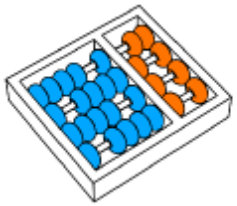
Avaliação de montadores *de novo* de RNA-Seq para análise de expressão diferencial de transcritos

Lucas Miguel de Carvalho

Orientador: Zanoni Dias
Coorientador: Felipe Rodrigues da Silva

Instituto de Computação - Unicamp
Laboratório Multiusuário de Bioinformática - Embrapa Informática Agropecuária



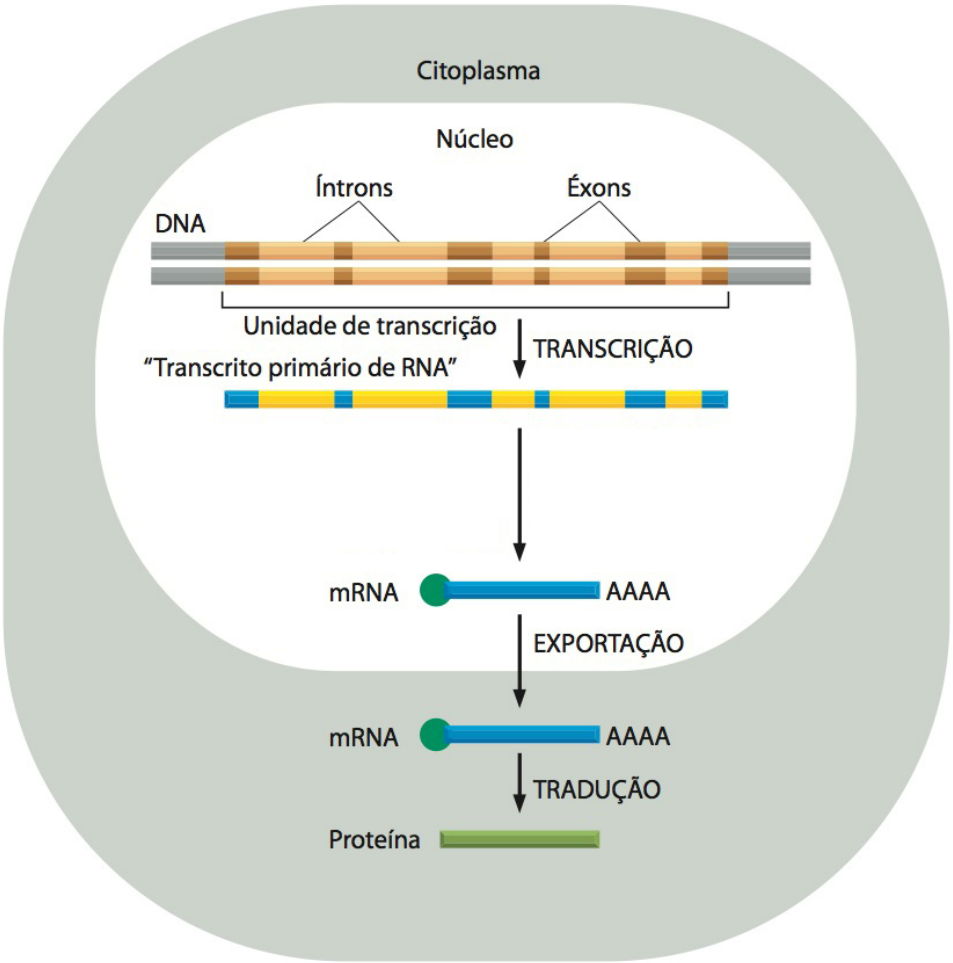
Embrapa
Informática Agropecuária

Laboratório Multiusuário de Bioinformática

10 de novembro de 2015

Roteiro

Introdução



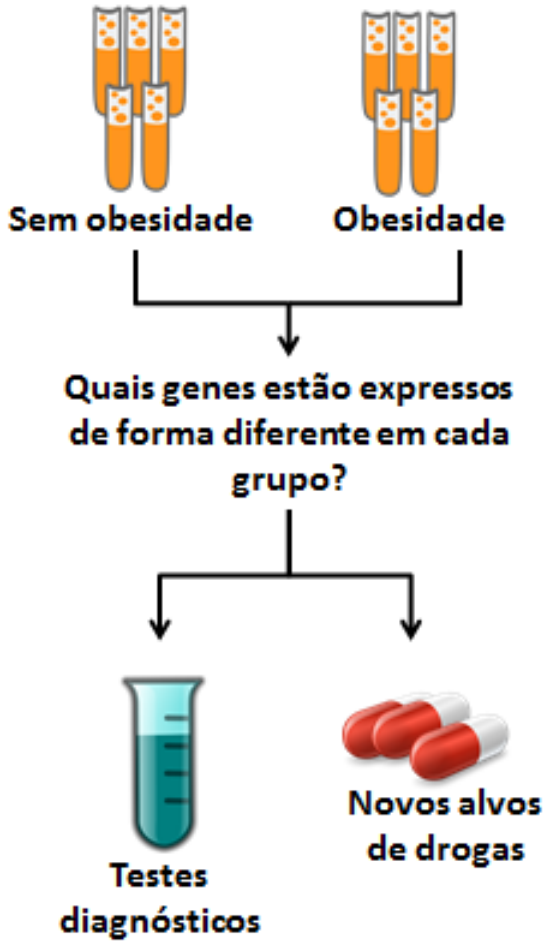


Figura: Dogma central da biologia. Figura extraída de "Biologia Molecular da Célula". Bruce Alberts, 2010.

Introdução

Transcriptômica

O transcriptoma é o conjunto completo de transcritos em uma célula. Seu estudo é denominado transcriptômica.

- Os principais objetivos da transcriptômica são:
  - Catalogar todas as espécies de transcrição, incluindo mRNAs, RNAs não-codificantes e pequenos RNAs;
  - Determinar a estrutura da transcrição de genes, em termos dos seus locais de início, 5' e 3', padrões de splicing e de outras modificações pós-transcricionais;
  - Quantificar alteração dos níveis de expressão em cada transcrição sob diferentes condições.

Introdução

- Várias tecnologias foram desenvolvidas para quantificar e deduzir o transcriptoma, incluindo tecnologias que utilizam hibridação ou abordagens baseadas em sequência.
  - Hibridação: Microarray
  - Baseado em sequência: SAGE e CAGE.
- Desvantagens do Microarray:
  - Dependência de um conhecimento existente sobre sequência do genoma;
  - Saturação de sinais.
- Desvantagens de abordagens baseadas em sequências:
  - Apenas uma porção dos transcritos são analisados.
  - Alto custo monetário;

Introdução

- Recentemente, o desenvolvimento de novos métodos de sequenciamento de alto rendimento de DNA forneceu um novo método tanto para o mapeamento quanto para a quantificação do transcriptoma.
- Este método, denominado RNA-Seq, tem vantagens claras sobre as abordagens existentes e vem revolucionando a maneira em que são analisados os transcriptomas eucarióticos e procarióticos.

RNA-Seq - Conceito e metodologia

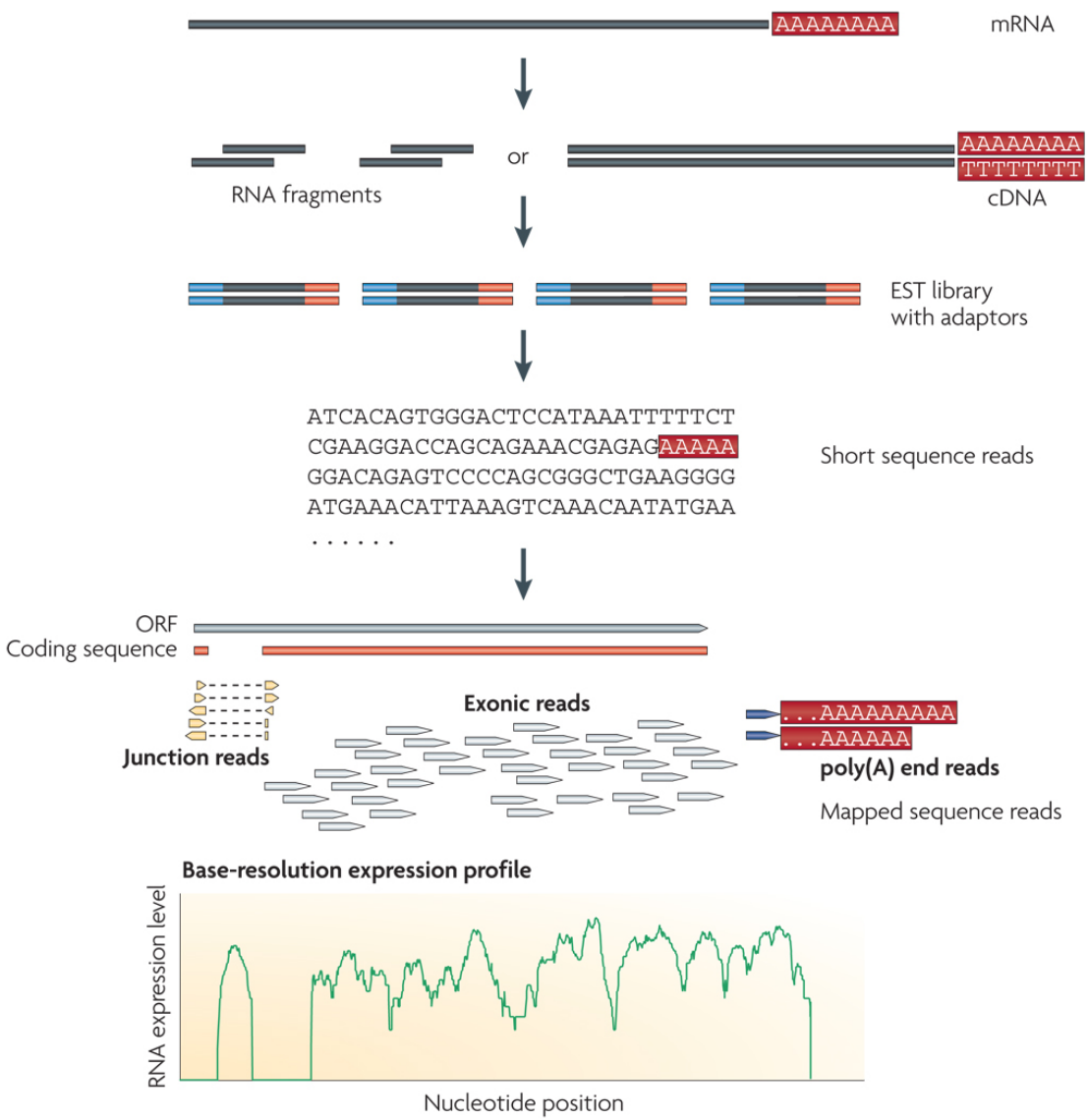


Figura: Experimento de RNA-Seq. Figura extraída de Wang et al., 2009.

RNA-Seq - Conceito e metodologia

- Vantagens do uso do RNA-Seq:
  - Não está limitado a encontrar transcritos descritos somente na sequência genômica. Isso faz com que o RNA-Seq seja atraente para organismos não modelos;
  - Possibilita a análise da conectividade de múltiplos éxons com sequenciamentos paired-end ou de reads longas, favorecendo o estudo de transcriptomas complexos;
  - Possui níveis elevados de reprodutibilidade, tanto para replicatas técnicas ou biológicas;
  - Como não possui etapas de clonagem, a tecnologia RNA-Seq requer um volume menor de mRNA;
  - Menor custo monetário.

RNA-Seq - Conceito e metodologia

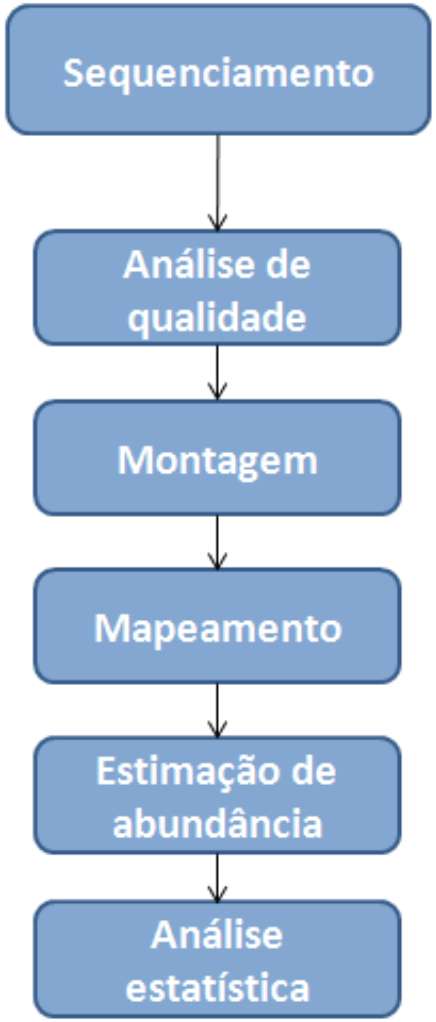
- Desafios em RNA-Seq:
  - Reconstrução do transcriptoma;
  - Armazenamento dos dados;
  - Análises computacionais intensivas;
  - Preparação da biblioteca de RNA degradado.

RNA-Seq - Conceito e metodologia

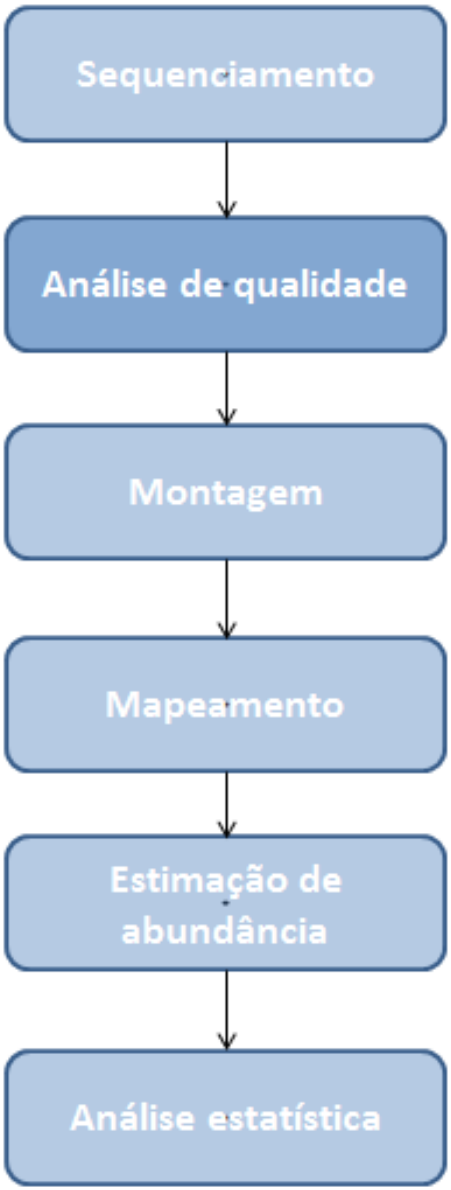
- Utilização de estudos de RNA-Seq:
  - Detecção de microsatélites;
  - Detecção de SNPs;
  - Montagem do transcriptoma;
  - Encontrar novos genes;
  - Análise de variantes de splicing;
  - **Encontrar genes ou transcritos diferencialmente expressos entre condições (estudos funcionais)**.

RNA-Seq - Conceito e metodologia

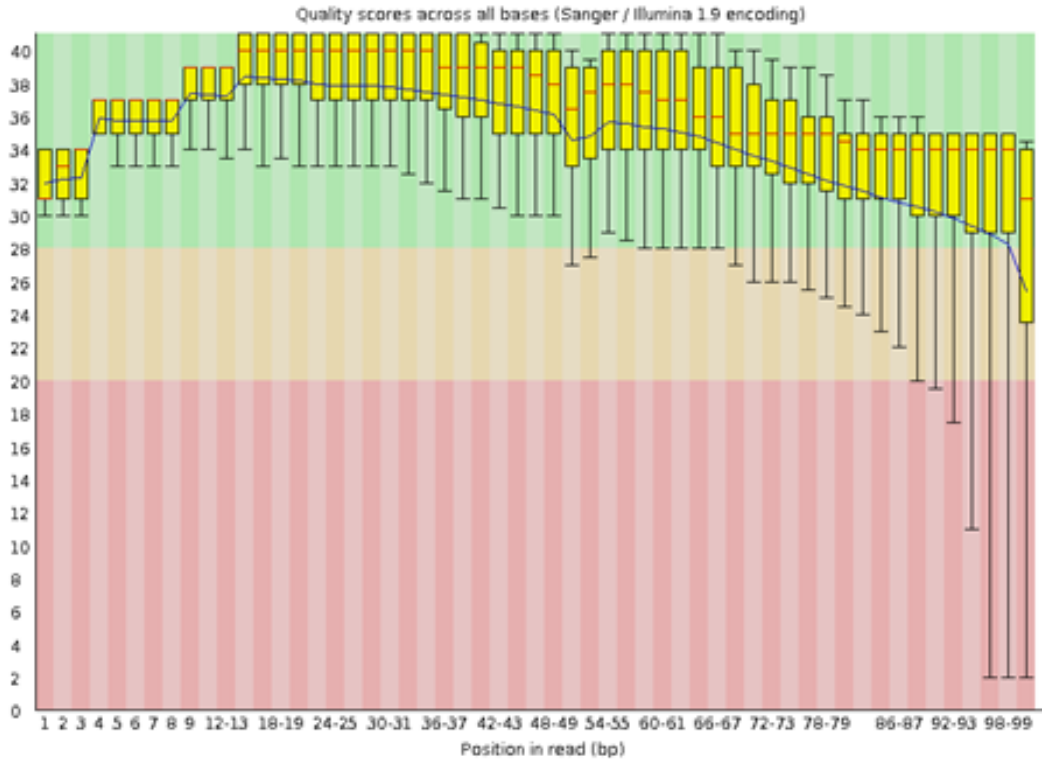
- Fluxograma de análise de RNA-Seq para transcritos diferencialmente expressos



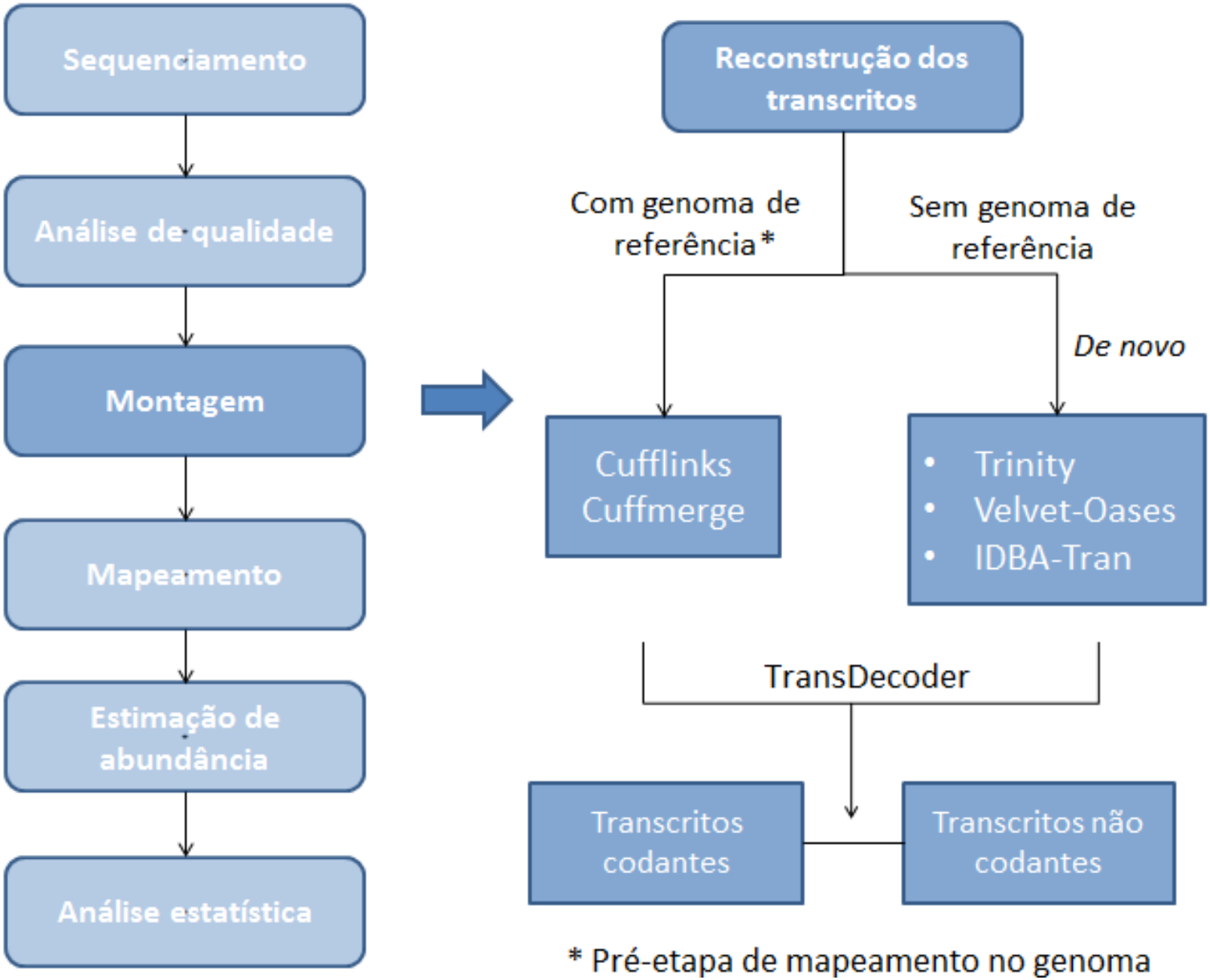
RNA-Seq - Conceito e metodologia



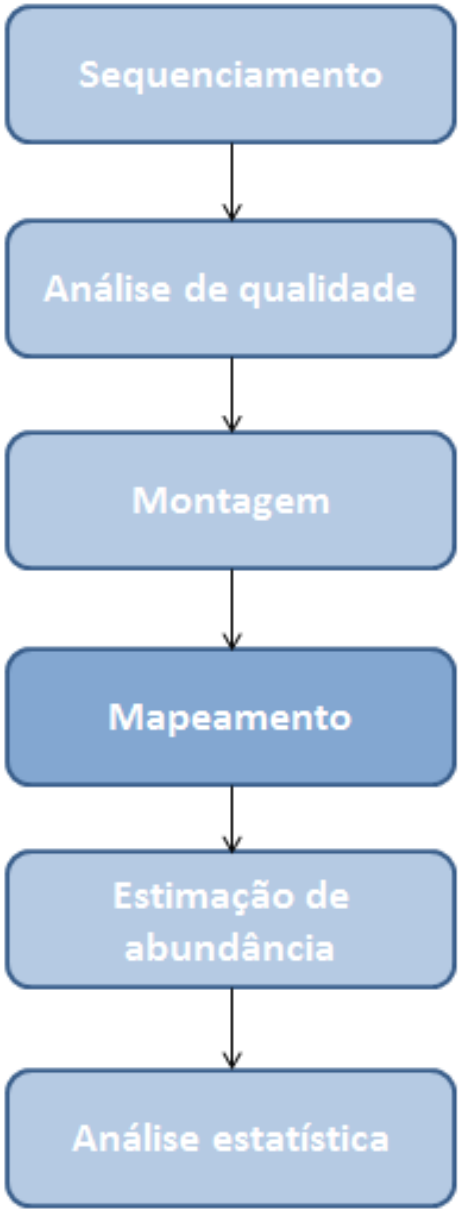
- Eliminação de adaptadores;
- Eliminação de ribossomais;
- Verificação da qualidade das bases e sequenciamento (**FASTQC**).

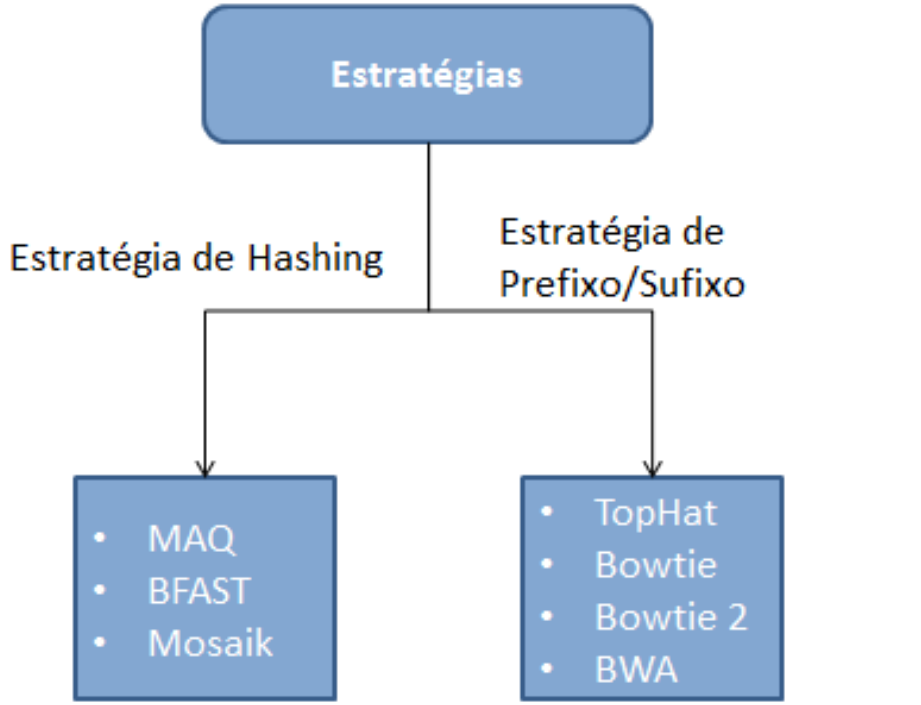


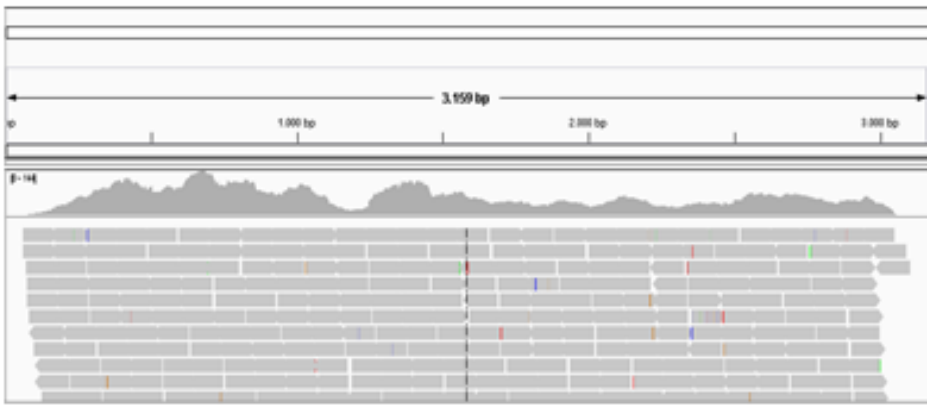
RNA-Seq - Conceito e metodologia



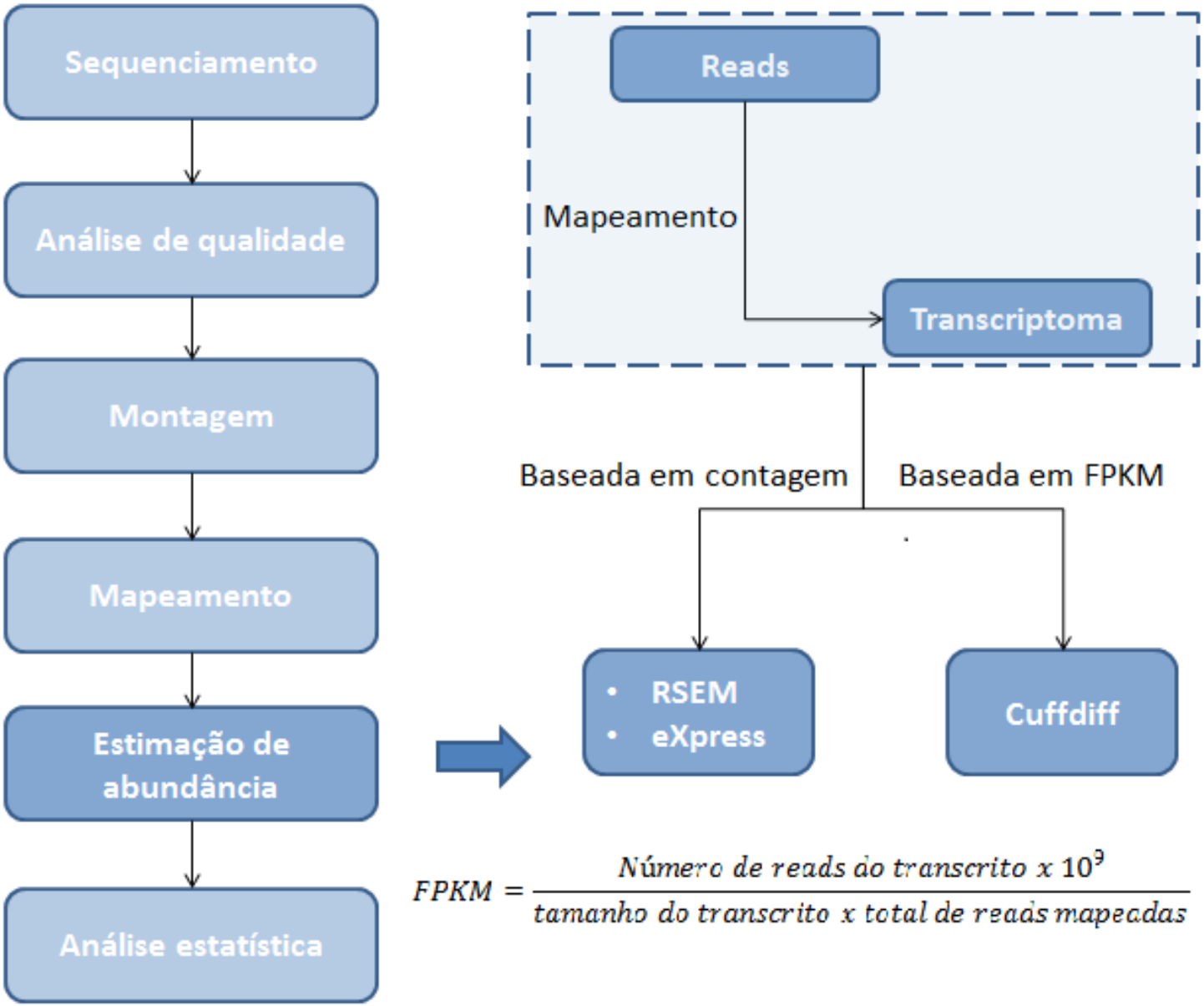
RNA-Seq - Conceito e metodologia





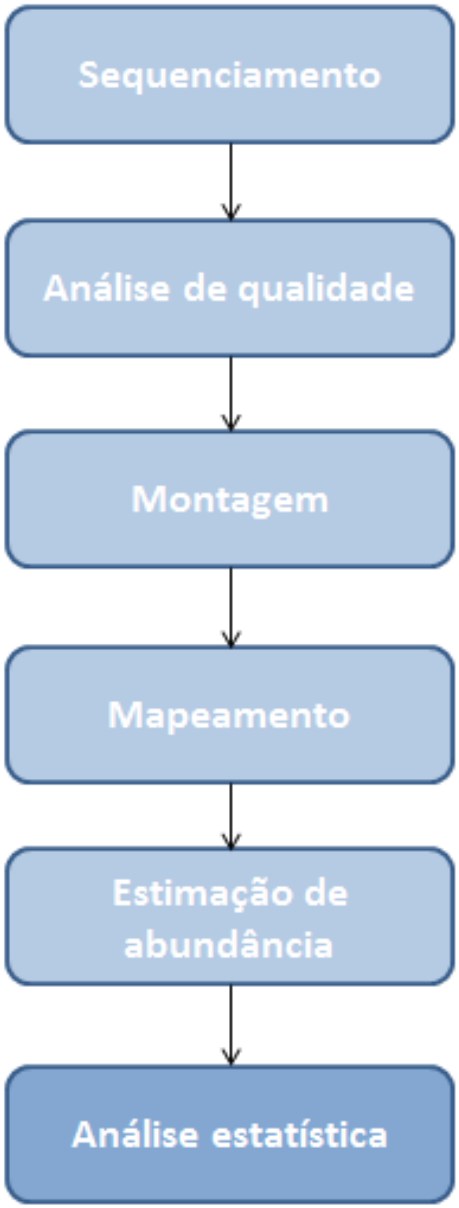


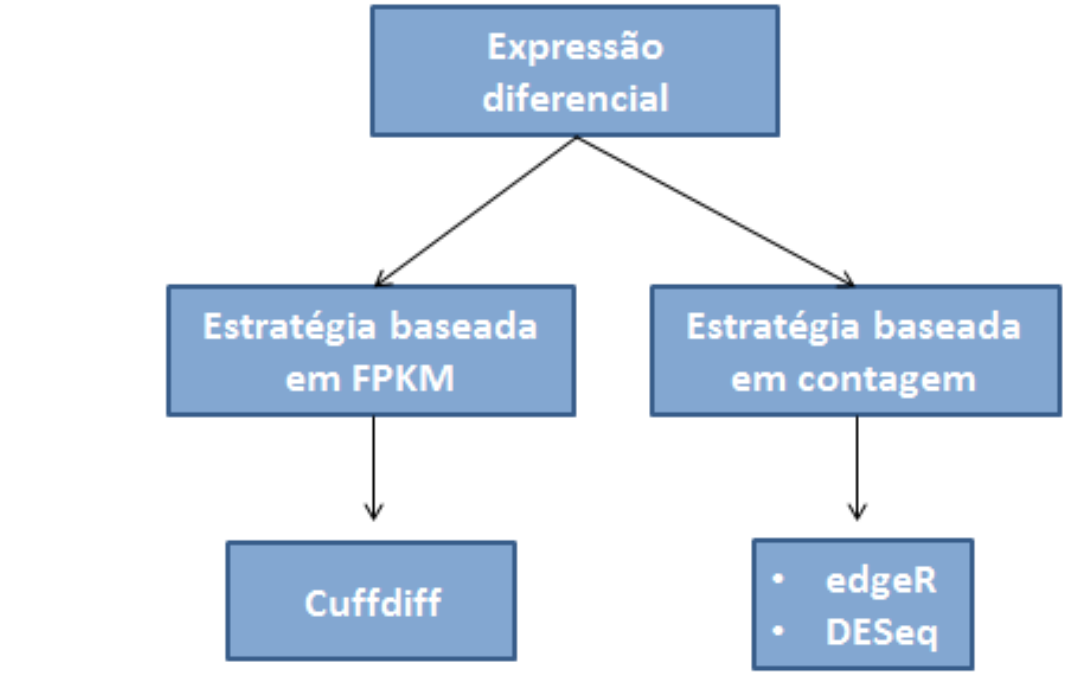
RNA-Seq - Conceito e metodologia



$$FPKM = \frac{Número\ de\ reads\ do\ transcrito\ x\ 10^9}{tamanho\ do\ transcrito\ x\ total\ de\ reads\ mapeadas}$$

RNA-Seq - Conceito e metodologia





Variam conforme o tipo de normalização dos dados

Saída:

Identificador	Log(FC)	P-value	FDR

Motivação

- Apesar de haver muitos transcriptomas *de novo* publicados, ainda não há um montador referência.
- Montagens *de novo* ainda possuem erros e os montadores *de novo* são sensíveis à eles.
- Testes estatísticos retornam muitos transcritos falsos positivos.

Objetivo

- Avaliar montadores *de novo* e sua influência na identificação de transcritos diferencialmente expressos.
- Avaliar como os montadores *de novo* se comportam devido à alteração do volume de dados.
- Propor critérios de seleção que diminuam o número de transcritos falsos positivos em uma análise de expressão diferencial.

Trabalhos relacionados

- Três trabalhos anteriores exploram a performance dos montadores *de novo* em relação a certas abordagens, como a métrica N50, conteúdo da sequência GC, a profundidade das taxas de cobertura, erros de *base-calling*, porcentagem de quimeras geradas e memória RAM.

Ano	Autores	Título
2011	Zhang *et al.*	A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies
2011	Lin *et al.*	Comparative studies of *de novo* assembly tools for next-generation sequencing technologies
2013	Lu *et al.*	Comparative study of *de novo* assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq

Trabalhos relacionados

- Fazem a comparação de estratégias que utilizam genoma de referência e sem genoma de referência (*de novo*).
- Utiliza métricas que dependem e não dependem do uso da referência.
- Propõem metodologias de montagem.

	Arabidopsis thaliana	*Canis familiaris*
Tipo de read	paired-end	paired-end
Condições	efeitos de substâncias	fases de ciclos estrais
Replicatas por condição	4	3
Total de reads	244.167.802	440.494.292
Tamanho da read (bp)	100	100

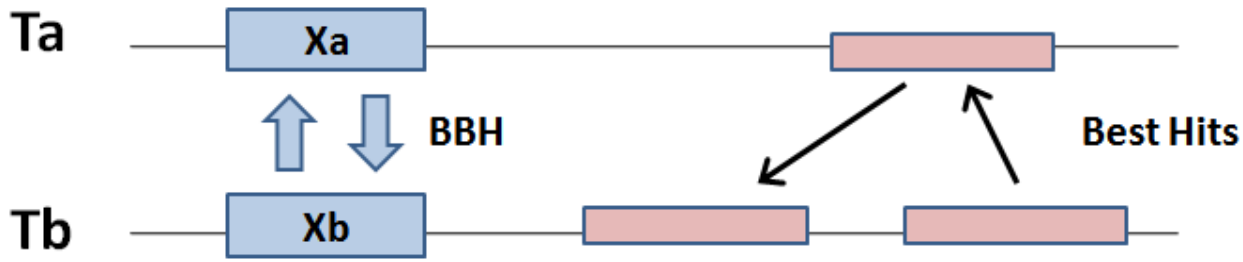
Tabela: Resumo dos dados utilizados no desenvolvimento do projeto.

- Análise de qualidade: FastQC e SeqClean.
- Montagem:
  - Com genoma de referência: Pipeline Cufflinks
  - *De novo*: Trinity, Velvet-Oases e IDBA-Tran
- Mapeamento: TopHat, Bowtie e Bowtie2
- Estimação de abundância: Cuffdiff, RSEM e eXpress
- Análise estatística: Cuffdiff e edgeR

- Foram realizadas as montagens *de novo* utilizando:
  - Trinity: k-mer = 25
  - Velvet-Oases e IDBA-Tran:
    - Variação de k-mer entre 21 e 31
    - Estratégia SAMP (Single Assembler Multiple-Parameters), que consiste na junção de várias montagens utilizando, por exemplo, o CAP3.
- Utilizando genoma de referência, a reconstrução dos transcritos foi feita pelo pipeline do TopHat-Cufflinks.

- Pensando em como os montadores *de novo* reagiriam à diminuição da quantidade de dados disponíveis para as montagens, foram gerados casos de testes:
  - Todos os reads da biblioteca do organismo (Tr).
  - Metade dos reads da biblioteca do organismo (Mr).
  - Somente reads da extremidade R1 da biblioteca do organismo (single-end) (Sr).

- Métricas de livre referência, tais como tamanho médio dos transcritos, N50, desvio padrão e mediana;
- Best Bidirectional Hit (BBH) e o fator k;
- Porcentagem de transcritos verdadeiramente diferencialmente expressos (fator de decisão *d*);
- Reconstrução de genes de cópias únicas (GCUs).

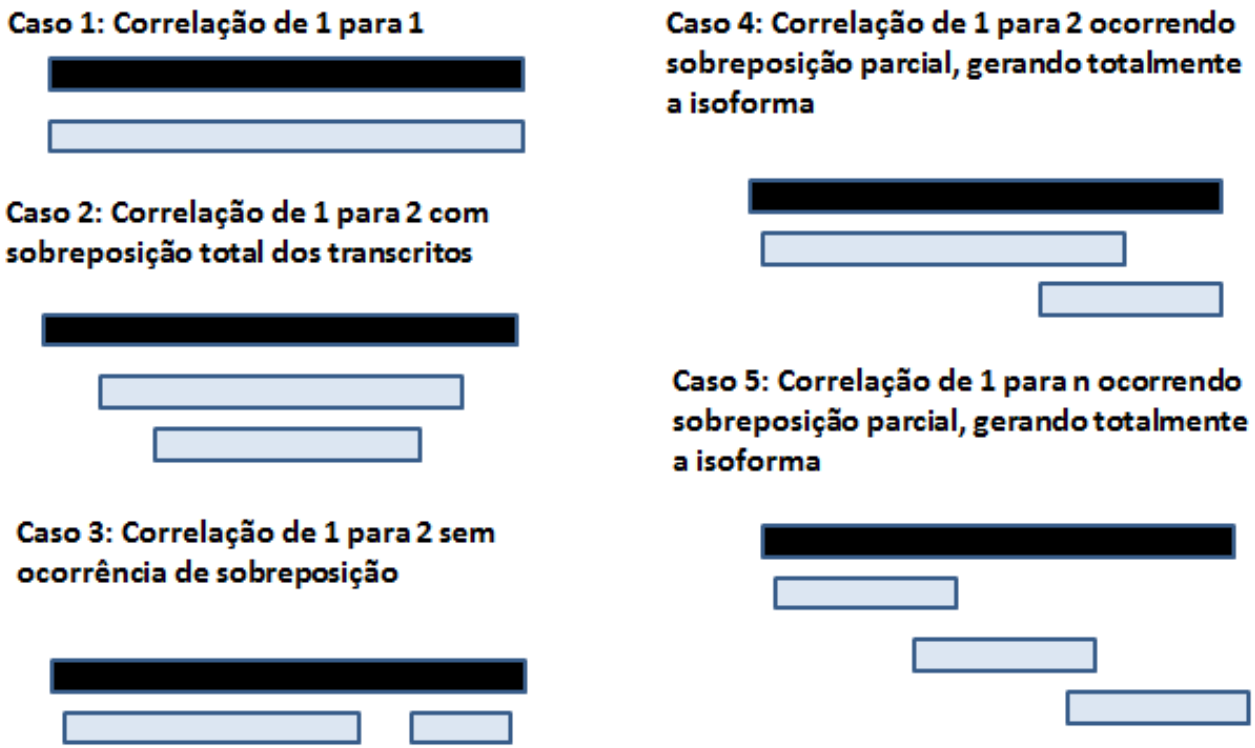


- Escolha da referência:
  - *Arabidopsis thaliana*: comparação entre transcritos anotados do RefSeq[1].
  - *Canis familiaris*: comparação entre transcritos gerados pelo Cufflinks.

[1] `ftp://ftp.arabidopsis.org/Sequences/blast_datasets/TAIR10_blastsets/`

Materiais e Métodos

BBH - Best Bidirectional Hit



- Modificação: considerar como BBH todos os transcritos com o mesmo e-value resultantes do blastn.
- fator k $= \frac{\text{transcritos montados com BBH}}{\text{total de transcritos montados}}$

- Se um transcrito *de novo* considerado diferencialmente expresso possui um BBH em relação aos transcritos diferencialmente expressos pelo Cuffdiff, então ele é considerado verdadeiramente diferencialmente expresso.

Fator de decisão d

$$\text{fator de decisão } d = \frac{\text{transcritos verdadeiramente diferencialmente expressos}}{\text{total de transcritos diferencialmente expressos}}$$

- Quanto mais próximo de 1 o fator d, menos transcritos falsos positivos determinada montagem gerou.

Materiais e Métodos

GCUs

- Genes conservados entre os organismos eucariotos.
- Utilizamos o banco de dados BUSCO[2], composto por 429 genes.
- Passos:
  1. Gerar ORFs pelo TransDecoder.
  2. Extrair as ORFs no frame correto de cada transcrito.
  3. Traduzir as ORFs.
  4. Realizar um alinhamento global utilizando o ClustalW2 entre os GCUs e a maior ORF do transcrito.
  5. Guardar o valor do score de alinhamento (valor W).

[2] http://busco.ezlab.org/

- Análises de expressão diferencial retornam muitos transcritos diferencialmente expressos falsos positivos.
- Propor critérios se seleção dos transcritos ditos diferencialmente expressos que maximizassem a chance dele ser verdadeiramente diferencialmente expresso.

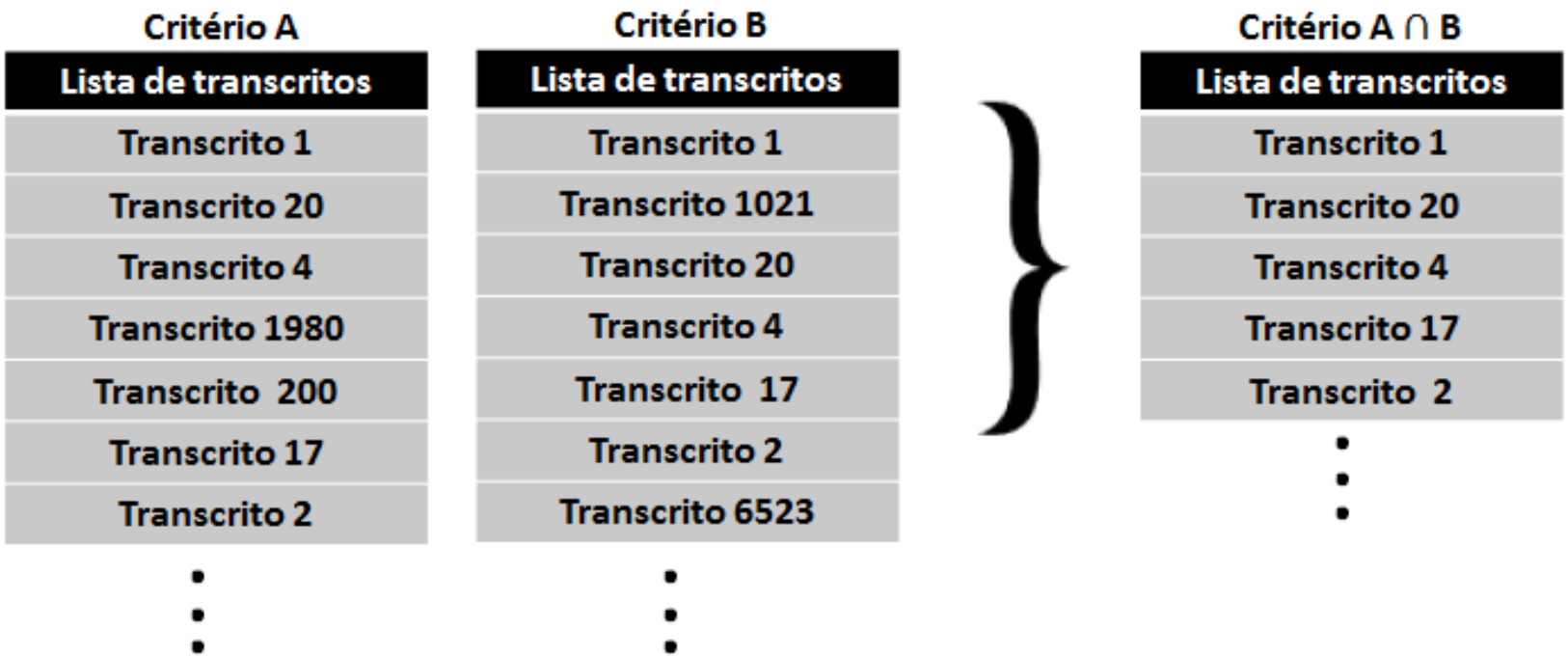
Materiais e Métodos

Critérios de seleção

	Critérios	Ordem
Critério 1	Número de reads	Decrescente
Critério 2	Fold-change	Decrescente
Critério 3	Fold-change	Crescente
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 / Decrescente
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	Crescente / Crescente
Critério 13	Fold-change e ORFs	Decrescente / Decrescente
Critério 14	Fold-change e P-value	Decrescente / Crescente
Critério 15	P-value e ORFs	Crescente / Decrescente

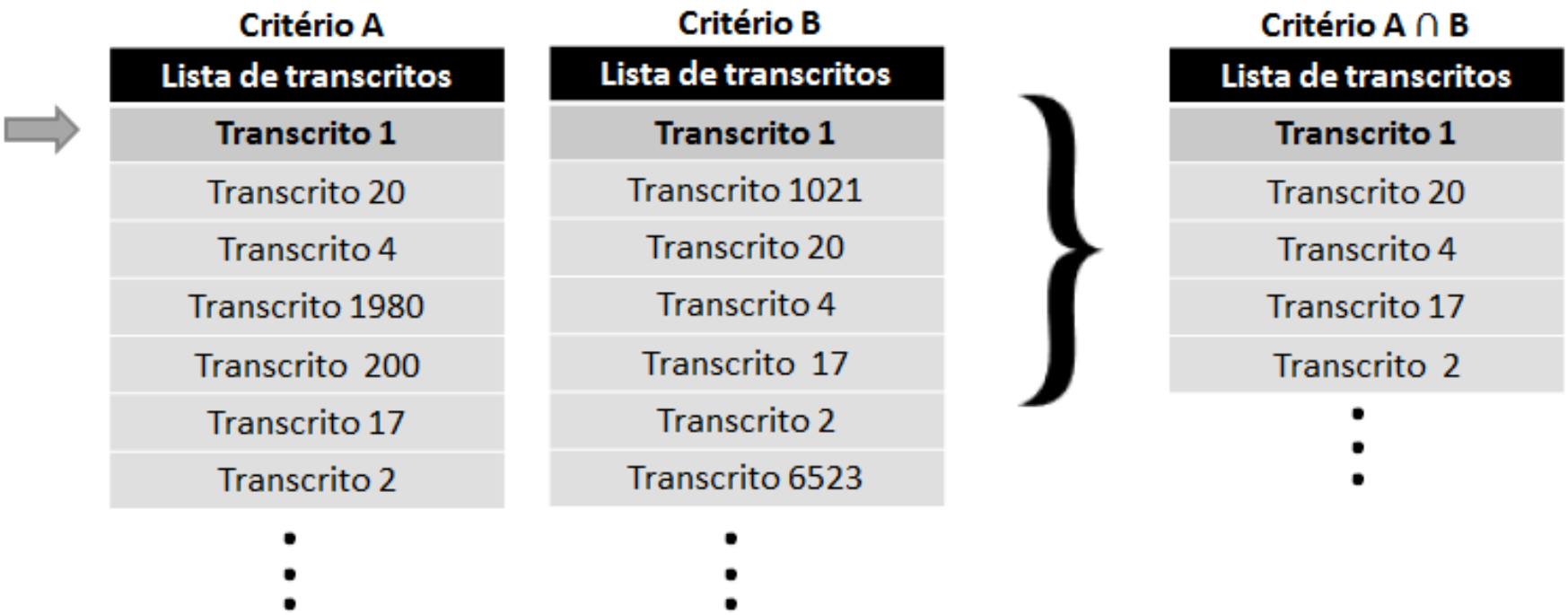
Materiais e Métodos

Critérios de seleção



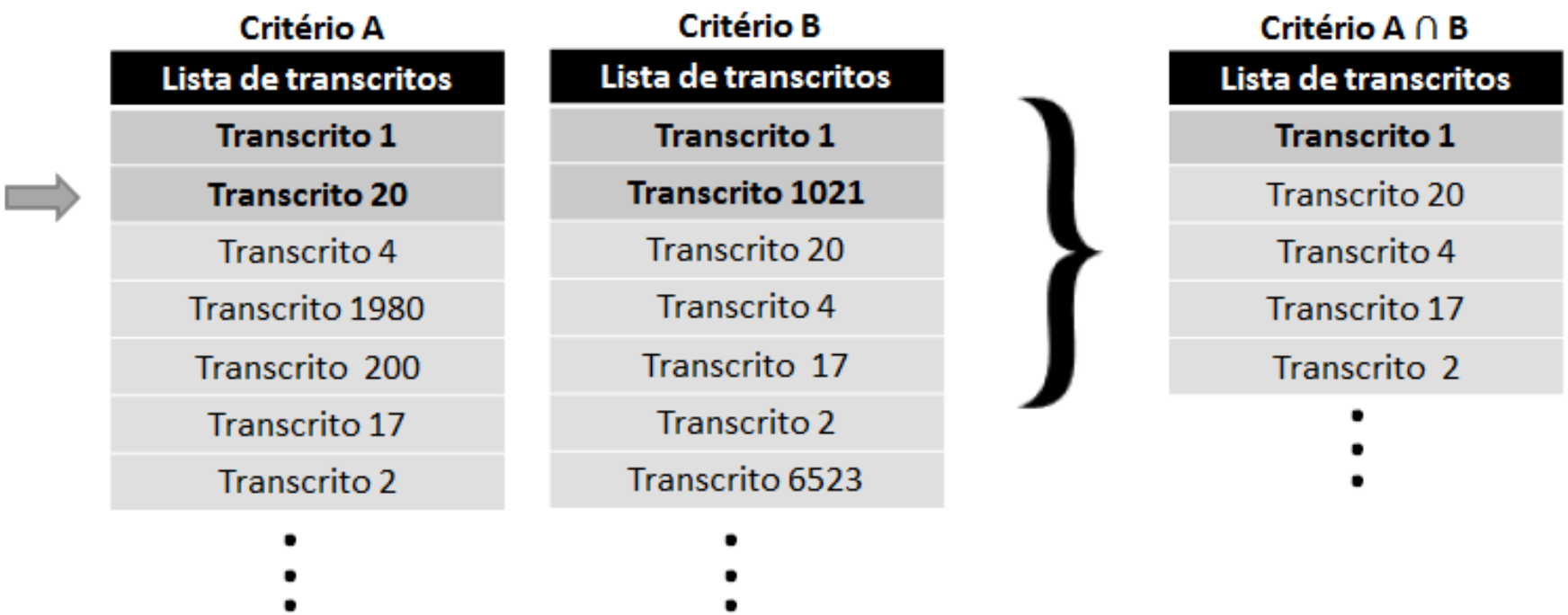
Materiais e Métodos

Critérios de seleção



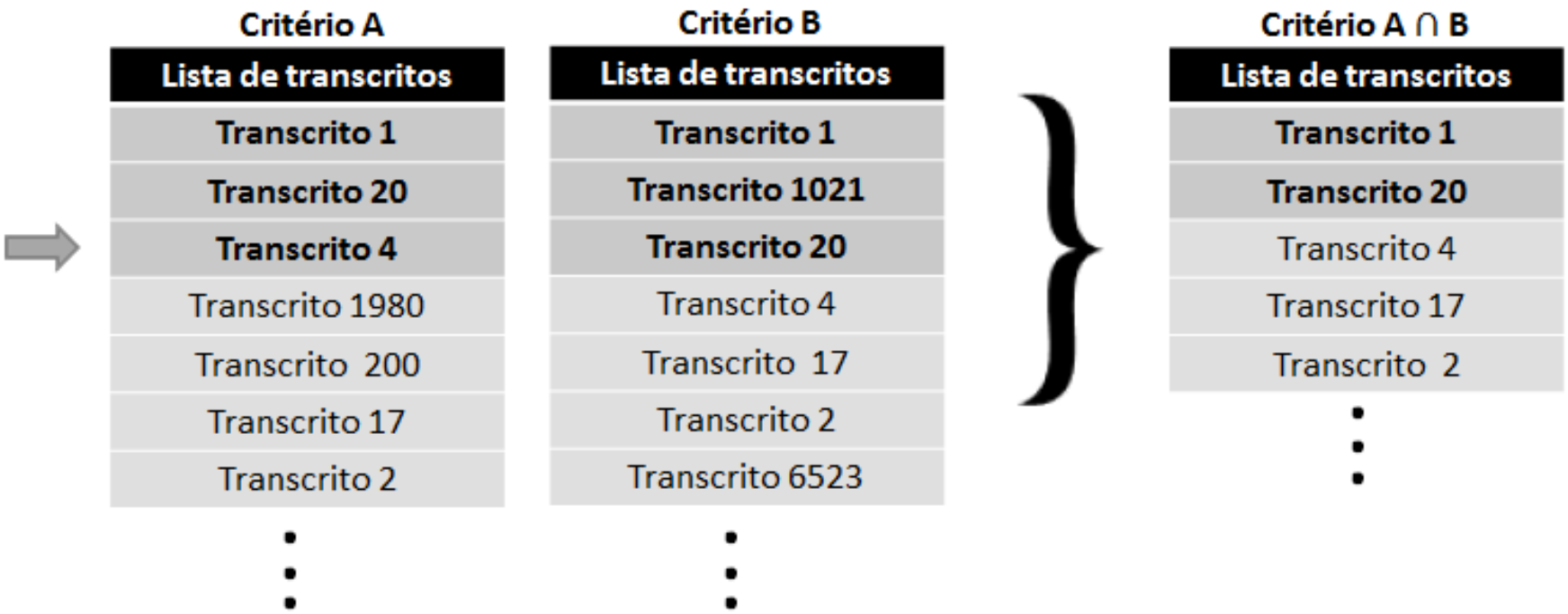
Materiais e Métodos

Critérios de seleção



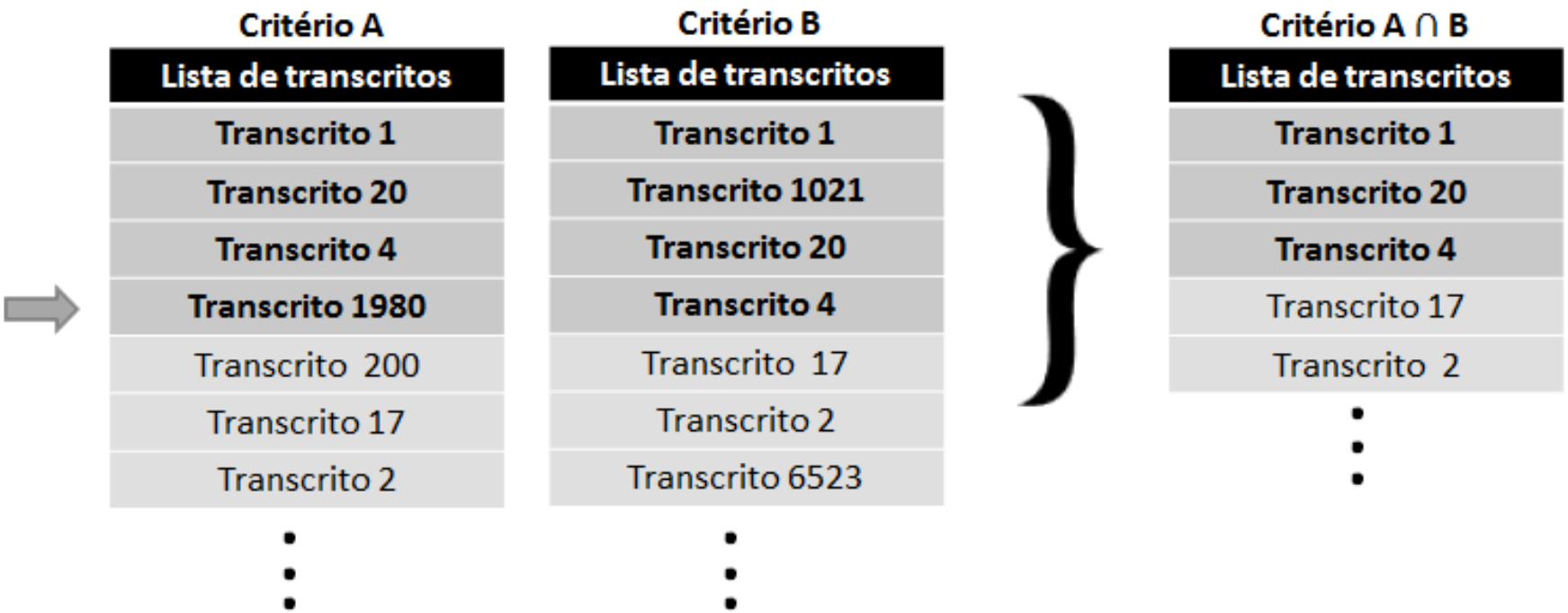
Materiais e Métodos

Critérios de seleção



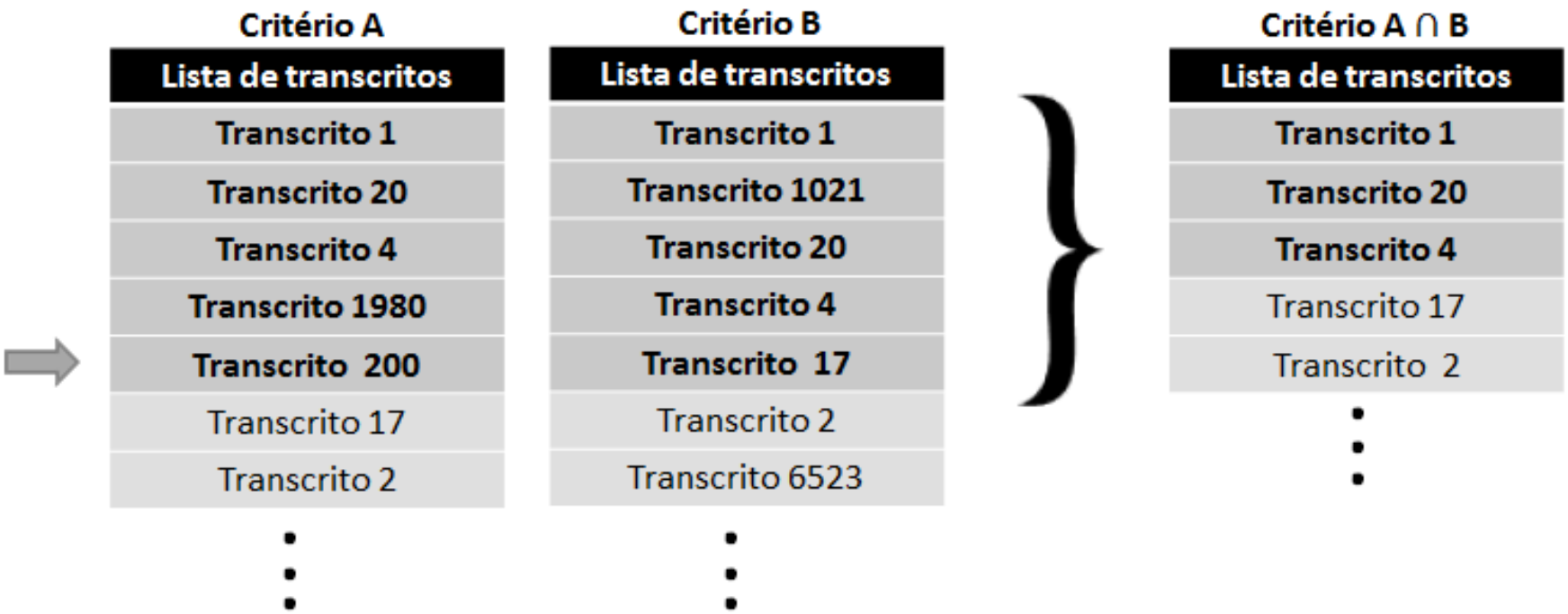
Materiais e Métodos

Critérios de seleção



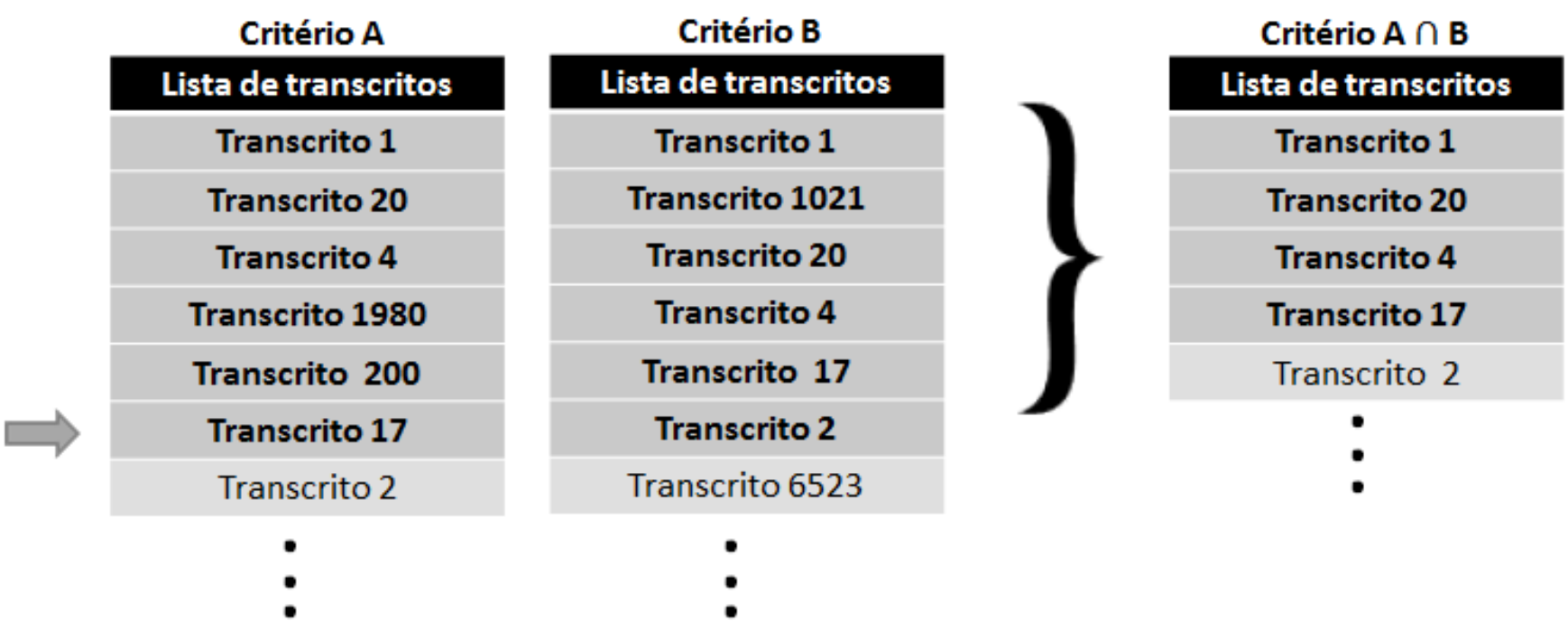
Materiais e Métodos

Critérios de seleção



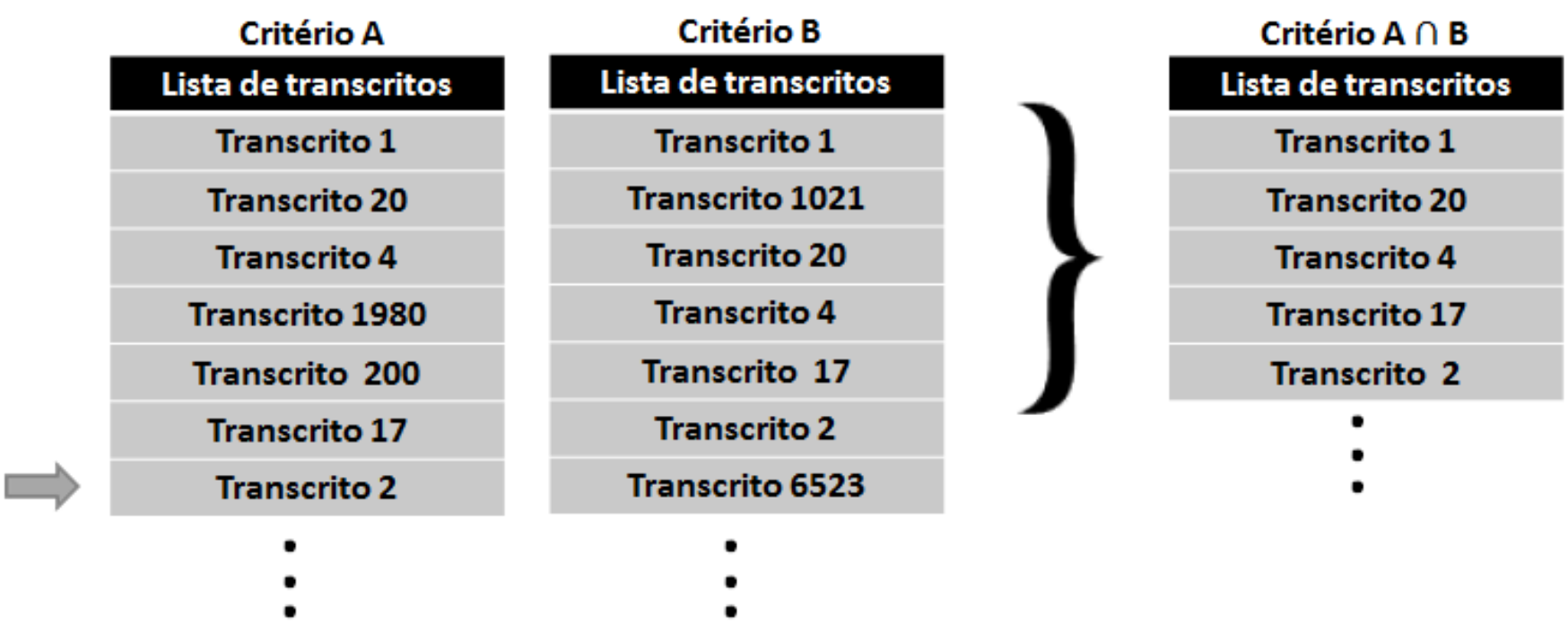
Materiais e Métodos

Critérios de seleção



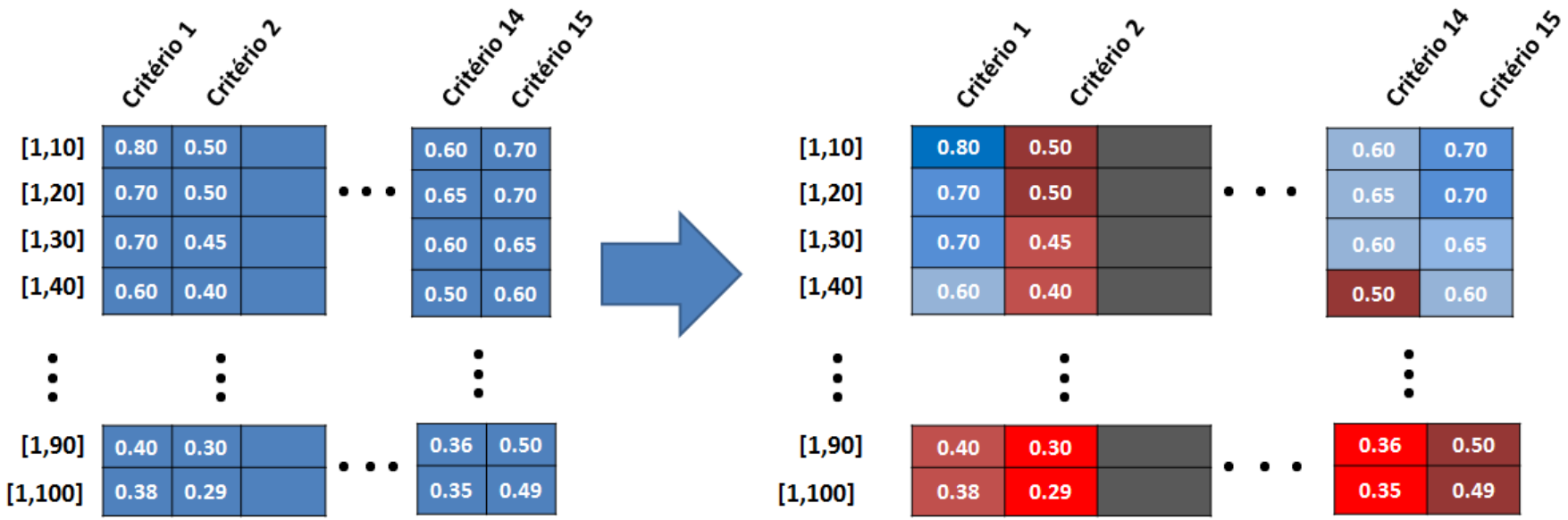
Materiais e Métodos

Critérios de seleção



- A comparação entre dois critérios foi dividida por intervalos.
- 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 a 10 transcritos e calculamos a porcentagem deles serem verdadeiramente diferencialmente expressos, em seguida, no intervalo de 1 a 20, e assim por diante.

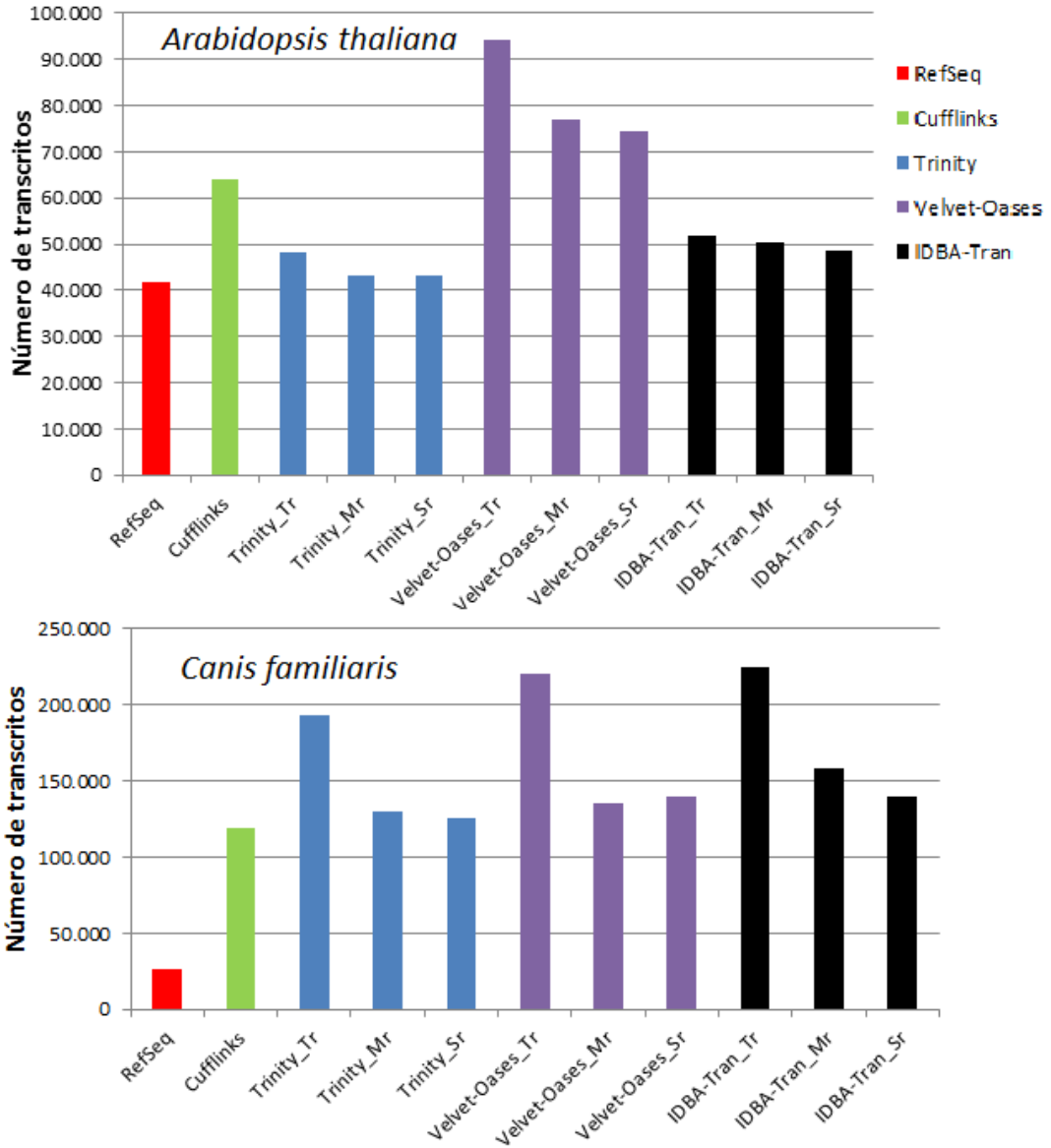
- Transformamos o resultado em uma matriz 15x10 (15 critérios por 10 intervalos).



- Validação: teste pareado de Wilcoxon.

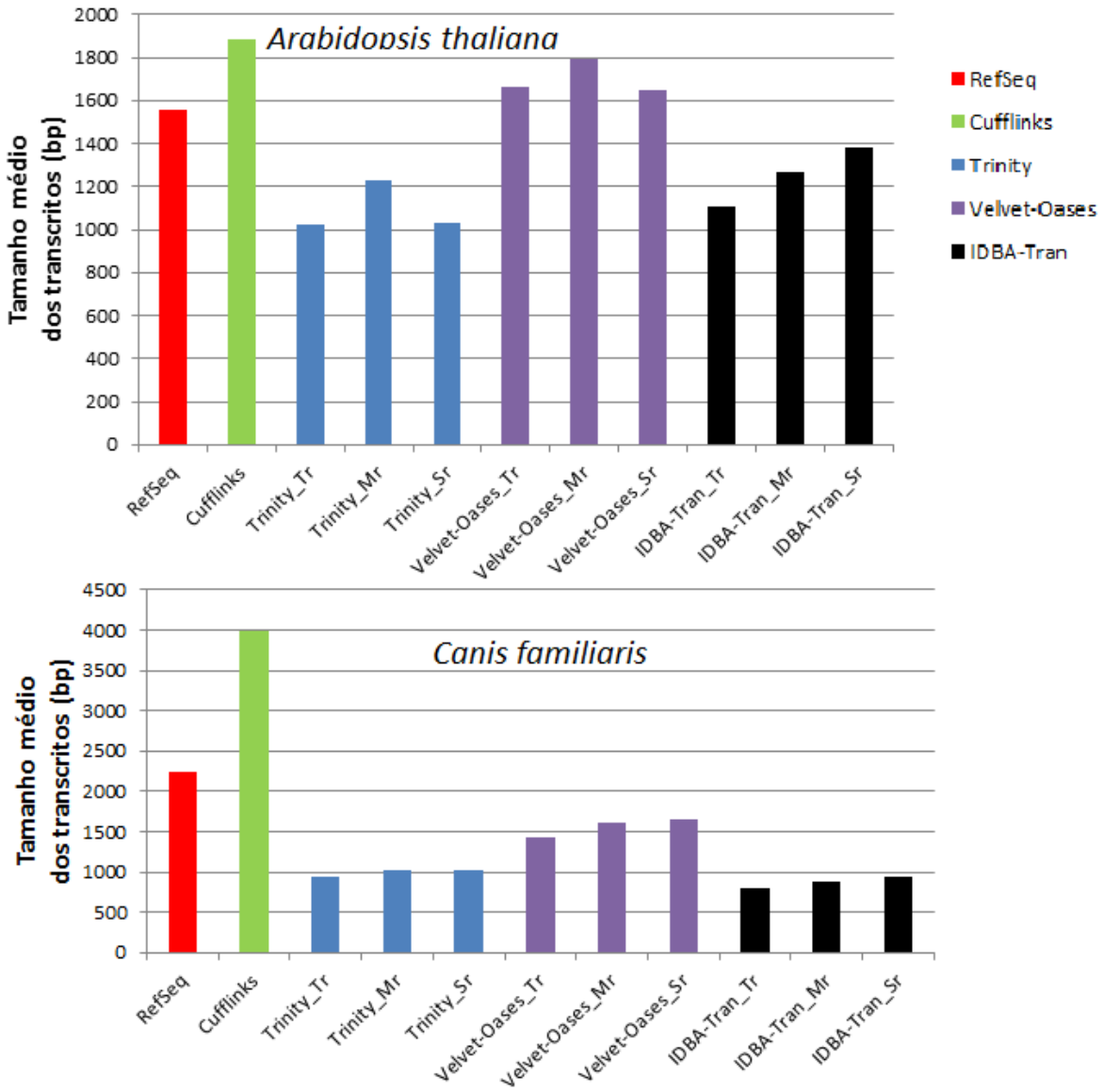
Resultados e discussões

Métricas de livre referência



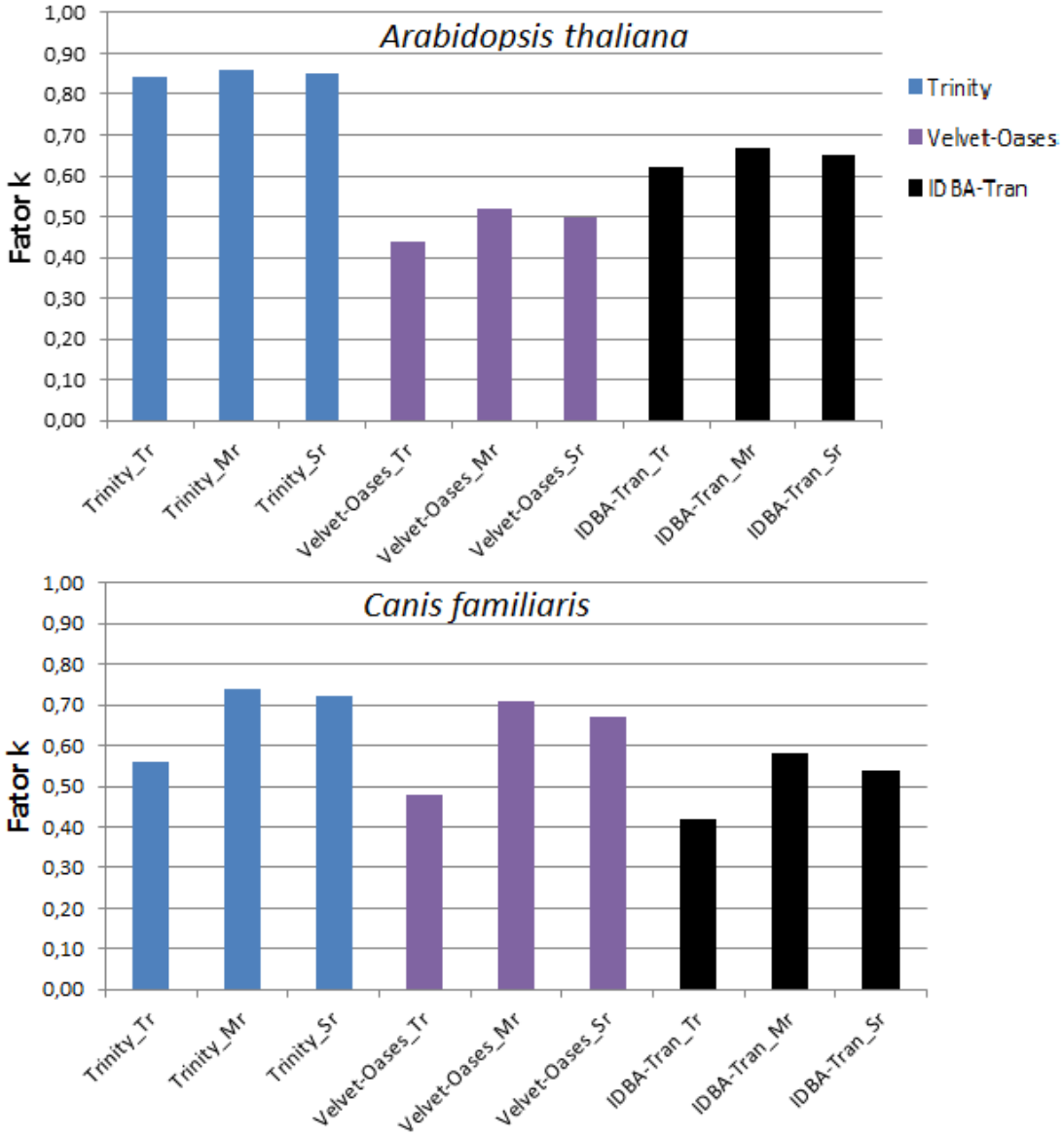
Resultados e discussões

Métricas de livre referência



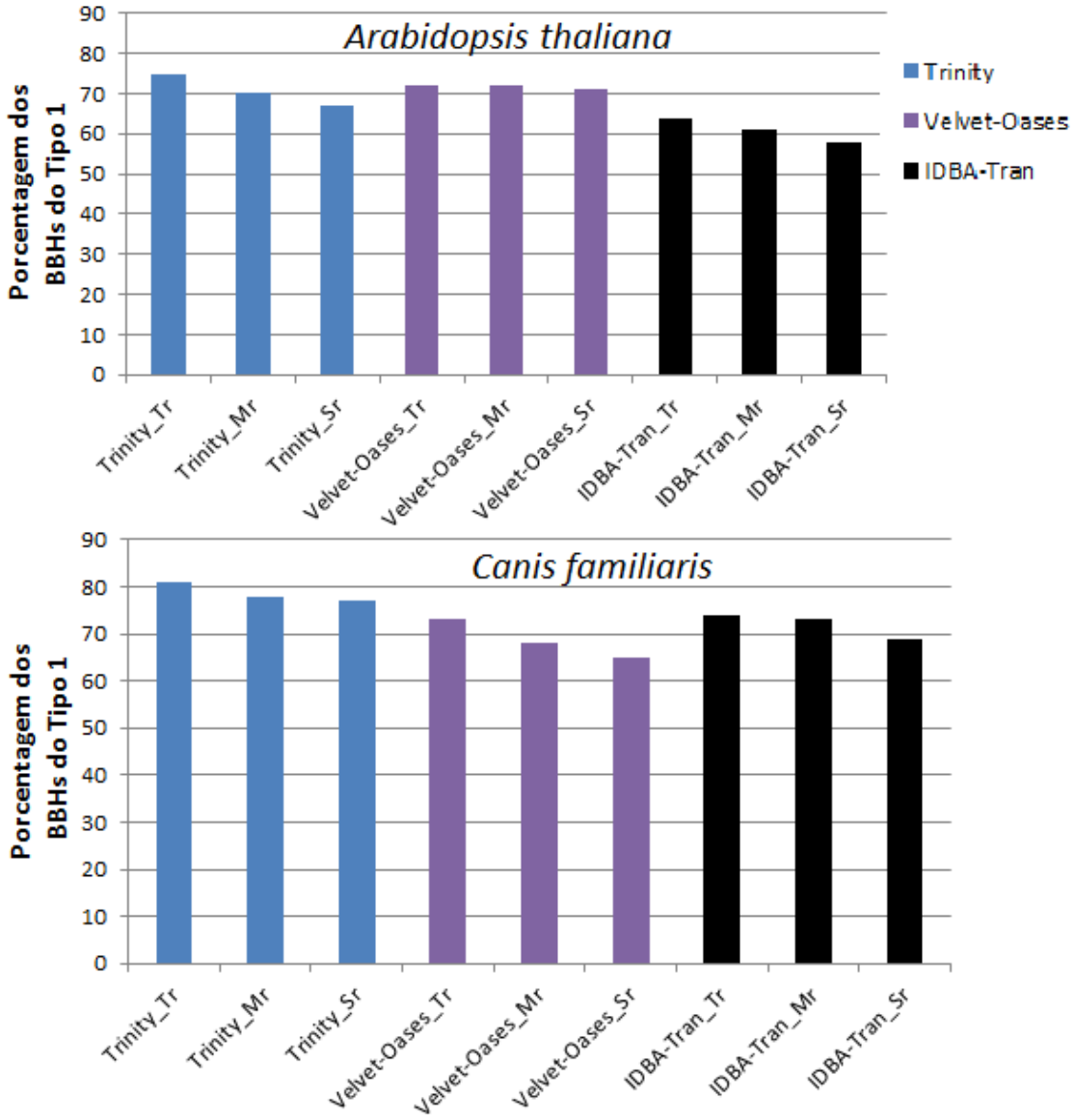
Resultados e discussões

BBH - Best Bidirectional Hit



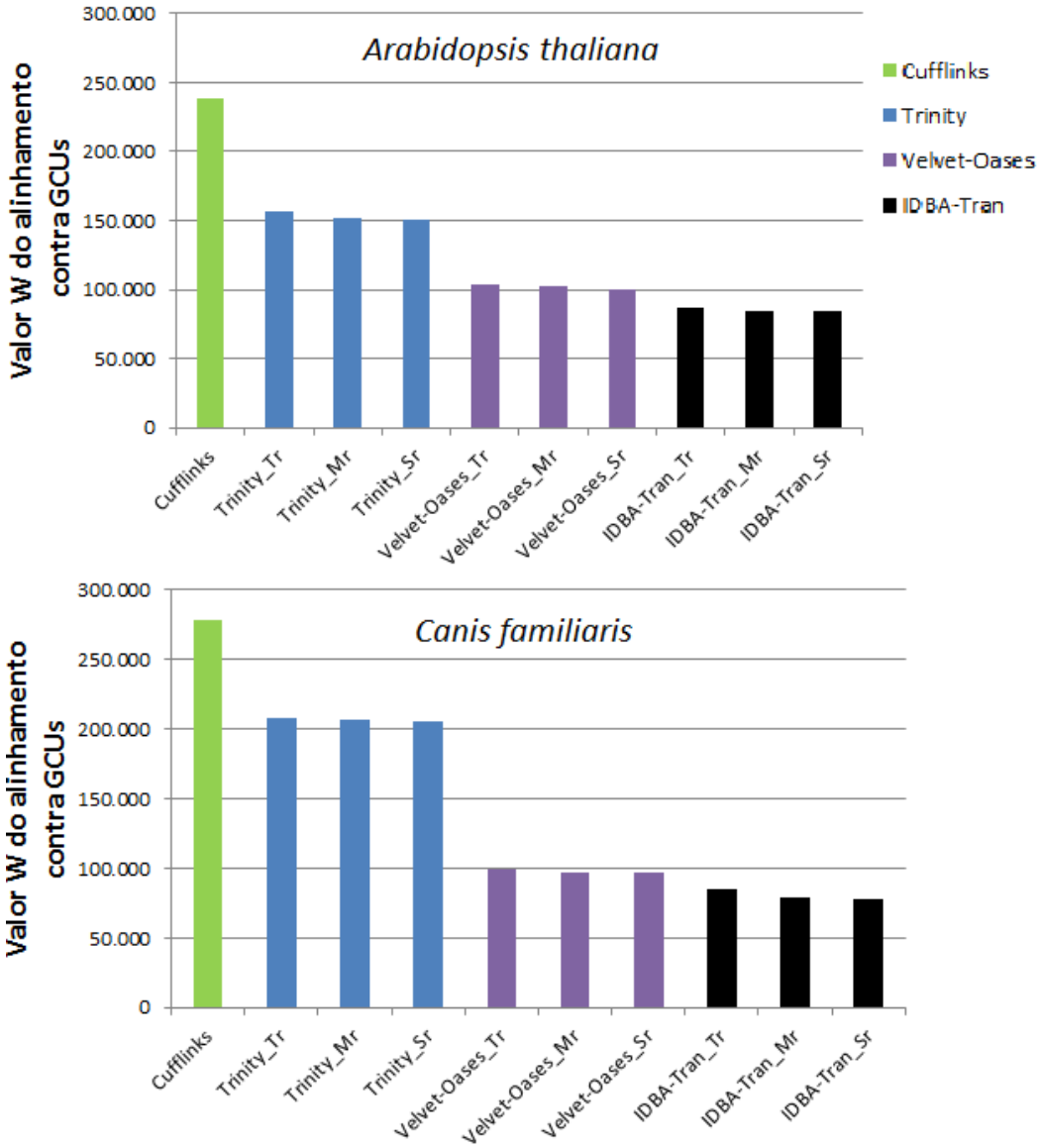
Resultados e discussões

BBH - Best Bidirectional Hit



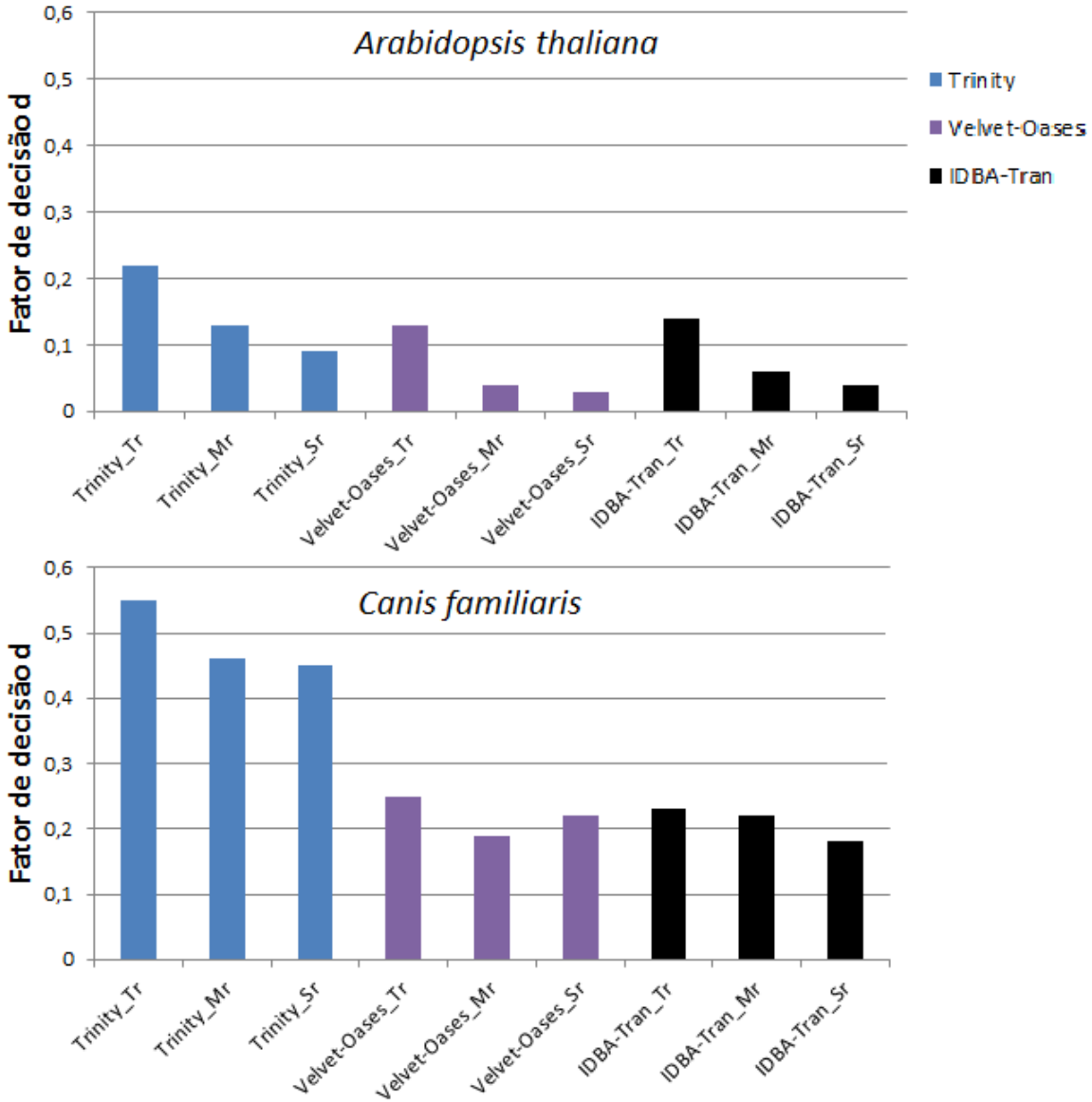
Resultados e discussões

Alinhamento contra GCUs (Genes de Cópias Únicas)



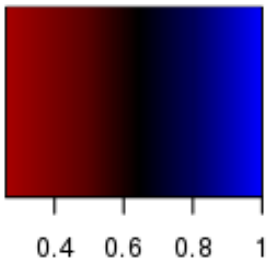
Resultados e discussões

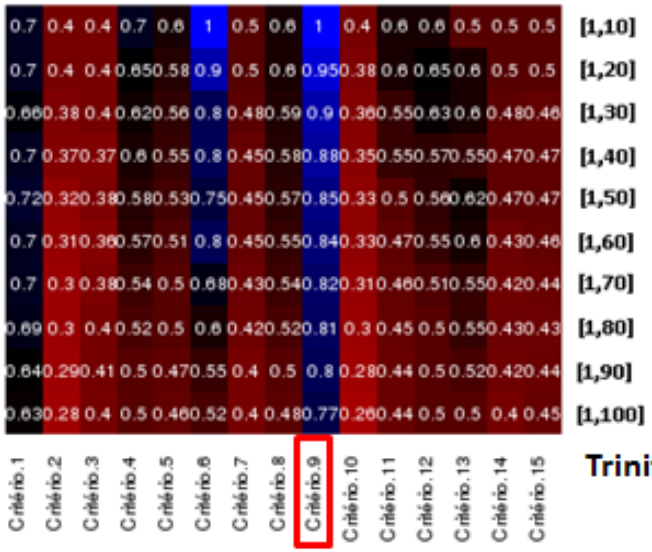
Fator de decisão *d*



Critérios de seleção

Caso de teste Tr - *Canis familiaris*

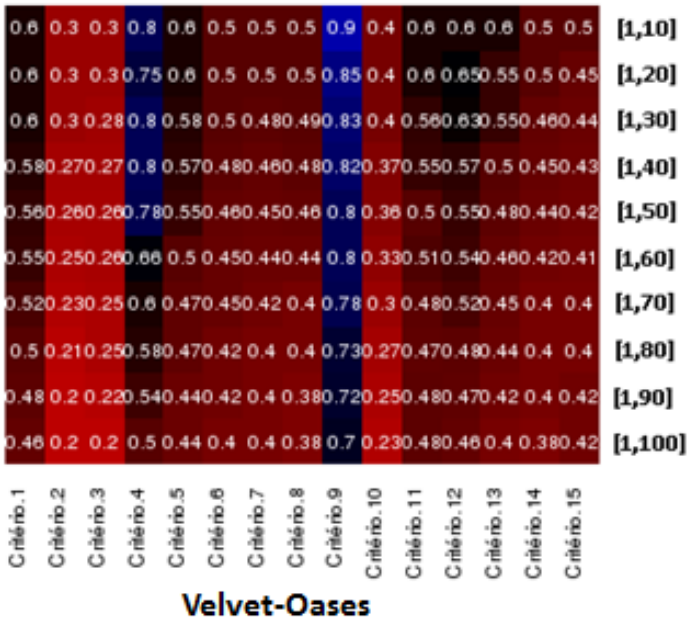




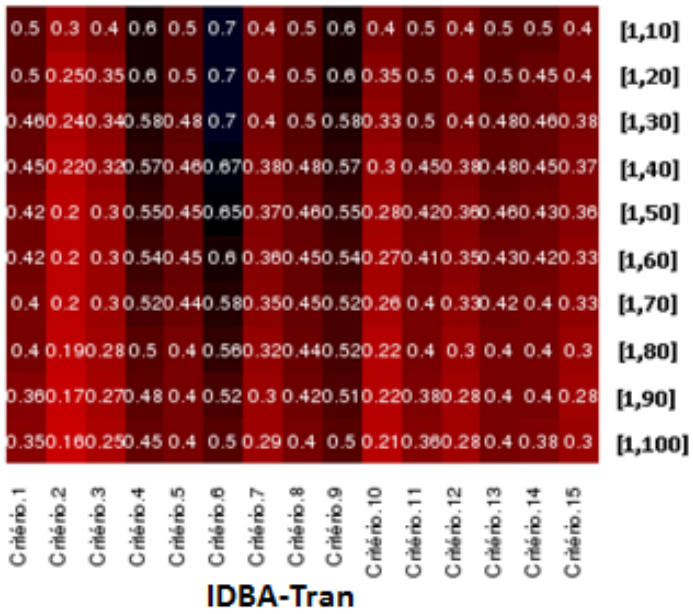
Critério 9

Número de reads e P-value/FDR

Trinity



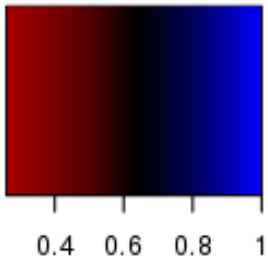
Velvet-Oases

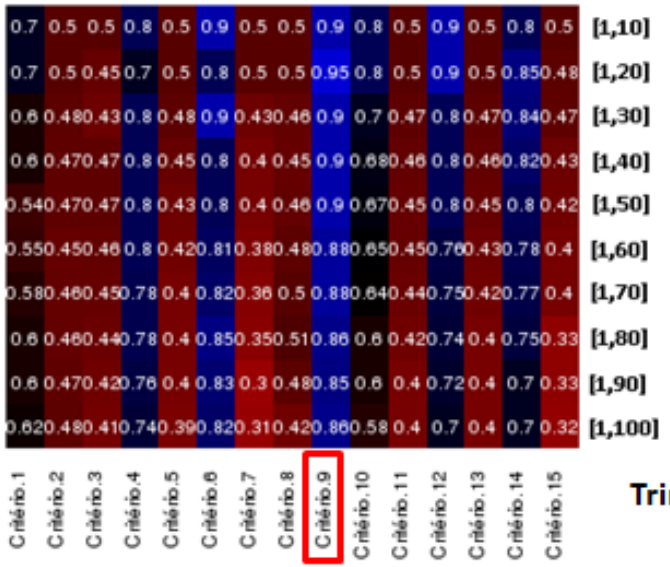


IDBA-Tran

Critérios de seleção

Caso de teste Mr - *Arabidopsis thaliana*

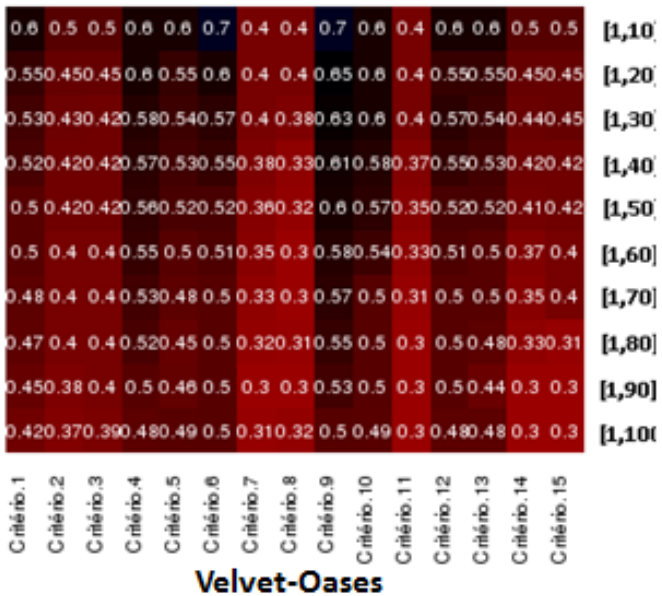




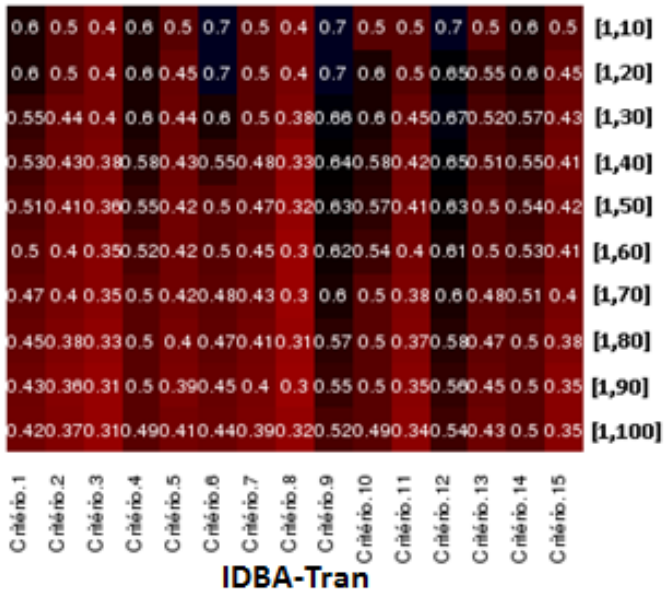
Trinity

Critério 9

Número de reads e P-value/FDR



Velvet-Oases



IDBA-Tran

- Após gerar todos os heatmaps de todos os casos de teste, verificamos qual o montador encontrou o maior número de transcritos verdadeiramente diferencialmente expresso e qual critério sobressaiu sobre os outros.
- *Arabidopsis thaliana*
  - Tr: Trinity com o Critério 1 (Número de reads).
  - Mr: Trinity com o Critério 9 (Número de reads e P-value/FDR).
  - Sr: Trinity com o Critério 9 (Número de reads e P-value/FDR).
- *Canis familiaris*
  - Tr: Trinity com o Critério 9 (Número de reads e P-value/FDR).
  - Mr: Trinity com o Critério 9 (Número de reads e P-value/FDR).
  - Sr: Trinity com o Critério 9 (Número de reads e P-value/FDR).

- Caso de teste Tr do montador Trinity de *Arabidopsis thaliana*:

	C_9	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}
[1,10]	0.90	0.90	0.40	0.50	0.80	0.50	0.80	0.90	0.60	0.50	0.40	0.60	0.60	0.80	0.50
[1,20]	0.95	0.95	0.40	0.50	0.70	0.50	0.90	0.85	0.55	0.55	0.40	0.60	0.60	0.80	0.50
[1,30]	0.90	0.92	0.38	0.50	0.65	0.44	0.75	0.8	0.50	0.60	0.40	0.58	0.56	0.80	0.46
[1,40]	0.90	0.95	0.36	0.48	0.70	0.43	0.72	0.80	0.47	0.60	0.38	0.57	0.54	0.80	0.45
[1,50]	0.88	0.95	0.34	0.46	0.70	0.41	0.70	0.82	0.47	0.60	0.40	0.56	0.52	0.75	0.44
[1,60]	0.86	0.90	0.33	0.43	0.75	0.40	0.67	0.80	0.47	0.55	0.38	0.55	0.50	0.75	0.41
[1,70]	0.85	0.88	0.30	0.42	0.75	0.40	0.64	0.78	0.45	0.50	0.38	0.54	0.47	0.70	0.42
[1,80]	0.85	0.95	0.28	0.40	0.80	0.40	0.63	0.75	0.41	0.50	0.40	0.51	0.46	0.70	0.38
[1,90]	0.85	0.90	0.28	0.40	0.75	0.38	0.61	0.73	0.40	0.44	0.37	0.50	0.45	0.66	0.38
[1,100]	0.80	0.86	0.27	0.40	0.75	0.38	0.59	0.70	0.40	0.40	0.38	0.59	0.47	0.63	0.41
Wilcoxon		-	X	X	X	X	X	X	X	X	X	X	X	X	X

- Caso de teste Tr do montador Trinity de *Canis familiaris*:

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

Conclusões

- Ao final deste trabalho podemos concluir que ainda há muitas dificuldades em analisar uma montagem *de novo*.
- Ao analisarmos as montagens *de novo*, a quantidade de verdadeiros positivos (transcritos verdadeiramente diferencialmente expressos) se altera devido a alteração do volume de dados.
- A abordagem que leva em consideração a eficiência de uma montagem *de novo* em reconstruir genes de cópias únicas possibilita a exploração de uma métrica intrínseca de montagem.
- Os testes comparativos realizados neste trabalho, como a porcentagem de BBHs do Tipo 1 e o valor W na busca de genes de cópias únicas (GCUs) em eucariotos, revela que o Trinity é o melhor montador *de novo*.

Conclusões

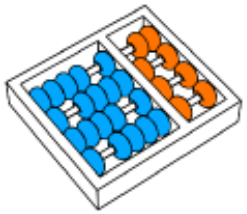
- Pela análise do fator de decisão *d* notamos que o edgeR retorna muitos falsos positivos, assim, critérios de seleção devem ser aplicados à lista final maximizando a escolha de um transcrito verdadeiramente diferencialmente expresso.
- Ao realizarmos testes inciais de escolha de transcritos diferencialmente expressos em uma análise de RNA-Seq, notamos que a porcentagem de acerto era em torno de 40%, utilizando o critério Fold-change. Após este estudo, o critério que se destacou foi o de ordenação por Número de Reads e P-value/FDR, alcançando um acerto de 100% em alguns casos de teste.
- O montador que mais se aproximou de encontrar transcritos verdadeiramente diferencialmente expressos, ordenado pelo Critério 9 e sem gerar muitos falsos positivos, foi o Trinity.

Trabalhos futuros

- Validar a seleção de transcritos verdadeiramente diferencialmente expressos, utilizando o critério de seleção 9, aplicando a dados publicados na literatura.
- Realizar todos os testes deste estudo com novos montadores *de novo*;
- Analisar a sensibilidade dos montadores *de novo*, em relação ao volume de dados, para novos casos de teste.

Agradecimentos

- Ao Instituto de Computação.
- À Embrapa Informática Agropecuária.
- Ao Laboratório Multiusuário de Bioinformática.
- Ao CNPq pela bolsa de mestrado durante o desenvolvimento do projeto.



Embrapa
Informática Agropecuária

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Laboratório Multiusuário de Bioinformática

Referências

B. Alberts (2010)

Biologia Molecular da Célula

Editora Artmed 199

Wang,Z. and Gerstein, M. and Snyder, M. (2009)

RNA-Seq: a revolutionary tool for transcriptomics

Nature Reviews Genetics 10, 57-63

Lu, B. and Zeng, Z. and Shi, T. (2013)

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

Zhang, W. and Chen, J. and Yang, Y. and Tang, Y. and Shang, J. and Shen, B. (2011)

A Practical Comparison of *De Novo* Genome Assembly Software Tools for Next-Generation Sequencing Technologies

PLoS ONE 6(3), e17915

Referências

Simão, A. F. and Waterhouse, R. M. and Ionnadis, P. and Kriventseva, E. V. and Zdobnov, E. M. (2015)

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.

Bioinformatics 31(19).

Trapnell, C. and Williams, B. A. andv Pertea, G. and Mortazavi, A. and Kwan, G. and van Baren, M. J. and Salzberg, S. L. and Wold, B. J. and Pachte, L.(2009)

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Nature Reviews Genetics 10,57–63.

Grabherr,M. G. and Haas, B. J. and Yassour, M. and Levin, J. Z. and Thompson, D. A. and Amit, I. and Adiconis, X. and Fan, L. and Raychowdhury, R. and Zeng, Q. and Chen, Z. and Mauceli, E. and Hacohen, N. and Gnirke, A. and Rhind,N. and di Palma, F. and Birren, B. W. and Nusbaum, C. and Lindblad-Toh, K. and Friedman, N. and Regev, A.

Full-length transcriptome assembly from RNA-Seq data without a reference genom

Nature Biotechnology 29, 644-652.

Referências

Schulz, M. H. and Zerbino, D. R. and Vingron, M. and Birney, E. (2012)

Oases: robust de novo RNA-Seq assembly across the dynamic range of expression levels

Bioinformatics 28,1086-92.

Peng, Y. and Leung, H. C. and Yiu, S. M. and Lv, M. J. and Zhu, X. G. and Chin, F. Y. (2013)

IDBA-Tran: a more robust *de novo* De Bruijn graph assembler for transcriptomes with uneven expression levels,

Bioinformatics 29, i326-34.

Lin, Y. and Li, J. and Shen, H. and Zhang, L. and Papasian, C. J. and Deng, H. W.(2011)

Comparative studies of *de novo* assembly tools for next-generation sequencing technologies

Bioinformatics 27(15), 2031-2037