia em determinadas condições ambientais, e tendo em vista que o RNA é a evidência direta de que uma parte do genoma é expressa e pode, portanto, codificar um gene, tecnologias para sequenciar o cDNA começaram a ser desenvolvidas.

O primeiro artigo que descreve a aplicação de vetores de cDNA, para análise da expressão, foi publicado em 1992 [26]. Vários artigos descreveram a utilização desta tecnologia para

a medição da expressão simultânea de milhares de genes no momento em que os primeiros microarrays foram publicados [13]. A princípio a tecnologia de microarray, que surgiu em meados da década de 90, era utilizada fortemente para a análise em níveis de expressão do genoma. Esses experimentos passaram por inovações, e novas abordagens surgiram, como, por exemplo, o Sequênciamento de Nova Geração (tradução livre do inglês NGS - Next Generation Sequence).

Devido à grande demanda de sequenciamentos genômicos e seu alto custo, houve a necessidade da criação de novas tecnologias de sequenciamento que permitiram a geração de dados de forma mais rápida e de baixo custo, esses equipamentos deram origem ao termo NGS. O sequenciamento por esta técnica ocorre de forma paralela, ou seja, otimiza o tempo de processamento em que milhões de sequencias de DNA são geradas por amostra. Os recentes avanços da tecnologia de nova geração também causaram uma revolução no sequenciamento nas moléculas de RNA-mensageiro (cDNA) ampliando muito as aplicações ao estudo de transcriptomas - conjunto dos transcritos de um organismo - em especial, na detecção de transcritos diferencialmente expressos (RNA-Seq).

Após se estabelecer como promissora, várias estratégias que utilizam dados de NGS para geração de dados de transcriptoma e genoma foram comparadas entre si. Existem estudos [10, 20, 45] que usaram critérios de desempenho para comparar os montadores *de novo* (montador que não utiliza um genoma de referência em sua montagem), como a métrica N50, viés de sequência GC, a profundidade das taxas de cobertura, erros de *base-calling* e memória RAM, mas, não houve um estudo mais profundo de como esses montadores *de novo* reagem se forem aplicados sobre suas montagens um *pipeline* de expessão diferencial, ou seja, como essas montagens interferem na busca de transcritos diferencialmente expressos.

As métricas chamadas livre de referência aplicadas a transcriptomas não dependem de um *benchmark* pré-selecionado na análise, ou seja, essas métricas dependem exclusivamente da montagem feita com os dados de RNA-Seq. Como exemplo de tais métricas podemos citar o N50 e afins, tamanho médio de transcritos e número de transcritos gerados. Estudos como [20,27] citam que as métricas livres de referência podem ser muito bem manipuladas, como por exemplo a métrica N50, que diz que melhores transcritos resultarão em um maior número de sobreposições identificadas entre os *reads* de entrada, portanto, terá mais transcritos longos montados. No entanto, é fácil ver que um conjunto trivial construído pela concatenação de todos os *reads* de entrada em uma única entrada maximizará o tamanho do transcrito e consequentemente do N50.

Existe uma metodologia criada para avaliar a qualidade com que uma biblioteca de RNA-Seq é criada e medir o uso efetivo do reagente no experimento, chamada *spike-ins*. As transcrições e o tamanho das sequências *spike-ins* são conhecidas, uma vez que eles são adicionados no começo da amostra e servem para controle. Essa metodologia é utilizada para medir a eficiência do montador *de novo*, já que a montagem final do montador deve conter os *spike-ins* colocados inicialmente.

Ao final da análise de um experimento de RNA-Seq, para obtenção de transcritos diferencialmente expressos, são utilizadas ferramentas estatísticas que verificam a quantidade de

dados que um transcrito apresenta em uma condição em relação à outra. Os transcritos com diferentes padrões de expressão são selecionados para serem testados em bancada, objetivando sua real validação. Existem estudos que comparam ferramentas de análise de expressão diferencial, como será explicado na Seção 3.5, mas ambas as análises são feitas sobre dados simulados ou já depositados no Genbank, sem levar em consideração a montagem do genoma ou o transcriptoma.

Pensando na decisão de escolha do melhor montador *de novo* para uma análise de expressão diferencial em RNA-Seq, que tenha como objetivo analisar sua capacidade de identificá-los, usamos uma metodologia, que se baseia não somente em métricas de livre referência, mas em outras que utilizem um transcriptoma de referência para a comparação. Ao final da análise poderemos notar qual montagem se aproxima da referência com maior precisão na identificação de transcritos diferencialmente expressos e qual métrica maximiza a escolha de um transcrito montado *de novo* ser diferenciamente expresso.

Após a introdução e a análise dos desafios existentes em uma montagem de RNA-Seq, neste trabalho tomamos como motivação três aspectos. O primeiro dele seria a não padronização de um montador *de novo* referência. O segundo aspecto é em relação aos resultados de um montagem *de novo*, já que elas ainda possuem erros e são sensíveis a eles. O terceiro aspecto está associado a testes estatísticos retornarem em seus resultados muitos transcritos falsos positivos.

Tomando essas motivações como um princípio, este trabalho tem como objetivos:

1. Avaliar os montadores *de novo* descritos na literatura e sua influência na identificação de transcritos verdadeiramente diferencialmente expressos;
2. Avaliar como os montadores *de novo* se comportam à alteração do volume de dados;
3. Propor critérios de seleção que identifiquem o menor número de transcritos falsos positivos em uma análise de expressão diferencial que posteriormente podem ser analisados em bancada por metodologias como *Real Time PCR (Real Time Polimerase Chain Reaction)*.

O texto da dissertação está organizado da seguinte maneira: o Capitulo 2 apresenta alguns conceitos básicos a fim de introduzir o contexto do trabalho; o Capítulo 3 descreve a metodologia de montagem de dados de RNA-Seq desenvolvida ao longo do mestrado; o Capítulo 4 descreve os métodos utilizados; o Capítulo 5 discorre sobre o aprimoramento da metodologia de RNA-Seq e os resultados obtidos; Capítulo 6 apresenta as conclusões finais da dissertação e propõe alguns trabalhos futuros.

Capítulo 2

Conceitos básicos

Este capítulo contém conceitos básicos a serem utilizados ao longo do trabalho. Na Seção 2.1 é feita uma introdução a genética; a Seção 2.2 diz a respeito de síntese proteica; a Seção 2.3 sobre transcriptoma; a Seção 2.4 discorre sobre diferenciação celular; a Seção 2.5 introduz o conceito de bioinformática; a Seção 2.6 a métrica N50 a ser utilizada no trabalho, e a Seção 2.7 sobre grafos.

2.1 Genética

A genética é o estudo dos genes de um organismo e a transferência de características físicas e biológicas de geração para geração, chamada de hereditariedade, promovendo variação celular. Entender como os genes de um organismo funcionam tem aplicações em diversas áreas, tais como biotecnologia, medicina e agricultura. Nesta seção serão abordados assuntos relacionados a genética de eucariotos que irão servir de introdução a conhecimentos descritos ao longo deste dissertação.

2.1.1 DNA - Ácido Desoxirribonucléico

O DNA (Figura 2.1) consiste em duas longas cadeias nucleotídicas, compostas por subunidades nucleotídicas. Cada cadeia denomina-se de fita de DNA.

A organização de seus nucleotídeos compõe o genoma de um organismo, carregando suas informações genéticas na forma de genes (sequências específicas de nucleotídeos que podem ou não codificar proteínas, transmitidos por hereditariedade).

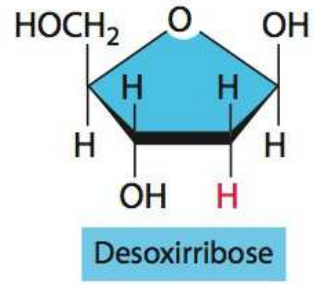

Figura 2.1: Esquema estrutural de uma desoxirribose – o açúcar de cinco carbonos presente no DNA. Imagem extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

Nucleotídeos do DNA

São constituídos de um açúcar de cinco carbonos, a desoxirribose, um grupo fosfato e uma base nitrogenada, sendo elas a timina (T), citosina (C), guanina (G) ou adenina (A) – os símbolos A, T, C e G são usados para representar os quatro nucleotídeos.

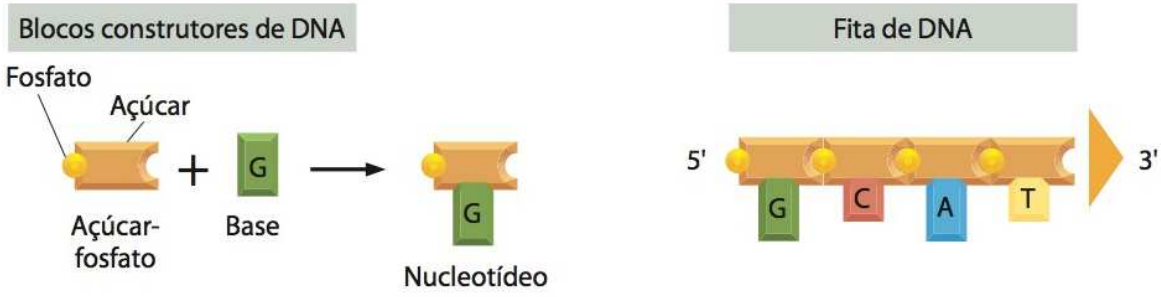

Figura 2.2: Esquema representando a estrutura dos quatro nucleotídeos constituintes do DNA. O grupo fosfato encontra-se ligado ao carbono 5 do açúcar. Imagem extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

O esqueleto da estrutura principal é composto por uma cadeia de açúcar e fosfato, ligados covalentemente entre si. O que irá diferenciar um monômero do outro é a presença das diferentes bases nitrogenadas (Figura 2.2).

A fita de DNA apresenta uma polaridade química, devido à forma a qual as bases estão dispostas; sendo que a extremidade 5' apresenta o grupo fosfato e a extremidade 3' uma hidroxila do carbono 3 do açúcar seguinte.

Para manter o esqueleto coeso, forma-se uma ligação fosfodiéster entre a extremidade 5' e a 3'. Na formação da dupla fita as formas e estruturas químicas das bases permitem que ligações de hidrogênio sejam formadas, de modo eficiente, somente entre o par A e T ou o par C e G. Uma diferença relevante entre tais pares é que se formam duas pontes de hidrogênio entre A e T e três pontes entre C e G, sendo considerados complementares. Observa-se que as bases estão localizadas mais internamente, e o esqueleto, externamente.

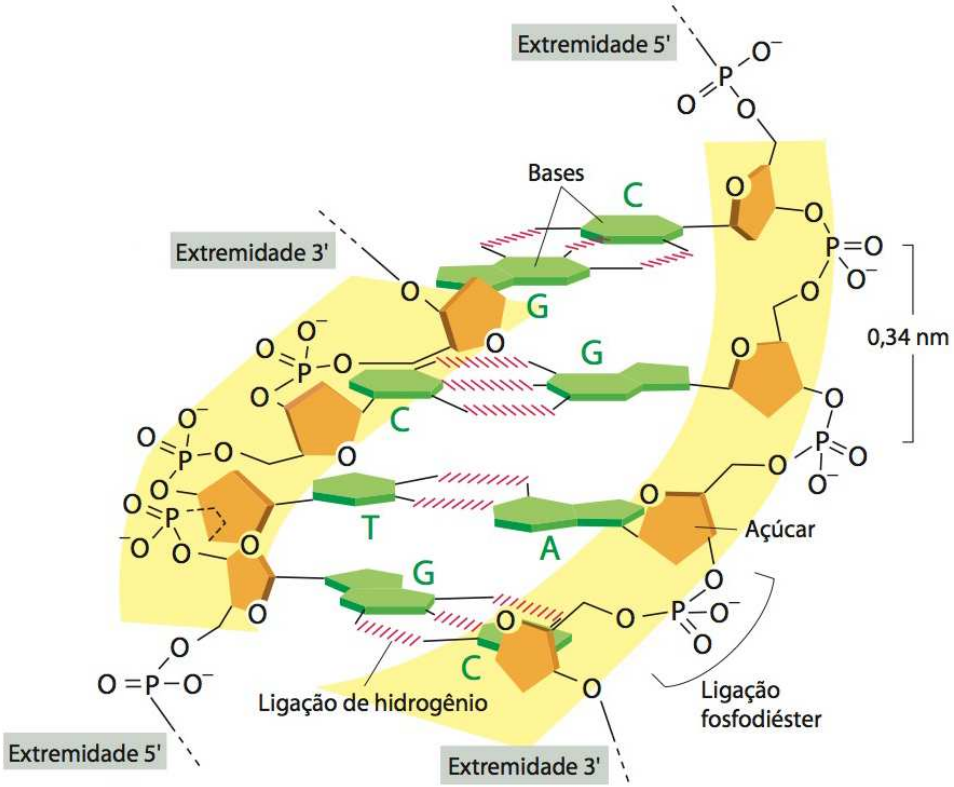

Figura 2.3: Estrutura da molécula de DNA ilustrando as ligações de hidrogênio entre as bases nitrogenadas e as fosfodiésteres, entre os nucleotídeos. Imagem extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

Mecânismo de replicação do DNA

O processo de replicação consiste na elaboração de uma cópia da molécula de DNA a partir uma um DNA-molde, na qual a sequência de nucleotídeos recém-sintetizada será complementar ao molde. Para a ocorrência da replicação, faz-se necessário a abertura da dupla hélice - pela DNA helicase – e polimerização de uma nova cadeia – pela DNA polimerase. Os nucleotídeos livres que servem de substrato para essa enzima são trifosfatos de desoxirribonucleotídeo. Cada uma das duas fitas de DNA serve de molde para as novas fitas polimerizadas (Figura 2.4).

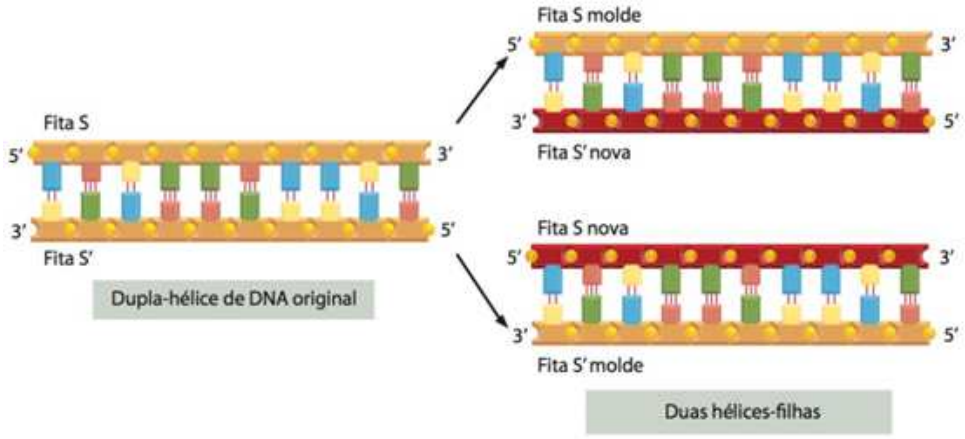

Figura 2.4: Representação da complementaridade das fitas de DNA, a partir de uma fita DNA-molde, e a demonstração da natureza semiconservativa da replicação. Imagem de extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

2.1.2 RNA - Ácido Ribonucléico

Estrutura da molécula de RNA

De modo semelhante ao DNA, o RNA é constituído por uma cadeia polipeptídica de nucleotídeos, os ribonucleotídeos, compostos por um açúcar de 5 carbonos, um grupo fosfato e uma base nitrogenada. No entanto, existem algumas diferenças fundamentais entre os ácidos: o açúcar presente é uma ribose (Figura 2.5); ao invés de conter uma timina, o RNA apresenta uma uracila (U); ele é constituído por uma fita simples.

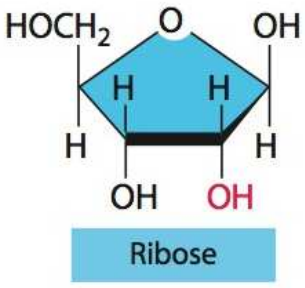

Figura 2.5: Esquema estrutural de uma ribose – o açúcar de cinco carbonos presente no RNA. Imagem extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

Transcrição do RNA

O RNA é confeccionado por meio da transcrição de DNA. Ela se inicia com a abertura e desespiralização de um pequeno segmento da dupla fita do DNA, sendo este, a partir de então, o molde para a síntese do RNA. Do mesmo modo como ocorre na replicação, a

transcrição é baseada na complementaridade de bases, sendo os ribonucleotídeos incorporados, covalentemente, à cadeia de RNA recém-sintetizada – o transcrito -, devida a uma ação enzimática - RNA polimerase.

A base U (do RNA) é complementar a A (do DNA), sendo ambas pareadas durante a transcrição. No entanto, quando a síntese acaba, as pontes de hidrogênio são desfeitas (Figura 2.6), conferindo ao RNA sua característica de fita simples.

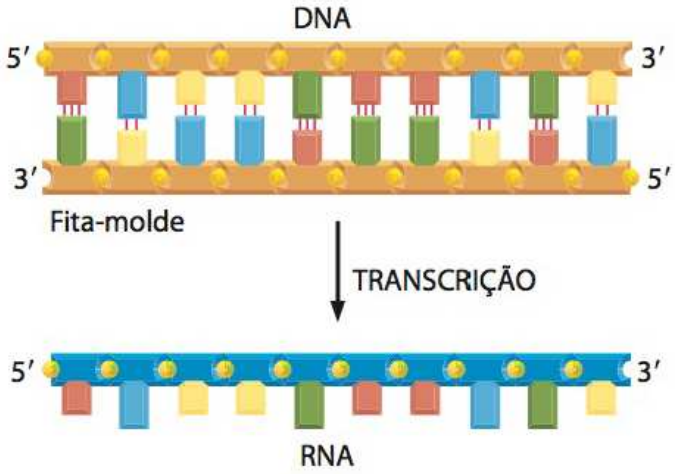

Figura 2.6: Representação esquemática da transcrição, mostrando que a fita simples de RNA é complementar à fita molde de DNA. Imagem extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

Existem vários tipos de RNA, tais como: RNA mensageiro (mRNA), RNA transportador (tRNA), RNA ribossômico (rRNA) e microRNAs.

Síntese do mRNA

As moléculas de mRNA são transcritas de uma sequência de genes presentes no DNA. Primeiro, há a síntese de um pré-RNA (transcrito primário de RNA), o qual contêm regiões de éxons e íntrons (regiões codificadoras e não codificadoras de proteínas, respectivamente). Pelo mecanismo de *splicing*, ocorre a excisão de íntrons. A sequência somente de éxons compõe o RNA maduro, portanto, ela apresenta uma série de códigos capazes de sintetizar uma proteína. Em seguida, ocorre a adição do Quepe, na extremidade 5', e da calda de Poli-A, na extremidade 3', dando ao RNA a devida identidade de mRNA (Figura 2.7).

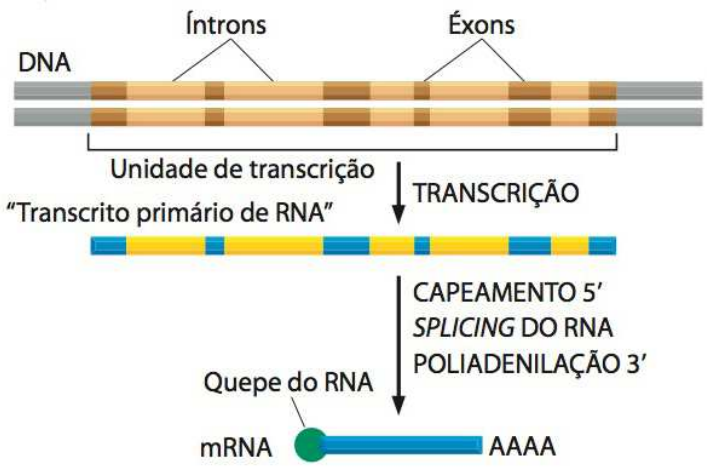

Figura 2.7: Representação do processo de síntese do mRNA, desde a transcrição até a adição do Quepe e da calda de Poli-A. Todo esse procedimento ocorre no núcleo celular. Imagem extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

2.2 Síntese Protéica

O mRNA, após sua síntese, é exportado para o meio citosólico. A leitura das informações genéticas, contidas no mRNA, necessárias para a síntese de uma proteína (processo denominado de tradução) se dá pelo ribossomo, dispersos no meio citosólico e na membrana do retículo endoplasmático rugoso.

A cada três nucleotídeos, temos um códon; e cada códon, a representação de seu aminoácido correspondente. O código genético padrão é constituído por 20 aminoácidos e é considerado degenerado, uma vez que alguns aminoácidos podem ser representados por mais de um códon (Figura 2.8).

GCA	AGA						GGA			UUA					AGC				
GCC	AGG						GGC		AUA	UUG				CCA	AGU				
GCG	CGA	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUA	AAA		UUC	CCC	UCA	ACA			GUA
GCU	CGC	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUC	AAG	AUG	UUU	CCG	UCC	ACC	UGG	UAC	GUC	UAA
	CGG									CUG				CCU	UCG	ACG		UAU	GUG	UAG
	CGU									CUU					UCU	ACU			GUU	UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Término
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Figura 2.8: Código genético padrão, representando os 20 aminoácidos e seus códons correspondentes. Imagem de extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

Portanto, o trânsito informacional genético é do DNA para o RNA, e deste para proteína. Tal fluxo é denominado Dogma Central da Biologia (Figura 2.9), sendo encontrado em todas as células, desde bactérias até seres humanos.

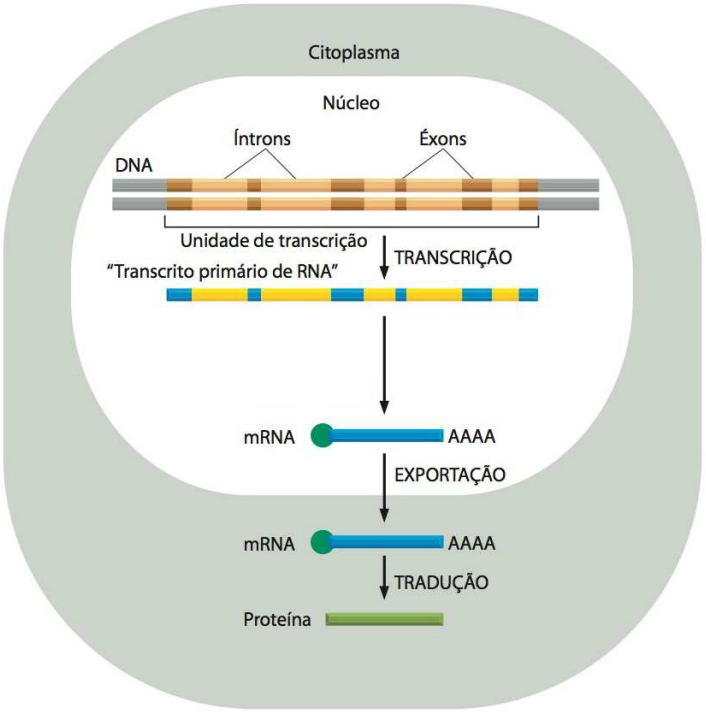

Figura 2.9: Esquema da ideia chave do Dogma Central da Biologia como observado em células eucarióticas. Imagem extraída de Biologia Molecular da Célula (Bruce Alberts) [1].

2.3 Transcriptoma

O transcriptoma é o conjunto de transcritos de uma célula, e sua quantidade amostrada depende da condição física do organismo e de vários fatores externos. A interpretação do transcriptoma é essencial para interpretar os elementos funcionais do genoma, amostrando os constituintes das células e tecidos e também, para a compreensão de doenças. As principais utilidades de um transcriptoma, segundo Wang et al. 2009, [43] são catalogar todos os tipos de transcritos, incluindo mRNAs, RNA não-codantes e pequenos RNAs; determinar a estrutura da transcrição de genes, padrões de splicing; e quantificar as mudanças em níveis de transcrição de expressão sob condições diferentes de desenvolvimento, exposição a fatores bióticos ou abióticos etc.

2.4 Diferenciação celular

A mudança do tipo celular que ocorre em uma célula é chamado de diferenciação celular. A diferenciação celular ocorre desde a fase de desenvolvimento, fase em que são gerados tecidos complexos, até a fase final adulta, na renovação e reparação celular. O tamanho da célula, o tipo celular, a atividade metabólica e a resposta a diferentes sinais são causados pela diferenciação celular. Tais mudanças são controladas pela modificação da expressão gênica, sem representação diferente da sequencia de DNA. Logo, as células podem

apresentar características diferentes, mesmo sendo representadas pelo mesmo genoma.

O comportamento das células em diferentes tecidos de um genoma pode ser analisada fazendo uma análise diferencial dos genes, que causam essa diferenciação celular, podendo usar uma técnica recentemente criada (a menos de uma década), chamada RNA-seq, na qual se utiliza o transcriptoma (RNA maduro) para se obter informações dos genes.

2.5 Bioinformática

O desenvolvimento de novas ferramentas computacionais e novas metodologias de bioinformática permitem uma análise rápida e precisa de sequências de ácidos nucleicos (DNA e RNA) e proteínas, possibilitando a detecção de mutações genéticas e seu efeito na estrutura e função dessas protínas. Por isso, a bioinformática é considerada uma ciência interdisciplinar que envolve conhecimentos nas áreas de biologia molecular, química molecular, física, matemática, ciência da computação, entre outros [34]. Dentre as diversas aplicações da bioinformática nas diversas áreas da biologia e medicina, destacam-se sua utilização (1) na biotecnologia, no estudo da constituição genética de diversos organismos com potencial de serem aplicados como biofábricas em processos industriais, e (2) na área de medicina personalizada nos quais genomas humanos (indivíduos específicos) são sequenciados e suas mutações relacionadas a fenótipos como doenças ou características de interesse. Deste modo, ferramentas computacionais de análises de sequenciamento de DNA, expressão gênica, proteômica e metabolômica se tornaram essenciais para compreender quais alterações no genoma ou no metabolismo dos diversos organismos de interesse podem estar relacionadas às suas características.

2.6 Métricas N50

A métrica N50 é o tamanho do transcrito cuja soma do tamanho de transcritos menores que ele é menor ou igual a metade da soma do tamanho de todos os transcritos. Essa métrica é utilizada, por exemplo, para calcular o tamanho médio do conjunto de sequências, além de, em genômica, ele ser utilizado para medir o valor médio de contigs em um genoma. Por exemplo, imaginemos que temos um transcriptoma com transcritos de tamanhos 90kbp, 70kbp, 50kbp, 40 kbp, 30 kbp, 20 kbp e 10 kbp. O somatório de todos o tamanho é 310 kbp, e o N50 é 70kbp, pois 90 kbp + 70 kbp já é maior que a metade da soma de todos os transcritos (ver Figura 2.10).

Para encontrar o N50 de uma determinada montagem, primeiro se ordena de forma decrescente a lista de transcritos por tamanho, e posteriomente identifica-se o tamanho do transcrito que, a partir dele, a soma de todos os tamanhos é pelo menos 50% do total de transcritos.

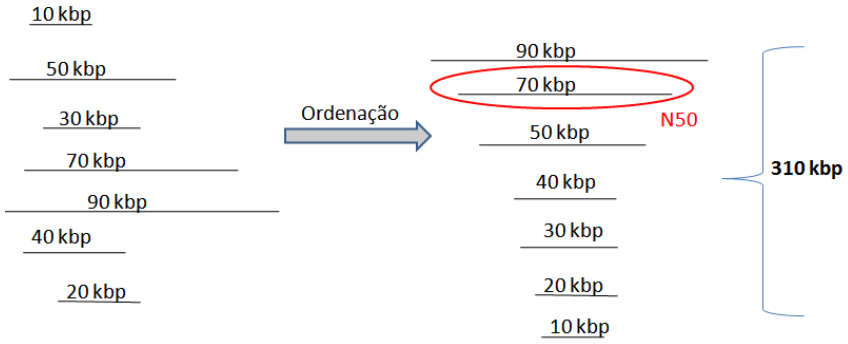

Figura 2.10: Exemplo de cálculo da métrica N50.

A determinação do N90, ou outra métrica, de um conjunto possuí a mesma ideia da métrica N50, sendo alterado somente a porcentagem da soma dos tamanhos dos transcritos sobre o conjunto total. Por exemplo, o valor do N90 é o tamanho do transcrito cuja soma do tamanho de transcritos menores que ele é menor ou igual à 90% da soma do tamanho de todos os transcritos.

O conceito de N50 será utilizado neste trabalho na apresentação dos resultados das montagens *de novo* obtidas descorridas no capítulo 5.

2.7 Grafos

Grafo é uma estrutura matemática utilizada para modelar as relações entre um conjunto de objetos. Em um grafo existem os vértices ou nós, e as arestas, que os conectam. Se um grafo é não-direcionado, quer dizer que não existe distinção entre a aresta (i, j) ou (j, i), já quando ele é direcionado, cada aresta possui um nó de saída e outro de chegada. Denotamos um grafo por G(V,E), onde ele possui um conjunto V de vértices e um conjunto E de arestas, como por exemplo na Figura 2.11.

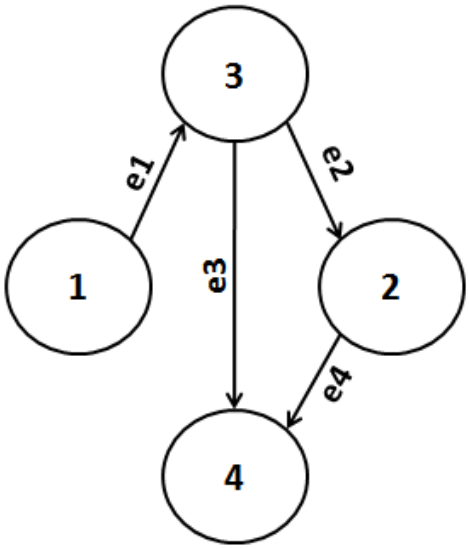

Figura 2.11: Exemplo de um grafo direcionado. Esse grafo G(V,E), possui um conjunto de nós V={1,2,3,4} e uma coleção de arestas E={e1,e2,e3,e4}.

Duas aplicações bem comuns de grafos são redes sociais, onde cada nó representaria uma pessoa e uma aresta não-direcionada representando a amizade da pessoa i com a pessoa j, e proteômica - estudo das proteínas de um organismo, no qual cada nó do grafo representaria uma proteína do organismo e as arestas representariam as interações entre elas.

Em um grafo, um caminho é uma sequência de vértices, no qual a partir de cada vértice existe uma aresta para o próximo vértice. Por exemplo, na Figura 2.11 um possível caminho no grafo seria a sequência de vértices (1,3,4).

Um caminho em um grafo é dito hamiltoniano se ele passa por todos os vértices do grafo sem repeti-los, ou seja, passa por todos os vértices uma única vez. Por exemplo, na Figura 2.11 o seu caminho hamiltoniano é dado pela sequência de vértices (1,3,2,4).

Um caminho em um grafo é dito euleriano se ele passa por todas as arestas do grafo sem repeti-las, ou seja, passa pro todas as arestas uma única vez.

Capítulo 3

Análise de RNA-Seq e sua metodologia

Os experimentos de RNA-Seq ao longo do tempo vêm se tornando uma abordagem muito usada por pesquisadores em experimentos de expressão gênica comparativa entre grupos. Uma melhor visualização da preparação dos dados de RNA-Seq pode ser observada na Figura 3.1. Para gerar as bibliotecas para análises posteriores, primeiramente as sequências longas de RNA são isoladas e purificadas. Posteriormente elas são convertidas em bibliotecas de cDNA (DNA - complementar). Adaptadores de sequenciamento são adicionados em cada fragmento de cDNA para obter sequências curtas somente com uma extremidade (chamada single end) ou ambas as extremidades (chamada paired-end). As sequências geradas (*reads*) tem, normalmente, tamanhos entre 30-400 bp, dependendo da tecnologia utilizada. Posteriormente as sequências curtas geradas são alinhadas com um genoma ou transcriptoma e então classificadas em três tipos: *reads* exônicos (*exonic reads*), *reads* de junção (*junction reads*) e caudas de *poly(A)* (*poly(A) end reads*). A partir desses três tipos é gerado um perfil de expressão por base nitrogenada (*Base-resolution expression profile*) de cada gene.

Para quantificar a expressão de um gene ou um transcrito, os *reads* de RNA-Seq precisam, primeiramente, passar por uma análise de qualidade e remoção de artefatos de sequenciamento. Posteriormente, os *reads* são mapeados num genoma de referência (no caso de um genoma disponível) ou mapeados num transcriptoma de referência que pode ser obtido a partir da montagem *de novo* dos próprios *reads*. Logo após o mapeamento, é feita a quantificação relativa de cada gene/transcrito e então é aplicado um teste estatístico sobre os dados com intuito de eliminar ruídos entre as réplicas experimentais. O fluxograma de uma análise básica de RNA-Seq para identificação de genes/transcritos diferencialmente expressos é mostrado na Figura 3.2. Cada passo do fluxograma será descrito ao longo deste capítulo.

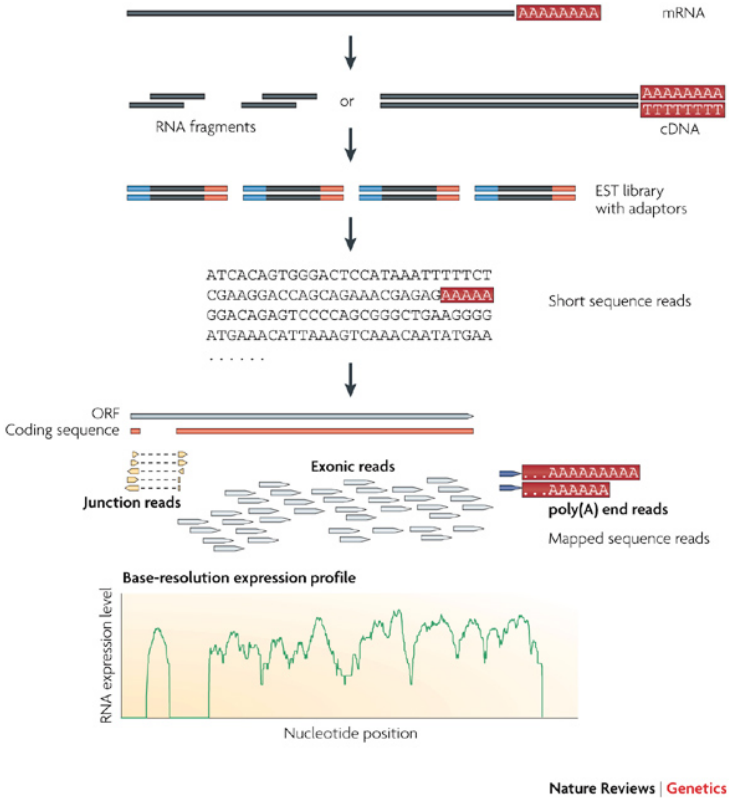

Figura 3.1: Preparação da biblioteca de RNA-Seq. Imagem extraída de Wang et al [43].

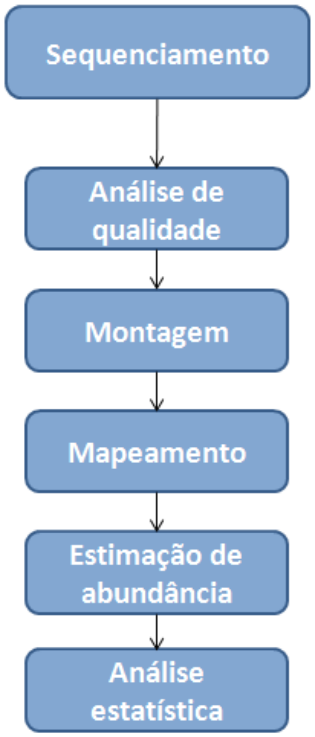

Figura 3.2: Fluxograma de uma análise de RNA-Seq para identificação de genes/transcritos diferencialmente expressos.

3.1 Análise de Qualidade

Avaliar a qualidade dos *reads* sequenciados é uma prática importante para garantir um um bom resultado em uma análise de RNA-Seq. Os *reads* gerados pelos sequenciadores possuem artefatos anexados às cadeias de DNA, por exemplo, adaptadores (sequências idênticas de aproximadamente 10 bp adicionado a cadeia de DNA), que, após o sequênciamento, precisam ser retirados. Existem algumas ferramentas que executam este procedimento, uma delas é o SeqyClean, e para analisar a qualidade da sequência utilizamos o programa o FastQC [4]. O programa que primeiro foi utilizado para analisar qualidade de um sequenciamento, mais especificamente de DNA, foi o Phred [7].
Para analisar se as *reads* tem alta qualidade, algumas métricas podem ser utilizadas como a probabilidade de ocorrer um erro na leitura de uma base no sequenciamento. A probabilidade de ocorrer um erro de sequenciamento (P), sendo que cada base possui uma qualidade Q, gerada pelo sequenciador, é dada por $P = 10^{\frac{-Q}{10}}$. Por exemplo, supondo que uma base tem qualidade Q=20, então a probabilidade desta base estar errada é de $P = 10^{\frac{-20}{10}} = 0.01$, ou seja, pela probabilidade ser muito baixa, a chance de ter ocorrido um erro de sequenciamento naquela base é baixa também. A qualidade média do fragmento (P_m) é dada por:

$$P_m = \frac{10^{\frac{-Q_1}{10}} + 10^{\frac{-Q_2}{10}} + ... + 10^{\frac{-Q_n}{10}}}{n} = \frac{\sum_{i=1}^{n} 10^{\frac{-Q_i}{10}}}{n}$$

onde Q_i é a base i e n o tamanho do fragmento.

3.2 Montagem

A principal e mais difícil etapa da análise de dados em um experimento de RNA-Seq é a montagem de seu transcriptoma de referência a partir das suas *reads*. Uma das principais dificuldades na etapa de montagem no processo de análise de RNA-Seq é o fato de acúmulo de erros de sequenciamento. Para gerar os transcritos a partir dos seus dados, podemos realizar uma montagem sem genoma de referência, chamada *de novo*, ou gerá-los através do mapeamento das *reads* no genoma de referência.

3.2.1 Montadores

Todos os algoritmos de montagem utilizados neste trabalho são reduzidos a problemas em grafos. A redução de montagens em grafos é complexa, e ainda as montagens reais dependem de heurísticas e algoritmos de aproximação para obter resultados aproximados, como, por exemplo, eliminar redundâncias, corrigir erros, descartar incertezas e reduzir sua complexidade [22].

Cufflinks [42]

Cufflinks resolve o problema de montagem encontrando um emparelhamento máximo (maximum matching) (Figura 3.3-b) em um grafo bipartido derivado de uma sobreposição de alinhamentos (Figura 3.3-a). Ele encontra os transcritos produzidos pelos eventos de splicing nas sequências utilizadas, no qual serão usados para estimar abundância. Para

montar todos os transcritos de cada locus, o algoritmo levaria um tempo razoável, assim, o uso do genoma favorece a montagem.

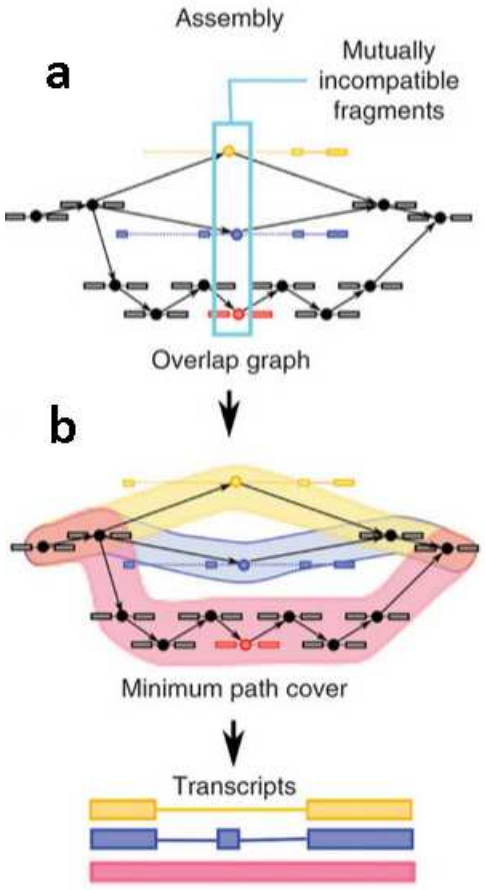

Figura 3.3: Algoritmo de amostragem de transcritos do Cufflinks. (a) montagem efeituada a partir dos *reads* de entrada. O algoritmo identifica possíveis *splicings* na montagem. (b) Cufflinks trata cada *splicing* de forma independente e efetua o algoritmo de encontrar o caminho mínimo em um grafo. (c) reconstrução final dos transcritos, obtidos por cada caminho mínimo no grafo. Imagem editada de Trapnell *et al.*. [42].

Cufflinks é composto por sub-programas que intensificam o pipeline de análise, possibilitando alcançar todas as etapas de uma análise de RNA-Seq. Após a análise de mapeamento feita pelo TopHat [41], o Cufflinks lê os mapeamentos e monta os transcritos; posteriormente o Cuffmerge, a partir de todas as montagens, fornece uma tabela final contendo todas as posições de cada transcrito montado. No próximo passo, que consiste na análise diferencial, a análise é feita pelo Cuffdiff, que fornece uma tabela contendo valores de expressão entre as condições fornecidas. Por fim, CummeRbund é utilizado para visualização de resultados de forma eficiente e dinâmica.

Trinity [8]

Trinity é um montador que possui a opção de gerar dados com um genoma de referência ou não, mas ele é amplamente utilizado para montagem de novo de transcriptomas, quando

não há um genoma de referência. Ele é composto por três módulos: Inchworm, Chrysalis e Butterfly (Figura 3.4). Primeiramente, o Inchworm monta os *reads* de forma gulosa e encontra caminhos mais abundantes em um grafo de De Bruijn. Os transcritos de variantes de splicing são encontrados, assim como transcritos em outras formas de baixa abundância. Posteriormente Chrysalis aglomera a saída dos dados do Ichworm em componentes que representam várias variantes de splicing e parálogos estritamente relacionados e constrói o grafo de Bruijn para cada componente. Finalmente no módulo do Butterfly, são gerados os transcritos e isoformas de splicing dos genes parálogos. Como citado por Lu *et al.* [22], existem módulos do Trinity que podem ser substituidos para uma melhor eficiência, como, por exemplo, substituir parte do Inchworm pelo Jellyfish, que aumenta o processamento de k-mers em paralelo. Existem artigos que fornecem pipelines de uso eficaz do montador Trinity, um deles é o artigo de Hass *et al.* [9].

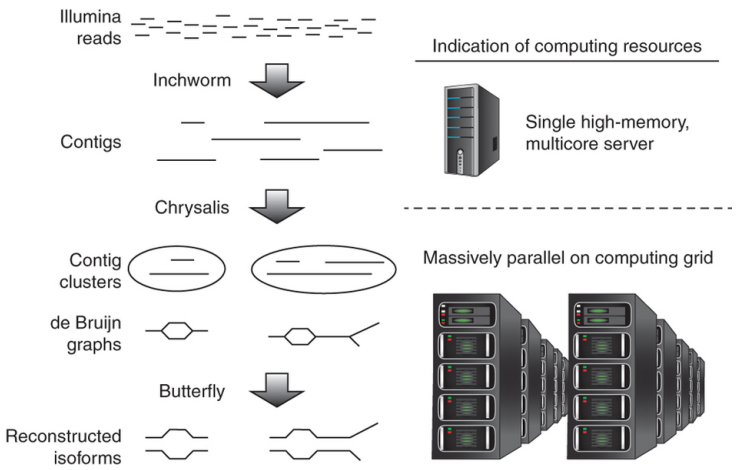

Figura 3.4: Principais etapas sequenciais do Trinity (à esquerda) e os recursos computacionais associados (à direita). O Trinity toma como entrada *reads* (superior esquerdo) e primeiro usa o módulo Inchworm para construir os contigs. Esta fase requer um único servidor de alta de memória (1 GB de RAM por 1 milhão *paired reads*, mas varia de acordo com complexidade do *read*(canto superior direito). O Chrysalis (Meio na esquerda) agrupa os clusters de contigs do Inchworm, muitas vezes gerando dezenas à centenas de milhares de clusters, cada um dos quais se processa a uma componente do grafo de De Bruijn, de forma independente e em paralelo, em um grid computing (canto inferior direito). O Butterfly (canto inferior esquerdo), em seguida, extrai todas as possíveis sequências de cada componente do grafo, que também pode ser paralelizado (Figura extraída de Hass *et al.* [9]).

Oases [38]

Oases combina estratégias de multiplos k-mers, mas com uma análise topológica semelhante ao Trinity, tentando lidar com níveis de expressão e variantes de splicing. Ele é processado em cima dos contigs gerados pelo Velvet e pós-processados usando um único k-mer, pré-definido. Ele constrói um grafo de Bruijn e posteriormente é feita uma análise topológica que extrai as isoformas montadas. Existe a possibilidade de executar o Oases com múltiplos k-mers, ocorrendo a mesclagem de várias montagens. Uma grande

vantagem do Oases é sua remoção dinâmica de erros, que contribui para sua robustez.

IDBA-Tran [29]

Semelhante ao Oases, IDBA-Tran também adota a ideia de múltiplos k-mers para lidar com transcrições em diferentes níveis de expressão. No entanto, em vez de gerar um grafo de De Bruijn e encontrar transcritos para cada valor de k, um grafo acumulado de De Bruijn é construído para juntar toda a informação de ambos os transcritos, com alta e baixa expressão. Durante cada iteração, um grafo acumulado de De Bruijn H_k para um k fixo é construído a partir da entrada. Os contigs construídos das iterações anteriores, isto é, os contigs construídos em H_{k-s}, onde s é o tamanho do passo de variação do k-mer, são utilizados como entrada na construção de H_k. A informação de profundidade é usado para separar as componentes do grafo de De Bruijn. Em uma forma ideal, os transcritos de genes diferentes são decompostos em componentes diferentes. Em cada componente, o *splicing* alternativo pode ser detectado e os transcritos recontruídos. Para acumular informações, todas as transcrições reconstruídas são utilizadas como entrada na próxima iteração.
O problema de descoberta de transcritos (TD-*Transcript Discovery*) é um problema NP-Hard: dado um grafo de De Bruijn G(V,E) com um conjunto de vértices V e arestas E, um conjunto de *reads paired-end* $P = (v_i, v_j)$, v_i, $v_j \in V$, com distância d e erro s, queremos encontrar t caminhos em G com um número máximo de *reads paired-ends* $P' \subset P$. O caminho p é dito suportado pelos *reads* paired-ends v_i e v_j, se a distância entre v_i e v_j em p está entre $d - s$ e $d + s$.

3.2.2 Montagem *de novo*

Montadores de novo como o Trinity, Oases, IDBA-Tran montam suas respectivas sequências de referência sem usar um genoma de referência. Essa estratégia é dificultada, por exemplo, pela qualidade da *read*, contaminações nas amostras, erros e qualidade do sequenciamento, complexidade do organismo e famílias multigênicas. A principal dificuldade de pesquisadores hoje em dia é avaliar quão boa ficou uma montagem de novo levando em conta fatores realmente relevantes. Para analisar uma montagem, são comparados valores de N50, N90, tamanho médio de um transcrito, mediana e desvio padrão dos dados, fatores que podem ser manipulados em montagens para obter valores ótimos, mas fora da realidade.
Esses montadores usam como estratégia principal duas metodologias: grafo de De Bruijn ou grafos consensos de sobreposições. Montadores como o Trinity, Oases e IDBA-Tran utilizam a estratégia de grafos de De Bruijn para suas montagens, já outros montadores, como o Cufflinks, utilizam grafos consensos de sobreposições. Pelo uso dessas diferenças de estratégias, existem prós e contras do uso de cada montador, listados na Tabela 3.1.

Tabela 3.1: Comparação entre algumas características relevantes de montadores de novo que utilizam estratégias de grafos de De Bruijn e grafos consensos de sobreposição.

Características	Grafos De Bruijn	Grafos de sobreposição
Sensibilidade a erros	Alta	Baixa
Possuem variação de k-mer	Sim	Sim
Lidam bem com repetição	Não	Não
Caminhos	Euleriano	Hamiltoniano
Sensibilidade ao k-mer	Alta	Baixa
Uso de memória	Elevado	Mediano
Tempo de máquina	Elevado	Baixo

Grafo consenso de sobreposição

Construímos um grafo de sobreposição a partir de um certo número de fragmentos. Cada fragmento é colocado no nó do grafo e então criamos arestas com direções entre eles, sendo o valor de cada aresta o número de bases sobrepostas entre cada fragmento (O usuário pode deteminar um número mínimo de sobreposição entre os fragmentos). Posteriormente encontramos o caminho hamiltoniano maximal, aquele caminho hamiltoniano de maior pesos nas arestas, sobre o grafo construído. A última etapa consiste em fazer a sobreosição entre os fragmentos compostos no caminho hamiltoniano maximal (ver Figura 3.5). Existem programas de montagens de genomas e transcriptomas que utilizam essa abordagem em seus algoritmos, é o caso do Phrap e o Cufflinks.

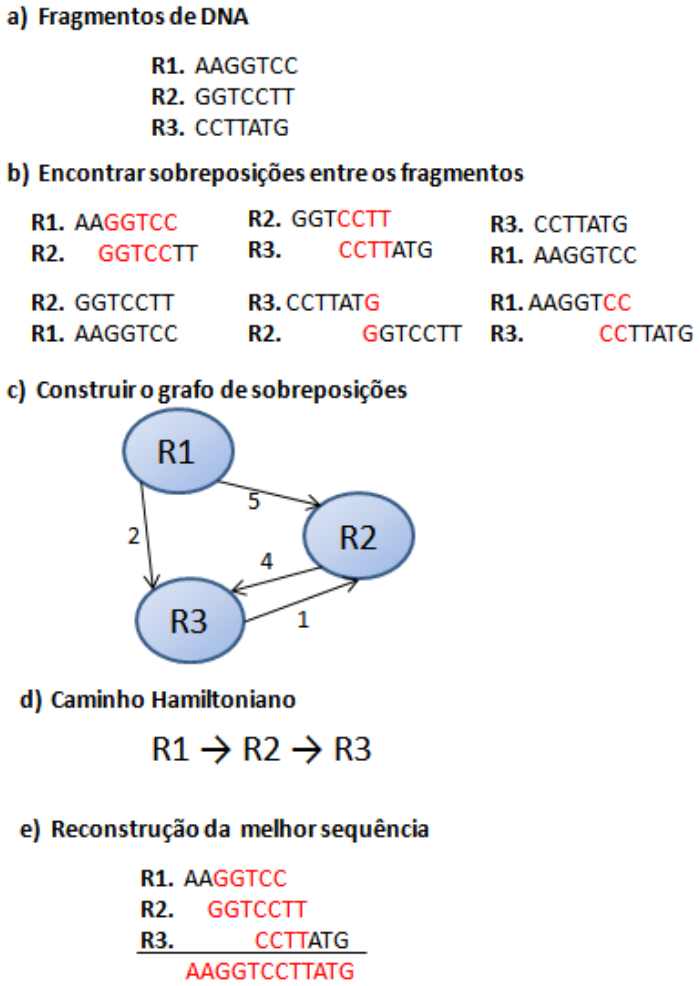

Figura 3.5: (a) Fragmentos de DNA a serem utilizados. (b) Encontrar todas as sobreposições (em vermelho) entre todos os fragmentos. (c) Com os fragmentos como nós, cada aresta drecionada do grafo tem como valor o número de bases sobrepostas entre cada fragmento encontrado na etapa (b). Nessa fase o usuário pode definir o valor mínimo de sobreposição entre dois fragmentos. (No nosso exemplo não constão as arestas de sobreposição com valor 0).(d) A partir do grafo encontramos o caminho hamiltoniano maximal. Em caso de genoma circular, encontraríamos o ciclo hamiltoniano no grafo. (e) Representação do overlap dos fragmentos compostos, em ordem, no caminho hamiltoniano, e obtenção da sequência consenso.

Grafo de *De Bruijn*

Utiliza-se grafo de *De Bruijn* para montar sequências de DNA a partir de subsequências menores. Dada uma coleção de sequências, criamos sequências menores ou iguais de tamanho k, essas chamadas de $k-mer$. O grafo de De Bruijn de ordem k é um grafo orientado cujo os vértices são todos os $k-mers$ e criamos uma aresta entre dois vértices x e y se existe uma sobreposição de $k-1$ caracteres entre o sufixo de x e o prefixo de y. Seja sua coleção de sequências (*reads*) definidas por ATGG, CTCG, GGCT. Vamos dividi-las em subsequências de tamanho 3 (k=3), ou seja, escrevemos todas as possibilidades de sequências para cada *read*.

read 1: ATGG	3-mers: ATG, TGG
read 2: CTCG	3-mers: CTC, TCG
read 3: GGCT	3-mers: GGC, GCT

Cada nó terá $(k-1)-mers$ e cada aresta k-mers. Todos os $(k-1)-mers$ gerados a partir de todos os $k-mers$, sem repetição, são AT, TG, GG, TC, GC, CT, CG. O grafo gerado é ilustrado na Figura 3.6.

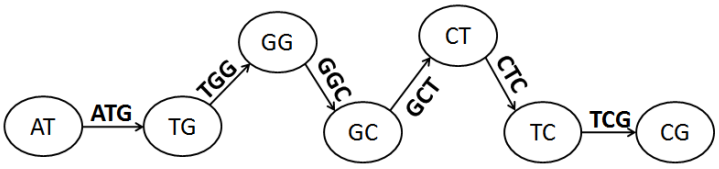

Figura 3.6: Exemplificação de um grafo de De Bruijn.

Reconstrução de um grafo de De Bruijn

Para obtermos a sequência que gerou os *reads* utilizados para gerar um grafo de De Bruijn, basta construir um caminho euleriano no grafo gerado, ou seja, um caminho que passa por todas as arestas uma única vez. No exemplo utilizado na seção anterior, o caminho euleriano é dado pela sequência do vértices:

$$AT \rightarrow TG \rightarrow GG \rightarrow GC \rightarrow CT \rightarrow TC \rightarrow CG.$$

A sequência final é gerada pela sobreposição de $(k-2)-mers$ dos vértices, gerado pelo caminho euleriano, logo ela é ATGGCTCG.

Alguns fatores podem ocasionar problemas na reconstrução de um grafo de *De Bruijn*, entre eles estão os erros de sequenciamento, que podem gerar erros de montagem, como sequências erradas (Figura 3.7-a) ou "bolhas", ocorrendo quando há mais de um caminho euleriano a ser percorrido no grafo (Figura 3.7-b), e, repetição de sequências, que formarão topologias diferenciadas no grafo, aumentando o uso de memória e o tempo de processamento dos dados (Figura 3.7-c).

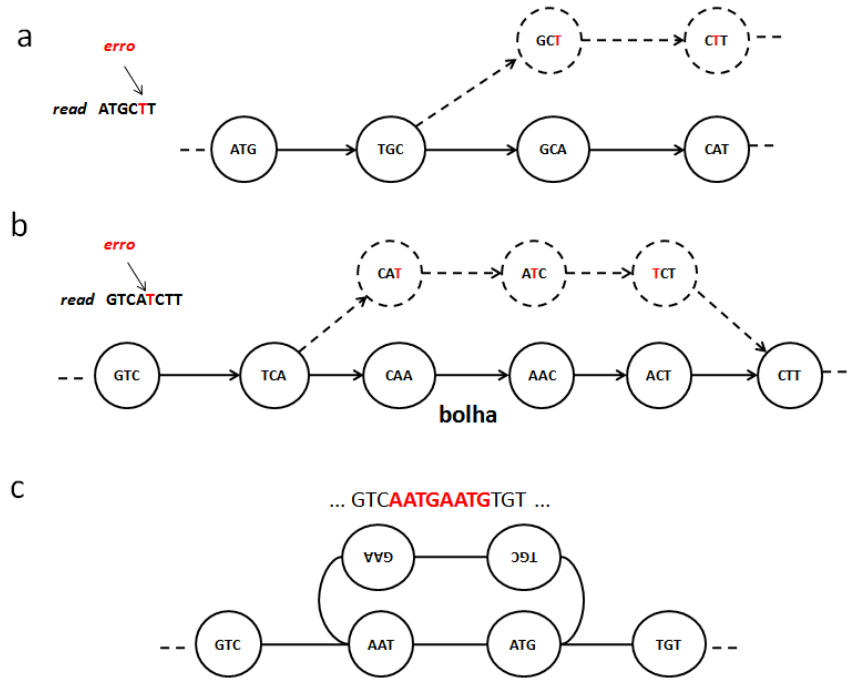

Figura 3.7: (a) Erro de sequenciamento causando uma reconstrução errada. (b) Erro de sequenciamento causando "bolhas"no grafo durante a reconstrução. (c) Repetições nas sequências geram topologias singulares e aumento no tempo de processamento dos dados.

Pensando em montagem de transcritomas, programas como Trinity, Oases, IDBA-Tran, utilizam a estratégia de grafo de *De Bruijn* para montar transcritos a partir dos *reads*.

3.2.3 Montagem com genoma de referência

Montagem com genoma de referência é uma estratégia com alto índice de qualidade, já que o genoma contribui com informações detalhadas sobre início e fim dos genes, como também informações sobre éxons e íntrons. Atualmente a estratégia mais utilizada para uma montagem com genoma de referência é usando a combinação de programas TopHat e o pacote Cufflinks. Estes programas fazem o alinhamento dos *reads* e montagem dos transcritos, calculando sua abundância e análises estatísticas de expressão diferencial.

3.2.4 Comparação de estratégias de montagens

Segundo Lu *et al.*, 2013 [22], geralmente montagens de novo requerem um tempo de processamento e uso de memória maior do que com genomas de referência. O programa Oases, por exemplo, consome muita memória, mas é executado de forma rápida. Ele ainda cita que montadores de novo encontram mais transcritos do que deveriam, em relação à anotação curada, sendo que muito deles são pequenos fragmentos. Além do mais, não existe uma maneira fácil de escolher o melhor método de montagem, sendo a melhor escolha feita dependendo do contexto concreto da análise. Lu *et al.* [22] ainda sugerem ainda que para a criação de um melhor transcriptoma de referência, deve-se combinar as montagens de novo com as de genoma de referência.

Para obter melhores montagens para um transcriptoma de referência, os parâmetros padrões devem ser desconsiderados. Segundo Li *et al.* [17] montagens com parâmetros padrões têm resultados piores do que montagens com parâmetros ajustados. A fim de gerar um transcriptoma mais completo, podem-se gerar montagens com vários $k-mers$ diferentes e então, posteriormente, fazer uma combinação dessas montagens.
Pensando neste aspecto, He *et al.* [10] sugerem novas nomenclaturas a respeito de tipos de montagens que foram usadas neste projeto. Existem três tipos de montagens pensando em gerar um transcriptoma de referência e possíveis parâmetros: SASP (Single-Assembler Single-Parameter), SAMP (Single-Assembler Multiple-Parameters) e CDTA (Combined De novo Transcriptome Assembly). A estratégia SASP consiste em montar um transcriptoma *de novo* usando somente um determinado montador, a estratégia SAMP consiste em realizar diversas montagens, variando os possíveis parâmetros de um montador, e posteriormente combiná-las utilizando, por exemplo, o CAP3 [11], e a estratégia CDTA consiste em combinar várias montagens de vários montadores e posteriormente juntá-las em uma única montagem, utilizando,por exemplo, o programa CAP3.

3.3 Mapeamento

Em uma análise, com o objetivo de encontrar transcritos diferencialmente expressos, o primeiro passo antes de estimar sua abundância é saber a posição de cada *read* no genoma ou no transcriptoma. Para tal, o processo de procurar a melhor posição de cada *read* em um genoma ou transcriptoma de referência é chamado de mapeamento.
O processo de mapeamento geralmente é lento, devido ao tamanho do genoma em que devemos procurar a melhor posição para todas os *reads* do experimento. Para medir a eficiência de um mapeador, programa que faz o mapeamento, leva-se em consideração o seu algoritmo, tempo de processamento e uso de memória.
Segundo Schbath *et al.*, 2012 [36]), os algoritmos mais utilizados por mapeadores são: *Hashing* e Transformação Burrows-Wheeler (BWT) [37]. A estratégia de Hashing se baseia em indexar os *reads* utilizados ou o genoma de referência. A estratégia BWT é dividida em dois grandes algoritmos, a árvore de sufixos e a matriz de sufixos.
A maneira mais simples de encontrar todas as ocorrências de uma leitura, se nenhum gap, consiste em "deslizar"o *read* ao longo da sequência do genoma e observar as posições em que existe um match. Infelizmente, apesar do conceito simples, este algoritmo tem complexidade $O(L_G L_r N_r)$ onde L_G é o tamanho da sequência do genoma, L_r o tamanho do *read* e N_r é o número de *reads*.
Entretanto, para ser eficaz, todos os métodos devem conter com uma etapa de pré-processamento. Por exemplo, é teoricamente praticável compilar uma lista de todas as palavras de comprimento 36 (36-mers) que são encontrados no genoma e determinar de uma vez por todas as suas posições. Então, podemos usar um algoritmo de *hashing* para transformar uma sequência de caracteres em uma chave que permite uma pesquisa rápida. A estratégia de *hashing* é baseada em indexar os *reads* no genoma de referência.

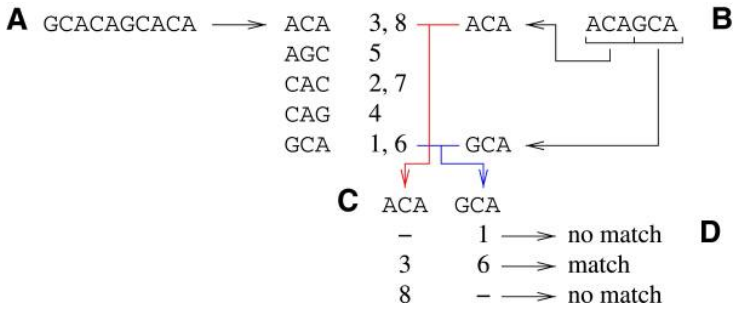

Figura 3.8: O algoritmo de *hashing*. (A) O genoma é dividido em 3-mers, e as suas respectivas posições no genoma são armazenadas. (B) O read é dividido em 3-mers. Os 3-mers a partir dos *reads* são comparados com os 3-mers a partir do genoma, utilizando um procedimento de *hashing*. (C) Posições para cada *read* são ordenadas e comparadas com a das outras sementes. (D) posições compatíveis são mantidas.(Figura extraída de Schbath *et al.*, 2012 [36]).

Uma desvantagem do método de *hashing* é que as sementes podem ser altamente repetidas no genoma. Como consequência, muitos hits devem ser verificadas na fase de "estender", que é demorado. Outra abordagem é dividir cada *read* em quatro substrings de 9 nucleótidos de comprimento. Então, como anteriormente, a cada substring de um *read* pode ser combinado utilizando a lista de 9-mers (Figura 3.8-B). Se as quatro subsequências de um *read* são encontradas na lista, na ordem correta e adjacentes uns aos outros, existe o *read* no genoma (Figura 3.8-C, D). No entanto, este algoritmo não permite considerar a desemparelhamentos.
Uma árvore de sufixos é uma árvore em que existe uma correspondência de um-para-um entre os caminhos a partir da raiz para as folhas e os sufixos existentes em uma cadeia de caracteres, em outras palavras, para todos os sufixos desta cadeia existe um caminho da raiz a uma folha em árvore (Figura 3.9). Observe que alguns espaço são salvos uma vez que todos os sufixos não são explicitamente escritos. Na verdade, os algoritmos atuais constroem árvores de sufixo cujo tamanho é proporcional ao do genoma, e o tempo também é proporcional ao tamanho do genoma.

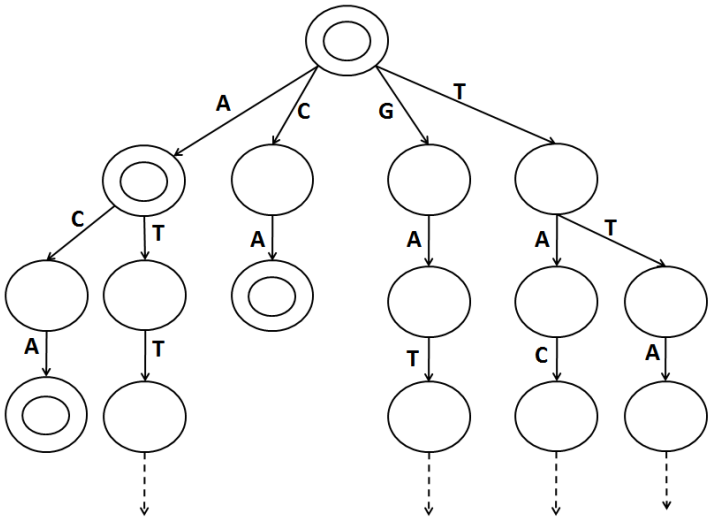

Figura 3.9: Árvore de sufixo do genoma GATTACA. Os círculos duplos representam que o sufixo, existente no genoma, termina o ramo da árvore. (Figura extraída de Schbath *et al.*,2012 [36]).

A estratégia de matriz de sufixo foi desenvolvida a fim de melhorar o processamento dos sufixos quando o genoma é grande, dificuldade encontrada na estratégia de árvores de sufixos. A sua estratégia é baseada na geração de uma matriz de sufixos, ordenados lexicograficamente. Primeiramente são gerados todos os sufixos do genoma e suas posições no genoma, e posteriormente a matriz é gerada a partir das posições dos sufixos lexicograficamente ordenados.
Segundo Langmead *et al.* [14], mapeadores que usam a estratégia de BWT são mais eficientes do que as que usam Hashing. Os mapeadores de *reads* curtas mais utilizados de estratégias BWT são TopHat [41], Bowtie [15], Bowtie2 [14] e BWA [18]. A diferença entre os mapeadores Bowtie, Bowtie2 e BWA estão no modo de operação da *read* no genoma, o tamanho maximo aceito do *read* e principalmente seu tempo de processamento. O tempo de processamento do mapeador Bowtie2 supera em três ordens de grandeza o BWA e ainda é uma otimização do Bowtie, tendo como principal diferença a permissão de indels. Há estudos [14, 16, 21, 35] que mostram que a acurácia entre os três mapeadores são bem parecidas, e muito altas, favorecendo o número de *reads* verdadeiramente mapeadas. Os experimentos que possuem *reads single-ends* são melhores mapeados utilizando Bowtie2, já experimentos que utilizam *reads* paired-ends são melhores mapeados tanto por Bowtie2 ou BWA.
O mapeamento genômico, como o transcriptômico, possuem certas vantagens e desvantagens. No genômico, uma vantagem seria, além do genoma já anotado, a possível descoberta de novos genes e isoformas; já as desvantagens se voltam para a dificuldade de lidar com *splicings* e interpretar íntrons, e a necessidade do genoma para decidir as coordenadas dos transcritos. No transcriptômico *de novo*, que não utiliza genoma de referência, pensando nas vantagens, estão a facilidade de predizer a abundância dos transcritos, e a não necessidade de lidar com íntrons; já as desvantagens seguem para a descoberta de genes, que possivelmente pode conter erros.

3.4 Determinação da abundância

A partir do mapeamento, temos a possibilidade de quantificar quantas *reads* foram mapeadas em cada transcrito ou gene de interesse. Para visualização de mapeamentos podemos utilizar ferramentas como o *samtools tview* [19] ou o IGV [12] (Figura 3.10).

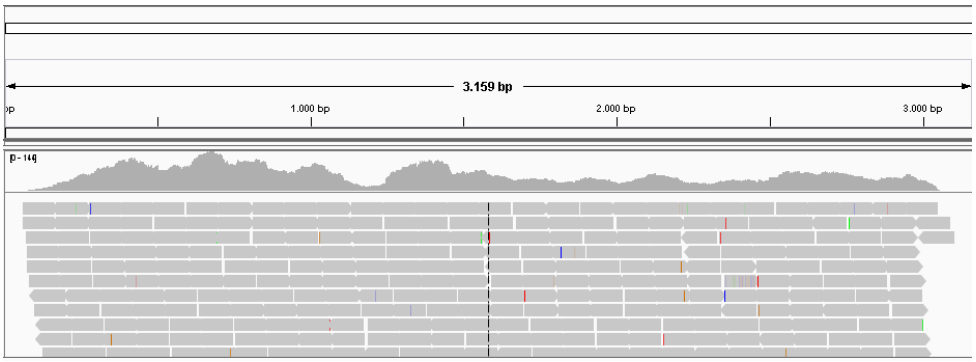

Figura 3.10: Exemplo da visualização de um mapeamento através do programa IGV. No visualizador é mostrado o tamanho do transcrito pelo tamanho da seta horizontal, no exemplo, o tamanho seria de 3.159 bp. As barras cinzas horizontais, apenas parte delas visíveis na figura, descrevem cada *read* e sua posição estabelecida no mapeamento para um determinado transcrito. Os picos em cinza representam a quantidade de *reads* mapeados naquela região, sendo que picos maiores indicam um maior número de *reads* naquela posição.

Em programas que estimam a abundância de transcritos utilizando genoma de referência, como no caso do Cufflinks (Figura 3.11), são obtidas abundâncias mais precisas em relação a referência, pois com o genoma, já se obtém a posição de genes, éxons e íntrons. Nos programas de montagem de novo, como, por exemplo no pipeline do Trinity, a abundância é estimada utilizando uma abordagem chamada RSEM (*RNA-Seq by Expectation Maximization*) [11] e outra chamada eXpress[1].
O RSEM é um estimador de abundância que se utiliza de evidências estatísticas para prever seus modelos de abundância. Uma análise utilizando RSEM consiste em apenas dois passos. No primeiro passo, um conjunto de referências dos transcritos é gerado para análises posteriores. Em seguida, um conjunto de *reads* de RNA-Seq são alinhados nos transcritos de referência e os alinhamentos resultantes são usados para estimar abundâncias e seus intervalos de confiança (pipeline observado na Figura 3.12).

[1] `http://bio.math.berkeley.edu/eXpress`

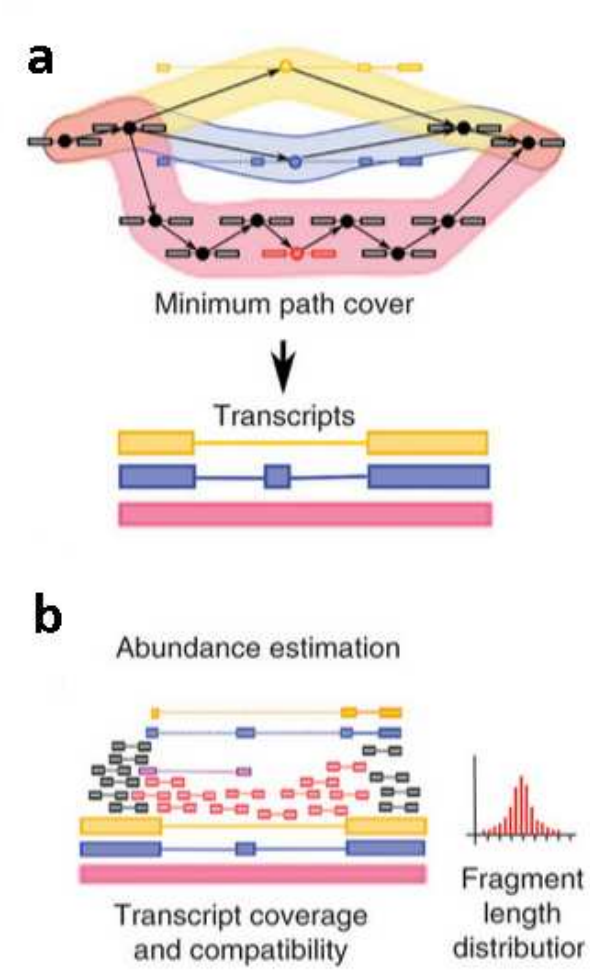

Figura 3.11: Estimação de abundância de transcritos do Cufflinks. (a) Determinação dos transcritos a partir de sobreposição de grafos. (b) Após o mapeamento dos *reads* estima-se a abundância de cada isoforma gerada na montagem (Figura editada de Trapnell *et al.* [8]).

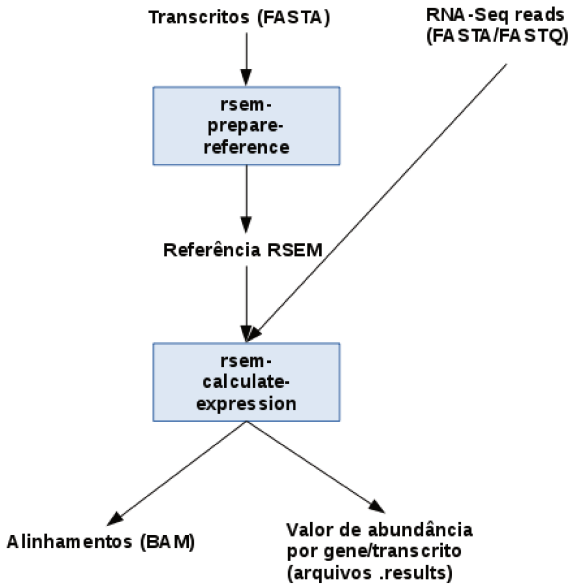

Figura 3.12: Pipeline utilizado na análise do RSEM para uma montagem *de novo*. (Figura editada de Li e Dewey [16].

O estimador de abundância eXpress utiliza espaço de memória proporcional ao tamanho do número de fragmentos utilizados. A partir do resultado de um mapeador, como o Bowtie, pode-se acoplá-lo ao eXpress sem precisar carregar o resultado do mapeamento na memória, sendo tal função uma das suas principais vantagens. Uma vantagem do eXpress, que o torna apropriado a montagens de novo, é a sua capacidade de resolver multimapeamentos entre famílias gênicas, problema que montadores de novo possuem certa dificuldade de resolver. O algoritmo utilizado para propor abundâncias de isoformas é uma variação do algoritmo utilizado pelo RSEM.

3.5 Análise de expressão diferencial

Uma das possíveis utilizações de experimentos de RNA-Seq é na busca e identificação de genes e/ou transcritos diferencialmente expressos entre duas condições. Tais genes são selecionados a partir de uma combinação de um limiar de mudança de expressão e corte de valores significativos, que são geralmente baseados em p-values gerados por um modelo estatístico.

A fim de modelar experimentos de RNA-Seq, os modelos estatísticos se utilizam de distribuições discretas (Poisson, Binomial Negativa etc.) em vez de contínuas (por exemplo, Normal) para normalização dos dados. As ferramentas mais utilizadas na análise de expressão diferencial de genes e transcritos são edgeR [32], Cufdiff e DESeq [2]. Segundo Zhang *et al.* [45], a fim de avaliar uma ferramenta de análise de expressão, devemos considerar três fatores: número de replicatas, a cobertura da sequência e o desbalanceamento entre os grupos de comparação.

Segundo Robles *et al.* [33] e Rapaport *et al.* [31], quanto maior o número de replicatas e a cobertura utilizada, melhor a acurácia desenvolvida pela ferramenta de análise de expres-

são diferencial. Ambos também citam que não existe uma ferramenta que se sobressaia sobre a outra em todos os critérios de qualidade. A comparação entre métodos de análises diferenciais já renderam muitos artigos [25, 31, 33, 46], e todos citam algumas características equivalentes, como a alta precisão da normalização de todos os métodos; a baixa consistência do Cuffdiff ao se utilizar baixa cobertura; o alto tempo de processamento do Cuffdiff, que fornece resultados após horas de processamento, tempo maior que os outros métodos, que fornecem em minutos; ambos métodos suportam análises sem replicatas e a dificuldade do DESeq em lidar com desbalanço entre replicatas.
Para desmistificar essas comparações, os desenvolvedores dos pacotes edgeR e DESeq publicaram um artigo [3] citando que nenhum método se sobressai em relação ao outro entre todas as condições do estudo. A Tabela 3.5 resume a comparação dos métodos utilizados por cada pacote utilizado neste estudo.

Pacote	Versão	Opções de normalização	Suposta distribuição dos read-counts	Teste de expressão diferencial
edgeR	3.8.5	TMM/Upper-Quantil/RLE/nenhum	Binomial negativa	teste exato
DESeq	1.18	DESeq size-factors	Binomial negativa	teste exato
Cuffdiff2	2.0	Geométrica/Uper-quartil/FPKM	Binomial Beta negativa	teste-t

Tabela 3.2: Resumo dos métodos utilizados por cada pacote de expressão diferencial.

Capítulo 4

Materiais e métodos

Neste capítulo, abordaremos a metodologia utilizada durante o trabalho na análise dos dados de RNA-Seq e, posteriormente, a seleção de transcritos verdadeiros positivos, diferencialmente expressos, gerados por montadores *de novo*.

4.1 Pipeline de RNA-Seq

Uma análise de RNA-seq é utilizada na identificação dos transcritos diferencialmente expressos entre duas condições, realizada a partir dos passos ilustrados na Figura 3.2 e descritos no Capítulo 3. Em suma, a análise requer um alto tempo de processamento e uso de memória, assim como *softwares* robustos de análise.

4.1.1 Dados utilizados

Os dados utilizados neste trabalho foram obtidos, a partir de projetos vinculados com a Embrapa (Empresa Brasileira de Pesquisas Agropecuárias). As bibliotecas utilizadas no projeto estão resumidas na Tabela 4.1.1.

	Organismo	Tipo de *read*	Condições	Replicatas por condição	Total de *reads*
Biblioteca 1	*Canis familiaris*	paired-end	ciclos estrais de ovários	4	440.494.292
Biblioteca 2	*Arabidopsis thaliana*	paired-end	Tratado Controle	4	244.167.802

Tabela 4.1: Resumo dos dados utilizados no desenvolvimento do projeto.

4.1.2 Análise de qualidade

O primeiro passo para uma análise em RNA-Seq é a análise da qualidade da biblioteca. Para a extração dos adaptadores foi utilizado o programa SeqyClean v 1.8.10, para dados Illumina, nas três bibliotecas. Em seguida, com o objetivo da análise da qualidade de cada biblioteca, utilizou-se o programa FastQC v.0.10.0. Para analisar a qualidade média da sua biblioteca é gerado um gráfico de Whisker, como mostrado na Figura 4.1.

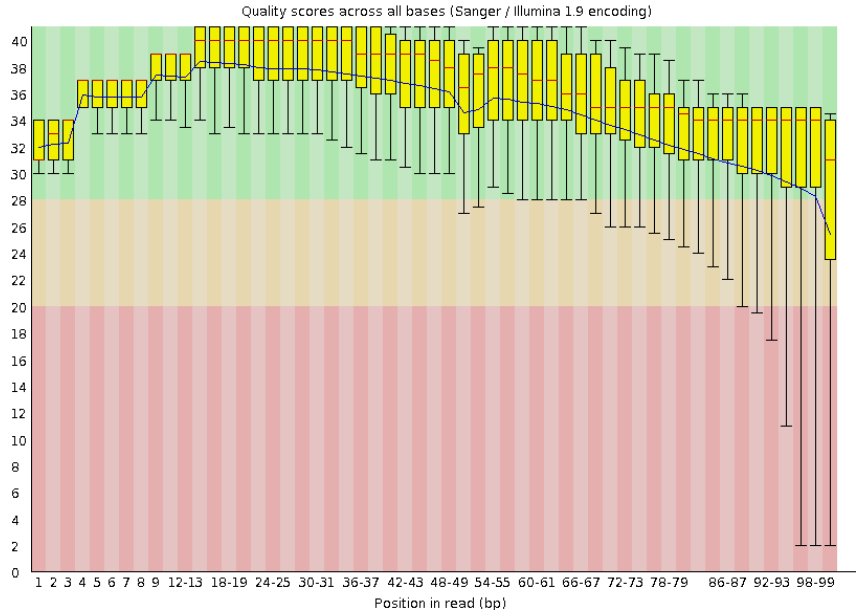

Figura 4.1: Exemplo de uma das análises de qualidade feitas através do software FastQC de uma biblioteca de *Arabidopsis thaliana*.

Podemos observar que existem algumas linhas representadas na Figura 4.1. A linha vermelha ao centro representa a mediana da qualidade da amostra das bases. A linha azul representa a média da qualidade da base. O tamanho da caixa amarela é dado pela diferença entre o primeiro quartil e o terceiro quartil, sendo assim, quanto maior a diferença entre as qualidades das bases, maior o tamanho da caixa, sendo que neste caso, queremos minimizar o tamanho. A linha contínua vertical sobre cada caixa amarela representa a dispersão dos dados. Ela representa a menor qualidade entre as bases e o maior valor entre as bases, então a fim de maximizar a qualidade da biblioteca, queremos linhas contínuas que não cheguem a qualidades baixas. Este eixo de qualidade é definido pelo eixo vertical esquerdo; já o eixo horizontal representa a variação das bases da sequência.
A análise de qualidade das bibliotecas de *Arabidopsis thaliana* e *Canis familiaris* verificamos que a qualidade média das bibliotecas eram em torno de 30. A porcentagem de *reads*, que foram retirados adaptadores Illumina de algumas bibliotecas, foi em torno de 1,5% para *Canis familiaris* e de 2% para *Arabidopsis thaliana*.

4.1.3 Montagem

A etapa de montagem é essencial para o estabelecimento de um bom transcriptoma de referência. Vários programas já foram utilizados para gerar um transcriptoma de novo ou fizeram parte de pesquisas acadêmicas que compararam cada montador. A Tabela 4.1.3 resume as citações contabilizadas no site Web Of Science[1] até julho de 2015 dos montadores *de novo* utilizados neste trabalho, sendo previamente escolhidos pelo grande número de citações (Trinity e Oases) e outro por ser um montador recém lançado (IDBA-Tran).

[1] `http://apps.webofknowledge.com/`

Montador	Número de citações	Ano de publicação
Trinity	2004	2011
Oases	478	2012
IDBA-Tran	15	2013

Tabela 4.2: Resumo do número de citações contabilizadas no site Web of Science de cada montador (Data de acesso: 26/08/2015).

Para a montagem, utilizando genoma de referência, utilizou-se o programa Cufflinks v 2.2.1.0, dentro da plataforma Galaxy [2], uma plataforma que gerencia a utilização de vários programas, de modo a montar um *pipeline*.
Para as montagens de novo, utilizaram-se os programas Trinity v2.0.4, Velvet-Oases v0.2.08 e IDBA-Tran v.1.1.1. Nas montagens em que se fez o uso do software Trinity, adotaram-se, além das opções padrão, as opções `-JM 150G` e `-min_kmer_cov` 2, sendo responsáveis pela memória utilizada no módulo Jellyfish e pelo número mínimo de cobertura do kmer, para ele ser utilizado na montagem, respectivamente. Já nas montagens utilizando o sofware Velvet-Oases, foi utilizado as linhas de comando, em sua respectiva ordem:

```
velveth <diretório> <kmer> <arquivo>;
velvetg <diretório> -read_trkg yes;
oases <diretório>
```

Já nas montagens em que se utilizou o programa IDBA-Tran, escolheu-se, além das opções padrão, as opções `-mink` , `-maxk` e `-step`, de modo a estabelecer o valor mínimo do kmer a ser utilizado (com –mink 23 e –maxk 31) e o passo de variação do kmer (com –step 2), respectivamente.
Como a saída do Cufflinks é um arquivo .gtf,que estabelece cada posição de genes, éxons e UTRs, então para gerar o arquivo do transcriptoma de referência, a partir das posições do mapeamento no genoma, usou o programa *gffread* com a opção `-w`, que representa a opção de extrair as sequências dos transcritos a partir de um arquivo gtf. Ao longo do trabalho, o transcriptoma gerado, por meio da saída do programa Cufflinks, será utilizado como um benchmark, uma vez que ele é o mais próximo da referência que se pode obter.

4.1.4 Mapeamento

O mapeamento neste trabalho foi feito utilizando o genoma de referência das espécies estudadas ou o transcriptoma gerado pela montagem *de novo*. Seguindo o *pipeline* estabelecido pelo Cufflinks, a partir do seu genoma de referência, utilizamos o programa TopHat para o mapeamento.
Nas montagens *de novo*, uma vez estabelecido o transcriptoma de referência, devemos mapear os *reads* de volta nele, visando o estabelecimento do transcrito de origem para os mesmos. Os mapeadores utilizados foram o Bowtie e Bowtie2. Nas montagens que foram feitas utilizando o programa Trinity, usamos o ao programa Bowtie com a opção `-all` , que retorna todas as possíveis posições do *read* no transcriptoma. Já nas montagens utilizando

[2] `http://galaxy-project.org`

os programas Velvet-Oases e IDBA-Tran, utilizou-se o mapeador Bowtie2 também com a opção `-all`. Tendo em vista que um *read* pode ser representada em um transcriptoma de referência por mais em um lugar, a estratégia de permitir *reads* multi-mapeadas foi usada. Vale a pena ressaltar que, antes de fazer o mapeamento, deve-se indexar o transciptoma a ser mapeado. O indexamento do Bowtie e do Bowtie2 foram deitos pelos programas bowtie-build e bowtie2-build, respectivamente, ambos disponibilizados juntos com os seus pacotes específicos.

4.1.5 Determinação da abundância e análise diferencial

Estimar a abundância é, de maneira básica, observar o mapeamento dos *reads* sobre o transcriptoma montado e dizer, por transcrito ou gene, quantos *reads* mapearam naquela região. Os estimadores de abundância levam em consideração o número de réplicas, o tamanho da biblioteca e o tamanho do transcriptoma.
Para estimar a abundância de transcritos gerados, a partir do Cufflinks, utilizou-se o programa Cuffdiff, pacote fornecido junto com o Cufflinks, que além de estimar a abundância dos transcritos por FPKM (fragmentos por quilobase de éxons por milhões de fragmentos mapeados) também faz a análise de expressão diferencial. Para estimar a abundância dos transcritos gerados pelo Trinity, utilizou-se o script `align_and_estimate_abundance.pl`, disponibilizado pelo pacote Trinity, com a opção de `-method RSEM`, que utiliza o RSEM como estimador de abundância. Para a abundância dos transcritos gerados pelo Velvet-Oases e IDBA-Tran, utilizou-se o eXpress.
Para a análise de expressão diferencial, usando genoma de referência, neste caso transcritos gerados a partir do Cufflinks, utilizou-se o programa Cuffdiff, pacote do Cufflinks. Para a análise dos programas de montagem *de novo*, em todos os casos, utilizou-se o programa edgeR, já que existia uma baixa intersecção entre a lista de transcritos diferencialmente expressos do DESeq e a do Cuffdiff, com significância estatística menor que 5% ($FDR \leq 0.05$).

4.2 Comparação de montagens

A determinação de uma melhor montagem de um transcriptoma ainda é um grande desafio para a comunidade científica. Como a montagem usando genoma de referência possui uma validação maior, já que ela utiliza informações do genoma, o objetivo deste trabalho, em relação a montagem, seria descobrir qual montagem *de novo* se "aproxima", com maior fidelidade, da montagem com genoma de referência.
A lista gerada pelos programas de análise de expressão diferencial se difere pelo nome do transcrito, sendo assim, precisamos fazer a designação correta de cada nome, comparando as sequências geradas em cada montagem. Para designar o nome que um transcrito possui na outra montagem, foi feito um blastn entre cada transcriptoma gerado e utilizou-se a metodologia BBH (Bidirectional Best Hit) [28]: dado um transcrito Xa do transcriptoma Ta e um outro transcrito Xb do transcriptoma Tb, Xa e Xb são ditos BBH um do outro se não existe ninguém mais similar a Xb do que Xa em Ta, e respectivamente, não existe ninguém mais similar a Xa do que Xb em Tb (Figura 4.2).

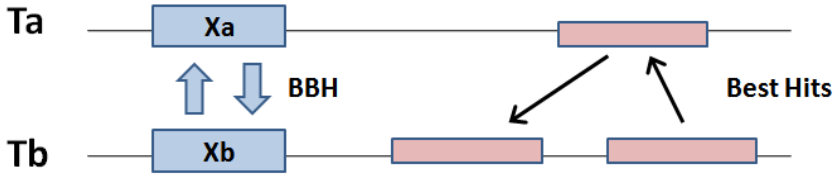

Figura 4.2: Exemplo de um alinhamento entre dois transcriptomas e seus possíveis resultados referentes ao melhor hit.

Espera-se que uma montagem consistente tenha um maior número de transcritos com BBH. Pensando nisso comparamos a porcentagem de BBH gerada em cada montagem *de novo*. A referência para comparação das montagens utilizando o organismo Arabidopsis thaliana foi extraída do site RefSeq, já a referência utilizada na comparação das montagens de *Canis familiaris* foram os transcritos gerados pelo nosso reconstrutor Cufflinks, que utiliza genoma de referência, já que nosso experimento utilizou cerca de 400 milhões de *reads*, muito superior ao número de *reads* utilizado pelo RefSeq para montar sua referência, que foi de cerca de milhões.
Como a maioria dos transcritos montados a partir de dados de RNA-Seq são derivadas de sequências de RNA poliadenilado, é esperado que eles gerem proteínas, logo, um bom transcriptoma teria a capacidade de conservar regiões que dessem origem a certas proteínas. Pensando nisso, neste trabalho foi analisado a qualidade da montagens finais de transcriptoma utilizando genes de cópia única (GCU), a partir do banco de dados BUSCO [39], utilizando 429 GCU.
Primeiramente fizemos um alinhamento tblastn entre os 429 GCU e a montagem feita em cada caso de teste. Pelo fato do blast produzir alinhamentos locais, não podemos utilizar uma critério binária (se existe um GCU na montagem ou não), pois sequências podem produzir *scores* de alinhamento melhores que outras. Pensando nisso, extraímos a ORF correta segundo o frame de alinhamento e utilizamos o programa ClustalW2 [40], com parâmetros padrão, para a execução de um alinhamento global entre a ORF traduzida e o GCU correspondente.
Ao final, para cada montagem, contamos o número de ORFs que obtiveram cobertura de no mínimo 70% no alinhamento, em relação a cada GCU, e somamos os valores do *score* de alinhamento global, produzidos pelo ClustalW2, destas ORFs, que chamaremos de W, ou seja, quanto maior for o valor de W, melhor o alinhamento, implicando em uma melhor a montagem.
Após a análise de expressão diferencial, e tendo convertido os nomes de cada transcrito entre duas montagens, é feito uma comparação entre as listas finais de cada análise de expressão diferencial e observamos a quantidade de transcritos em comum. Para analisar se certa montagem trouxe como resultado transcritos verdadeiros, criamos um fator de decisão d, representado pela razão entre o número de transcritos em comum entre as montagens sobre o total de transcritos ditos como diferencialmente expressos. Espera-se que, quanto melhor a intersecção entre as listas, mais perto de reconstruir os transcritos verdadeiros diferencialmente expressos o montador *de novo* está, ou seja, mais próximo de 1 é o valor de d.

4.3 Classificação dos transcritos diferencialmente expressos

Os programas de análise de expressão diferencial retornam como saída uma lista de transcritos, podendo ou não conter quimeras geradas por problemas de montagem. Tendo em vista que sem a obtenção do genoma a montagem fica sujeita a mais erros, um experimento de RNA-Seq, com o objetivo de validar a expressão de um determinado transcrito em uma determinada condição, fica sujeito a validações incorretas ao se selecionar um transcrito para ser validado em bancada. Ao realizarmos testes inciais de escolha de transcritos diferencialmente expressos em uma análise de RNA-Seq, notamos que a porcentagem de acerto era em torno de 40%, utilizando o critério Fold-change (FC).
Pensando nestes aspectos, uma classificação por determinados critérios, que aumentariam a chance de escolher um transcrito dito verdadeiramente positivo, seria de grande ajuda. Os critérios empregados neste trabalho para a classificação estão resumidas na Tabela 4.3. O critério 17 é o critério sugerido pela literatura, na qual a seleção é feita após gerar o gráfico do Volcano plot [6]. O gráfico do Volcano plot exibe os genes/transcritos que são muito expressos (possuem um p-value baixo e um Fold-Change alto, são ditos *up-regulated*) e pouco expressos (possuem um p-value baixo e um Fold-Change baixo, são ditos *down-regulated*). Para geração do critério 17 aplicamos o módulo do log FC, assim o topo da lista seria formado tanto pelos *up-reguated* e *down-regulated*.

	Critérios	Ordem (crescente/decrescente)
Critério 1	Número de *reads*	decrescente
Critério 2	Fold-change	crescente
Critério 3	Fold-change	decrescente
Critério 4	P-value/FDR	crescente
Critério 5	ORFs	decrescente
Critério 6	Número de *reads* e Fold-change	(decrescente/crescente)*
Critério 7	Número de *reads* e Fold-change	(decrescente/crescente)*
Critério 8	Número de *reads* e ORFs	(decrescente/decrescente)*
Critério 9	Número de *reads* e P-value/FDR	(decrescente/crescente)*
Critério 10	Fold-change e Fold-change	(crescente/crescente)*
Critério 11	Fold-change e ORFs	(crescente/decrescente)*
Critério 12	Fold-change e P-value/FDR	(crescente/crescente)*
Critério 13	Fold-change e ORFs	(decrescente/decrescente)*
Critério 14	Fold-change e P-value/FDR	(decrescente/crescente)*
Critério 15	P-value/FDR e ORFs	(crescente/decrescente)*
Critério 16	Módulo do log(FC)	decrescente
Critério 17	P-value/FDR e Módulo do log(FC)	(crescente/decrescente)*

*As ordens são respectivas aos critérios. Por exemplo, o critério Número de *reads* e P-value/FDR (Critério 9) possui ordem (decrescente/crescente), significando que a lista do criério Número de *reads* foi ordenada de forma decrescente e a do critério P-value/FDR foi ordenada de forma crescente.

Tabela 4.3: Critérios de classificação utilizados.

Cada critério descrito na tabela acima é calculado de uma forma diferente. O número de *reads* é calculdo pela soma de todos os *reads* mapeados por determinado transcrito em todas as condições da análise; o fold-change é calculado pela razão entre o total de *reads* de um determinado transcrito em uma condição e o total de *reads* desse mesmo transcrito em outra condição; o p-value é a probabilidade de um valor tão ou mais extremo quanto o observado ocorrer dentro da hipótese nula. Após alguns testes, observou-se que o valor de FDR (p-value ajustado) se comportava de forma igualitária ao valor de p-value, logo, colocamos ambos juntos no teste; e para cada transcrito foi extraído o tamanho da sua maior ORF e esse valor utilizado no critério ORFs.
A ordenação por dois critérios funcionou da segunte forma: primeiro, ordenamos ambos os critérios pela forma descrita na Tabela 4.3. Posteriormente, percorremos ambas as listas até encontrarmos intersecções entre elas, e essa intersecção seria a lista definida para ambos os critéros utilizados. A Figura 4.3 exemplifica este método.

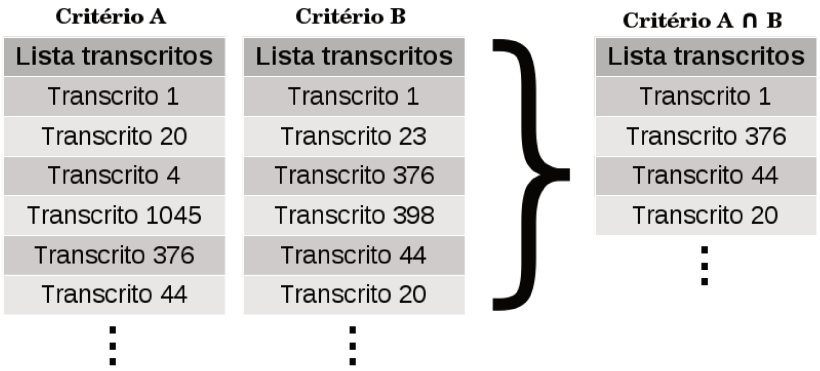

Figura 4.3: Exemplo do método de ordenação com dois critérios. Dado a lista de transcritos ordenadas de uma forma pré-estabelecida, para dois critérios distintos, percorremos ambas as listas e encontramos suas intersecções. A lista gerada pelas intersecções será utilizada para comparação nas etapas posteriores deste trabalho.

Após a lista ordenada (tanto por um ou dois critérios), verificou-se, em intervalos de 10 em 10 transcritos, quantos eram verdadeiramente diferencialmente expressos. Por exemplo, analisamos o intervalo de 1 à 10 transcritos e contamos quantos eram verdadeiramente diferencialmente expressos, em seguida, no intervalo de 1 à 20, e assim por diante. Um transcrito, nestes intervalos, foi dito como verdadeiramente diferencialmente expresso se ele tinha correspondência com algum transcrito dito diferencialmente expresso pelo Cuffdiff, gerado pelo pipeline do Cufflinks.
Para dizer se um critério possui mais significância estatística em relação a outro critério, foi realizado um teste pareado de Wilcoxon [44], utilizando o pacote *stats* da plataforma R [30] versão 3.1.0, um teste estatístico não paramétrico, que possui como hipótese nula que ambas as distribuições são estocasticamente idênticas. Já que todos os resultados estão no mesmo intervalo, eles podem então ser emparelhados. Assumimos que n objetos estão dispostos em duas observações x_i e y_i para cada objeto i, no qual resultados estão dispostos em duas amostras $x_i, \ldots, x_n$ e $y_i, \ldots, y_n$. Para ser executado o teste pareado de Wilcoxon, com hipótese H_0 de que as diferenças entre as amostras seja zero, deve ter

como requisitos que valor $d_i = x_i - y_i$ seja independente e que as observações estejam distribuídas em intervalos. O algoritmo é dado por:

1. Para cada par de amostra (x_i, y_i) é calculada sua diferença $d_i = x_i - y_i$
2. O valor absoluto das diferenças $|d_i|$ é calculado e ranqueado. O valor da sua posição de ranqueamento ganha o sinal, sendo positivo se $d_i > 0$ e negativo se $d_i < 0$. Valores que tenham $d_i = 0$ não entram no ranqueamento.
3. É calculado o valor de significância W, dado pela soma dos valores positivos da sua posição de ranqueamento ($W = \sum_{i=1}^{n} d_i \times R$, onde R tem valor $+1$ se $d_i > 0$, caso contrário, $R = 0$)
4. Se o valor de $W > W_{critico}$[3], então certo critério tem significância estatística em relação a outro. Em nossos testes utilizamos $W_{critico} = 31$, já que $n = 15$ [23].

A diferença entre o test-t emparelhado para o teste pareado de Wilcoxon é o tipo de distribuição assumida pelos testes. No teste-t emparelhado, as diferenças entre os pares devem apresentar uma distribuição normal, já no teste pareado de Wilcoxon, por ser não-paramétrico, não leva essa premissa. Isso é observável ao compará-los entre os testes estatísticos na presença de *outliers* (valores extremos), assim, os *outliers* desta distribuição distorceriam o resultado do teste-t emparelhado, mas tem impacto limitado no teste pareado de Wilcoxon.

[3]Ver Apêndice A.3

Capítulo 5

Resultados e discussões

Neste capítulo serão apresentados os resultados obtidos após as montagens feitas a partir das bibliotecas e a análise dos critérios de ordenação, ambos citados no Capítulo 4.

5.1 Casos de testes

Para a execução da metodologia de análise de RNA-seq foram estabelecidos três conjuntos de dados para cada organismo do estudo. Pensando em como os montadores *de novo* reagiriam à diminuição do uso de dados disponíveis para as montagens, foram gerados casos de testes, a partir do mesmo volume de dados, para verificar como os montadores reagiriam a certa quantidade de dados, de modo a ocasionar uma grande dispersão entre as montagens, assim, poderia se determinar uma quantidade de dados mínima para a manutenção de uma boa montagem *de novo*. A análise de como cada montador se comporta durante a alteração dos dados utilizados pode ser respondida pela análise das tabelas de resultados de cada montador. Os conjuntos de testes utilizados foram os seguintes:

1. Todos os *reads* da biblioteca do organismo (Tr).
2. Metade dos *reads* da biblioteca do organismo (Mr).
3. Somente *reads* da extremidade R1 da biblioteca do organismo (single-end) (Sr).

Note que o caso Sr e Mr tem o mesmo número de *reads*, mas características diferentes.

5.2 Montagens

A partir dos casos de testes gerados para a execução deste trabalho, citado na Seção 5.1, foram realizadas montagens com genoma de referência e montagens *de novo*. Como as montagens feitas a partir dos programas Velvet-Oases e IDBA-Tran foram feitas a partir da variação de k-mers, as métricas média, mediana, N50, desvio padrão e número de transcritos gerados foram resumidas como a média aritmética entre todas as montagens para estes montadores. Nesta seção também são comparadas as montagens através da análise do BBH, como descrito na Seção 4.2.

A Tabela 5.2 mostra o valor das mesmas métricas aplicadas em cada montagem *de novo* ao transcriptoma curado de cada genoma. As métricas geradas com genoma de referência utilizando o Cufflinks para cada caso de teste estão resumidas na Tabela 5.2. Analisando

a intersecção dos transcritos gerados pelo Cufflinks e os de referência, notou-se uma intersecção de 98% proporcionada pelos dados de *Arabidopsis thaliana* e de 100% pelos dados de *Canis familiaris*. A anotação gerada pelo Refseq de *Canis familiaris* possui muito menos transcritos que os gerado pelo Cufflinks em nossa análise (26.107 e 118.712 transcritos, respectivamente). Este fato se dá pelo modo da geração de dados, sendo que o Refseq utilizou menos sequências, pelo ano de publicação dos dados, feita em 2011, e por usarem dados simulados. Pensando nisso, para o caso de teste usando dados de *Canis familiaris* iremos considerar a montagem do Cufflinks como a referência para as análises.

	Número de transcritos	Tamanho médio (bp)	Mediana (bp)	Desvio Padrão	N50 (bp)
*Arabdopsis Thaliana**	41.671	1.556,65	1.371	1.105,13	1.912
*Canis familiares***	26.107	2.248,23	1.749	2.234,70	3.680

Tabela 5.1: Resumo das métricas de montagem referente aos cDNAs de referência anotados.

(*) Dados disponíveis em `ftp://ftp.arabidopsis.org/Sequences/blast_datasets/TAIR10_blastsets/`

(**) Dados disponíveis em `ftp://ftp.ensembl.org/pub/release-81/fasta/canis_familiaris/`

	Número de transcritos	Tamanho médio (bp)	Mediana (bp)	Desvio Padrão	N50 (bp)
Arabidopsis thaliana	65.658	1.883,45	1.583	1.373,82	2.335
Canis familiaris	118.712	3.995,69	3.275	3.292,81	5.782

Tabela 5.2: Resumo das métricas de montagem referente ao Cufflinks

5.2.1 *Arabidopsis thaliana*

Arabidopsis thaliana é um organismo modelo no estudo da genética, já que trabalhos com seu genoma vem sendo desenvolvidos desde o início de seu sequenciamento no ano de 1996. Ela foi utilizada nos testes pela disponibilização dos dados e por ser um organismo modelo. Abaixo estão descritas as métricas de livre referência geradas a partir dos montadores *de novo*.

	Número de transcritos	Tamanho médio (bp)	Mediana (bp)	Desvio Padrão	N50 (bp)
Tr	48.103	1.024,83	789	1.182,11	1958
Mr	43.114	1.228,52	872	1.144,51	1911
Sr	43.094	1.208,57	849	1.134,68	1883

Tabela 5.3: Resumo das métricas de montagem referente ao montador *de novo* Trinity.

	Número de transcritos	Tamanho médio (bp)	Mediana (bp)	Desvio Padrão	N50 (bp)
Tr	94.361	1.661,14	1.318	1.349,84	2.319
Mr	76.907	1.792,96	1.482	1.396,80	2.387
Sr	74.521	1.646,83	1.354	1.255,48	2.183

Tabela 5.4: Resumo das métricas de montagem referente ao montador *de novo* Velvet-Oases.

	Número de transcritos	Tamanho médio (bp)	Mediana (bp)	Desvio Padrão	N50 (bp)
Tr	51.820	1.110,55	769	968,14	1.578
Mr	50.424	1.267,28	949	1.049,84	1.772
Sr	48.574	1.384,34	1.077	1.120,08	1.912

Tabela 5.5: Resumo das métricas de montagem referente ao montador *de novo* IDBA-Tran

A análise de comparação das métricas de livre referência para o caso de teste de *Arabidopsis thaliana* tomou como referência os dados descritos na Tabela 5.2. Os transcritos gerados pelo Cufflinks foram em maior número em relação aos dados de referência, mostrando que o Cufflinks encontrou mais transcritos, sendo estes verdadeiros ou não. O montador *de novo* Trinity foi o que mais se aproximou da referência nas métricas número de transcritos, desvio padrão e N50. O montador Velvet-Oases se aproximou mais em relação à referência na métrica mediana, e o montador IDBA-Tran na métrica tamanho médio.
Em ambos os montadores *de novo* houve uma diminuição no número de transcritos gerados alterando o tamanho da biblioteca, mas, vale ressaltar que o tamanho dos dados das bibliotecas Mr e Sr são iguais, mudando somente o tipo de dado (*paired-end* e *single-end*), mostrando a partir das tabelas que, apesar de próximas, as montagens são diferentes pelas métricas apresentadas.

5.2.2 *Canis familiaris*

Canis familiaris, mais conhecido como cachorro, é um organismo muito útil na área de pesquisa médica devido a sua variação morfológica e genética. A criação de cães e a combinação de indivíduos geraram raças caninas que são mais sucetíveis a doenças como câncer, cegueira e problemas cardíacos, tal fato que leva cientistas a usarem tais raças em pesquisas. Utilizamos neste estudo dados de cachorro devido a sua complexidade, já que ele é pertencente do reino *Animalia*. Abaixo estão descritas as métricas de livre referência obtido por cada montador *de novo*.

	Número de transcritos	Tamanho médio (bp)	Mediana (bp)	Desvio Padrão	N50 (bp)
Tr	193.227	942,63	427	1.258,75	1978
Mr	130.383	1.014,80	449	1.296,80	2144
Sr	126.008	1.029,56	467	1.289,58	2123

Tabela 5.6: Resumo das métricas de montagem referente ao montador *de novo* Trinity.

	Número de transcritos	Tamanho médio (bp)	Mediana (bp)	Desvio Padrão	N50 (bp)
Tr	220.508	1.418,20	789	1.636,86	2.423
Mr	135.146	1.606,75	913	1.757,62	2.802
Sr	139.882	1.655,48	959	1.775,62	2.849

Tabela 5.7: Resumo das métricas de montagem referente ao montador *de novo* Velvet-Oases.

	Número de transcritos	Tamanho médio (bp)	Mediana (bp)	Desvio Padrão	N50 (bp)
Tr	225.065	800,02	530	765,56	995
Mr	158.358	883,13	547	914,17	1.200
Sr	140.291	944,39	579	982,6	1.334

Tabela 5.8: Resumo das métricas de montagem referente ao montador *de novo* IDBA-Tran

Os montadores de novo Trinity, Velvet-Oases e IDBA-Tran para ambos os casos de testes conseguiram manter o valor das métricas tamanho médio, mediana, desvio padrão e N50, independente da variação de dados, observando uma menor sensibilidade para o volume de dados utilizados em sua montagem.
Como já mencionado, as bibliotecas Mr e Sr possuem o mesmo tamanho, mudando somente o tipo de dado (*paired-end* e *single-end*), mostrando a partir das tabelas que, apesar de próximas, as montagens são diferentes pelas métricas apresentadas.
Em relação ao número de transcritos gerados, ambos os montadores geraram transcritos a mais que o de referência descrito na Tabela 5.2, fato que mostra a tendência dos montadores *de novo* gerarem transcritos falsos positivos.

5.2.3 Análise do BBH

A análise de BBH descreve se duas montagens possuem transcritos em comum através da similaridade da sequência. Avaliar o BBH entre duas montagens pode definir uma comparação, neste caso uma montagem *de novo* em relação à referência, e dizer se ela foi bem sucedida ou não. Uma montagem é dita bem sucedida se ela consegue reconstruir uma maior porcentagem de transcritos em relação à referência. As tabelas 5.2.3 e 5.2.3 mostram a razão entre o total de transcritos com BBH e o total de transcritos montados por cada montador *de novo* em seu respetivo caso de teste.

	Trinity	Velvet-Oases	IDBA-Tran
Tr	0.84	0.44	0.62
Mr	0.86	0.52	0.67
Sr	0.85	0.50	0.65

Tabela 5.9: Razão entre o total de transcritos com BBH e o total de transcritos montados para os dados de *Arabidopsis thaliana*.

	Trinity	Velvet-Oases	IDBA-Tran
Tr	0.56	0.48	0.42
Mr	0.74	0.71	0.58
Sr	0.72	0.67	0.54

Tabela 5.10: Razão entre o total de transcritos com BBH e o total de transcritos montados para os dados de *Canis familiaris*.

Analisando as tabelas 5.2.3 e 5.2.3 notamos que quanto maior o volume de dados, mais falsos positivos podemos encontrar, ocasionado pela diferença entre o caso de teste Tr e os Mr e Sr. Além disso, o montador *de novo* Trinity conseguiu uma maior porcentagem de reconstrução dos transcritos montados, representando uma melhor montagem em relação à anotação. Ainda podemos notar que as porcentagens de reconstrução entre os casos de teste Mr e Sr, apesar de próximos, sofrem alteração devido ao tipo de dado estudado.
Para o organismo *Canis familiaris*, o caso de teste Tr teve uma menor porcentagem de reconstrução, pois o número de transcritos com BBH em sua montagem (56 % da ordem de 193 mil) é maior que a dos outros casos de teste Mr e Sr (74 % da ordem de 130 mil e 72% da ordem de 126 mil, respectivamente), mostrando que apesar de uma maior eficiência, o número de falsos positivos montados pelo caso Tr foi maior que o Mr e Sr.

5.2.4 Avaliação intrínseca de montadores *de novo*

Pensando em uma métrica intrínseca que avaliasse um montador *de novo*, ou seja, que não utilizasse um genoma de referência para validação, verificaremos sua capacidade de encontrar genes de cópia única (chamaremos de GCUs) em eucariotos, utilizando 429 GCUs do banco de dados BUSCO. As tabelas 5.2.4, 5.2.4 e 5.2.4 abaixo resumem esta análise, juntamente com o valor de W, e os resultados do Cufflinks. No Apêndice A.1 deste trabalho foram colocados todos os GCUs e quais foram encontrados em cada caso de teste.

Montador	Caso de teste	Total de GCUs encontrados	Porcentagem representativa (%)	Valor W
Trinity	Tr	319	74.35	156.719
	Mr	312	72.72	152.415
	Sr	310	72.22	151.120
Velvet-Oases	Tr	296	68.99	104.038
	Mr	291	67.83	102.182
	Sr	287	66.89	100.345
IDBA-Trans	Tr	246	57.34	86.832
	Mr	243	56.64	85.112
	Sr	240	55.94	84.016

Tabela 5.11: Porcentagem de GCUs encontrados em cada análise por cada montador *de novo* para o organismo *Arabidopsis thaliana* dividido por casos de teste. O total de GCUs encontrados são aqueles que obtiveram alinhamentos com cobertura de pelo menos 70%.

Montador	Caso de teste	Total de GCUs encontrados	Porcentagem representativa (%)	Valor W
Trinity	Tr	337	78.55	208.318
	Mr	333	77.62	206.845
	Sr	332	77.38	205.776
Velvet-Oases	Tr	212	49.41	99.558
	Mr	208	48.48	97.110
	Sr	208	48.48	96.530
IDBA-Trans	Tr	191	44.52	84.625
	Mr	182	42.42	79.234
	Sr	180	41.95	78.212

Tabela 5.12: Porcentagem de GCUs encontrados em cada análise por cada montador *de novo* para o organismo *Canis familiaris* dividido por casos de teste. O total de GCUs encontrados são aqueles que obtiveram alinhamentos com cobertura de pelo menos 70%.

Organismo	Total de GCUs encontrados	Porcentagem representativa (%)	Valor W
Arabidopsis thaliana	371	86.48	238.465
Canis familiaris	361	84.14	278.688

Tabela 5.13: Porcentagem de GCUs encontrados em cada organismo pelo Cufflinks. O total de GCUs encontrados são aqueles que obtiveram alinhamentos com cobertura de pelo menos 70%.

Ao analisarmos as tabelas acima, considerando o somatório dos *scores* de alinhamento global entre a ORF do transcrito montado e a proteína correspondente (Valor W), notamos que o montador *de novo* Trinity se sobressai em relação aos outros em todos os casos de teste, e em ambos os organismos. Inclusive o montador Trinity é o que mais encontra GCUs em ambos os casos de teste, em relação aos outros montadores.
Em relação ao melhor montador nesta análise, neste caso o Trinity, podemos notar que os casos de teste Mr e Sr, que são diferentes somente pelo tipo de dado e não pelo volume de dado, possuem Valor W diferentes, mesmo sofrendo pouca alteração.
O Cufflinks, que utiliza genoma de referência para a geração de transcritos, para o organismo *Arabidopsis thaliana* encontra um número de GCUs próximo ao montador *de novo* Trinity, mas o valor W é muito superior, certa de 52% maior, fato devido ao alinhamento global produzir um maior *score*. Para organismo *Canis familiaris* apesar do montador *de novo* Trinity encontrar quase o mesmo número de GCUs que o Cufflinks, o valor W para o Cufflinks é superior, cerca de 33% maior. Fatos esperados, já que o Cufflinks utiliza genoma de referência em sua geração de transcritos.
Essa métrica intrínseca proposta neste trabalho leva em consideração fatores apenas presentes na montagem e, sem depender de referência, pode dizer se um transcriptoma foi corretamente montado. Os GCUs devem aparecer em análises com genomas eucariotos (se os mesmos tiverem expressos) e se o montador *de novo* conseguir reconstruir uma maior porcentagem desses genes, pode-se dizer que melhor foi a montagem realizada por tal.

5.3 Tempo de processamento e memória utilizada nas montagens *de novo*

Um fator que deve ser levado em consideração para montagens é o alto custo de processamento dos dados, quando não se tem o genoma de referência. Pensando neste fato, e na sensibilidade do algoritmo de cada montador em relação aos dados utilizados, foram gerados gráficos mostrando tanto o tempo de processamento utilizado em cada caso de teste por cada montador (Figura 5.1), quanto a memória utilizada no processo (Figura 5.2). Em cada figura os casos de teste referente ao organismo *Arabidopsis thaliana* contém um sufixo '-A' e os referente ao organismo *Canis familiaris* contém um sufixo '-C'. Os testes foram realizados em um servidor, contendo as seguintes configurações:

- Um Head Node contendo 2 processadores AMD Opteron Sixteen-core 6376 de 2.3 Ghz, 16 Mb de cache com 256 GB de memória RAM DDR3 de 1600 MHz.
- Um Service/Data Node contendo 2 processadores AMD Opteron Sixteen-core 6376 de 2.3 Ghz, 16 Mb de cache com 128 GB de memória RAM DDR3 de 1600 MHz.
- Cinco Compute Nodes contendo 4 processadores AMD Opteron Sixteen-core 6378 de 2.4 Ghz, 16 Mb de cache com 512 GB de memória RAM DDR3 de 1600 MHz.
- Um Compute Nodes contendo 16 processadores Intel(R) Xeon(R) de 2.4 GHz, 16 MB de cache (160 núcleos) com 2 TB de memória RAM DDR3 1600 MHz.
- Dois Compute Nodes contendo 8 processadores Xeon 6-core de 2.00 GHz com 18 Mb de cache com 512 GB e 1 TB de memória RAM.
- Um Compute Node com 4 processadores Six-Core AMD Opteron(tm) de 2.4 GHz com 256 GB de memória RAM.

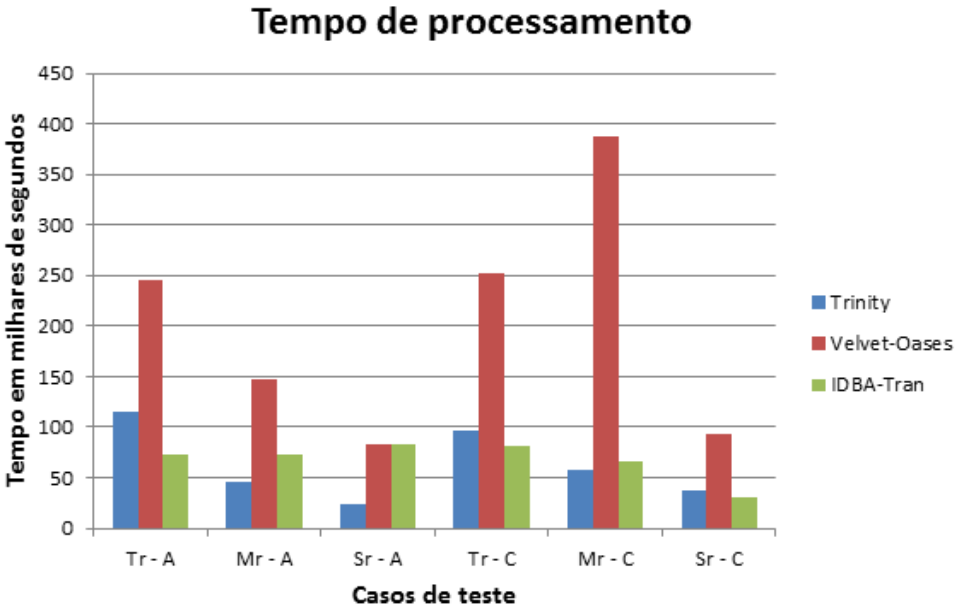

Figura 5.1: Resumo do tempo de processamento das montagens de novo dividido por caso de testes e montadores.

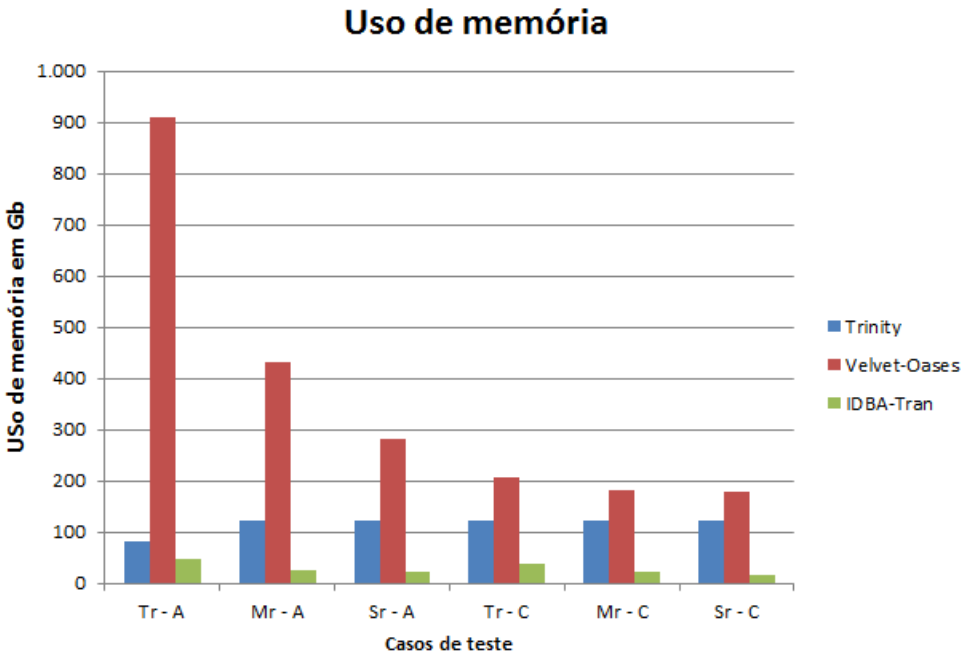

Figura 5.2: Resumo da memória utilizada durante as montagens de novo dividido por caso de testes e montadores.

Analisando o gráfico na Figura 5.1 observamos que os algoritmos dos montadores *de novo* levam um tempo razoável para seu processamento, variando entre 24 e 168 horas. Os algoritmos dos montadores são sensíveis a alteração do volume de dados utilizados na montagem, assim como o tipo de dado. Para os casos de teste que utilizam o volume total de dados (casos de teste Tr) o tempo de processamento em relação aos que utilizam metade do volume (casos de teste Mr) e os com tipo de dados diferentes (casos de teste Sr) o tempo de processamento é muito maior. Ainda a partir desse resultado pode-se observar que o tipo de dado interfere na sensibilidade do algoritmo em relação ao tempo em ambos os montadores *de novo*, com exceção do Velvet-Oases.
Analisando a Figura 5.2, percebemos claramente que computadores pessoais, que possuem tipicamente até 8GB de memória RAM, não possuem memória suficiente para realizar uma montagem *de novo*, que necessita em média de 116 GB para o Trinity, 365 GB para o Velvet-Oases e 29 GB para o IDBA-Tran, sendo possível somente com o uso de servidores, por exemplo. O montador *de novo* Velvet-Oases é quem mais utiliza memória para a sua montagem, principalmente na fase de gerar o grafo. Pelo gráfico podemos notar que o montador Trinity não altera seu uso de memória pela variação do volume de dados utilizados ou o tipo de dado. O montador *de novo* que menos requer memória e tempo de processamento é o IDBA-Tran, já que seu algortimo de montagem se utiliza da montagem anterior para realizar uma nova montagem com *k-mers* diferentes, diminuindo o tempo de processamento para um novo grafo ser gerado.

5.4 Transcritos diferencialmente expressos gerados

A partir da lista de transcritos ditos como diferencialmente expressos por cada montador *de novo* foi feita a comparação com a lista dos transcritos gerados pelo Cuffdiff, considerados verdadeiros, depois da análise por BBH correspondente. Um fator chamado de decisão (d) foi criado a partir da razão entre o número de transcritos contidos na intersecção entre os métodos sobre os transcritos totais ditos como diferencialmente expressos, que avaliaria a porcentagem de transcritos verdadeiros e a quantidade de falsos positivos. Para os montadores que geraram montagens com diferentes *k-mers* foi feita uma média dos fatores de decisão para cada caso de teste, como mostra a Tabela 5.4.

	Organismo	Trinity	Velvet-Oases	IDBA-Tran
Arabidopsis thaliana	Tr	0.22	0.13	0.14
	Mr	0.13	0.04	0.06
	Sr	0.09	0.03	0.04
Canis familiaris	Tr	0.55	0.25	0.23
	Mr	0.46	0.19	0.22
	Sr	0.45	0.22	0.18

Tabela 5.14: Fatores de decisão por montador e casos de teste.

Pela Tabela 5.4, podemos notar que os fatores de decisão são baixos, fato que mostra que os testes de expressão diferencial descrevem muitos transcritos falsos positivos. Para casos de teste onde a intersecção entre os métodos de montagem foi alta, houve destaque para o montador Trinity em casos de teste com alto volume de dados (caso de teste Tr). Pensando em escolher uma determinada lista de transcritos para ser testado em bancada em uma análise de *Real-Time PCR*, por exemplo, necessitamos encontrar um modo de escolher transcritos verdadeiramente diferencialmente expressos em uma lista que pode conter transcritos falsos positivos. Para executar tal procedimento, escolhemos alguns critérios de seleção que ordenados corretamente, fornecessem uma lista com uma maior chance de um determinado transcrito ser realmente diferencialmente expresso. Tal análise segue na próxima seção.

5.5 Análise dos critérios de seleção

Tendo como objetivo do trabalho analisar critérios que maximizam a seleção de transcritos verdadeiros positivos após uma análise de expressão diferencial em experimentos de RNA-Seq descritos na Seção 4.3, foram feitas as análises de expressão diferencial de cada montador para cada caso de teste, e analisado para cada critério qual montador seria mais eficiente e qual critério maximiza a escolha de um transcrito verdadeiro para o teste de bancada.

Foram determinados 5 critérios com testes iniciais e suas possíveis combinações, totalizando um total de 17 critérios. Além disso, foram analisados a quantidade, em porcentagem, do número de transcritos verdadeiros em 10 intervalos, variando de 10 a 100, do

total de transcritos ditos como verdadeiros após a análise de expressão diferencial. Foi gerada então uma matriz 17x10, em que cada posição continha qual a porcentagem de acerto naquele intervalo e qual montador tinha gerado aquela porcentagem. Após essa análise conseguimos determinar qual montador *de novo* se sairia melhor em encontrar uma maior porcentagem de transcritos verdadeiros e qual o melhor critério de seleção destes transcritos.
As montagens feitas com o programa Trinity utilizou k-mer 25. Já as montagens SASP com o programa Velvet-Oases foram realizadas uma progressão de k-mers, de 23 a 31, mesma variação utilizada com o programa IDBA-Trans. Ao final de cada montagem, foi realizada uma outra montagem SAMP-Oases e SAMP-IDBA, onde o montador CAP3 é utilizado para condensar os resultados das montagens anteriores. Destas montagens são liberadas pelo CAP3 arquivos .singlets e .contigs, que representam os transcritos não utilizados na montagem consenso e os contigs finais da montagem, respectivamente. A partir destes arquivos é gerado um arquivo .uniqs, que é a junção dos arquivos .singlets e .contigs.
Com a ideia de que transcritos montados em todas as montagens, independente da variação do k-mer, estarão representados pelo arquivo .contigs, foram feitas mais dois tipos de montagens: uma montagem utilizando somente o arquivo .contigs, chamada de Contigs-SAMP, e outra montagem utilizando o arquivo .uniqs, chamada de Uniqs-SAMP. Para os montadores Velvet-Oases e IDBA-Tran, foram geradas sete montagens (k-mer23, k-mer25, k-mer27, k-mer29, k-mer31, Contigs-SAMP e Uniqs-SAMP), sendo elas utilizadas para a escolha do melhor tipo de montagem para cada tipo de dado.

5.5.1 Análise por montador

Para cada montador, em cada caso de teste, foi gerado uma matriz 17x10 (17 critérios x 10 intervalos), na qual cada posição consta a melhor porcentagem de transcritos verdadeiramente diferencialmente expressos em cada intervalo i e critério j. Em seguida, geramos uma matriz que representa cada caso de teste, onde cada posição da matriz representa a melhor porcentagem obtida por um montador *de novo* dado tal critério naquele intervalo. Por exemplo, se utilizando o Critério 1 (Número total de *reads*) no intervalo [1,10] o montador Trinity encontrou 70% dos transcritos verdadeiramente diferencialmente expressos, o Velvet-Oases encontrou 60% e o IDBA-Tran encontrou 50%, nesta posição da matriz que representa o caso de teste constará a melhor porcentagem encontrada, neste caso a do Trinity de 70%, que será representado por 0.7.
A partir da matriz que representa cada caso de teste, foi gerado um heatmap onde cada posição representa a maior porcentagem de transcritos verdadeiramente diferncialmente expressos encontrados por um montador *de novo*. Na Figura 5.3 é representado o heatmap gerado a partir da matriz para o caso de teste Tr de *Arabidopsis thaliana*. A Figura 5.4, a Figura 5.5 e a Figura 5.6 representam o heatmap gerado a partir da matriz para os montadores Trinity, Velvet-Oases e IDBA-Tran, respectivamente. Abaixo estão representados os heatmaps do caso de teste Tr tanto para *Arabidopsis thaliana* como para *Canis familiaris*.

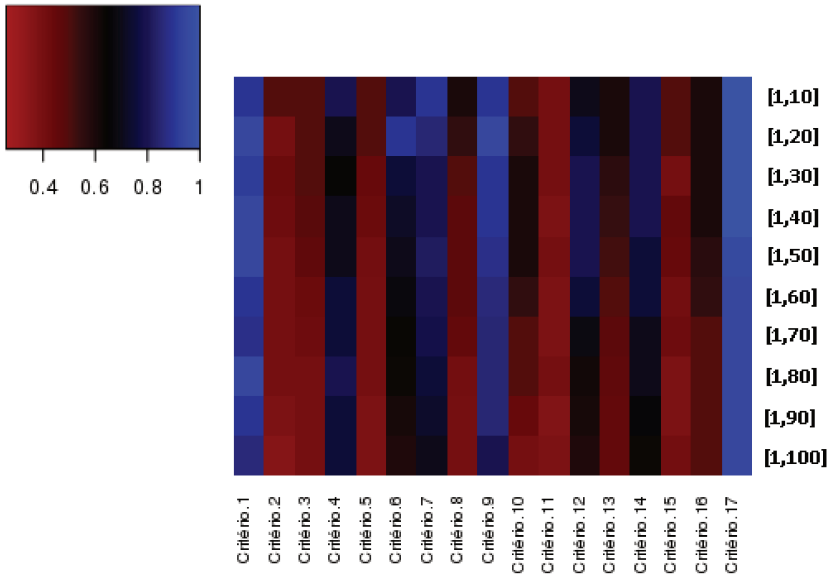

Figura 5.3: Heatmap representando o caso de teste Tr de *Arabidopsis thaliana*. Cada posição do heatmap é representando pela melhor porcentagem obtida por um montador *de novo* dado tal critério em tal intervalo.

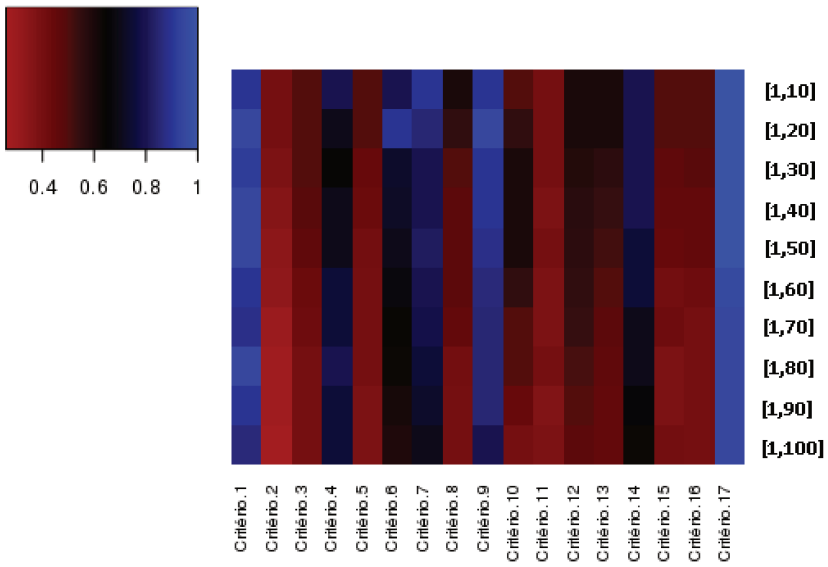

Figura 5.4: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Trinity no caso de teste Tr de *Arabidopsis thaliana*.

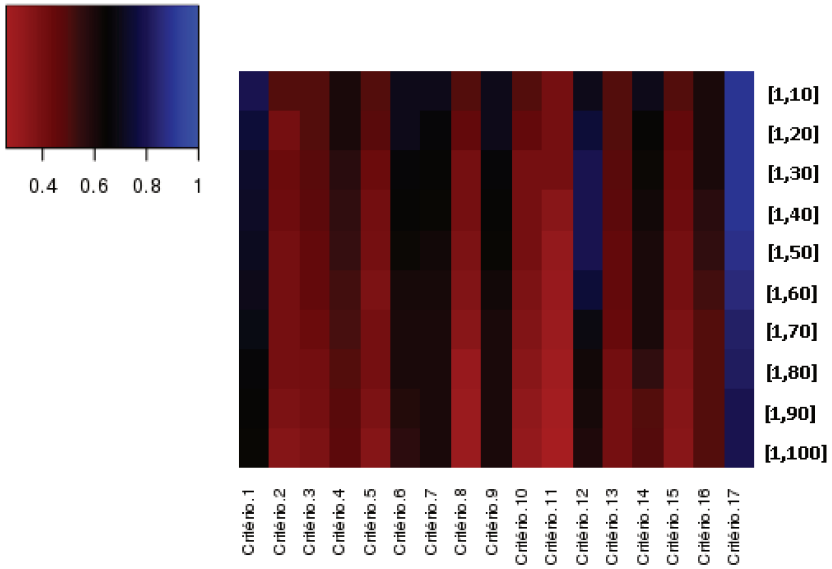

Figura 5.5: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Velvet-Oases no caso de teste Tr de *Arabidopsis thaliana*.

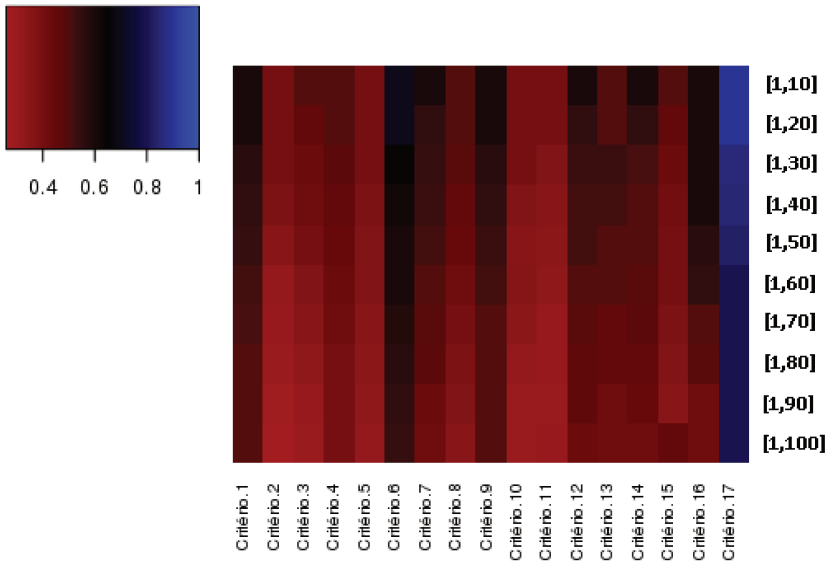

Figura 5.6: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* IDBA-Tran no caso de teste Tr de *Arabidopsis thaliana*.

A comparação dos resultados obtidos para o caso de teste Tr de *Arabidopsis thaliana* mostra que o Trinity é o montador *de novo* que encontra mais transcritos verdadeiaramente diferencialmente expressos. O heatmap gerado na Figura 5.3 tem mais posições geradas pelo heatmap do Trinity, como mostrado na Figura 5.4. Os outros heatmaps para cada caso de teste e cada montador *de novo* estão no Apêndice A.2. É possível ainda notar que o critério que gerou a maior porcentagem de transcritos diferencialmente expressos

em todos os casos de testes foi o Critério 17 (P-value/PDR e Módulo de log(FC)).

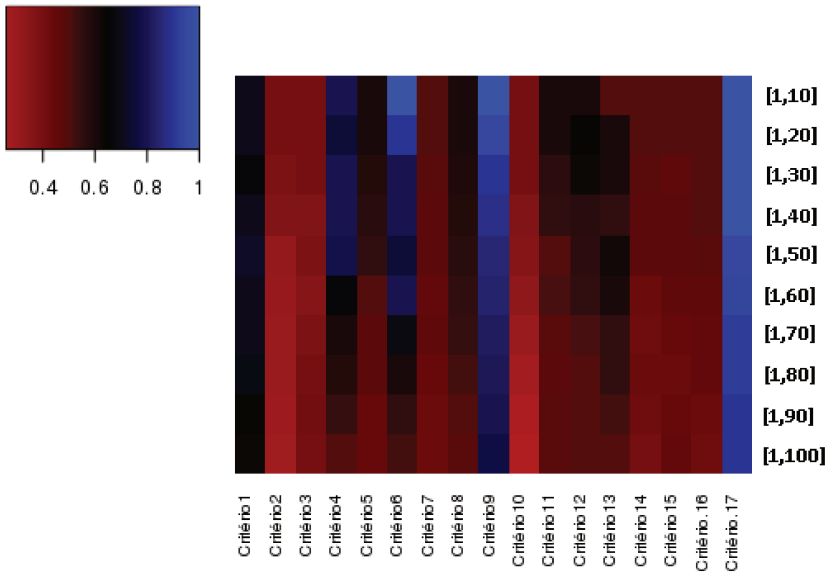

Figura 5.7: Heatmap representando o caso de teste Tr de *Canis familiaris*. Cada posição do heatmap é representando pela melhor porcentagem obtida por um montador *de novo* dado tal critério em tal intervalo.

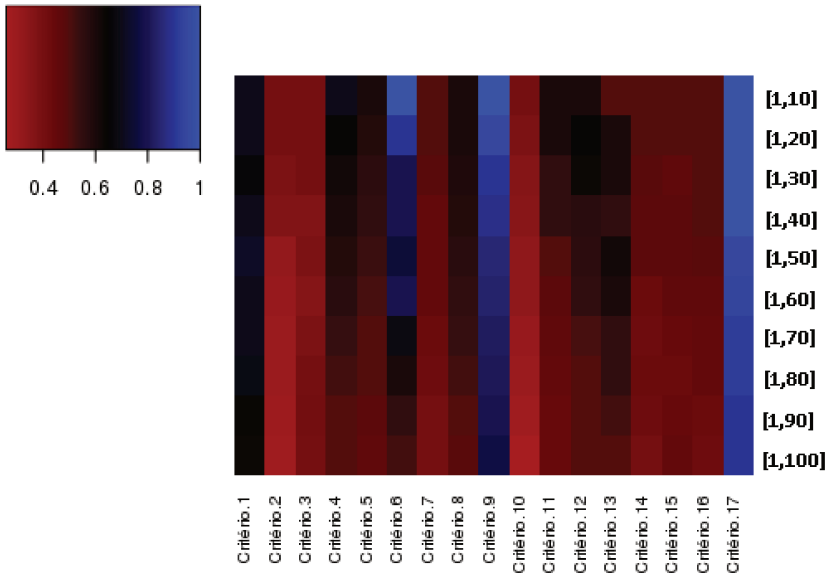

Figura 5.8: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Trinity no caso de teste Tr de *Canis familiaris*.

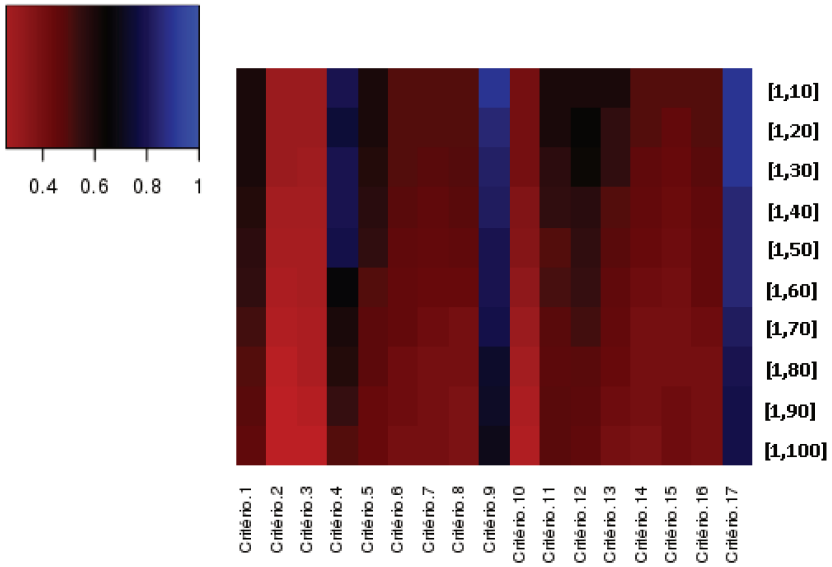

Figura 5.9: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Velvet-Oases no caso de teste Tr de *Canis familiaris*.

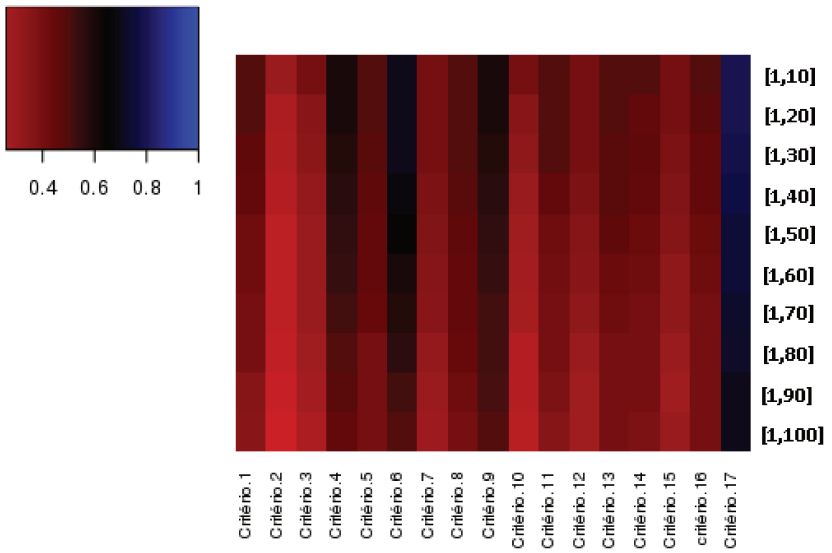

Figura 5.10: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* IDBA-Tran no caso de teste Tr de *Canis familiaris*.

Como observado no caso de teste Tr de *Arabidosis thaliana*, os resultados mostram que o Trinity é o montador *de novo* que encontra mais transcritos verdadeiaramente diferencialmente expressos para o caso de teste Tr de *Canis familiaris* e, o Critério 17 (P-value/PDR e Módulo do log(FC)) é o critério que gerou a maior porcentagem de acerto por intervalo. O critério 17 é composto por dois critérios que conseguem tirar dos dados informações importantes sobre o experimento e montagem. O critério P-value/FDR mostra a probabi-

lidade de determinado transcrito ser diferencialmente expresso, fato que é complementado pela grande diferencia de abundância entre as condições, mostrada pelo log(FC).
O critério 9 que também teve um acerto elevado acabou perdendo para o critério 17 devido ao fato de levar em consideração o número de *reads*. Por exemplo, em um caso onde o número de *reads* mapeados em ambas as condições são altos, mas não com uma significante diferença, podemos dizer que ele é expresso, mas não diferencialmente expresso. Então a seleção com o critério 9 o deixaria no topo da sua lista ordenada, já o critério 17 não.

5.5.2 Validação estatística

Para a validação dos critérios que se destacaram em relação aos outros, foi feito de teste pareado de Wilcoxon, com confiança de 95%. A partir das tabelas que geraram os *heatpamp*, fica evidente que o Critério 17 (P-value/PDR e Módulo de log(FC)) se sobresai em relação aos demais. Por exemplo, nas tabelas 5.5.2 e 5.5.2, sobre o Critério 17 foram realizados os testes de classificação de Wilcoxon e ele obteve significância sobre os outros critérios. Os demais casos de teste estão disponíveis nas tabelas no Apêndice A.3.

Tabela 5.15: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr de *Arabidopsis thaliana* para o montador Trinity. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	$C16$
[1,10]	1	0.90	0.40	0.50	0.80	0.50	0.80	0.90	0.60	0.90	0.50	0.40	0.60	0.60	0.80	0.50	0.5
[1,20]	1	0.95	0.40	0.50	0.70	0.50	0.90	0.85	0.55	0.95	0.55	0.40	0.60	0.60	0.80	0.50	0.5
[1,30]	1	0.92	0.38	0.50	0.65	0.44	0.73	0.80	0.50	0.90	0.60	0.40	0.58	0.56	0.80	0.46	0.48
[1,40]	1	0.95	0.36	0.48	0.70	0.43	0.72	0.80	0.47	0.90	0.60	0.38	0.57	0.54	0.80	0.45	0.45
[1,50]	1	0.95	0.34	0.46	0.70	0.41	0.70	0.82	0.47	0.88	0.60	0.40	0.56	0.52	0.75	0.44	0.45
[1,60]	0.96	0.90	0.33	0.43	0.75	0.40	0.67	0.80	0.47	0.86	0.55	0.38	0.55	0.50	0.75	0.41	0.42
[1,70]	0.95	0.88	0.30	0.42	0.75	0.40	0.64	0.78	0.45	0.85	0.50	0.38	0.54	0.47	0.70	0.42	0.40
[1,80]	0.95	0.95	0.28	0.40	0.80	0.40	0.63	0.75	0.41	0.85	0.50	0.40	0.51	0.46	0.70	0.38	0.40
[1,90]	0.95	0.90	0.28	0.40	0.75	0.38	0.61	0.73	0.40	0.85	0.44	0.37	0.50	0.45	0.66	0.38	0.40
[1,100]	0.95	0.86	0.27	0.40	0.75	0.38	0.59	0.70	0.40	0.80	0.40	0.38	0.47	0.45	0.63	0.41	0.40
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela 5.16: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr de *Canis familiaris* para o montador Trinity. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.70	0.40	0.40	0.70	0.60	1.00	0.50	0.60	1.00	0.40	0.60	0.60	0.50	0.50	0.50	0.50
[1,20]	1	0.70	0.40	0.40	0.65	0.58	0.90	0.50	0.60	0.95	0.38	0.60	0.65	0.60	0.50	0.50	0.50
[1,30]	1	0.66	0.38	0.40	0.62	0.56	0.80	0.48	0.59	0.90	0.36	0.55	0.63	0.60	0.48	0.46	0.50
[1,40]	1	0.70	0.37	0.37	0.60	0.55	0.80	0.45	0.58	0.88	0.35	0.55	0.57	0.55	0.47	0.47	0.50
[1,50]	0.95	0.72	0.32	0.38	0.58	0.53	0.75	0.45	0.57	0.85	0.33	0.50	0.56	0.62	0.47	0.47	0.48
[1,60]	0.94	0.70	0.31	0.36	0.57	0.51	0.80	0.45	0.55	0.84	0.33	0.47	0.55	0.60	0.43	0.46	0.46
[1,70]	0.92	0.70	0.30	0.38	0.54	0.50	0.68	0.43	0.54	0.82	0.31	0.46	0.51	0.55	0.42	0.44	0.45
[1,80]	0.92	0.69	0.30	0.40	0.52	0.50	0.60	0.42	0.52	0.81	0.30	0.45	0.50	0.55	0.43	0.43	0.45
[1,90]	0.90	0.64	0.29	0.41	0.50	0.47	0.55	0.40	0.50	0.80	0.28	0.44	0.50	0.52	0.42	0.44	0.43
[1,100]	0.90	0.63	0.28	0.40	0.50	0.46	0.52	0.40	0.48	0.77	0.26	0.44	0.50	0.50	0.40	0.45	0.42
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Ao analisarmos as tabelas acima e no Apêndice A.3 , podemos concluir que utilizando o montador *de novo* Trinity e aplicarmos a seleção pelo Critério 17 (P-value/PDR e Módulo de log(FC)), obtemos um maior número de transcritos verdadeiramente diferencialmente expressos.

Capítulo 6

Conclusões e Trabalhos futuros

Ao final deste trabalho podemos concluir que ainda há muitas dificuldades em analisar uma montagem *de novo* e em encontrar isoformas que sejam próximas da referência. Sobre a análise de montagens *de novo*, devido a alteração do volume de dados, a quantidade de verdadeiros positivos (transcritos verdadeiramente diferencialmente expressos) se altera. Os testes comparativos realizados neste trabalho, como a porcentagem de BBHs por montador e o valor W na busca de genes de cópias únicas (GCUs) em eucariotos, revela que o Trinity é o melhor montador *de novo*.

Este trabalho sugere uma métrica de validação que pode ser aplicada a reconstrução de transcriptomas sob determinadas condições, sendo ela a métrica de alinhamento contra os genes de cópias únicas (GCUs). Nesse caso o montador que melhor reconstruir genes conservados, mais convicto é o transcriptoma gerado.

Ao realizarmos testes inciais de escolha de transcritos diferencialmente expressos em uma análise de RNA-Seq, notamos que a porcentagem de acerto era em torno de 40%, utilizando o critério Fold-change. Com intuito de fazer uma análise de expressão diferencial maximizando a chance de acerto na escolha de um transcrito verdadeiramente diferencialmente expresso, testamos outros critérios se seleção, inclusive um já sugerido pela literatura, e o critério que se destacou foi o de ordenação por P-value/FDR e Módulo do log(FC). O montador que mais se aproximou de encontrar transcritos verdadeiramente diferencialmente expressos sem gerar muitos falsos positivos foi o Trinity.

Para trabalhos futuros, pode se determinar novos critérios de seleção para a análise de maximização de transcritos verdadeiros e, novos casos de testes para a análise do melhor montador *de novo*, podendo ainda analisar a sua sensibilidade em relação ao volume de dados para novos casos de teste.

Referências Bibliográficas

[1] Bruce Alberts. *Biologia Molecular da Célula*. Editora Artmed, 2010.

[2] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[3] S. Anders, D.J. McCarthy, Y. Chen, M. Okoniewski, G.K. Smyth, W. Huber, and Robinson M.D. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, 8:1765–1786, 2013.

[4] Simon Andrews. FastQC: A Quality Control tool for High Throughput Sequence Data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. Acessado: 10-10-2014.

[5] Marcelo Falsarella Carazzolle, Lucas Miguel de Carvalho, Hugo Henrique Slepicka, Ramon Oliveira Vidal, Gonçalo Amarante Guimarães Pereira, Jörg Kobarg, and Gabriela Vaz Meirelles. Iis – integrated interactome system: A web-based platform for the annotation, analysis and visualization of protein-metabolite-gene-drug interactions by integrating a variety of data sources and tools. *PLoS ONE*, 9(6):e100385, 06 2014.

[6] X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4:210, 2003.

[7] B. Ewing, L. Hillier, MC. Wendl, and Green P. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Research*, 3:175–185, 1998.

[8] M.G Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, and Q. Zeng. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29:644–52, 2011.

[9] B. Hass, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, et al. De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8:1494–1512, 2013.

[10] B. He, S. Zhao, Y. Chen, Q. Cao, C. Wei, X. Cheng, and Y. Zhang. Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. *BMC Genomics*, 16:65, 2015.

[11] X. Huang and A. Madan. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9:868–877, 1999.

[12] T. James, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29:24–26, 2011.

[13] B. Jordan. Historical background and anticipated developments. *Annals of the New York Academy of Sciences*, 975:24–32, 2002.

[14] B Langmead and SL. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9:357–359, 2012.

[15] B Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25, 2009.

[16] B Li and C.N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, 2011.

[17] B. Li, M. Fiomore, Y. Bai, M. Collins, J. A. Thomson, R. Stewart, and C. N. Dewey. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, 15:553, 2014.

[18] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–60, 2009.

[19] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25:2078–9, 2009.

[20] Y. Lin, J. Li, H. Shen, L. Zhang, C.J. Papasian, and H.W. Deng. Comparative studies of <italic>de novo</italic> assembly tools for next-generation sequencing technologies s. *Bioinformatics*, 27(15):2031–2037, 2011.

[21] Robert Lindner and Caroline C. Friedel. A comprehensive evaluation of alignment algorithms in the context of RNA-Seq. *PLoS ONE*, 7(12):e52403, 2012.

[22] B. Lu, Z. Zeng, and T. Shi. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on rna-Seq. *Science China Life Sciences*, 56(2):143–155, 2013.

[23] Prem S. Mann. *Introdução à Estatística*. LTC, 2006.

[24] NCBI. An introduction to genetic analysis. `http://www.ncbi.nlm.nih.gov/books/NBK22071/`. Acessado: 01-07-2014.

[25] I. Nookaew, M. Papini, N. Pornputtapong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research*, 40:1–14, 2012.

[26] K. Okubo, N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. Large scale cdna sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genetics*, 2:173–179, 1992.

[27] S.T. O'Neil and S. J. Emrich. Assessing *de novo* transcriptome assembly metrics for consistency and utility. *BMC Genomics*, 14:465, 2013.

[28] R. Overbeek, M. Fonstein, G. D. D'Souza, M. Puch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, page 2896–2901, 1999.

[29] Y. Peng, H.C.M. Leung, S-M. Yiu, M-J. Lv, X-G. Zhu, and F.Y.L. Chin. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, 29:i326–34, 2013.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[31] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, CE. Mason, ND.

Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-Seq data. *Genome Biology*, 14:R95, 2013.

[32] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–40, 2010.

[33] JA. Robles, SE. Qureshi, SJ. Stephen, SR. Wilson, CJ. Burden, and JM. Taylor. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *Genome Biology*, 13:484, 2012.

[34] Shen S. and Tuszynski J.A. *Theory and Mathematical Methods for Bioformatics.* Springer, 1st edition, 2008.

[35] S Schbath et al. Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. *Journal of Computational Biology*, 19:796–813, 2012.

[36] S. Schbath, V. Martin, M. Zytnicki, J. Fayolle, V. Loux, and J-F. Gibrat. Mapping reads on a genomic sequence: An algorithmic overview and a practical comparative analysis. *Journal of Computational Biology*, 19:796–813, 2012.

[37] S. Schbath, V. Martin, M. Zytnicki, J. Fayolle, V. Loux, and J.F. Gibrat. Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. *Journal of Computational Biology*, 19:796–813, 2012.

[38] M.H. Schulz, D.R. Zerbino, M. Vingron, and E. Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28:1086–92, 2012.

[39] A.F. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 2015. published online.

[40] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[41] C Trapnell, L Pachter, and SL. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25:1105–11, 2009.

[42] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Reviews Genetics*, 10:57–63, 2009.

[43] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 2009.

[44] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

[45] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen. A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLoS ONE*, 6(3):e17915, 2011.

[46] Z.H Zhang. A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. *PLoS ONE*, 9:e103207, 2014.

Apêndice A

Resultados obtidos

A.1 Tabelas da análise dos GCUs

Os GCU contido nas tabelas abaixo foram extraídos do banco de dados do BUSCO [1]. O número de cada GCU foi mantido, para manter a relação com o banco do BUSCO.

A.1.1 *Arabidopsis thaliana*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
10084	X	X	X	X	X	X	X	X	X
101066	X	X	X	X	X	X	X	X	X
101455	X	X	X	X	X	X	X	X	X
101494	X	X	X	X	X	X	X	X	X
101792									
102027	X	X	X	X	X	X	X	X	X
102374									
10237	X	X	X	X	X	X	X	X	X
102673	X	X	X	X	X	X	X	X	X
103500	X	X	X						
103567	X	X	X	X	X	X	X	X	X
103854	X	X	X	X	X	X	X	X	X
104000									
104072	X	X	X	X	X	X	X	X	X
104490	X	X	X	X	X	X	X	X	X
105787	X	X	X	X	X	X	X		
10656	X	X	X	X			X		
10666	X	X	X	X	X	X			
107021	X	X	X	X	X	X	X	X	X

Continua na próxima página

[1] `http://busco.ezlab.org/`

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
107122	X	X	X	X	X	X	X	X	X
10724	X	X	X	X	X	X	X	X	X
107984	X	X	X	X	X	X	X	X	X
107993									
108031	X	X	X	X	X	X	X	X	X
108196	X	X	X	X	X	X	X	X	X
108225	X	X	X	X	X	X	X	X	X
108277	X	X	X	X	X	X	X	X	X
108369	X	X	X	X	X	X		X	X
108604	X	X	X	X	X	X	X	X	X
108655	X	X	X	X	X	X	X	X	X
108683	X	X	X	X	X	X	X	X	X
109520	X	X	X	X	X	X	X	X	X
110661	X	X	X	X	X	X			
110988	X	X	X	X	X	X	X	X	X
111118	X	X	X	X	X	X	X	X	X
111440	X	X	X	X	X	X	X	X	X
11190									
111911	X	X	X	X	X	X	X	X	X
111967	X	X	X	X	X	X	X	X	X
112611	X	X	X	X	X	X			
11390	X	X	X	X	X	X			
114281	X	X	X	X	X	X	X	X	X
11533	X	X	X	X	X	X	X	X	X
115762	X	X	X	X	X	X	X	X	X
115797	X	X	X	X	X	X	X		X
116058	X	X	X	X	X	X	X	X	X
118570									
119017	X	X	X	X	X	X	X	X	X
119372	X	X	X	X	X	X	X	X	X
120109	X	X	X				X	X	X
12171									
12273	X	X	X	X	X	X	X	X	X
12446	X	X	X	X	X	X	X	X	X
12503									
12632	X	X	X	X	X	X	X	X	X
12776	X	X	X	X	X	X	X	X	X
12786	X	X	X	X	X	X	X	X	X
12890									

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
13177	X	X	X	X	X	X	X	X	X
13653	X	X	X	X	X	X	X	X	X
13721	X	X	X	X	X	X	X	X	X
14004	X	X	X	X	X	X	X	X	X
14025	X	X	X	X	X	X	X	X	X
14187	X	X	X	X	X	X	X	X	X
14507	X	X	X	X	X	X	X	X	X
14782	X	X	X	X	X	X	X	X	X
15263	X	X	X	X	X	X	X	X	X
15336	X	X	X	X	X	X	X	X	X
15489	X	X	X	X	X	X	X	X	X
15858	X	X	X	X	X	X	X	X	X
16098	X	X	X	X	X	X	X	X	X
16404	X	X	X	X	X	X	X	X	X
16677	X	X	X	X	X	X	X	X	X
16703	X	X	X	X	X	X	X	X	X
17219	X	X	X	X	X	X	X	X	X
17368	X	X	X	X	X	X	X	X	X
18012	X	X	X	X	X	X	X	X	X
18235	X	X	X	X	X	X	X	X	X
18676	X	X	X	X	X	X	X	X	X
18841	X	X	X	X	X	X	X	X	X
18979	X	X	X	X					
19189	X	X	X	X	X	X	X	X	X
19339	X	X	X	X	X	X	X	X	X
19380	X	X	X	X	X	X	X	X	X
19530	X	X	X	X	X	X	X	X	X
19798	X	X	X	X	X	X	X	X	X
1997									
20005	X	X	X	X	X	X	X	X	X
20110	X	X	X	X	X	X	X	X	X
20575	X	X	X	X	X	X	X	X	X
20636	X	X	X	X	X	X	X	X	X
20889	X	X	X	X	X	X	X	X	X
20937	X	X	X	X	X	X	X	X	X
20938	X	X	X	X	X	X	X	X	X
21113	X	X	X	X	X	X	X	X	X
21220	X	X	X	X	X	X	X	X	X
21580	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
21650	X	X	X						
22064									
22571	X	X	X	X	X	X	X	X	X
22825	X	X	X	X	X	X	X	X	X
23109	X	X	X	X	X	X	X	X	X
23168	X	X	X	X	X	X	X	X	X
23244	X	X	X	X	X	X	X	X	X
23270	X	X	X	X	X	X	X	X	X
23533	X	X	X	X	X	X	X	X	X
23548	X	X	X	X	X	X	X	X	X
23551	X	X	X	X	X	X	X	X	X
23603	X	X	X	X	X	X	X	X	X
23681	X	X	X						
23743	X	X	X	X	X	X	X	X	X
24411	X	X	X	X	X	X	X	X	X
24617	X	X	X	X	X	X	X	X	X
24645	X	X	X	X	X	X	X	X	X
24712	X	X	X	X	X	X	X	X	X
25367	X	X	X	X	X	X	X	X	X
25643									
2567	X	X	X	X	X	X	X	X	X
25901	X	X	X	X	X	X	X	X	X
26000	X	X	X	X	X	X	X	X	X
26021	X	X	X	X	X	X			
26044	X	X	X	X	X	X	X	X	X
26453	X	X	X	X	X	X	X	X	X
26551	X	X	X	X	X	X	X	X	X
26641	X	X	X	X	X	X	X	X	X
26657	X	X	X	X	X	X	X	X	X
26905	X	X	X	X	X	X	X	X	X
27773	X	X	X	X	X	X	X	X	X
27929	X	X	X	X	X	X	X	X	X
27931									
28124	X	X	X	X	X	X	X	X	X
28189	X	X	X	X	X	X	X	X	X
28258	X	X	X	X	X	X	X	X	X
28286	X	X	X	X	X	X	X	X	X
28385	X	X	X	X	X	X	X	X	X
28421	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
28665	X	X	X	X	X	X	X	X	X
28859	X	X	X	X	X	X	X	X	X
30346	X	X	X	X	X	X	X	X	X
30694	X	X	X	X	X	X	X	X	X
31019									
31103	X	X	X	X	X	X	X	X	X
31291	X	X	X	X	X	X	X	X	X
31322	X	X	X	X	X	X	X	X	X
31431	X	X	X	X	X	X	X	X	X
3146									
31476	X	X	X	X	X	X	X	X	X
31511	X	X	X	X	X	X	X	X	X
31859	X	X	X	X	X	X	X	X	X
32000	X	X	X	X	X	X	X	X	X
32086	X	X	X	X	X	X	X	X	X
32598	X	X	X	X	X	X	X	X	X
32812	X	X	X	X	X	X	X	X	X
33317	X	X	X	X	X	X	X	X	X
33643	X	X	X	X	X	X	X	X	X
33782	X	X	X	X	X	X	X	X	X
33798	X	X	X	X	X	X	X	X	X
34104	X	X	X	X	X	X	X	X	X
34106	X	X	X	X			X		
34335	X	X	X	X	X	X	X	X	X
35347	X	X	X	X	X	X	X	X	X
35370	X	X	X	X	X	X	X	X	X
35676	X	X	X	X	X	X	X	X	X
35796	X	X	X	X	X	X	X	X	X
36153	X	X	X	X	X	X	X	X	X
36329	X	X	X	X	X	X	X	X	X
36399	X	X	X	X	X	X		X	X
36409	X	X	X	X	X	X	X	X	X
36494	X	X	X	X	X	X	X	X	X
36637									
3664	X	X	X	X	X	X	X	X	X
3720	X	X	X	X	X	X	X	X	X
37480	X	X	X	X	X	X	X	X	X
37784	X	X	X	X	X	X	X	X	X
37793	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
38002									
38372	X	X	X	X	X	X	X	X	X
38531	X	X	X	X	X	X	X	X	X
38726	X	X	X	X	X	X	X	X	X
38874	X	X	X	X	X	X	X	X	X
38877	X	X	X	X	X	X	X	X	X
39240	X	X	X	X	X	X	X	X	X
39448	X	X	X	X	X	X	X	X	X
39556	X	X	X	X	X	X	X	X	X
40266	X	X	X	X	X	X	X	X	X
40337	X	X	X	X	X	X	X	X	X
40482	X	X	X	X	X	X			
40771	X	X	X	X	X	X	X	X	X
40793	X	X	X	X	X	X	X	X	X
40853	X	X	X	X	X	X	X	X	X
41083	X	X	X	X	X	X	X	X	X
41165									
41324	X	X	X	X	X	X	X	X	X
41414	X	X	X	X	X	X	X	X	X
41555	X	X	X	X	X	X	X	X	X
41596	X	X	X	X	X	X	X	X	X
41608	X	X	X	X	X	X	X	X	X
41638	X	X	X						
4174	X	X	X	X	X	X	X	X	X
41859	X	X	X	X	X	X	X	X	X
42401	X	X	X	X	X	X	X	X	X
42408	X	X	X	X	X	X	X	X	X
42619	X	X	X	X	X	X	X	X	X
42968	X	X	X	X	X	X	X	X	X
43132	X	X	X	X	X	X	X	X	X
43265	X	X	X	X	X	X	X	X	X
4338	X	X	X	X	X	X	X	X	X
43393	X	X	X	X	X	X	X	X	X
43652									
43752	X	X	X	X	X	X	X	X	X
43976	X	X	X	X	X	X	X	X	X
44022	X	X	X	X	X	X	X	X	X
44116									
44633	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
44714	X	X	X	X	X	X	X	X	X
44787	X	X	X	X	X	X	X	X	X
44961	X	X	X	X	X	X	X	X	X
45012									
45275	X	X	X	X	X	X	X	X	X
46033	X	X	X	X	X	X	X	X	X
46157	X	X	X	X	X	X	X	X	X
46335	X	X	X	X	X	X	X	X	X
46376	X	X	X	X	X	X	X	X	X
46423									
46740	X	X	X	X	X	X	X	X	X
46794	X	X	X	X	X	X	X	X	X
47192	X	X	X	X	X	X	X	X	X
47983	X	X	X	X	X	X	X	X	X
48422	X	X	X	X	X	X	X	X	X
484	X	X	X	X	X	X	X	X	X
48783	X	X	X	X	X	X	X	X	X
49963	X	X	X	X	X	X	X	X	X
50215									
50336	X	X	X	X	X	X	X	X	X
50408	X	X	X	X	X	X	X	X	X
5074	X	X	X	X	X	X	X	X	X
51307	X	X	X	X	X	X	X	X	X
51347	X	X	X	X	X	X	X	X	X
5171	X	X	X	X	X	X	X	X	X
51857	X	X	X	X	X	X	X	X	X
51874	X	X	X	X	X	X	X	X	X
52007	X	X	X	X	X	X	X	X	X
52088	X	X	X	X	X	X	X	X	X
52229	X	X	X	X	X	X	X	X	X
52278	X	X	X	X	X	X	X	X	X
52639	X	X	X	X	X	X	X	X	X
52725	X	X	X	X	X	X	X	X	X
52909	X	X	X	X	X	X	X	X	X
5300	X	X	X	X	X	X	X	X	X
53121	X	X	X	X	X	X	X	X	X
53341									
53371	X	X	X	X	X	X	X	X	X
53899	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
5390	X	X	X	X	X	X			
53973	X	X	X	X	X	X	X	X	X
54689	X	X	X	X	X	X	X	X	X
54848	X	X	X	X	X	X	X	X	X
55084	X	X	X	X	X	X	X	X	X
55388	X	X	X	X	X	X	X	X	X
55553									
5560	X	X	X	X	X	X	X	X	X
55947	X	X	X	X	X	X	X	X	X
55970	X	X	X	X	X	X	X	X	X
56438	X	X	X	X	X	X	X	X	X
57208	X	X	X	X	X	X			
57491	X	X	X	X	X	X	X	X	X
57900	X	X	X	X	X	X			
58153	X	X	X	X	X	X	X	X	X
58169	X	X	X	X	X	X	X	X	X
58444	X	X	X	X	X	X	X	X	X
58648	X	X	X	X	X	X	X	X	X
58966	X	X	X	X	X	X	X	X	X
59214	X	X	X	X	X	X	X	X	X
59321	X	X	X	X	X	X	X	X	X
5992	X	X	X	X	X	X	X	X	X
60021	X	X	X	X	X	X	X	X	X
60479	X	X	X	X	X	X	X	X	X
60589	X	X	X	X	X	X	X	X	X
610	X	X	X	X	X	X	X	X	X
61341	X	X	X	X	X	X	X	X	X
61383	X	X	X	X	X	X	X	X	X
62016	X	X	X	X	X	X	X	X	X
62559	X	X	X	X	X	X	X	X	X
62598	X	X	X	X	X	X	X	X	X
62604	X	X	X	X	X	X	X	X	X
63015	X	X	X	X	X	X			
63486	X	X	X	X	X	X	X	X	X
63582	X	X	X	X	X	X	X	X	X
63670	X	X	X	X	X	X	X	X	X
6367	X	X	X	X	X	X	X	X	X
63890	X	X	X	X	X	X	X	X	X
64129	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
64334	X	X	X	X	X	X	X	X	X
64505	X	X	X	X	X	X	X	X	X
64602	X	X	X	X	X	X	X	X	X
64707	X	X	X	X	X	X	X	X	X
65239	X	X	X	X	X	X	X	X	X
65290	X	X	X	X	X	X	X	X	X
65570	X	X	X	X	X	X	X	X	X
66609	X	X	X				X	X	X
66860	X	X	X	X	X	X	X	X	X
67316	X	X	X	X	X	X	X	X	X
67848	X	X	X	X	X	X	X	X	X
67940	X	X	X	X	X	X			
67997	X	X	X	X	X	X	X	X	X
68029	X	X	X	X	X	X	X	X	X
68108	X	X	X	X	X	X	X	X	X
68322	X	X	X	X	X	X	X	X	X
68638	X	X	X	X	X	X	X	X	X
6863	X	X	X	X	X	X	X	X	X
6877	X	X	X	X	X	X	X	X	X
69365	X	X	X	X	X	X	X	X	X
70306	X	X	X	X	X	X	X	X	X
70662	X	X	X	X	X	X	X	X	X
7080	X	X	X	X	X	X	X	X	X
71025	X	X	X	X	X	X	X	X	X
71165	X	X	X	X	X	X	X	X	X
7127	X	X	X	X	X	X	X	X	X
71430	X	X	X	X	X	X	X	X	X
71582	X	X	X				X	X	X
7192	X	X	X	X	X	X	X	X	X
71936	X	X	X	X	X	X	X	X	X
72009	X	X	X	X	X	X	X	X	X
72039									
72882	X	X	X	X	X	X	X	X	X
73019	X	X	X	X	X	X	X	X	X
73146									
73202	X	X	X	X	X	X	X	X	X
73262	X	X	X	X	X	X	X	X	X
73336	X	X	X	X	X	X			
7363	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
74245	X	X	X	X	X	X	X	X	X
75497	X	X	X	X	X	X	X	X	X
75810	X	X	X	X	X	X	X	X	X
76117									
76164	X	X	X	X	X	X	X	X	X
76234									
76375	X	X	X	X	X	X	X	X	X
76469	X	X	X						
76532	X	X	X	X	X	X	X	X	X
76601	X	X	X	X	X	X	X	X	X
77026	X	X	X	X	X	X	X	X	X
77356									
77491	X	X	X	X	X	X	X	X	X
77538	X	X	X	X	X	X	X	X	X
77568	X	X	X	X	X	X	X	X	X
77808	X	X	X	X	X	X	X	X	X
77841	X	X	X	X	X	X			
77991	X	X	X	X	X	X	X	X	X
78028	X	X	X	X	X	X	X	X	X
78139	X	X	X	X	X	X	X	X	X
78293	X	X	X				X	X	X
79105	X	X	X				X	X	X
79481	X	X	X	X	X	X	X	X	X
79596	X	X	X	X	X	X		X	X
7959	X	X	X	X	X	X	X	X	X
79949	X	X	X	X	X	X	X	X	X
7999									
80337	X	X	X	X	X	X	X	X	X
80490	X	X	X	X	X	X	X	X	X
8059	X	X	X				X	X	X
80607									
80665	X	X	X	X	X	X	X	X	X
81405	X	X	X	X	X	X	X	X	X
81496	X	X	X	X	X	X	X	X	X
81841	X	X	X	X	X	X	X	X	X
81876	X	X	X	X	X	X	X	X	X
82126	X	X	X	X	X	X	X	X	X
82148									
82322	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
82369	X	X	X	X	X	X	X	X	X
82570	X	X	X	X	X	X	X	X	X
82580	X	X	X	X	X	X	X	X	X
82694	X	X	X	X	X	X	X	X	X
8413	X	X	X	X	X	X	X	X	X
84185	X	X	X	X	X	X	X	X	X
84231	X	X	X	X	X	X	X	X	X
84276	X	X	X	X	X	X	X	X	X
84291	X	X	X	X	X	X	X	X	X
84363									
84699	X	X	X				X	X	X
84840	X	X	X	X	X	X			
84872	X	X	X	X	X	X	X	X	X
8568	X	X	X	X	X	X	X	X	X
85806	X	X	X	X	X	X	X	X	X
86467	X	X	X	X	X	X	X	X	X
87189	X	X	X	X	X	X	X	X	X
87390	X	X	X	X	X	X	X	X	X
87581	X	X	X	X	X	X	X	X	X
87828	X	X	X	X	X	X	X	X	X
88190									
88303	X	X	X	X	X	X	X	X	X
88540	X	X	X	X	X	X	X	X	X
88572	X	X	X	X	X	X			
88842	X	X	X	X	X	X	X	X	X
8904									
89090	X	X	X	X	X	X	X	X	X
89116	X	X	X	X	X	X	X	X	X
89715	X	X	X	X	X	X	X	X	X
89858	X	X	X	X	X	X	X	X	X
89947	X	X	X	X	X	X	X	X	X
90721	X	X	X				X	X	X
91487	X	X	X	X	X	X	X	X	X
93187	X	X	X	X	X	X	X	X	X
93191	X	X	X	X	X	X	X	X	X
93494	X	X	X	X	X	X	X	X	X
93624	X	X	X	X	X	X	X	X	X
93746									
94134	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.1 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
94766	X	X	X	X	X	X	X	X	X
94871	X	X	X	X	X	X	X	X	X
94998	X	X	X	X	X	X	X	X	X
9563	X	X	X	X	X	X	X	X	X
9609	X	X	X	X	X	X	X	X	X
96385	X	X	X	X	X	X	X	X	X
96430	X	X	X	X	X	X	X	X	X
9651	X	X	X	X	X	X	X	X	X
96703	X	X	X	X	X	X	X	X	X
96830	X	X	X	X	X	X	X	X	X
98066	X	X	X	X	X	X	X	X	X
98097	X	X	X	X	X	X			
98103	X	X	X						
98198	X	X	X	X	X	X	X	X	X
98672	X	X	X	X	X	X	X	X	X
98854	X	X	X	X	X	X	X	X	X
98871	X	X	X	X	X	X	X	X	X
99355	X	X	X	X	X	X			
99389	X	X	X	X	X	X	X	X	X
99736									

Tabela A.1: Tabela de GCUs para o organismo *Arabidopsis thaliana* por cada caso de teste. O símbolo X representa que certos GCUs estão representados no caso de teste.

A.1.2 *Canis familiaris*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
10084	X	X	X	X	X	X	X	X	X
101066	X	X	X	X	X	X	X	X	X
101455	X	X	X	X					
101494	X	X	X	X	X	X			
101792									
102027	X	X	X	X	X	X	X	X	X
102374									
10237	X	X	X	X	X	X	X	X	X
102673	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
103500	X	X	X	X	X	X			
103567	X	X	X	X	X	X	X	X	X
103854	X	X	X	X	X	X	X	X	X
104000									
104072	X	X	X	X	X	X	X	X	X
104490	X	X	X	X	X	X	X	X	X
105787	X	X	X	X	X		X		
10656	X	X	X	X	X	X	X	X	X
10666	X	X	X	X	X	X	X	X	X
107021	X	X	X	X	X	X	X	X	X
107122	X	X	X	X	X	X	X	X	X
10724	X	X	X						
107984	X	X	X	X	X	X	X	X	X
107993	X	X	X	X	X	X			
108031	X	X	X	X	X	X	X	X	X
108196	X	X	X	X	X	X	X	X	X
108225	X	X	X	X	X				
108277									
108369	X	X	X	X	X	X			
108604	X	X	X	X	X	X	X	X	X
108655	X	X	X	X	X	X			
108683	X	X	X						
109520	X	X	X	X	X	X	X	X	X
110661	X	X	X	X					
110988	X	X	X	X	X	X	X	X	X
111118									
111440	X	X	X	X	X	X	X	X	X
11190									
111911	X	X	X	X	X	X	X	X	X
111967	X	X	X	X	X	X	X	X	X
112611	X	X	X	X	X	X	X	X	X
11390	X	X	X	X	X	X			
114281	X	X	X	X	X	X	X	X	X
11533	X	X	X	X	X	X	X	X	X
115762	X	X	X	X	X	X	X	X	X
115797	X	X	X	X	X	X			
116058	X	X	X	X	X	X	X	X	
118570									
119017	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
119372	X	X	X	X	X	X	X	X	X
120109	X	X	X	X	X	X			
12171	X	X	X	X	X	X	X	X	X
12273	X	X	X	X	X	X	X	X	X
12446	X	X	X	X	X	X	X	X	X
12503	X	X	X	X	X	X	X	X	X
12632	X	X	X	X	X	X	X	X	X
12776	X	X	X	X	X	X	X	X	X
12786	X	X	X	X	X	X	X	X	X
12890									
13177	X	X	X	X	X	X	X	X	X
13653	X	X	X	X	X	X	X	X	X
13721	X	X	X	X	X	X	X	X	X
14004	X	X	X	X	X	X	X	X	X
14025	X	X	X	X	X	X	X	X	X
14187	X	X	X	X			X	X	X
14507	X	X	X	X	X	X	X	X	X
14782	X	X	X	X	X	X	X	X	X
15263	X	X	X	X	X	X	X	X	X
15336	X	X	X	X	X	X	X	X	X
15489	X	X	X	X	X	X	X	X	X
15858	X	X	X	X	X	X	X	X	X
16098	X	X	X	X	X	X	X	X	X
16404	X	X	X	X	X	X	X	X	X
16677	X	X	X	X	X	X	X	X	X
16703	X	X	X	X	X	X	X	X	X
17219	X	X	X	X	X	X	X	X	X
17368	X	X	X	X	X	X	X	X	X
18012	X	X	X	X	X	X	X	X	X
18235	X	X	X	X	X	X	X	X	X
18676	X	X	X	X	X	X	X	X	X
18841	X	X	X	X			X	X	X
18979									
19189	X	X	X	X	X	X	X	X	X
19339	X	X	X	X	X	X	X	X	X
19380	X	X	X	X	X	X	X	X	X
19530	X	X	X	X	X	X	X	X	X
19798	X	X	X	X	X	X	X	X	X
1997									

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
20005	X	X	X	X	X	X	X	X	X
20110	X	X	X	X	X	X	X	X	X
20575	X	X	X	X	X	X	X	X	X
20636	X	X	X	X	X	X	X	X	X
20889	X	X	X	X	X	X	X	X	X
20937	X	X	X	X	X	X	X	X	X
20938	X	X	X	X	X	X	X	X	X
21113	X	X	X	X	X	X	X	X	X
21220	X	X	X	X	X	X	X	X	X
21580	X	X	X	X	X	X	X	X	X
21650	X	X	X	X	X	X	X	X	X
22064				X	X	X	X	X	X
22571	X	X	X	X	X	X	X	X	X
22825	X	X	X	X	X	X	X	X	X
23109	X	X	X	X	X	X	X	X	X
23168	X	X	X	X	X	X	X	X	X
23244	X	X	X	X	X	X	X	X	X
23270	X	X	X	X	X	X	X	X	X
23533	X	X	X	X	X	X	X	X	X
23548	X	X	X	X	X	X	X	X	X
23551	X	X	X	X	X	X	X	X	X
23603	X	X	X	X	X	X	X	X	X
23681	X	X	X	X	X	X	X	X	X
23743	X	X	X	X	X	X	X	X	X
24411	X	X	X	X	X	X	X	X	X
24617	X	X	X	X	X	X	X	X	X
24645	X	X	X	X	X	X	X	X	X
24712	X	X	X	X	X	X	X	X	X
25367	X	X	X	X	X	X	X	X	X
25643									
2567	X	X	X	X	X	X	X	X	X
25901	X	X	X	X	X	X	X	X	X
26000	X	X	X	X	X	X	X	X	X
26021	X	X	X	X	X	X	X	X	X
26044	X	X	X	X	X	X	X	X	X
26453	X	X	X	X	X	X			
26551	X	X	X	X	X	X	X	X	X
26641	X	X	X	X	X	X	X	X	X
26657	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
26905	X	X	X	X	X	X	X	X	X
27773	X	X	X	X	X	X	X	X	X
27929	X	X	X	X	X	X	X	X	X
27931	X	X	X	X	X	X	X	X	X
28124	X	X	X	X	X	X	X	X	X
28189	X	X	X	X	X	X	X	X	X
28258	X	X	X	X	X	X	X	X	X
28286	X	X	X	X	X	X	X	X	X
28385	X	X	X	X	X	X	X	X	X
28421	X	X	X	X	X	X	X	X	X
28665	X	X	X	X	X	X	X	X	X
28859									
30346	X	X	X	X	X	X	X	X	X
30694	X	X	X	X	X	X	X	X	X
31019	X	X	X	X	X	X			
31103	X	X	X	X	X	X	X	X	X
31291	X	X	X	X	X	X	X	X	X
31322	X	X	X	X	X	X	X	X	X
31431	X	X	X	X	X	X	X	X	X
3146	X	X	X	X	X	X	X	X	X
31476	X	X	X	X	X	X	X	X	X
31511	X	X	X	X	X	X	X	X	X
31859	X	X	X	X	X	X	X	X	X
32000	X	X	X	X	X	X	X	X	X
32086	X	X	X	X	X	X	X	X	X
32598	X	X	X	X	X	X	X	X	X
32812	X	X	X	X	X	X	X	X	X
33317	X	X	X	X	X	X	X	X	X
33643	X	X	X	X	X	X	X	X	X
33782	X	X	X	X	X	X	X	X	X
33798	X	X	X	X	X	X	X	X	X
34104	X	X	X	X	X	X	X	X	X
34106	X	X	X	X	X	X	X	X	X
34335	X	X	X	X	X	X	X	X	X
35347	X	X	X	X	X	X	X	X	X
35370	X	X	X	X	X	X	X	X	X
35676	X	X	X	X	X	X	X	X	X
35796	X	X	X	X	X	X	X	X	X
36153	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
36329	X	X	X	X	X	X	X	X	X
36399	X	X	X	X	X	X	X	X	X
36409	X	X	X	X	X	X	X	X	X
36494	X	X	X	X	X	X	X	X	X
36637	X	X	X	X	X	X	X	X	X
3664	X	X	X	X	X	X	X	X	X
3720	X	X	X	X	X	X	X	X	X
37480	X	X	X	X	X	X	X	X	X
37784	X	X	X				X	X	X
37793	X	X	X	X	X	X	X	X	X
38002	X	X	X	X	X	X			
38372	X	X	X	X	X	X	X	X	X
38531	X	X	X	X	X	X	X	X	X
38726	X	X	X	X	X	X	X	X	X
38874	X	X	X	X	X	X	X	X	X
38877	X	X	X	X	X	X	X	X	X
39240	X	X	X	X	X	X	X	X	X
39448	X	X	X	X	X	X	X	X	X
39556	X	X	X	X	X	X	X	X	X
40266	X	X	X	X	X	X	X	X	X
40337	X	X	X	X	X	X	X	X	X
40482	X	X	X	X	X	X	X	X	X
40771	X	X	X	X	X	X	X	X	X
40793	X	X	X	X	X	X	X	X	X
40853	X	X	X	X	X	X	X	X	X
41083	X	X	X	X	X	X	X	X	X
41165									
41324	X	X	X	X	X	X	X	X	X
41414	X	X	X	X	X	X	X	X	X
41555	X	X	X	X	X	X	X	X	X
41596	X	X	X	X	X	X	X	X	X
41608	X	X	X	X	X	X	X	X	X
41638	X	X	X	X	X	X	X	X	X
4174	X	X	X	X	X	X	X	X	X
41859	X	X	X	X	X	X	X	X	X
42401	X	X	X	X	X	X	X	X	X
42408	X	X	X	X	X	X	X	X	X
42619	X	X	X	X	X	X	X	X	X
42968	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
43132	X	X	X	X	X	X	X	X	X
43265	X	X	X	X	X	X	X	X	X
4338	X	X	X	X	X	X	X	X	X
43393	X	X	X	X	X	X	X	X	X
43652	X	X	X	X	X	X	X	X	X
43752	X	X	X	X	X	X	X	X	X
43976	X	X	X	X	X	X	X	X	X
44022	X	X	X	X	X	X	X	X	X
44116									
44633	X	X	X	X	X	X	X	X	X
44714	X	X	X	X	X	X	X	X	X
44787	X	X	X	X	X	X	X	X	X
44961	X	X	X	X	X	X	X	X	X
45012	X	X	X	X	X	X	X	X	X
45275	X	X	X	X	X	X	X	X	X
46033	X	X	X	X	X	X	X	X	X
46157	X	X	X	X	X	X	X	X	X
46335	X	X	X	X	X	X	X	X	X
46376	X	X	X	X	X	X	X	X	X
46423	X	X	X				X		
46740	X	X	X	X	X	X	X	X	X
46794	X	X	X	X	X	X	X	X	X
47192	X	X	X	X	X	X	X	X	X
47983	X	X	X	X	X	X	X	X	X
48422	X	X	X	X	X	X	X	X	X
484	X	X	X	X	X	X	X	X	X
48783	X	X	X	X	X	X	X	X	X
49963	X	X	X	X	X	X	X	X	X
50215	X	X	X	X	X	X	X	X	X
50336	X	X	X	X	X	X	X	X	X
50408				X	X	X	X		
5074	X	X	X	X	X	X	X	X	X
51307	X	X	X	X	X	X	X	X	X
51347	X	X	X	X	X	X	X	X	X
5171	X	X	X	X	X	X	X	X	X
51857	X	X	X	X	X	X	X	X	X
51874	X	X	X	X	X	X	X	X	X
52007	X	X	X	X	X	X	X	X	X
52088	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
52229	X	X	X	X	X	X	X	X	X
52278	X	X	X	X	X	X	X	X	X
52639	X	X	X	X	X	X	X	X	X
52725	X	X	X	X	X	X	X	X	X
52909	X	X	X	X	X	X	X	X	X
5300	X	X	X	X	X	X	X	X	X
53121	X	X	X	X	X	X	X	X	X
53341									
53371	X	X	X	X	X	X	X	X	X
53899	X	X	X	X	X	X	X	X	
5390	X	X	X	X	X	X			
53973	X	X	X	X	X	X	X	X	X
54689	X	X	X	X	X	X	X	X	X
54848	X	X	X	X	X	X	X	X	X
55084	X	X	X	X	X	X	X	X	X
55388	X	X	X				X	X	X
55553	X	X	X				X	X	X
5560	X	X	X	X	X	X	X	X	X
55947	X	X	X	X	X	X	X	X	X
55970	X	X	X	X	X	X	X	X	X
56438	X	X	X	X	X	X	X	X	X
57208	X	X	X	X	X	X	X	X	X
57491	X	X	X	X	X	X	X	X	X
57900	X	X	X	X	X	X	X	X	X
58153	X	X	X	X	X	X	X	X	X
58169	X	X	X	X	X	X			
58444	X	X	X	X	X	X	X	X	X
58648	X	X	X	X	X	X	X	X	X
58966	X	X	X	X	X	X	X	X	X
59214	X	X	X	X	X	X	X	X	X
59321	X	X	X	X	X	X			
5992	X	X	X	X	X	X	X	X	X
60021	X	X	X	X	X	X	X	X	X
60479	X	X	X	X	X	X	X	X	X
60589	X	X	X	X	X	X	X	X	X
610	X	X	X	X	X	X	X	X	X
61341	X	X	X	X	X	X	X	X	X
61383	X	X	X	X	X	X	X	X	X
62016	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
62559	X	X	X	X	X	X	X	X	X
62598	X	X	X	X	X	X	X	X	X
62604	X	X	X	X	X	X	X	X	X
63015	X	X	X	X	X	X	X	X	X
63486	X	X	X	X	X	X	X	X	X
63582	X	X	X	X	X	X	X	X	X
63670	X	X	X	X	X	X	X	X	X
6367	X	X	X	X	X	X	X	X	X
63890	X	X	X	X	X	X	X	X	X
64129	X	X	X	X	X	X	X	X	X
64334	X	X	X	X	X	X	X	X	X
64505	X	X	X	X	X	X	X	X	X
64602	X	X	X	X	X	X			
64707	X	X	X	X	X	X	X	X	X
65239	X	X	X				X	X	X
65290	X	X	X	X	X	X	X	X	X
65570	X	X	X	X	X	X	X	X	X
66609	X	X	X	X	X	X			
66860	X	X	X	X	X	X	X	X	X
67316	X	X	X	X	X	X	X	X	X
67848	X	X	X	X	X	X	X	X	X
67940	X	X	X	X	X	X	X	X	X
67997	X	X	X				X	X	X
68029	X	X	X	X	X	X	X	X	X
68108	X	X	X	X	X	X	X	X	X
68322	X	X	X	X	X	X	X	X	X
68638	X	X	X	X	X	X	X	X	X
6863	X	X	X	X	X	X	X	X	X
6877	X	X	X	X	X	X	X	X	X
69365	X	X	X	X	X	X	X	X	X
70306				X	X	X	X	X	X
70662	X	X	X	X	X	X	X	X	X
7080	X	X	X	X	X	X	X	X	X
71025	X	X	X	X	X	X	X	X	X
71165	X	X	X	X	X	X	X	X	X
7127	X	X	X	X	X	X	X	X	X
71430	X	X	X	X	X	X	X	X	X
71582	X	X	X	X	X	X	X	X	X
7192	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
71936	X	X	X	X	X	X	X	X	X
72009	X	X	X	X	X	X	X	X	X
72039	X	X	X	X	X	X	X	X	X
72882	X	X	X	X			X		
73019	X	X	X	X	X	X	X	X	X
73146	X	X	X	X	X	X	X	X	X
73202	X	X	X	X	X	X	X	X	X
73262	X	X	X	X	X	X		X	X
73336	X	X	X	X	X	X			
7363	X	X	X	X	X	X	X	X	X
74245	X	X	X				X	X	X
75497	X	X	X	X	X	X	X	X	X
75810	X	X	X	X	X	X	X	X	X
76117	X	X	X	X	X	X			
76164	X	X	X	X	X	X	X	X	X
76234	X	X	X	X	X	X	X	X	X
76375	X	X	X	X	X	X	X	X	X
76469	X	X	X	X	X	X	X	X	X
76532	X	X	X	X	X	X	X	X	X
76601	X	X	X	X	X	X	X	X	X
77026	X	X	X	X	X	X	X	X	X
77356									
77491	X	X	X	X	X	X	X	X	X
77538	X	X	X	X	X	X	X	X	X
77568	X	X	X	X	X	X	X	X	X
77808	X	X	X	X	X	X	X	X	X
77841	X	X	X	X	X	X	X	X	X
77991	X	X	X	X	X	X	X	X	X
78028	X	X	X	X	X	X	X	X	X
78139	X	X	X	X	X	X	X	X	X
78293	X	X	X	X	X	X	X	X	X
79105	X	X	X	X	X	X	X	X	X
79481	X	X	X	X	X	X	X	X	X
79596	X	X	X	X	X	X	X	X	X
7959	X	X	X	X	X	X	X	X	X
79949	X	X	X	X	X	X	X	X	X
7999									
80337	X	X	X	X	X	X	X	X	X
80490	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
8059	X	X	X	X	X	X	X	X	X
80607	X	X	X	X	X	X	X	X	X
80665	X	X	X	X	X	X	X	X	X
81405	X	X	X	X	X	X	X	X	X
81496	X	X	X						
81841	X	X	X	X	X	X	X	X	X
81876	X	X	X	X	X	X	X	X	X
82126	X	X	X	X	X	X	X	X	X
82148	X	X	X	X	X	X	X	X	X
82322	X	X	X	X	X	X	X	X	X
82369	X	X	X						
82570	X	X	X	X	X	X	X	X	X
82580	X	X	X	X	X	X	X	X	X
82694	X	X	X	X	X	X	X	X	X
8413	X	X	X	X	X	X	X	X	X
84185	X	X	X	X	X	X	X	X	X
84231	X	X	X	X	X	X	X	X	X
84276	X	X	X	X	X	X	X	X	X
84291	X	X	X	X	X	X	X	X	X
84363	X	X	X	X	X	X	X	X	X
84699	X	X	X	X	X	X	X	X	X
84840	X	X	X	X	X	X	X	X	X
84872	X	X	X	X	X	X	X	X	X
8568	X	X	X						
85806	X	X	X	X	X	X	X	X	X
86467	X	X	X	X	X	X	X	X	X
87189	X	X	X	X	X	X	X	X	X
87390	X	X	X	X	X	X			
87581	X	X	X	X	X	X	X	X	X
87828	X	X	X	X	X	X	X	X	X
88190									
88303	X	X	X	X	X	X	X	X	X
88540	X	X	X	X	X	X	X	X	X
88572	X	X	X	X	X	X	X	X	X
88842	X	X	X	X	X	X	X	X	X
8904	X	X	X	X	X	X	X	X	X
89090	X	X	X	X	X	X	X	X	X
89116	X	X	X	X	X	X	X	X	X
89715	X	X	X	X	X	X	X	X	X

Continua na próxima página

Tabela A.2 – *Continua da página anterior*

GCU	Trinity			Velvet-Oases			IDBA-Trans		
	Tr	Mr	Sr	Tr	Mr	Sr	Tr	Mr	Sr
89858	X	X	X	X	X	X	X	X	X
89947	X	X	X	X	X	X	X	X	X
90721	X	X	X	X	X	X	X	X	X
91487	X	X	X	X	X	X	X	X	X
93187	X	X	X	X	X	X	X	X	X
93191	X	X	X	X	X	X	X	X	X
93494	X	X	X	X	X	X	X	X	X
93624	X	X	X	X	X	X	X	X	X
93746									
94134	X	X	X	X	X	X			
94766	X	X	X	X	X	X	X	X	X
94871	X	X	X	X	X	X	X	X	X
94998	X	X	X	X	X	X	X	X	X
9563	X	X	X						
9609	X	X	X	X	X	X			
96385	X	X	X	X	X	X	X	X	X
96430	X	X	X						
9651	X	X	X	X	X	X	X	X	
96703	X	X	X	X	X	X	X	X	X
96830	X	X	X	X	X	X	X	X	X
98066	X	X	X	X	X	X	X	X	X
98097	X	X	X	X	X	X			
98103	X	X	X	X	X	X	X	X	
98198	X	X	X	X	X	X	X	X	X
98672									
98854	X	X	X	X	X	X	X	X	X
98871	X	X	X	X	X	X	X	X	X
99355	X	X	X	X	X	X	X	X	X
99389	X	X	X	X	X	X			
99736									

Tabela A.2: Tabela de GCUs para o organismo *Canis Familiaris* por cada caso de teste. O símbolo X representa que certos GCUs estão representados no caso de teste.

A.2 Heatmaps dos critérios de seleção

O Apêndice A.2 é composto pelos heatmaps e gráficos gerados a partir deles que foram citados no Capítulo 5, Seção 5.4.2.

A.2.1 *Arabidopsis thaliana*

Caso de teste Tr

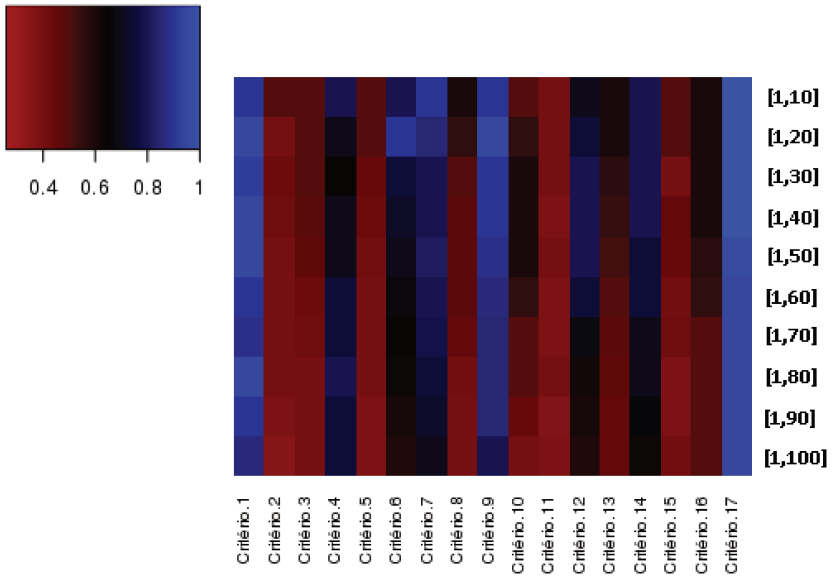

Figura A.1: Heatmap representando o caso de teste Tr de *Arabidopsis thaliana*. Cada posição do heatmap é representando pela melhor porcentagem obtida por um montador *de novo* dado tal critério por intervalo.

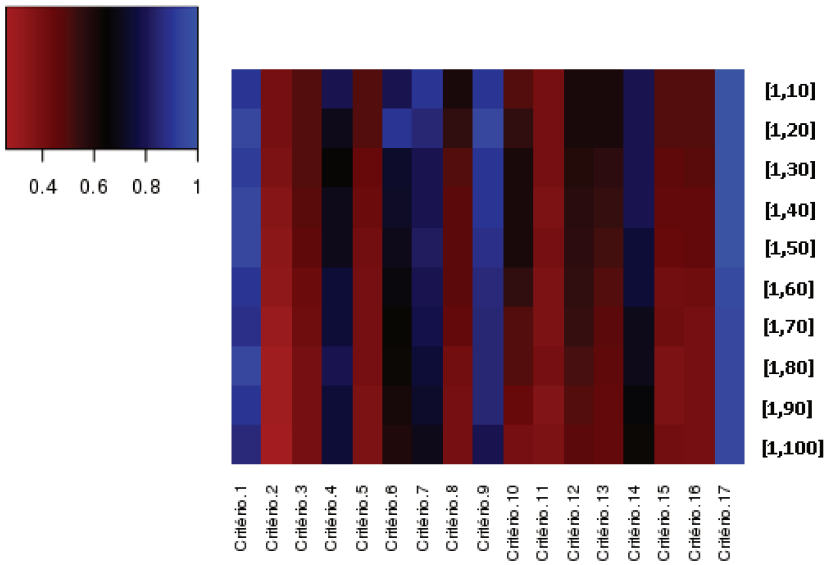

Figura A.2: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Trinity no caso de teste Tr de *Arabidopsis thaliana*.

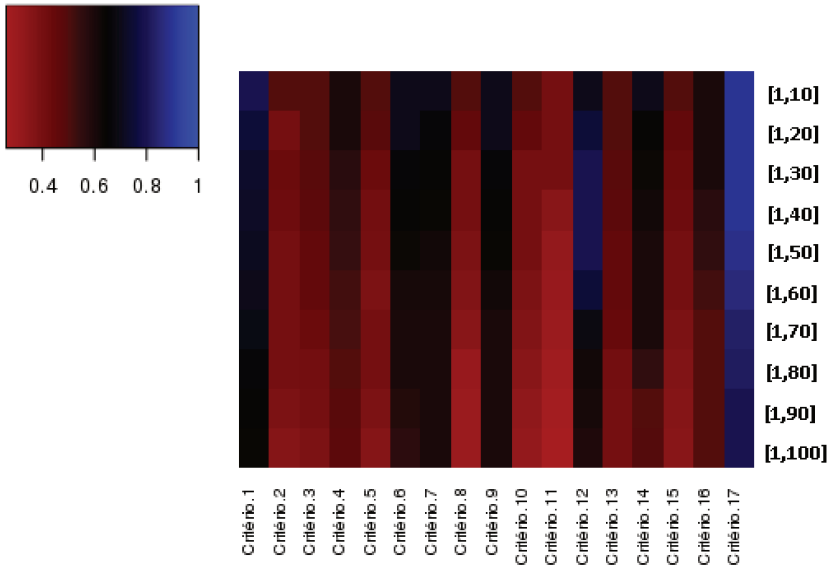

Figura A.3: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Velvet-Oases no caso de teste Tr de *Arabidopsis thaliana*.

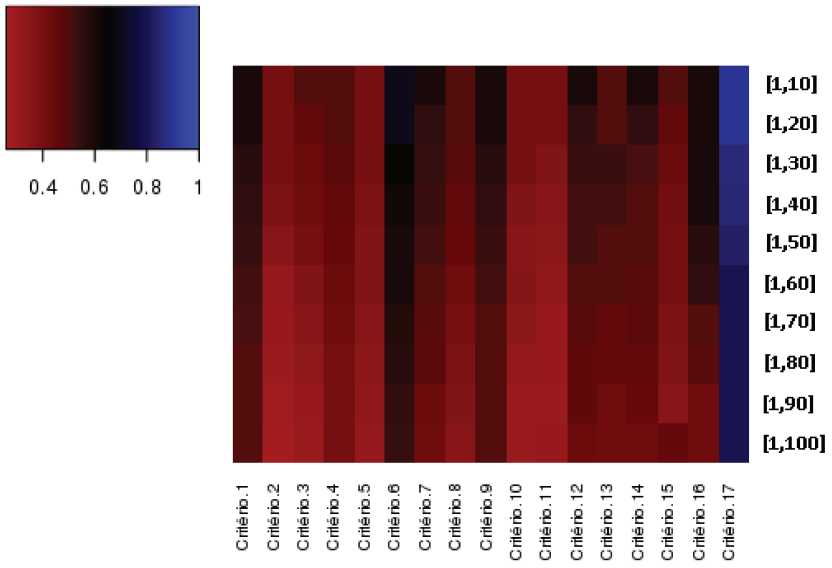

Figura A.4: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* IDBA-Tran no caso de teste Tr de *Arabidopsis thaliana*.

Caso de teste Mr

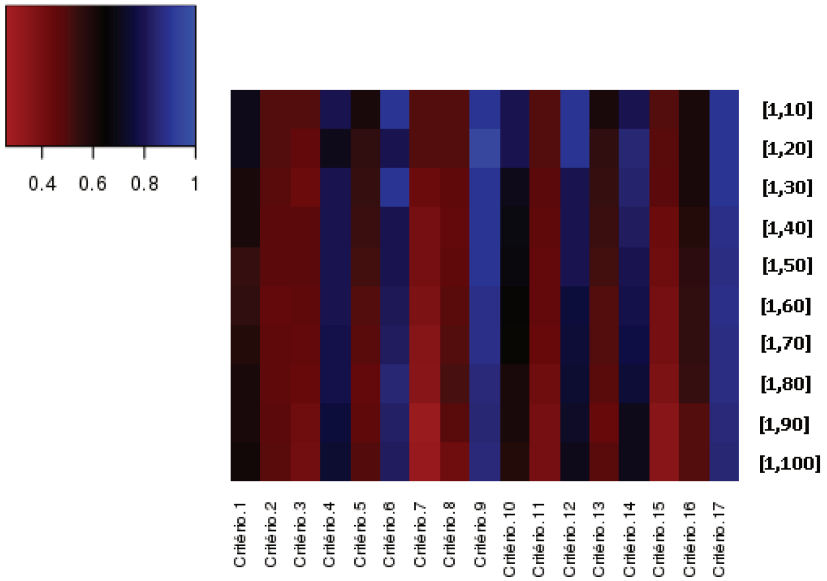

Figura A.5: Heatmap representando o caso de teste Mr de *Arabidopsis thaliana*. Cada posição do heatmap é representando pela melhor porcentagem obtida por um montador *de novo* dado tal critério por intervalo.

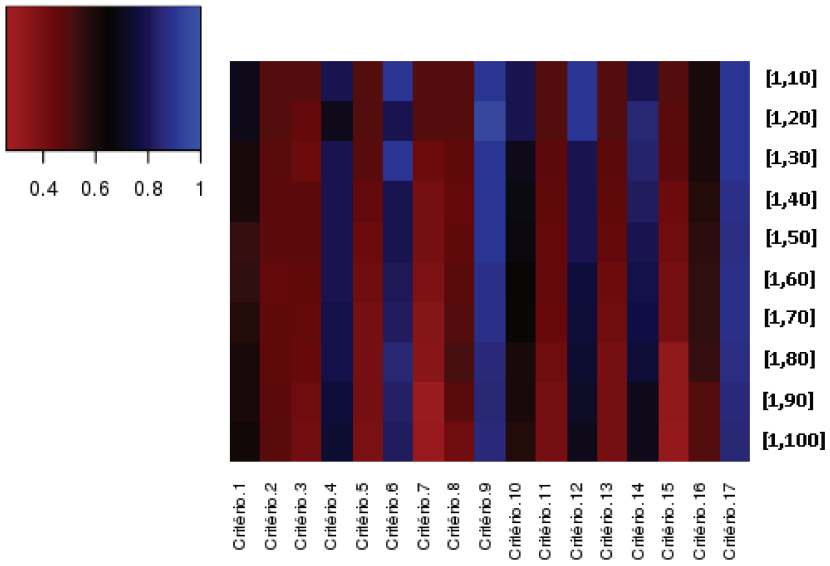

Figura A.6: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Trinity no caso de teste Mr de *Arabidopsis thaliana*.

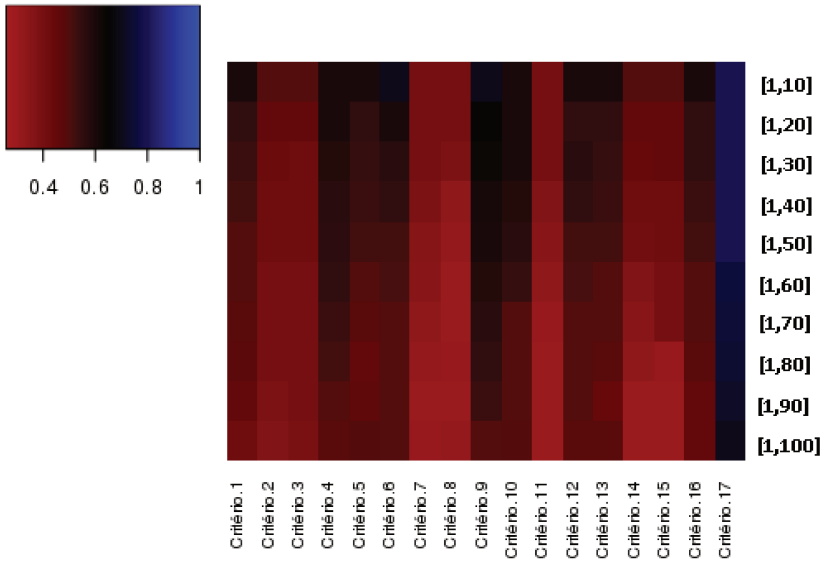

Figura A.7: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Velvet-Oases no caso de teste Mr de *Arabidopsis thaliana*.

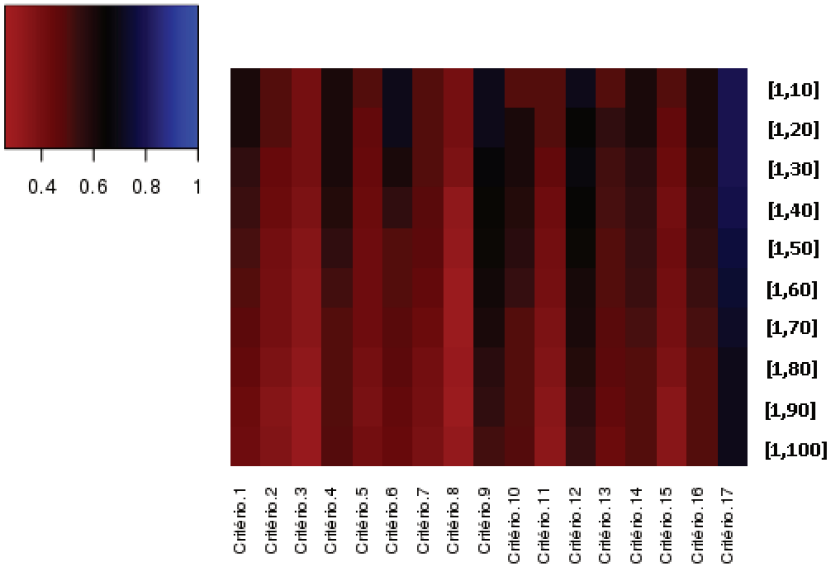

Figura A.8: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* IDBA-Tran no caso de teste Mr de *Arabidopsis thaliana*.

Caso de teste Sr

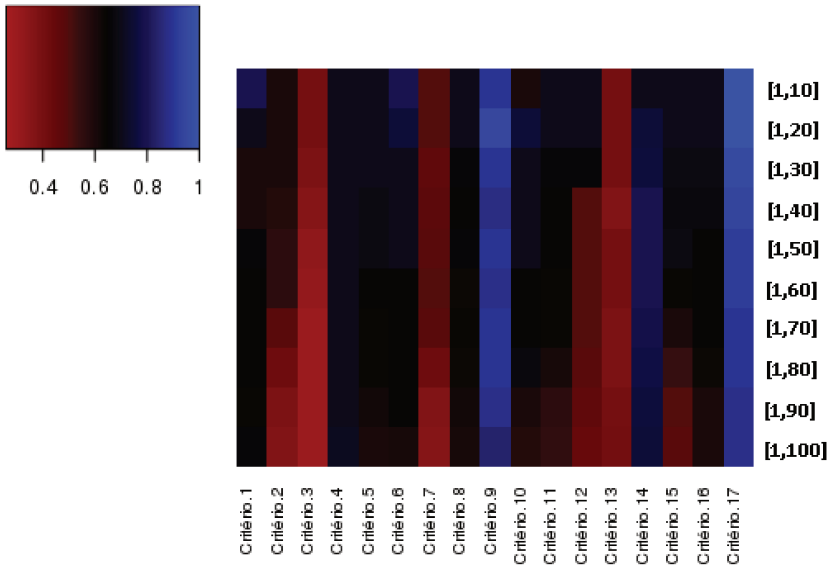

Figura A.9: Heatmap representando o caso de teste Sr de *Arabidopsis thaliana*. Cada posição do heatmap é representando pela melhor porcentagem obtida por um montador *de novo* dado tal critério por intervalo.

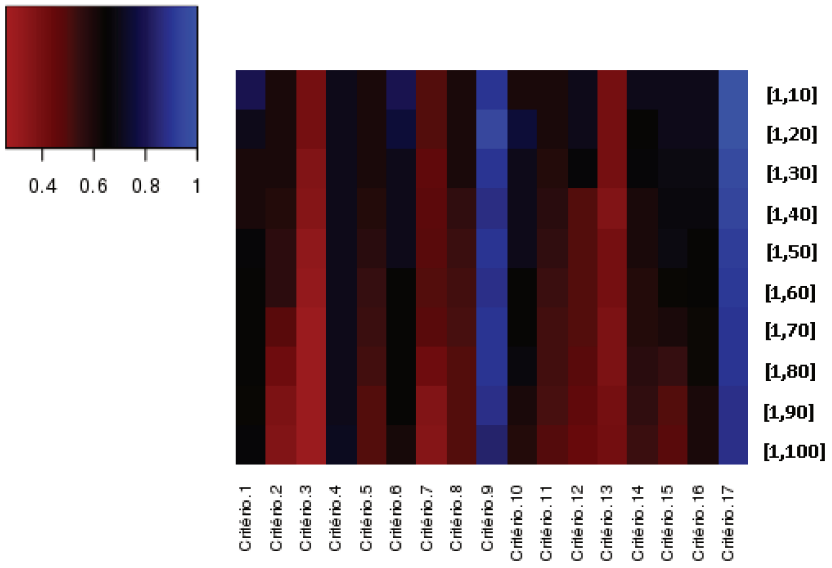

Figura A.10: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Trinity no caso de teste Sr de *Arabidopsis thaliana*.

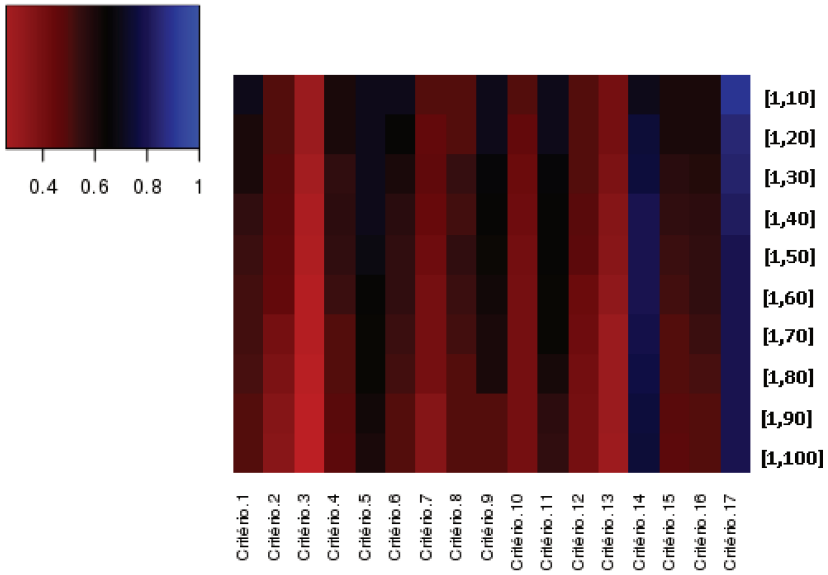

Figura A.11: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Velvet-Oases no caso de teste Sr de *Arabidopsis thaliana*.

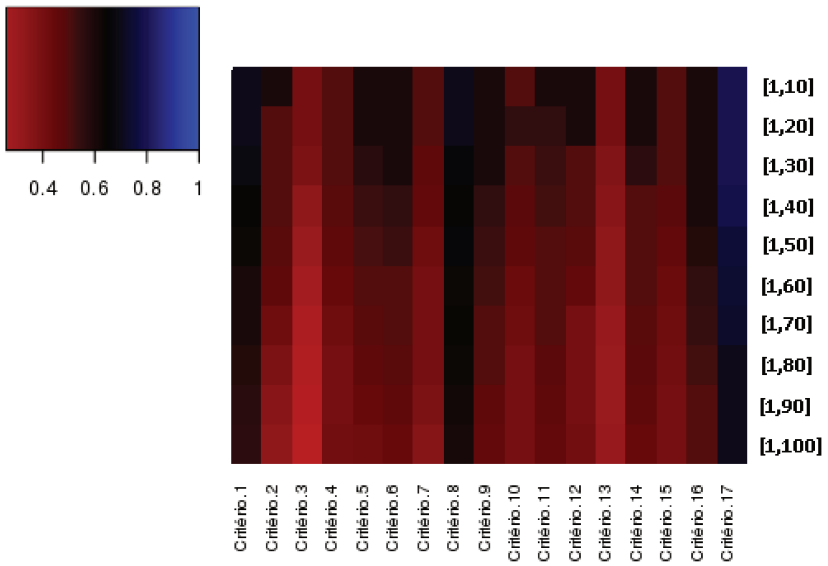

Figura A.12: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* IDBA-Tran no caso de teste Sr de *Arabidopsis thaliana*.

A.2.2 *Canis familiaris*

Caso de teste Tr

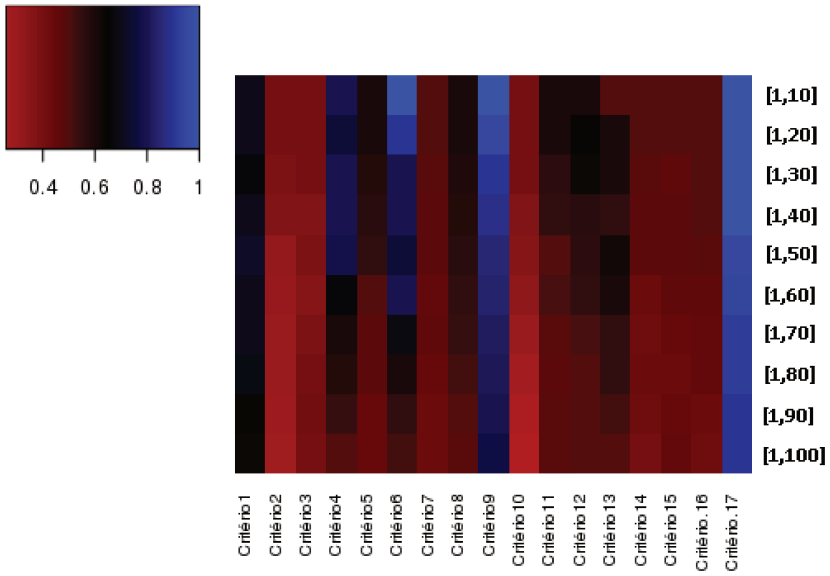

Figura A.13: Heatmap representando o caso de teste Tr de *Canis familiaris*. Cada posição do heatmap é representando pela melhor porcentagem obtida por um montador *de novo* dado tal critério por intervalo.

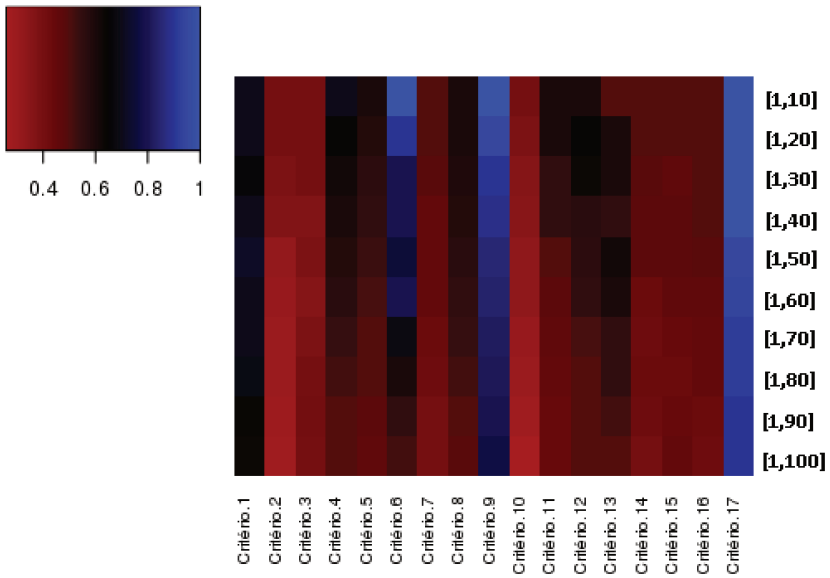

Figura A.14: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Trinity no caso de teste Tr de *Canis familiaris*.

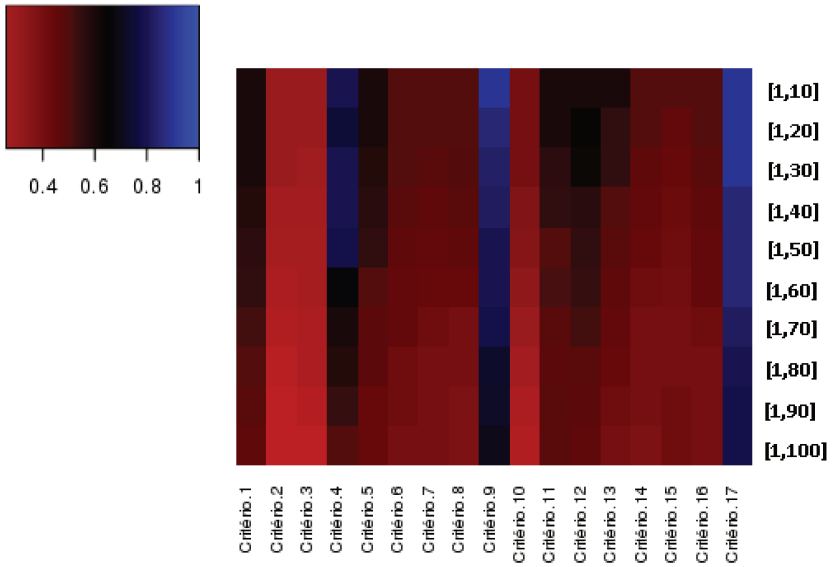

Figura A.15: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Velvet-Oases no caso de teste Tr de *Canis familiaris*.

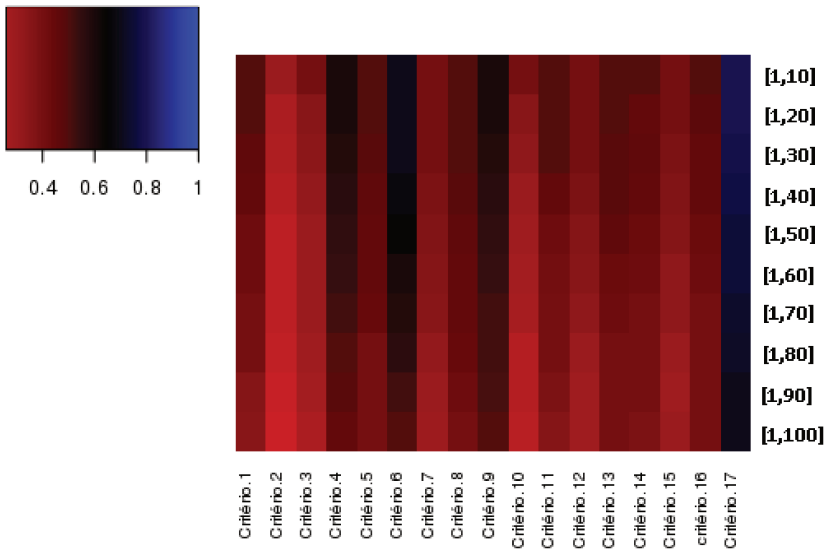

Figura A.16: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* IDBA-Tran no caso de teste Tr de *Canis familiaris*.

Caso de teste Mr

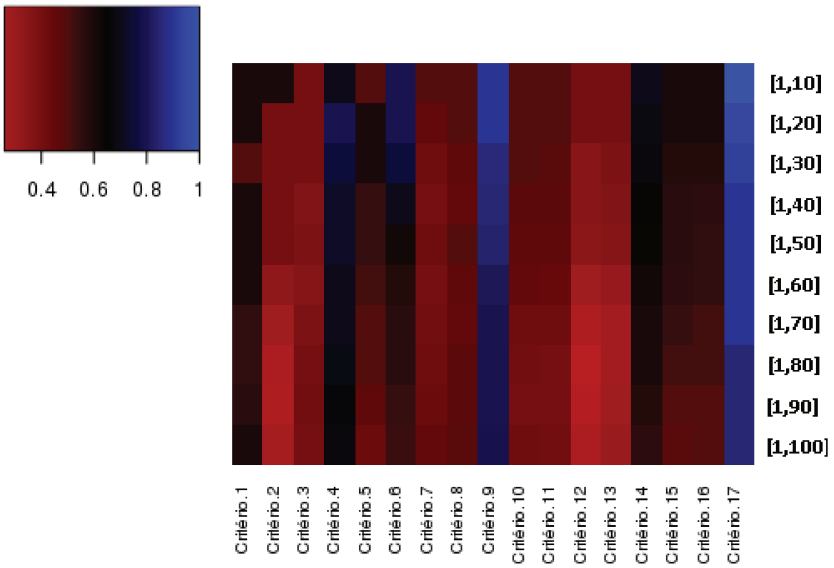

Figura A.17: Heatmap representando o caso de teste Mr de *Canis familiaris*. Cada posição do heatmap é representando pela melhor porcentagem obtida por um montador *de novo* dado tal critério por intervalo.

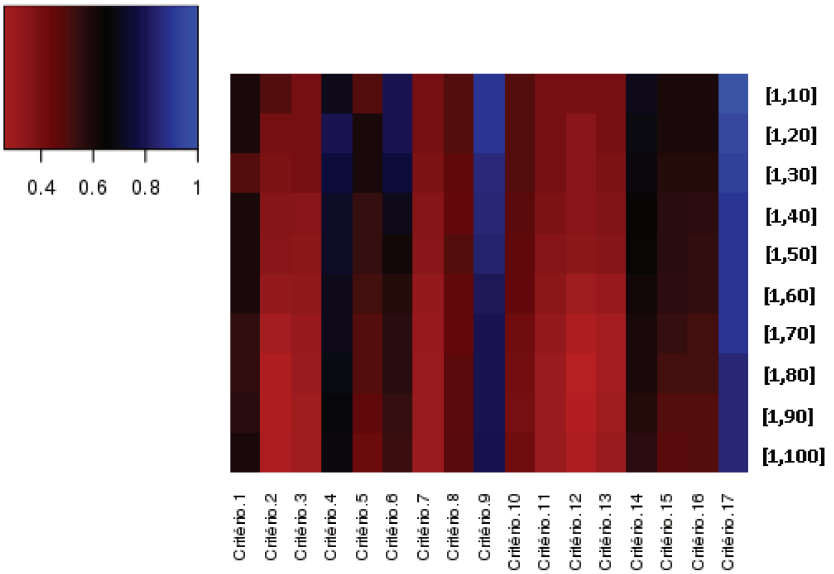

Figura A.18: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Trinity no caso de teste Mr de *Canis familiaris*.

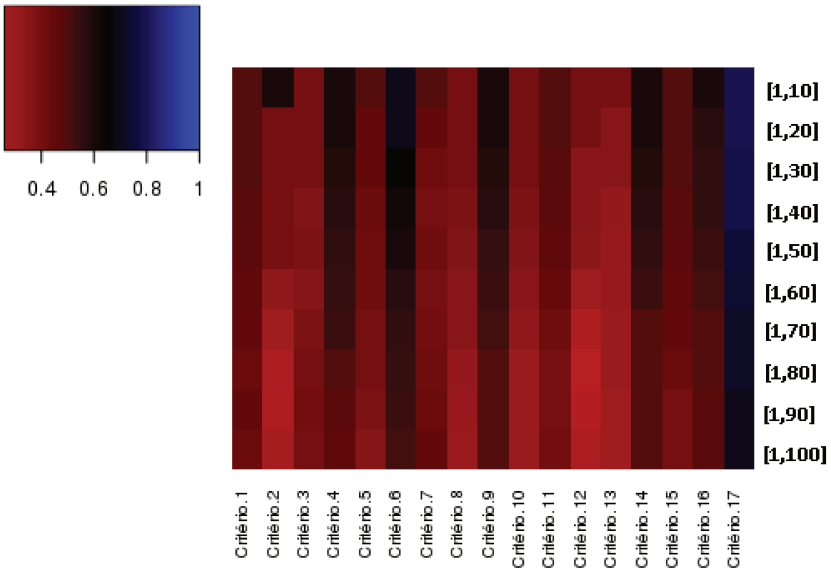

Figura A.19: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Velvet-Oases no caso de teste Mr de *Canis familiaris*.

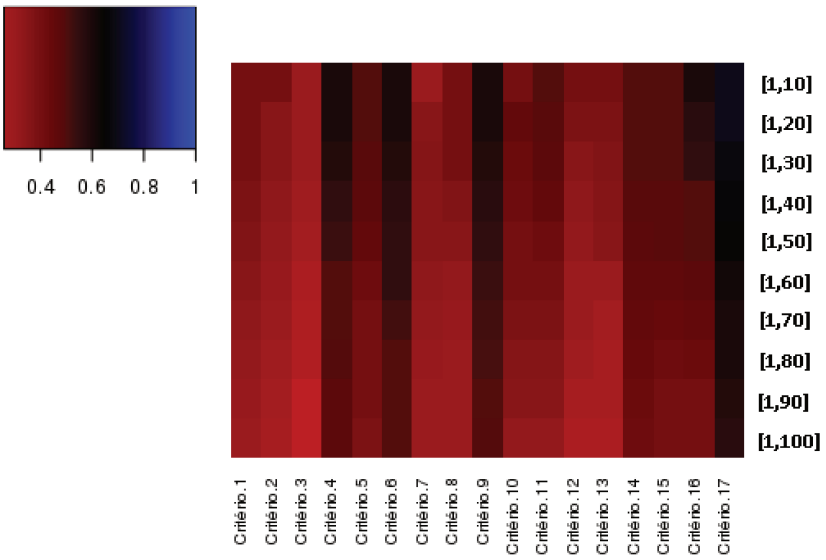

Figura A.20: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* IDBA-Tran no caso de teste Mr de *Canis familiaris*.

Caso de teste Sr

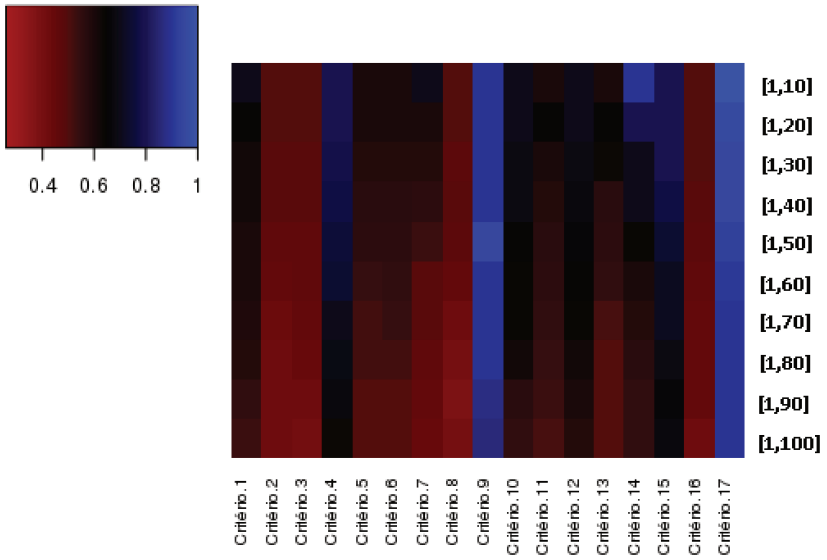

Figura A.21: Heatmap representando o caso de teste Sr de *Canis familiaris*. Cada posição do heatmap é representando pela melhor porcentagem obtida por um montador *de novo* dado tal critério por intervalo.

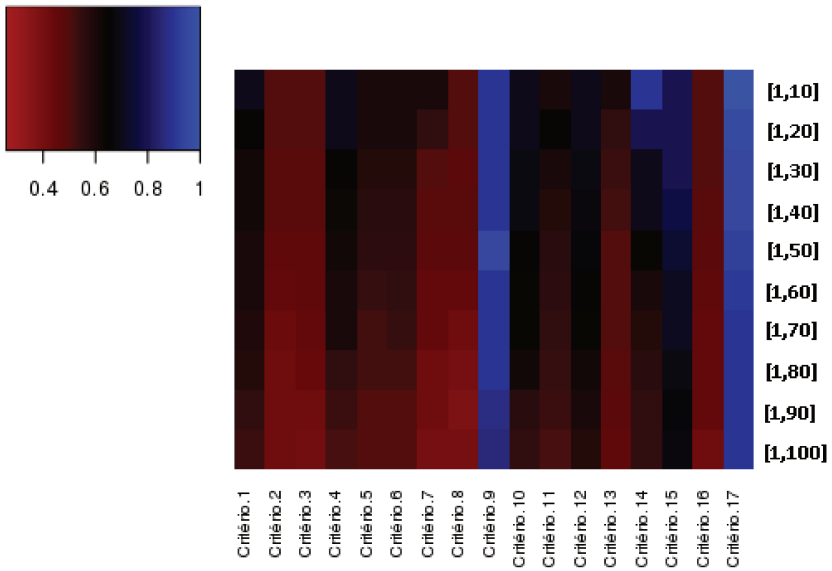

Figura A.22: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Trinity no caso de teste Sr de *Canis familiaris*.

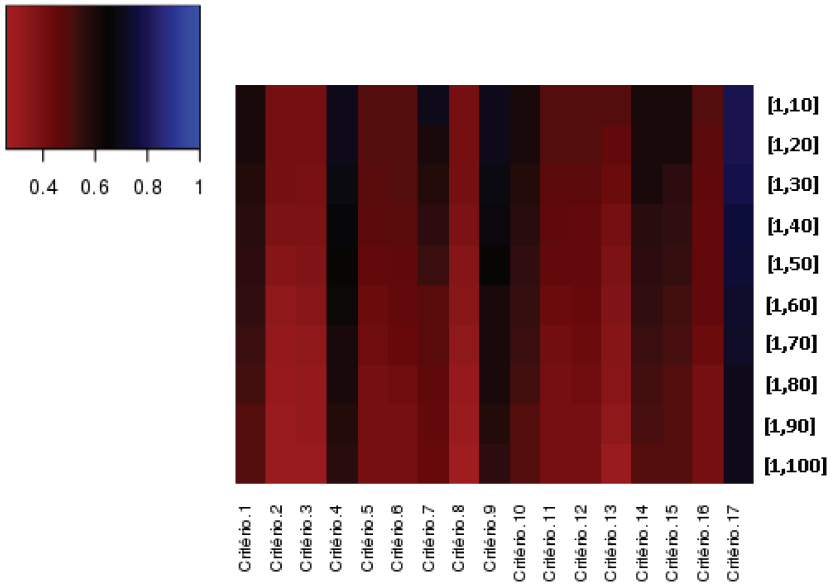

Figura A.23: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* Velvet-Oases no caso de teste Sr de *Canis familiaris*.

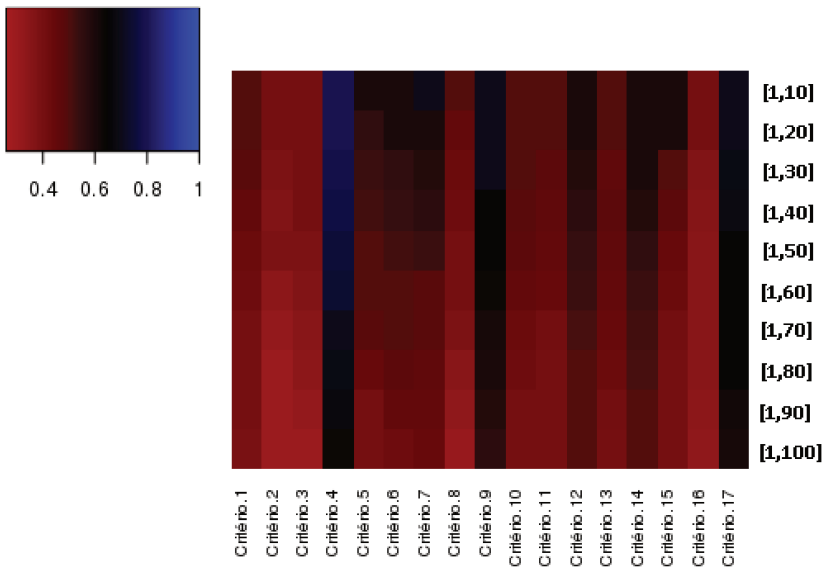

Figura A.24: Heatmap gerado a partir da matriz de porcentagem de transcritos verdadeiramente expressos, por cada critério dado certo intervalo, para o montador *de novo* IDBA-Tran no caso de teste Sr de *Canis familiaris*.

A.3 Tabelas de dados do teste pareados de Wilcoxon

Abaixo estão os resultados dos testes pareados de Wilcoxon. Os testes foram aplicados para observar a relevância estatística do Critério 9 sobre os outros, e estão apresentados por montador. O valor do $W_{critico}$, definido pelo usuário, dado um valor n (número de amostras a comparar) é resumido pela Figura A.25.

	$w_{0.005}$	$w_{0.01}$	$w_{0.025}$	$w_{0.05}$	$w_{0.10}$	$w_{0.20}$	$w_{0.30}$	$w_{0.40}$	$w_{0.50}$	$\frac{n(n+1)}{2}$
$n=4$	0	0	0	0	1	3	3	4	5	10
5	0	0	0	1	3	4	5	6	7.5	15
6	0	0	1	3	4	6	8	9	10.5	21
7	0	1	3	4	6	9	11	12	14	28
8	1	2	4	6	9	12	14	16	18	36
9	2	4	6	9	11	15	18	20	22.5	45
10	4	6	9	11	15	19	22	25	27.5	55
11	6	8	11	14	18	23	27	30	33	66
12	8	10	14	18	22	28	32	36	39	78
13	10	13	18	22	27	33	38	42	45.5	91
14	13	16	22	26	32	39	44	48	52.5	105
15	16	20	26	31	37	45	51	55	60	120
16	20	24	30	36	43	51	58	63	68	136
17	24	28	35	42	49	58	65	71	76.5	153
18	28	33	41	48	56	66	73	80	85.5	171
19	33	38	47	54	63	74	82	89	95	190
20	38	44	53	61	70	83	91	98	105	210
21	44	50	59	68	78	91	100	108	115.5	131
22	49	56	67	76	87	100	110	119	126.5	153
23	55	63	74	84	95	110	120	130	138	176
24	62	70	82	92	105	120	131	141	150	300
25	69	77	90	101	114	131	143	153	162.5	325
26	76	85	99	111	125	142	155	165	175.5	351
27	84	94	108	120	135	154	167	178	189	378
28	92	102	117	131	146	166	180	192	203	406
29	101	111	127	141	158	178	193	206	217.5	435
30	110	121	138	152	170	191	207	220	232.5	465
31	119	131	148	164	182	205	221	235	248	496
32	129	141	160	176	195	219	236	250	264	528
33	139	152	171	188	208	233	251	266	280.5	561
34	149	163	183	201	222	248	266	282	297.5	595
35	160	175	196	214	236	263	283	299	315	630
36	172	187	209	228	251	279	299	317	333	666
37	184	199	222	242	266	295	316	335	351.5	703
38	196	212	236	257	282	312	334	353	370.5	741
39	208	225	250	272	298	329	352	372	390	780
40	221	239	265	287	314	347	371	391	410	820
41	235	253	280	303	331	365	390	411	430.5	861
42	248	267	295	320	349	384	409	431	451.5	903
43	263	282	311	337	366	403	429	452	473	946
44	277	297	328	354	385	422	450	473	495	990
45	292	313	344	372	403	442	471	495	517.5	1035
46	308	329	362	390	423	463	492	517	540.5	1081
47	324	346	379	408	442	484	514	540	564	1128

Figura A.25: Tabela representativa dado um valor n de amostras a serem comparadas e o valor de $W_{critico}$, definido pelo usuário. O valor de $W_{0.05}$, por exemplo, representa uma confiança de 95%, ou seja, $p - value \leq 0.05$.

A.3.1 *Arabidopsis thaliana*

Caso de teste Tr

Tabela A.3: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.90	0.50	0.50	0.80	0.50	0.80	0.90	0.60	0.90	0.50	0.40	0.70	0.60	0.80	0.50	0.60
[1,20]	1	0.95	0.40	0.50	0.70	0.50	0.90	0.85	0.55	0.95	0.55	0.40	0.75	0.60	0.80	0.50	0.60
[1,30]	1	0.92	0.43	0.50	0.65	0.44	0.75	0.80	0.50	0.90	0.60	0.40	0.80	0.56	0.80	0.40	0.60
[1,40]	1	0.95	0.42	0.48	0.70	0.43	0.72	0.80	0.47	0.90	0.60	0.38	0.80	0.54	0.80	0.45	0.60
[1,50]	0.96	0.95	0.40	0.46	0.70	0.41	0.70	0.82	0.47	0.88	0.60	0.40	0.80	0.52	0.75	0.44	0.57
[1,60]	0.95	0.90	0.40	0.43	0.75	0.40	0.67	0.80	0.47	0.86	0.55	0.38	0.75	0.50	0.75	0.41	0.55
[1,70]	0.95	0.88	0.40	0.42	0.75	0.40	0.64	0.78	0.45	0.85	0.50	0.38	0.68	0.47	0.70	0.42	0.50
[1,80]	0.95	0.95	0.40	0.40	0.80	0.40	0.63	0.75	0.41	0.85	0.50	0.40	0.62	0.46	0.70	0.38	0.50
[1,90]	0.95	0.90	0.38	0.40	0.75	0.38	0.61	0.73	0.40	0.85	0.44	0.37	0.61	0.45	0.66	0.38	0.50
[1,100]	0.95	0.86	0.36	0.40	0.75	0.38	0.59	0.70	0.40	0.80	0.40	0.38	0.59	0.45	0.63	0.41	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.4: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr para o montador Trinity. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	$C16$
[1,10]	1	0.90	0.40	0.50	0.80	0.50	0.80	0.90	0.60	0.90	0.50	0.40	0.60	0.60	0.80	0.50	0.5
[1,20]	1	0.95	0.40	0.50	0.70	0.50	0.90	0.85	0.55	0.95	0.55	0.40	0.60	0.60	0.80	0.50	0.5
[1,30]	1	0.92	0.38	0.50	0.65	0.44	0.73	0.80	0.50	0.90	0.60	0.40	0.58	0.56	0.80	0.46	0.48
[1,40]	1	0.95	0.36	0.48	0.70	0.43	0.72	0.80	0.47	0.90	0.60	0.38	0.57	0.54	0.80	0.45	0.45
[1,50]	1	0.95	0.34	0.46	0.70	0.41	0.70	0.82	0.47	0.88	0.60	0.40	0.56	0.52	0.75	0.44	0.45
[1,60]	0.96	0.90	0.33	0.43	0.75	0.40	0.67	0.80	0.47	0.86	0.55	0.38	0.55	0.50	0.75	0.41	0.42
[1,70]	0.95	0.88	0.30	0.42	0.75	0.40	0.64	0.78	0.45	0.85	0.50	0.38	0.54	0.47	0.70	0.42	0.40
[1,80]	0.95	0.95	0.28	0.40	0.80	0.40	0.63	0.75	0.41	0.85	0.50	0.40	0.51	0.46	0.70	0.38	0.40
[1,90]	0.95	0.90	0.28	0.40	0.75	0.38	0.61	0.73	0.40	0.85	0.44	0.37	0.50	0.45	0.66	0.38	0.40
[1,100]	0.95	0.86	0.27	0.40	0.75	0.38	0.59	0.70	0.40	0.80	0.40	0.38	0.47	0.45	0.63	0.41	0.40
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.5: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr para o montador Velvet-Oases. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.9	0.80	0.50	0.50	0.60	0.50	0.70	0.70	0.50	0.70	0.50	0.40	0.70	0.50	0.70	0.50	0.6
[1,20]	0.9	0.75	0.40	0.50	0.60	0.48	0.70	0.66	0.45	0.70	0.45	0.40	0.75	0.50	0.65	0.45	0.6
[1,30]	0.9	0.73	0.43	0.48	0.57	0.43	0.66	0.65	0.40	0.66	0.40	0.40	0.80	0.48	0.63	0.43	0.6
[1,40]	0.9	0.72	0.42	0.47	0.55	0.41	0.65	0.64	0.40	0.65	0.40	0.35	0.80	0.47	0.62	0.42	0.57
[1,50]	0.88	0.71	0.40	0.45	0.54	0.40	0.63	0.62	0.38	0.64	0.40	0.32	0.80	0.45	0.60	0.40	0.55
[1,60]	0.86	0.70	0.40	0.45	0.52	0.38	0.61	0.61	0.37	0.62	0.38	0.31	0.75	0.45	0.60	0.40	0.52
[1,70]	0.83	0.69	0.40	0.43	0.51	0.40	0.60	0.60	0.35	0.60	0.37	0.30	0.68	0.44	0.60	0.38	0.50
[1,80]	0.82	0.66	0.40	0.41	0.50	0.40	0.60	0.60	0.31	0.60	0.35	0.28	0.62	0.41	0.55	0.37	0.50
[1,90]	0.80	0.65	0.38	0.40	0.48	0.38	0.58	0.60	0.30	0.60	0.33	0.27	0.61	0.40	0.50	0.36	0.50
[1,100]	0.80	0.64	0.36	0.38	0.47	0.36	0.56	0.60	0.30	0.60	0.32	0.26	0.59	0.40	0.49	0.35	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.6: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr para o montador IDBA-Tran. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.9	0.60	0.40	0.50	0.50	0.40	0.70	0.60	0.50	0.60	0.40	0.40	0.60	0.50	0.60	0.50	0.6
[1,20]	0.9	0.60	0.40	0.45	0.50	0.40	0.70	0.55	0.50	0.60	0.40	0.40	0.55	0.50	0.55	0.45	0.6
[1,30]	0.86	0.57	0.40	0.43	0.47	0.40	0.65	0.54	0.48	0.57	0.40	0.37	0.53	0.53	0.51	0.43	0.6
[1,40]	0.85	0.55	0.38	0.42	0.45	0.38	0.62	0.53	0.45	0.55	0.37	0.35	0.52	0.52	0.50	0.41	0.6
[1,50]	0.83	0.54	0.35	0.40	0.44	0.37	0.60	0.52	0.44	0.53	0.35	0.34	0.52	0.50	0.50	0.40	0.57
[1,60]	0.80	0.52	0.32	0.37	0.43	0.37	0.60	0.50	0.42	0.52	0.36	0.33	0.50	0.50	0.48	0.40	0.55
[1,70]	0.80	0.51	0.31	0.35	0.42	0.35	0.58	0.48	0.40	0.50	0.34	0.31	0.48	0.45	0.47	0.38	0.50
[1,80]	0.80	0.50	0.30	0.33	0.40	0.34	0.57	0.47	0.38	0.50	0.32	0.31	0.46	0.45	0.45	0.37	0.48
[1,90]	0.80	0.50	0.28	0.31	0.40	0.33	0.55	0.43	0.37	0.50	0.30	0.30	0.46	0.42	0.44	0.35	0.42
[1,100]	0.80	0.50	0.27	0.30	0.40	0.32	0.54	0.42	0.35	0.50	0.30	0.31	0.43	0.42	0.42	0.45	0.42
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Caso de teste Mr

Tabela A.7: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Mr. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.90	0.70	0.50	0.50	0.80	0.60	0.90	0.50	0.50	0.90	0.80	0.50	0.90	0.60	0.80	0.50	0.60
[1,20]	0.90	0.70	0.50	0.45	0.70	0.55	0.80	0.50	0.50	0.95	0.80	0.50	0.90	0.55	0.85	0.48	0.60
[1,30]	0.90	0.60	0.48	0.43	0.80	0.54	0.90	0.43	0.46	0.90	0.70	0.47	0.80	0.54	0.84	0.47	0.60
[1,40]	0.88	0.60	0.47	0.47	0.80	0.53	0.80	0.40	0.45	0.90	0.68	0.46	0.80	0.53	0.82	0.43	0.58
[1,50]	0.87	0.54	0.47	0.47	0.80	0.52	0.80	0.40	0.46	0.90	0.67	0.45	0.80	0.52	0.80	0.42	0.56
[1,60]	0.88	0.55	0.45	0.46	0.80	0.50	0.81	0.38	0.48	0.88	0.65	0.45	0.76	0.50	0.78	0.40	0.55
[1,70]	0.87	0.58	0.46	0.45	0.78	0.48	0.82	0.36	0.50	0.88	0.64	0.44	0.75	0.50	0.77	0.40	0.55
[1,80]	0.87	0.60	0.46	0.44	0.78	0.45	0.85	0.35	0.51	0.86	0.60	0.42	0.74	0.48	0.75	0.38	0.54
[1,90]	0.86	0.60	0.47	0.42	0.76	0.46	0.83	0.30	0.48	0.85	0.60	0.40	0.72	0.44	0.70	0.35	0.50
[1,100]	0.85	0.62	0.48	0.41	0.74	0.49	0.82	0.31	0.42	0.86	0.58	0.40	0.70	0.48	0.70	0.35	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.8: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Mr para o montador Trinity. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.90	0.70	0.50	0.50	0.80	0.50	0.90	0.50	0.50	0.90	0.80	0.50	0.90	0.50	0.80	0.50	0.60
[1,20]	0.90	0.70	0.50	0.45	0.70	0.50	0.80	0.50	0.50	0.95	0.80	0.50	0.90	0.50	0.85	0.48	0.60
[1,30]	0.90	0.60	0.48	0.43	0.80	0.48	0.90	0.43	0.46	0.90	0.70	0.47	0.80	0.47	0.84	0.47	0.60
[1,40]	0.88	0.60	0.47	0.47	0.80	0.45	0.80	0.40	0.45	0.90	0.68	0.46	0.80	0.46	0.82	0.43	0.58
[1,50]	0.87	0.54	0.47	0.47	0.80	0.43	0.80	0.40	0.46	0.90	0.67	0.45	0.80	0.45	0.80	0.42	0.56
[1,60]	0.88	0.55	0.45	0.46	0.80	0.42	0.81	0.38	0.48	0.88	0.65	0.45	0.76	0.43	0.78	0.40	0.55
[1,70]	0.88	0.58	0.46	0.45	0.78	0.40	0.82	0.36	0.50	0.88	0.64	0.44	0.75	0.42	0.77	0.40	0.55
[1,80]	0.87	0.60	0.46	0.44	0.78	0.40	0.85	0.35	0.51	0.86	0.60	0.42	0.74	0.40	0.75	0.33	0.54
[1,90]	0.86	0.60	0.47	0.42	0.76	0.40	0.83	0.30	0.48	0.85	0.60	0.40	0.72	0.40	0.70	0.33	0.50
[1,100]	0.85	0.62	0.48	0.41	0.74	0.39	0.82	0.31	0.42	0.86	0.58	0.40	0.70	0.40	0.70	0.32	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.9: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Mr para o montador Velvet-Oases. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.80	0.60	0.50	0.50	0.60	0.60	0.70	0.40	0.40	0.70	0.60	0.40	0.60	0.60	0.50	0.50	0.60
[1,20]	0.80	0.55	0.45	0.45	0.60	0.55	0.60	0.40	0.40	0.65	0.60	0.40	0.55	0.55	0.45	0.45	0.55
[1,30]	0.80	0.53	0.43	0.42	0.58	0.54	0.57	0.40	0.38	0.63	0.60	0.40	0.57	0.54	0.44	0.45	0.55
[1,40]	0.80	0.52	0.42	0.42	0.57	0.53	0.55	0.38	0.33	0.61	0.58	0.37	0.55	0.53	0.42	0.42	0.53
[1,50]	0.80	0.50	0.42	0.42	0.56	0.52	0.52	0.36	0.32	0.60	0.57	0.35	0.52	0.52	0.41	0.42	0.52
[1,60]	0.76	0.50	0.40	0.40	0.55	0.50	0.51	0.35	0.30	0.58	0.54	0.33	0.51	0.50	0.37	0.40	0.50
[1,70]	0.75	0.48	0.40	0.40	0.53	0.48	0.50	0.33	0.30	0.57	0.50	0.31	0.50	0.50	0.35	0.40	0.50
[1,80]	0.74	0.47	0.40	0.40	0.52	0.45	0.50	0.32	0.31	0.55	0.50	0.30	0.50	0.48	0.33	0.31	0.48
[1,90]	0.72	0.45	0.38	0.40	0.50	0.46	0.50	0.30	0.30	0.53	0.50	0.30	0.50	0.44	0.30	0.30	0.45
[1,100]	0.70	0.42	0.37	0.39	0.48	0.49	0.50	0.31	0.32	0.50	0.49	0.30	0.48	0.48	0.30	0.30	0.45
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.10: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Mr para o montador IDBA-Tran. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.80	0.60	0.50	0.40	0.60	0.50	0.70	0.50	0.40	0.70	0.50	0.50	0.70	0.50	0.60	0.50	0.60
[1,20]	0.80	0.60	0.50	0.40	0.60	0.45	0.70	0.50	0.40	0.70	0.60	0.50	0.65	0.55	0.60	0.45	0.60
[1,30]	0.80	0.55	0.44	0.40	0.60	0.44	0.60	0.50	0.38	0.66	0.60	0.45	0.67	0.52	0.57	0.43	0.58
[1,40]	0.78	0.53	0.43	0.38	0.58	0.43	0.55	0.48	0.33	0.64	0.58	0.42	0.65	0.51	0.55	0.41	0.57
[1,50]	0.76	0.51	0.41	0.36	0.55	0.42	0.50	0.47	0.32	0.63	0.57	0.41	0.63	0.50	0.54	0.42	0.55
[1,60]	0.74	0.50	0.40	0.35	0.52	0.42	0.50	0.45	0.30	0.62	0.54	0.40	0.61	0.50	0.53	0.41	0.53
[1,70]	0.72	0.47	0.40	0.35	0.50	0.42	0.48	0.43	0.30	0.60	0.50	0.38	0.60	0.48	0.51	0.40	0.51
[1,80]	0.70	0.45	0.38	0.33	0.50	0.40	0.47	0.41	0.31	0.57	0.50	0.37	0.58	0.47	0.50	0.38	0.50
[1,90]	0.70	0.43	0.36	0.31	0.50	0.39	0.45	0.40	0.30	0.55	0.50	0.35	0.56	0.45	0.50	0.35	0.50
[1,100]	0.70	0.42	0.37	0.31	0.49	0.41	0.44	0.39	0.32	0.52	0.49	0.34	0.54	0.43	0.50	0.35	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Caso de teste Sr

Tabela A.11: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Sr. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.80	0.60	0.40	0.70	0.70	0.80	0.50	0.70	0.90	0.60	0.70	0.70	0.40	0.70	0.70	0.70
[1,20]	1	0.70	0.60	0.40	0.70	0.70	0.75	0.50	0.70	0.95	0.75	0.70	0.70	0.40	0.75	0.70	0.70
[1,30]	0.96	0.60	0.60	0.38	0.70	0.70	0.70	0.46	0.66	0.90	0.70	0.66	0.66	0.40	0.76	0.68	0.68
[1,40]	0.94	0.60	0.58	0.36	0.70	0.68	0.70	0.47	0.65	0.87	0.70	0.65	0.50	0.37	0.80	0.67	0.67
[1,50]	0.92	0.66	0.56	0.33	0.70	0.68	0.70	0.48	0.66	0.90	0.70	0.65	0.50	0.40	0.80	0.68	0.65
[1,60]	0.92	0.65	0.56	0.32	0.70	0.65	0.65	0.50	0.63	0.88	0.65	0.64	0.50	0.40	0.80	0.64	0.65
[1,70]	0.90	0.65	0.48	0.30	0.70	0.64	0.65	0.48	0.64	0.90	0.65	0.64	0.50	0.38	0.78	0.60	0.65
[1,80]	0.90	0.65	0.42	0.30	0.70	0.64	0.65	0.42	0.63	0.90	0.67	0.61	0.48	0.38	0.77	0.54	0.63
[1,90]	0.88	0.64	0.38	0.30	0.70	0.62	0.65	0.37	0.62	0.88	0.60	0.56	0.46	0.40	0.76	0.50	0.60
[1,100]	0.88	0.66	0.37	0.30	0.71	0.60	0.61	0.36	0.61	0.84	0.58	0.55	0.44	0.41	0.75	0.48	0.60
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.12: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Sr para o montador Trinity. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.80	0.60	0.40	0.70	0.60	0.80	0.50	0.60	0.90	0.60	0.60	0.70	0.40	0.70	0.70	0.70
[1,20]	1	0.70	0.60	0.40	0.70	0.60	0.75	0.50	0.60	0.95	0.75	0.60	0.70	0.40	0.65	0.70	0.70
[1,30]	0.96	0.60	0.60	0.37	0.70	0.60	0.70	0.46	0.60	0.90	0.70	0.58	0.66	0.40	0.66	0.68	0.68
[1,40]	0.94	0.60	0.58	0.36	0.70	0.58	0.70	0.47	0.55	0.87	0.70	0.57	0.50	0.37	0.60	0.67	0.67
[1,50]	0.92	0.66	0.56	0.33	0.70	0.57	0.70	0.48	0.53	0.90	0.70	0.55	0.50	0.40	0.60	0.68	0.65
[1,60]	0.91	0.65	0.56	0.32	0.70	0.54	0.65	0.50	0.52	0.88	0.65	0.53	0.50	0.40	0.58	0.64	0.65
[1,70]	0.90	0.65	0.48	0.30	0.70	0.53	0.65	0.48	0.51	0.90	0.65	0.52	0.50	0.38	0.58	0.60	0.63
[1,80]	0.90	0.65	0.42	0.30	0.70	0.52	0.65	0.42	0.50	0.90	0.67	0.52	0.48	0.38	0.57	0.54	0.63
[1,90]	0.88	0.64	0.38	0.30	0.70	0.50	0.65	0.37	0.50	0.88	0.60	0.51	0.46	0.40	0.55	0.50	0.60
[1,100]	0.88	0.66	0.37	0.30	0.71	0.50	0.61	0.36	0.50	0.84	0.58	0.49	0.44	0.41	0.53	0.48	0.60
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.13: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Sr para o montador Velvet-Oases. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.90	0.70	0.50	0.30	0.60	0.70	0.70	0.50	0.50	0.70	0.50	0.70	0.50	0.40	0.70	0.60	0.60
[1,20]	0.85	0.60	0.50	0.30	0.60	0.70	0.65	0.45	0.50	0.70	0.45	0.70	0.50	0.40	0.75	0.60	0.60
[1,30]	0.84	0.60	0.48	0.27	0.55	0.70	0.60	0.46	0.54	0.66	0.43	0.66	0.50	0.38	0.76	0.57	0.58
[1,40]	0.82	0.55	0.47	0.25	0.56	0.70	0.57	0.44	0.52	0.65	0.42	0.65	0.48	0.36	0.80	0.55	0.56
[1,50]	0.80	0.53	0.46	0.24	0.55	0.68	0.55	0.42	0.55	0.63	0.41	0.65	0.47	0.35	0.80	0.53	0.55
[1,60]	0.80	0.52	0.45	0.22	0.53	0.65	0.55	0.40	0.53	0.62	0.40	0.64	0.43	0.33	0.80	0.52	0.55
[1,70]	0.80	0.52	0.40	0.22	0.50	0.64	0.53	0.40	0.52	0.60	0.40	0.64	0.42	0.30	0.78	0.50	0.53
[1,80]	0.80	0.51	0.38	0.21	0.50	0.64	0.52	0.40	0.50	0.60	0.40	0.61	0.41	0.30	0.77	0.50	0.51
[1,90]	0.80	0.50	0.36	0.20	0.48	0.62	0.50	0.36	0.50	0.50	0.40	0.56	0.40	0.30	0.76	0.48	0.50
[1,100]	0.80	0.50	0.35	0.20	0.47	0.60	0.50	0.36	0.50	0.50	0.40	0.55	0.40	0.29	0.75	0.47	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.14: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Sr para o montador IDBA-Tran. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.80	0.70	0.60	0.40	0.50	0.60	0.60	0.50	0.70	0.60	0.50	0.60	0.60	0.40	0.60	0.50	0.60
[1,20]	0.80	0.70	0.50	0.40	0.50	0.60	0.60	0.50	0.70	0.60	0.55	0.55	0.60	0.40	0.60	0.50	0.60
[1,30]	0.80	0.68	0.50	0.38	0.50	0.57	0.60	0.46	0.66	0.60	0.50	0.53	0.50	0.37	0.56	0.50	0.60
[1,40]	0.78	0.65	0.50	0.33	0.48	0.53	0.55	0.45	0.65	0.55	0.47	0.52	0.50	0.35	0.50	0.47	0.60
[1,50]	0.75	0.63	0.48	0.30	0.46	0.51	0.53	0.42	0.66	0.53	0.46	0.50	0.48	0.33	0.50	0.45	0.58
[1,60]	0.74	0.60	0.46	0.27	0.44	0.50	0.50	0.40	0.63	0.52	0.43	0.50	0.45	0.33	0.50	0.43	0.55
[1,70]	0.73	0.60	0.42	0.25	0.42	0.48	0.50	0.40	0.64	0.50	0.42	0.50	0.40	0.31	0.48	0.42	0.54
[1,80]	0.70	0.58	0.38	0.23	0.40	0.46	0.48	0.40	0.63	0.50	0.40	0.47	0.40	0.30	0.47	0.41	0.52
[1,90]	0.70	0.57	0.35	0.22	0.40	0.44	0.46	0.38	0.62	0.46	0.40	0.46	0.40	0.30	0.46	0.40	0.50
[1,100]	0.70	0.56	0.33	0.21	0.41	0.42	0.44	0.36	0.61	0.45	0.40	0.45	0.41	0.31	0.44	0.40	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

A.3.2 Canis familiaris

Caso de teste Tr

Tabela A.15: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.70	0.40	0.40	0.80	0.60	1.00	0.50	0.60	1.00	0.40	0.60	0.60	0.50	0.50	0.50	0.50
[1,20]	1	0.70	0.40	0.40	0.75	0.60	0.90	0.50	0.60	0.95	0.40	0.60	0.65	0.60	0.50	0.50	0.50
[1,30]	1	0.66	0.38	0.40	0.80	0.58	0.80	0.48	0.59	0.90	0.40	0.56	0.63	0.60	0.48	0.46	0.50
[1,40]	1	0.70	0.37	0.37	0.80	0.57	0.80	0.47	0.58	0.88	0.37	0.55	0.57	0.55	0.47	0.47	0.50
[1,50]	0.95	0.72	0.32	0.38	0.78	0.55	0.75	0.47	0.57	0.85	0.36	0.50	0.56	0.62	0.47	0.47	0.48
[1,60]	0.94	0.70	0.31	0.36	0.66	0.50	0.80	0.45	0.55	0.84	0.33	0.51	0.55	0.60	0.43	0.46	0.46
[1,70]	0.92	0.70	0.30	0.38	0.60	0.47	0.68	0.46	0.54	0.82	0.30	0.48	0.51	0.55	0.42	0.44	0.45
[1,80]	0.92	0.69	0.30	0.40	0.58	0.47	0.60	0.44	0.52	0.81	0.27	0.47	0.50	0.55	0.43	0.43	0.45
[1,90]	0.90	0.64	0.29	0.41	0.54	0.44	0.55	0.43	0.50	0.80	0.25	0.48	0.50	0.52	0.42	0.44	0.43
[1,100]	0.90	0.63	0.28	0.40	0.50	0.44	0.52	0.43	0.48	0.77	0.23	0.48	0.50	0.50	0.40	0.45	0.42
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.16: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr para o montador Trinity. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.70	0.40	0.40	0.70	0.60	1.00	0.50	0.60	1.00	0.40	0.60	0.60	0.50	0.50	0.50	0.50
[1,20]	1	0.70	0.40	0.40	0.65	0.58	0.90	0.50	0.60	0.95	0.38	0.60	0.65	0.60	0.50	0.50	0.50
[1,30]	1	0.66	0.38	0.40	0.62	0.56	0.80	0.48	0.59	0.90	0.36	0.55	0.63	0.60	0.48	0.46	0.50
[1,40]	1	0.70	0.37	0.37	0.60	0.55	0.80	0.45	0.58	0.88	0.35	0.55	0.57	0.55	0.47	0.47	0.50
[1,50]	0.95	0.72	0.32	0.38	0.58	0.53	0.75	0.45	0.57	0.85	0.33	0.50	0.56	0.62	0.47	0.47	0.48
[1,60]	0.94	0.70	0.31	0.36	0.57	0.51	0.80	0.45	0.55	0.84	0.33	0.47	0.55	0.60	0.43	0.46	0.46
[1,70]	0.92	0.70	0.30	0.38	0.54	0.50	0.68	0.43	0.54	0.82	0.31	0.46	0.51	0.55	0.42	0.44	0.45
[1,80]	0.92	0.69	0.30	0.40	0.52	0.50	0.60	0.42	0.52	0.81	0.30	0.45	0.50	0.55	0.43	0.43	0.45
[1,90]	0.90	0.64	0.29	0.41	0.50	0.47	0.55	0.40	0.50	0.80	0.28	0.44	0.50	0.52	0.42	0.44	0.43
[1,100]	0.90	0.63	0.28	0.40	0.50	0.46	0.52	0.40	0.48	0.77	0.26	0.44	0.50	0.50	0.40	0.45	0.42
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.17: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr para o montador Velvet-Oases. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.90	0.60	0.30	0.30	0.80	0.60	0.50	0.50	0.50	0.90	0.40	0.60	0.60	0.60	0.50	0.50	0.50
[1,20]	0.90	0.60	0.30	0.30	0.75	0.60	0.50	0.50	0.50	0.85	0.40	0.60	0.65	0.55	0.50	0.45	0.50
[1,30]	0.90	0.60	0.30	0.28	0.80	0.58	0.50	0.48	0.49	0.83	0.40	0.56	0.63	0.55	0.46	0.44	0.48
[1,40]	0.85	0.58	0.27	0.27	0.80	0.57	0.48	0.46	0.48	0.82	0.37	0.55	0.57	0.50	0.45	0.43	0.46
[1,50]	0.85	0.56	0.26	0.26	0.78	0.55	0.46	0.45	0.46	0.80	0.36	0.50	0.55	0.48	0.44	0.42	0.45
[1,60]	0.85	0.55	0.25	0.26	0.66	0.50	0.45	0.44	0.44	0.80	0.33	0.51	0.54	0.46	0.42	0.41	0.45
[1,70]	0.82	0.52	0.23	0.25	0.60	0.47	0.45	0.42	0.40	0.78	0.30	0.48	0.52	0.45	0.40	0.40	0.42
[1,80]	0.80	0.50	0.21	0.25	0.58	0.47	0.42	0.40	0.40	0.73	0.27	0.47	0.48	0.44	0.40	0.40	0.40
[1,90]	0.78	0.48	0.20	0.22	0.54	0.44	0.42	0.40	0.38	0.72	0.25	0.48	0.47	0.42	0.40	0.42	0.40
[1,100]	0.78	0.46	0.20	0.20	0.50	0.44	0.40	0.40	0.38	0.70	0.23	0.48	0.46	0.40	0.38	0.42	0.40
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.18: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Tr para o montador IDBA-Tran. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.80	0.50	0.30	0.40	0.60	0.50	0.70	0.40	0.50	0.60	0.40	0.50	0.40	0.50	0.50	0.40	0.50
[1,20]	0.80	0.50	0.25	0.35	0.60	0.50	0.70	0.40	0.50	0.60	0.35	0.50	0.40	0.50	0.45	0.40	0.47
[1,30]	0.78	0.46	0.24	0.34	0.58	0.48	0.70	0.40	0.50	0.58	0.33	0.50	0.40	0.48	0.46	0.38	0.45
[1,40]	0.77	0.45	0.22	0.32	0.57	0.46	0.67	0.38	0.48	0.57	0.30	0.45	0.38	0.48	0.45	0.37	0.45
[1,50]	0.75	0.42	0.20	0.30	0.55	0.45	0.65	0.37	0.46	0.55	0.28	0.42	0.36	0.46	0.43	0.36	0.43
[1,60]	0.75	0.42	0.20	0.30	0.54	0.45	0.60	0.36	0.45	0.54	0.27	0.41	0.35	0.43	0.42	0.33	0.42
[1,70]	0.73	0.40	0.20	0.30	0.52	0.44	0.58	0.35	0.45	0.52	0.26	0.40	0.33	0.42	0.40	0.33	0.40
[1,80]	0.72	0.40	0.19	0.28	0.50	0.40	0.56	0.32	0.44	0.52	0.22	0.40	0.30	0.40	0.40	0.30	0.40
[1,90]	0.70	0.36	0.17	0.27	0.48	0.40	0.52	0.30	0.42	0.51	0.22	0.38	0.28	0.40	0.40	0.28	0.40
[1,100]	0.70	0.35	0.16	0.25	0.45	0.40	0.50	0.29	0.40	0.50	0.21	0.36	0.28	0.40	0.38	0.30	0.40
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Caso de teste Mr

Tabela A.19: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Mr. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.60	0.60	0.40	0.70	0.50	0.80	0.50	0.50	0.90	0.50	0.50	0.40	0.40	0.70	0.60	0.60
[1,20]	0.95	0.60	0.40	0.40	0.80	0.60	0.80	0.45	0.50	0.90	0.50	0.50	0.40	0.40	0.68	0.60	0.60
[1,30]	0.93	0.50	0.40	0.40	0.76	0.60	0.76	0.42	0.46	0.86	0.50	0.48	0.35	0.38	0.67	0.58	0.58
[1,40]	0.90	0.60	0.40	0.37	0.72	0.54	0.70	0.40	0.45	0.85	0.47	0.47	0.35	0.37	0.65	0.57	0.56
[1,50]	0.90	0.60	0.40	0.38	0.72	0.54	0.62	0.42	0.50	0.84	0.46	0.46	0.34	0.36	0.64	0.57	0.55
[1,60]	0.90	0.60	0.33	0.36	0.70	0.52	0.58	0.40	0.46	0.81	0.45	0.44	0.28	0.31	0.62	0.56	0.55
[1,70]	0.90	0.55	0.28	0.38	0.70	0.50	0.57	0.41	0.45	0.80	0.42	0.42	0.24	0.27	0.60	0.54	0.52
[1,80]	0.85	0.55	0.25	0.40	0.69	0.50	0.57	0.42	0.47	0.80	0.41	0.40	0.21	0.27	0.60	0.52	0.52
[1,90]	0.85	0.57	0.24	0.41	0.66	0.46	0.54	0.43	0.47	0.80	0.40	0.40	0.22	0.28	0.58	0.50	0.50
[1,100]	0.85	0.60	0.26	0.40	0.67	0.43	0.53	0.45	0.48	0.79	0.42	0.41	0.25	0.30	0.56	0.48	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.20: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Mr para o montador Trinity. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.60	0.50	0.40	0.70	0.50	0.80	0.40	0.50	0.90	0.50	0.40	0.40	0.40	0.70	0.60	0.60
[1,20]	0.95	0.60	0.40	0.40	0.80	0.60	0.80	0.40	0.50	0.90	0.50	0.40	0.35	0.40	0.68	0.60	0.60
[1,30]	0.93	0.50	0.38	0.40	0.76	0.60	0.76	0.38	0.46	0.86	0.50	0.40	0.35	0.38	0.67	0.58	0.58
[1,40]	0.90	0.60	0.36	0.35	0.72	0.54	0.70	0.36	0.45	0.85	0.47	0.38	0.35	0.37	0.65	0.57	0.56
[1,50]	0.90	0.60	0.35	0.34	0.72	0.54	0.62	0.35	0.50	0.84	0.46	0.36	0.34	0.36	0.64	0.57	0.55
[1,60]	0.90	0.60	0.32	0.33	0.70	0.52	0.58	0.32	0.46	0.81	0.45	0.34	0.28	0.31	0.62	0.56	0.55
[1,70]	0.90	0.55	0.27	0.31	0.70	0.50	0.57	0.31	0.45	0.80	0.42	0.32	0.24	0.27	0.60	0.54	0.52
[1,80]	0.85	0.55	0.24	0.30	0.69	0.50	0.57	0.30	0.47	0.80	0.41	0.30	0.21	0.27	0.60	0.52	0.52
[1,90]	0.85	0.57	0.24	0.28	0.66	0.46	0.54	0.30	0.47	0.80	0.40	0.30	0.22	0.28	0.58	0.50	0.50
[1,100]	0.85	0.60	0.25	0.28	0.67	0.43	0.53	0.30	0.48	0.79	0.42	0.30	0.25	0.30	0.56	0.48	0.50
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.21: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Mr para o montador Velvet-Oases. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.80	0.50	0.60	0.40	0.60	0.50	0.70	0.50	0.40	0.60	0.40	0.50	0.40	0.40	0.60	0.50	0.60
[1,20]	0.80	0.50	0.40	0.40	0.60	0.45	0.70	0.45	0.40	0.60	0.40	0.50	0.40	0.35	0.60	0.50	0.57
[1,30]	0.78	0.50	0.40	0.40	0.58	0.45	0.65	0.42	0.40	0.58	0.40	0.48	0.35	0.35	0.58	0.50	0.55
[1,40]	0.78	0.48	0.40	0.37	0.57	0.43	0.62	0.40	0.38	0.57	0.38	0.47	0.35	0.32	0.57	0.48	0.55
[1,50]	0.75	0.47	0.40	0.38	0.55	0.42	0.60	0.42	0.37	0.54	0.37	0.46	0.34	0.31	0.55	0.47	0.53
[1,60]	0.74	0.46	0.33	0.36	0.54	0.42	0.57	0.40	0.35	0.53	0.35	0.44	0.28	0.31	0.53	0.46	0.52
[1,70]	0.72	0.45	0.28	0.38	0.53	0.40	0.55	0.41	0.35	0.52	0.33	0.42	0.24	0.30	0.50	0.45	0.50
[1,80]	0.72	0.43	0.25	0.40	0.50	0.40	0.54	0.42	0.32	0.50	0.30	0.40	0.21	0.30	0.50	0.43	0.50
[1,90]	0.70	0.45	0.24	0.41	0.47	0.38	0.53	0.43	0.31	0.50	0.30	0.40	0.22	0.28	0.50	0.40	0.48
[1,100]	0.70	0.43	0.26	0.40	0.46	0.36	0.52	0.45	0.30	0.50	0.30	0.41	0.25	0.28	0.50	0.40	0.48
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.22: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Mr para o montador IDBA-Tran. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.70	0.40	0.40	0.30	0.60	0.50	0.60	0.30	0.40	0.60	0.40	0.50	0.40	0.40	0.50	0.50	0.60
[1,20]	0.70	0.40	0.35	0.30	0.60	0.50	0.60	0.35	0.40	0.60	0.45	0.48	0.38	0.38	0.50	0.50	0.57
[1,30]	0.67	0.40	0.35	0.30	0.58	0.48	0.58	0.36	0.40	0.58	0.43	0.47	0.35	0.37	0.50	0.50	0.55
[1,40]	0.66	0.38	0.33	0.28	0.55	0.47	0.56	0.35	0.37	0.57	0.42	0.45	0.33	0.36	0.48	0.48	0.50
[1,50]	0.65	0.37	0.32	0.27	0.53	0.45	0.55	0.35	0.35	0.55	0.40	0.42	0.32	0.35	0.47	0.48	0.50
[1,60]	0.62	0.35	0.31	0.25	0.50	0.42	0.55	0.33	0.32	0.53	0.40	0.40	0.30	0.30	0.46	0.46	0.47
[1,70]	0.60	0.33	0.30	0.24	0.50	0.40	0.52	0.32	0.31	0.52	0.38	0.38	0.30	0.27	0.45	0.44	0.45
[1,80]	0.60	0.32	0.28	0.23	0.49	0.40	0.50	0.31	0.30	0.51	0.36	0.36	0.28	0.26	0.44	0.42	0.43
[1,90]	0.58	0.31	0.27	0.20	0.47	0.40	0.50	0.30	0.30	0.50	0.35	0.35	0.26	0.26	0.43	0.40	0.40
[1,100]	0.57	0.30	0.26	0.20	0.47	0.38	0.50	0.30	0.30	0.49	0.32	0.32	0.25	0.25	0.42	0.40	0.40
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Caso de teste Sr

Tabela A.23: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Sr. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.70	0.50	0.50	0.80	0.60	0.60	0.70	0.50	0.90	0.70	0.60	0.70	0.60	0.90	0.80	0.50
[1,20]	0.96	0.65	0.50	0.50	0.80	0.60	0.60	0.60	0.50	0.90	0.70	0.65	0.70	0.65	0.80	0.80	0.50
[1,30]	0.95	0.62	0.48	0.48	0.78	0.58	0.58	0.58	0.47	0.90	0.68	0.60	0.68	0.63	0.70	0.80	0.50
[1,40]	0.95	0.62	0.48	0.48	0.77	0.57	0.57	0.56	0.48	0.90	0.68	0.58	0.67	0.57	0.70	0.77	0.48
[1,50]	0.93	0.60	0.46	0.46	0.75	0.56	0.56	0.53	0.47	0.95	0.65	0.57	0.66	0.56	0.64	0.74	0.47
[1,60]	0.91	0.60	0.45	0.46	0.74	0.54	0.55	0.48	0.45	0.90	0.64	0.56	0.65	0.55	0.60	0.71	0.46
[1,70]	0.90	0.59	0.43	0.45	0.70	0.52	0.54	0.48	0.42	0.90	0.64	0.55	0.64	0.51	0.58	0.71	0.45
[1,80]	0.90	0.58	0.42	0.44	0.69	0.52	0.52	0.46	0.40	0.90	0.62	0.54	0.62	0.50	0.57	0.68	0.45
[1,90]	0.90	0.55	0.42	0.42	0.67	0.50	0.50	0.45	0.38	0.87	0.57	0.53	0.60	0.50	0.55	0.66	0.45
[1,100]	0.90	0.53	0.42	0.41	0.63	0.50	0.50	0.44	0.40	0.86	0.55	0.51	0.58	0.50	0.55	0.67	0.42
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.24: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Sr para o montador Trinity. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	1	0.70	0.50	0.50	0.70	0.60	0.60	0.60	0.50	0.90	0.70	0.60	0.70	0.60	0.90	0.80	0.50
[1,20]	0.96	0.65	0.50	0.50	0.70	0.60	0.60	0.55	0.50	0.90	0.70	0.65	0.70	0.55	0.80	0.80	0.50
[1,30]	0.95	0.62	0.48	0.48	0.65	0.58	0.58	0.50	0.47	0.90	0.68	0.60	0.68	0.53	0.70	0.80	0.50
[1,40]	0.95	0.62	0.48	0.48	0.63	0.57	0.57	0.48	0.48	0.90	0.68	0.58	0.67	0.52	0.70	0.77	0.48
[1,50]	0.93	0.60	0.46	0.46	0.62	0.56	0.56	0.47	0.47	0.95	0.65	0.57	0.66	0.50	0.64	0.74	0.47
[1,60]	0.91	0.60	0.45	0.46	0.60	0.54	0.55	0.45	0.45	0.90	0.64	0.56	0.65	0.50	0.60	0.71	0.46
[1,70]	0.90	0.59	0.43	0.45	0.60	0.52	0.54	0.45	0.42	0.90	0.64	0.55	0.64	0.50	0.58	0.71	0.45
[1,80]	0.90	0.58	0.42	0.44	0.55	0.52	0.52	0.42	0.40	0.90	0.62	0.54	0.62	0.48	0.57	0.68	0.45
[1,90]	0.90	0.55	0.42	0.42	0.53	0.50	0.50	0.42	0.38	0.87	0.57	0.53	0.60	0.47	0.55	0.66	0.45
[1,100]	0.90	0.53	0.42	0.41	0.51	0.50	0.50	0.40	0.40	0.86	0.55	0.51	0.58	0.46	0.55	0.67	0.42
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.25: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Sr para o montador Velvet-Oases. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.80	0.60	0.40	0.40	0.70	0.50	0.50	0.70	0.40	0.70	0.60	0.50	0.50	0.50	0.60	0.60	0.50
[1,20]	0.80	0.60	0.40	0.40	0.70	0.50	0.50	0.60	0.40	0.70	0.60	0.50	0.50	0.45	0.60	0.60	0.47
[1,30]	0.78	0.58	0.40	0.39	0.68	0.48	0.50	0.58	0.40	0.68	0.58	0.47	0.47	0.43	0.60	0.56	0.46
[1,40]	0.76	0.57	0.38	0.38	0.66	0.47	0.48	0.56	0.38	0.67	0.57	0.46	0.45	0.40	0.57	0.55	0.45
[1,50]	0.75	0.56	0.36	0.37	0.65	0.45	0.46	0.53	0.36	0.65	0.55	0.45	0.45	0.38	0.56	0.54	0.45
[1,60]	0.73	0.55	0.33	0.35	0.63	0.43	0.45	0.48	0.35	0.60	0.54	0.43	0.44	0.37	0.55	0.52	0.45
[1,70]	0.72	0.53	0.32	0.33	0.60	0.42	0.44	0.48	0.33	0.60	0.53	0.41	0.43	0.36	0.53	0.51	0.43
[1,80]	0.70	0.52	0.31	0.32	0.60	0.40	0.42	0.46	0.31	0.60	0.52	0.40	0.42	0.35	0.52	0.50	0.40
[1,90]	0.70	0.50	0.30	0.32	0.58	0.40	0.40	0.45	0.30	0.58	0.50	0.40	0.40	0.33	0.51	0.50	0.40
[1,100]	0.70	0.50	0.30	0.30	0.57	0.40	0.40	0.44	0.28	0.56	0.50	0.40	0.40	0.30	0.50	0.50	0.40
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabela A.26: Aplicação do teste pareado de Wilcoxon, sobre cada critério $C_i, 1 \leq i \leq 17$, no Caso de teste Sr para o montador IDBA-Tran. A comparação feita é a do Critério 17 (C_{17}) em relação aos demais. O X indica que há significância estatística na aplicação do teste, e o - indica que não há significância.

	C_{17}	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
[1,10]	0.70	0.50	0.40	0.40	0.80	0.60	0.60	0.70	0.50	0.70	0.50	0.50	0.60	0.50	0.60	0.60	0.40
[1,20]	0.70	0.50	0.40	0.40	0.80	0.55	0.60	0.60	0.45	0.70	0.50	0.50	0.60	0.50	0.60	0.60	0.40
[1,30]	0.69	0.48	0.38	0.40	0.78	0.53	0.55	0.58	0.43	0.70	0.50	0.47	0.58	0.46	0.60	0.50	0.37
[1,40]	0.68	0.45	0.37	0.40	0.77	0.52	0.54	0.56	0.42	0.65	0.48	0.46	0.56	0.47	0.58	0.47	0.36
[1,50]	0.65	0.43	0.38	0.38	0.75	0.50	0.52	0.53	0.40	0.65	0.47	0.45	0.54	0.46	0.55	0.44	0.35
[1,60]	0.65	0.42	0.34	0.37	0.74	0.50	0.50	0.48	0.40	0.63	0.45	0.44	0.53	0.45	0.53	0.43	0.35
[1,70]	0.65	0.40	0.32	0.35	0.70	0.48	0.50	0.48	0.38	0.61	0.43	0.41	0.51	0.44	0.52	0.41	0.35
[1,80]	0.65	0.40	0.30	0.34	0.69	0.44	0.47	0.46	0.35	0.60	0.42	0.40	0.50	0.43	0.51	0.40	0.35
[1,90]	0.62	0.40	0.30	0.32	0.67	0.40	0.45	0.45	0.33	0.58	0.40	0.40	0.50	0.41	0.50	0.40	0.34
[1,100]	0.61	0.39	0.30	0.30	0.63	0.40	0.42	0.44	0.31	0.56	0.40	0.40	0.50	0.40	0.50	0.40	0.33
Wilcoxon		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

A.4 Artigo extra publicado

O aluno Lucas Miguel de Carvalho fez parte do desenvolvimento, no ano de 2013, de alguns módulos de uma plataforma integrada com uma interface baseada em web para a anotação, análise e visualização dos perfis de interação de genes/proteínas, metabólitos e drogas de interesse chamada de IIS (Integrated Interactome System). O artigo resultante deste trabalho foi publicado na revista PLOS ONE [5].