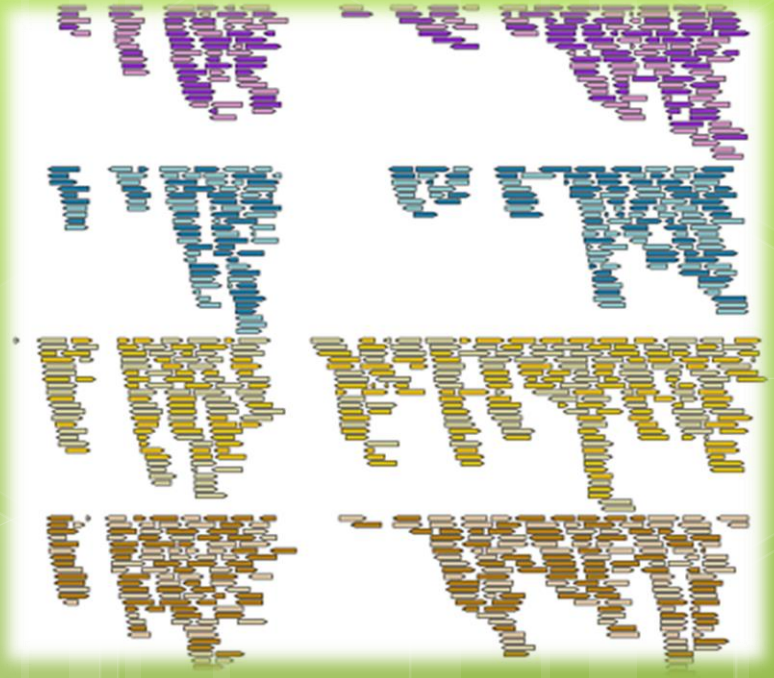
RNA sequencing



Diego R. Mazzotti, Ph.D.
Pesquisador
Laboratório de Biologia Molecular do Sono
Departamento de Psicobiologia
Universidade Federal de São Paulo

Supervisor
Molecular Core
Associação Fundo de Incentivo à Pesquisa (AFIP)

E-mail: mazzottidr@gmail.com

RNA-sequencing

- Uso de ferramentas de sequenciamento de nova geração (NGS) para avaliação da expressão gênica, por meio do sequenciamento do **transcriptoma**

- **Transcriptoma** é o conjunto de transcritos expressos em um determinado momento por uma célula ou tecido
  - Inclui todas as formas de RNA: mRNA, miRNA, rRNA, tRNA, ncRNA

RNA-sequencing

- Método mais atual para avaliação da expressão gênica em **larga escala**

- Antes, métodos de hibridação, como *microarrays* eram os mais utilizados

RNA-sequencing vs. Microarrays

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

RNA-sequencing

- Principal vantagem sobre o *microarray* → capacidade de detectar transcritos "novos", independente das sequências presentes

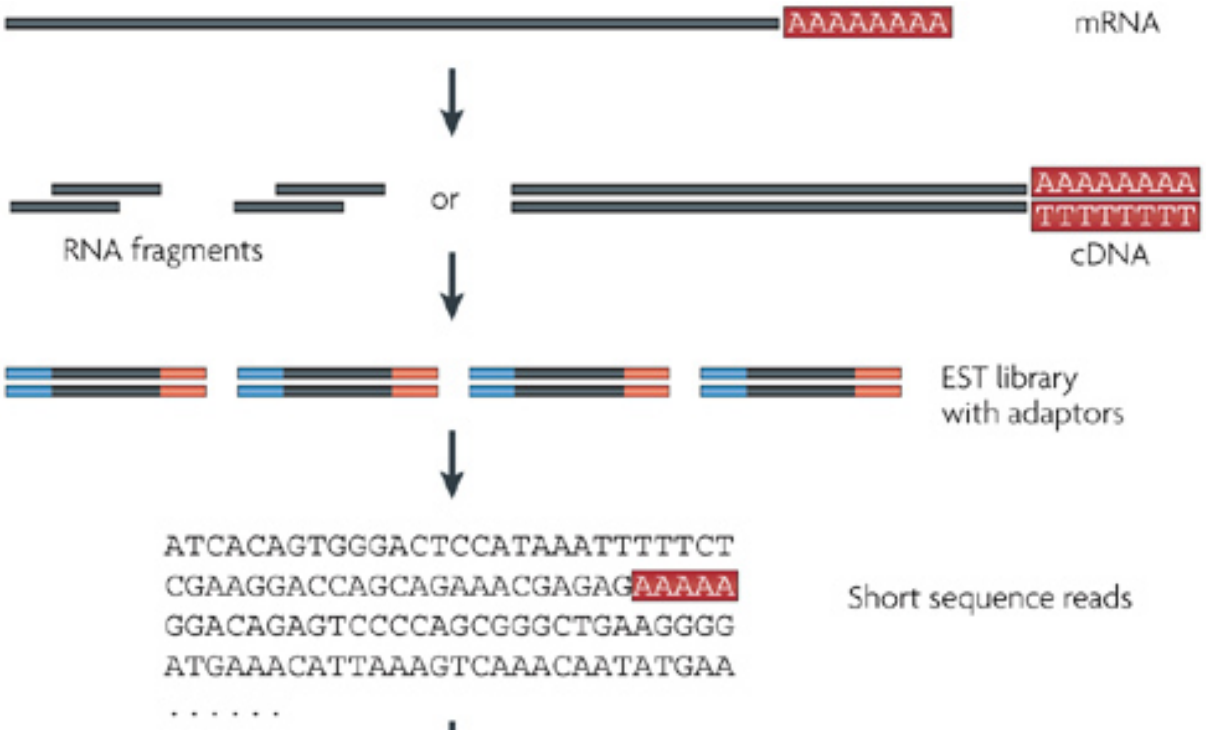
> "RNA-Seq [...] is expected to revolutionize the manner in which eukaryotic transcriptomes are analysed."

Wang et al, 2009 – Nature Review Genetics

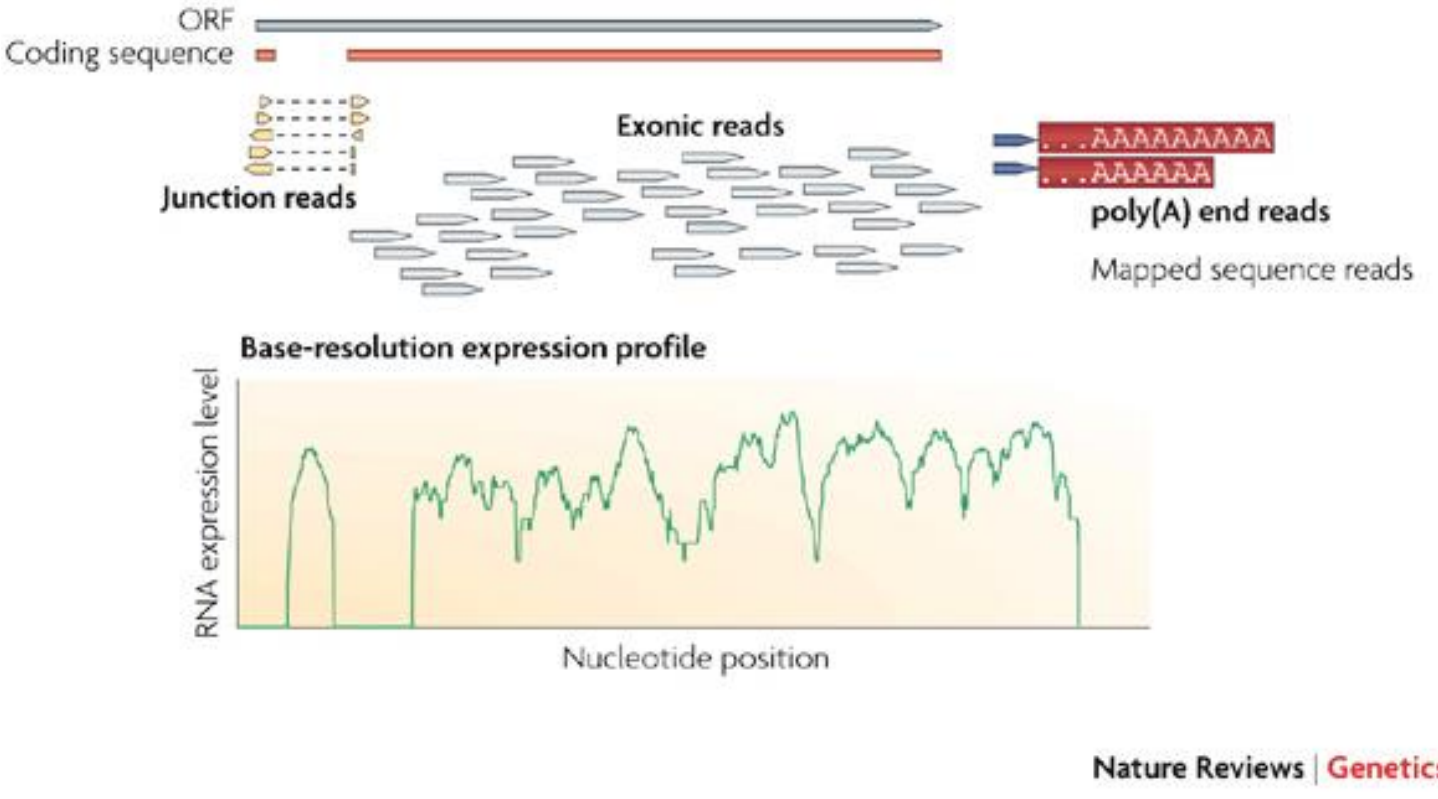
RNA-sequencing – Metodologia

- No geral, RNA é isolado, fragmentado e usado como molde para síntese de cDNA
- Fragmentos de cDNA são sequenciados e em seguida mapeados para um genoma ou transcriptoma de referência

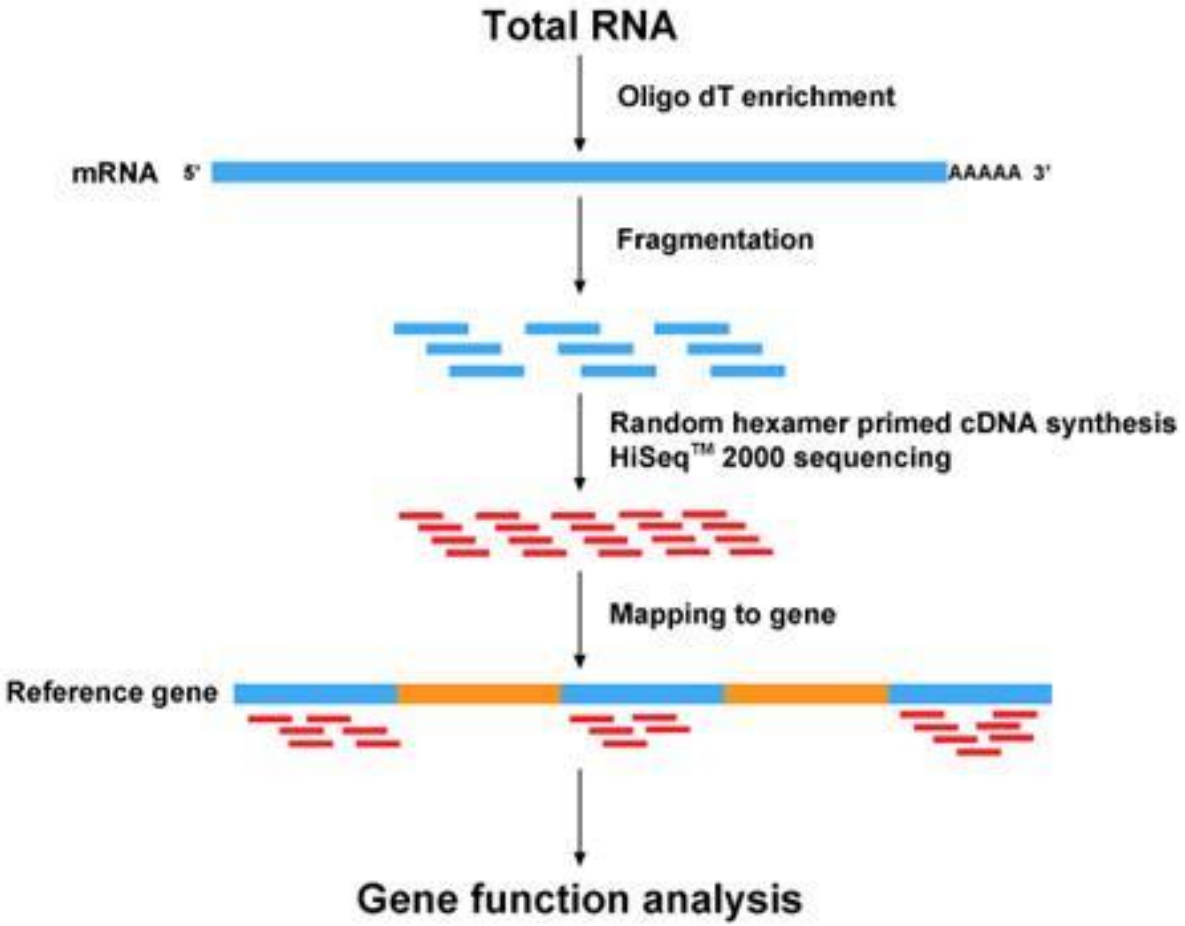
RNA-sequencing



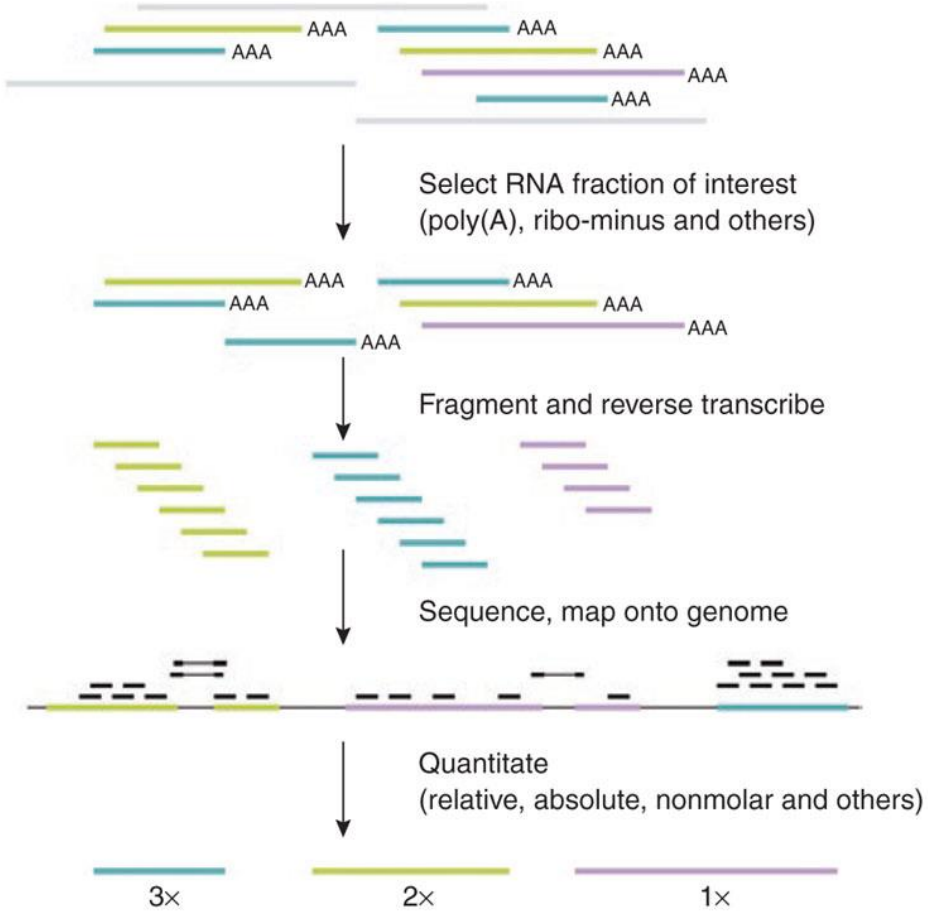
RNA-sequencing



RNA-sequencing



RNA-sequencing



RNA-sequencing – análise dos dados

- Planejamento da análise é determinante na informação que pode ser obtida
- "Racional" da análise é o mesmo para detecção de variantes – juntar os reads / alinhamento / *variant calling*
- Diferencial → **Cobertura** é o que precisamos saber

RNA-sequencing – análise dos dados

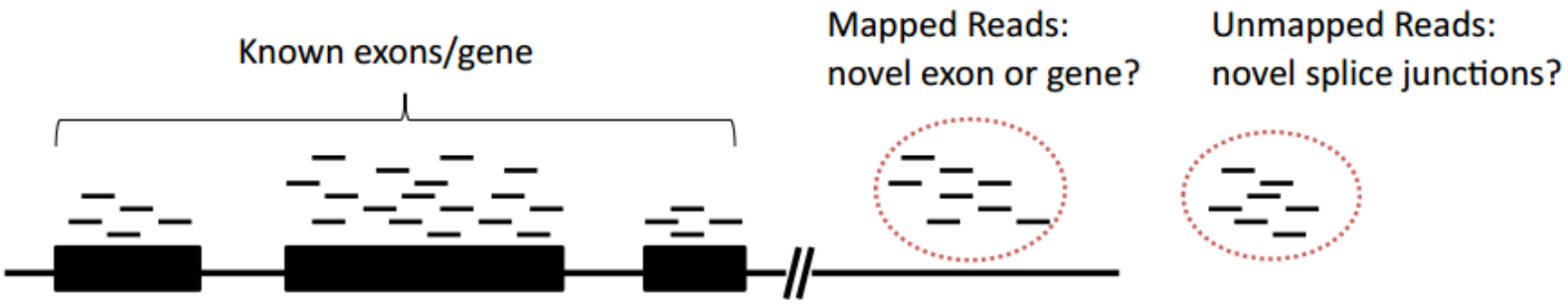
- Perguntas:
  - **Onde alinhar?** Genoma de referência / Transcriptoma de referência / montagem (*assembly*) dos reads
  - **O que é a expressão gênica?** Contagem dos reads / Cobertura dos éxons
  - **Normalização?** Como garantir que amostras são comparáveis e que diferenças de reads são os únicos fatores relacionados à expressão?
  - **Qual software usar?**
  - **Desenho experimental?**

RNA-sequencing – análise dos dados

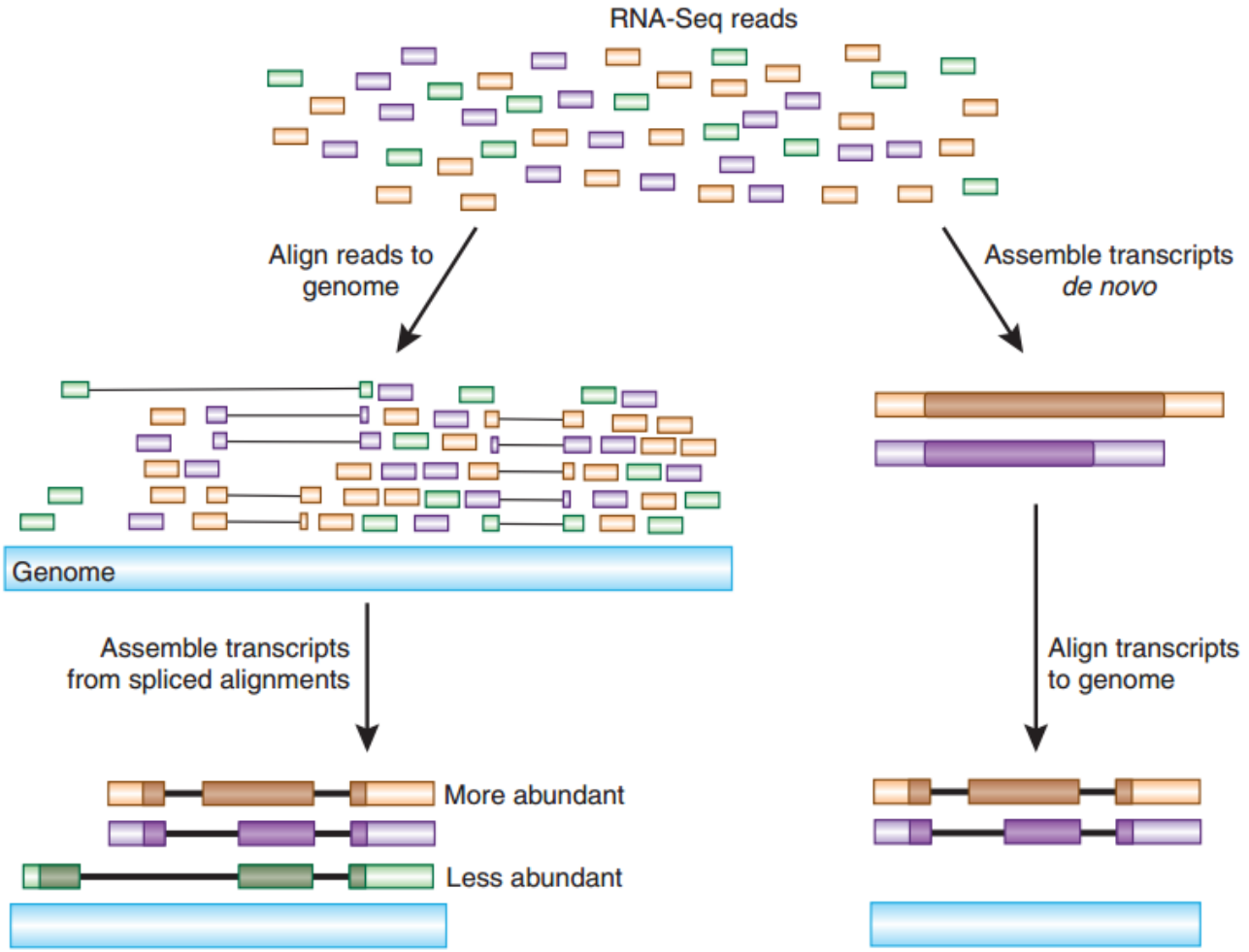
- **Onde alinhar?**
  - Genoma de referência ?
  - Transcriptoma de referência?
  - Montagem (*assembly*) dos reads?

RNA-sequencing – análise dos dados

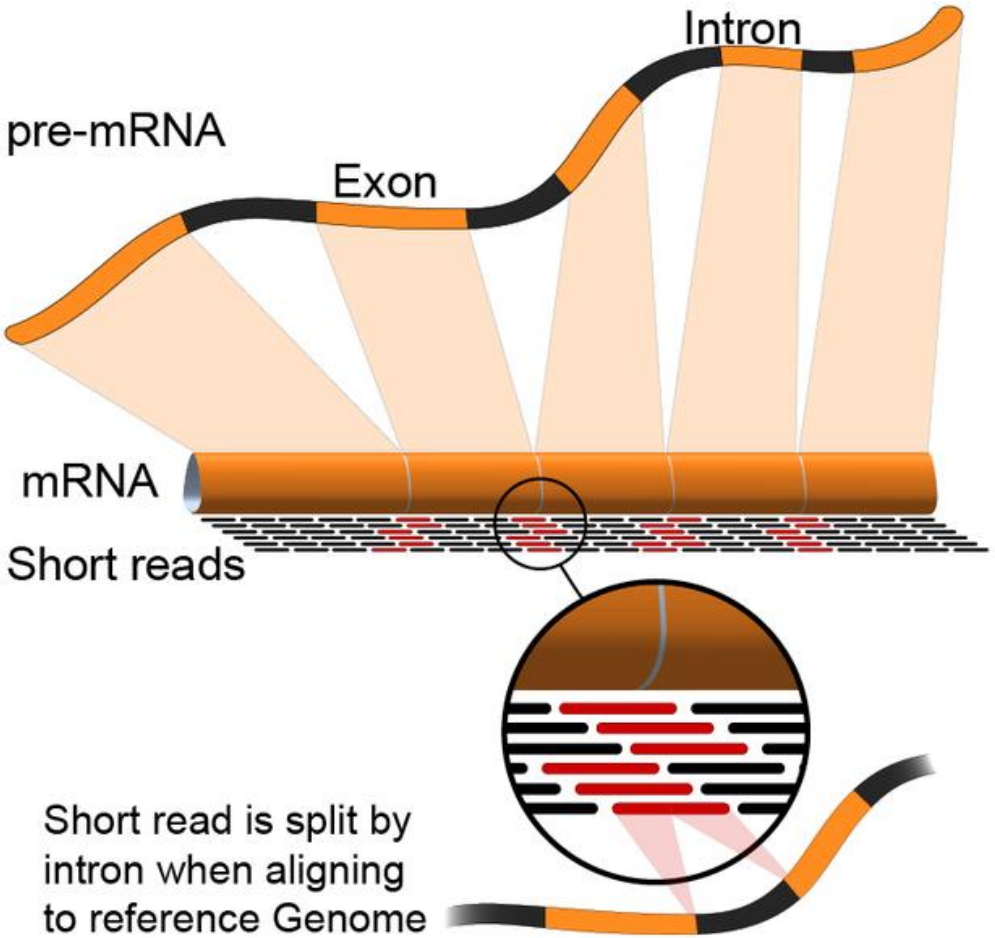
- **Possibilidades de alinhamento**



RNA-sequencing – análise dos dados



RNA-sequencing – análise dos dados

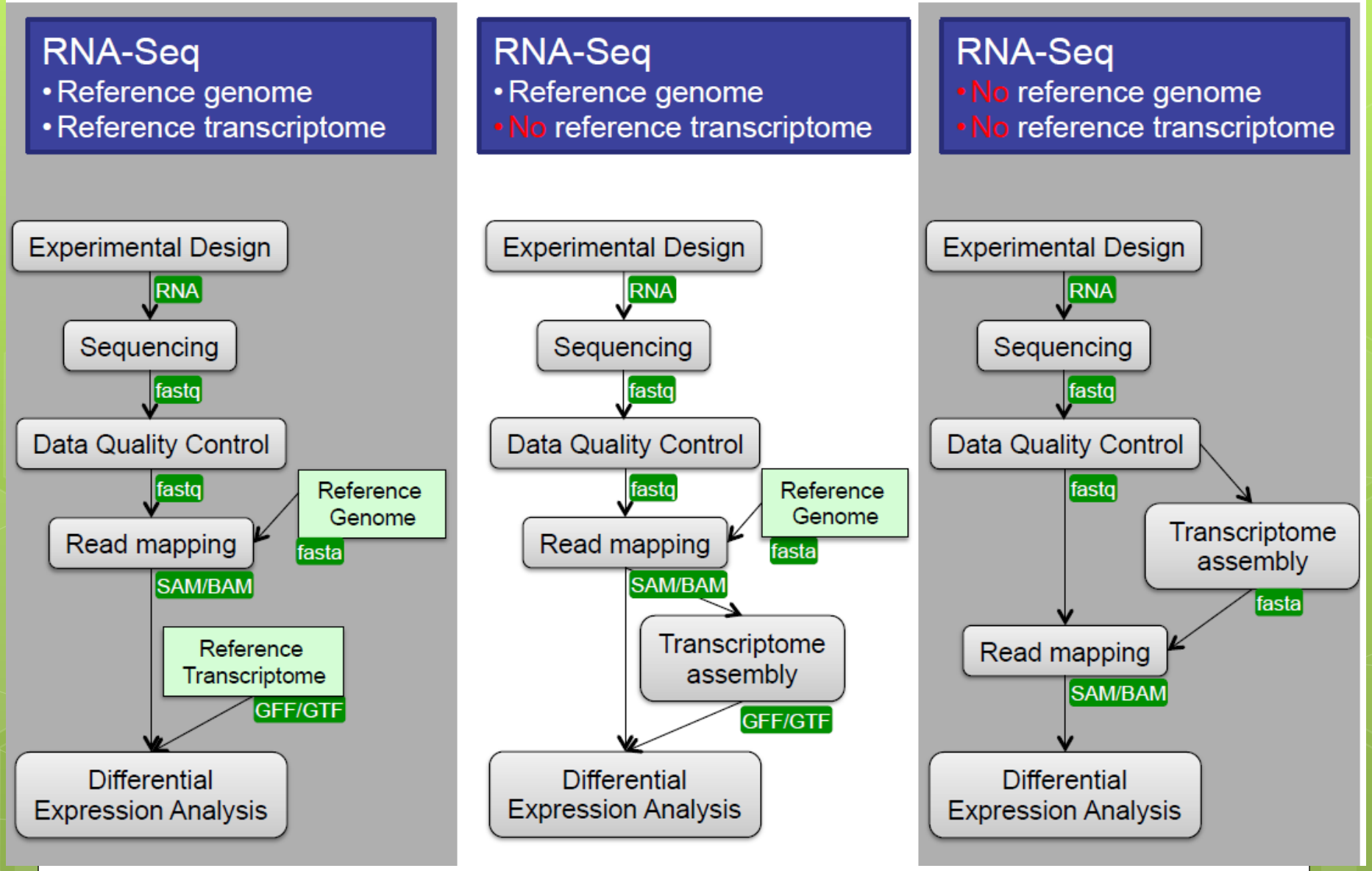


RNA-sequencing – análise dos dados

	Vantagens	Desvantagens
Genoma	• Menos alinhamentos múltiplos • Pode indicar novos éxons (não presentes no transcritptoma de referência)	• Limitado à expressão diferencial de éxons • Reads em junção éxon-éxon são perdidos • Não detecta fusão gênica • Necessita de um genoma de referência bem anotado
Transciptoma	• Consegue detectar diferentes transcritos, desde que conhecidos • Computacionalmente mais leve	• Aparecimento de possiveis alinhamentos mútilpos (diferentes isoformas / mesmos éxons) • Não detecta fusão gênica • Necessita de transcriptoma bem anotado
Assembly	• Não necessita de genoma/ transcriptoma de referência • Detecta fusão gênica • Detecta novas isoformas	• Computacionalmente intensivo para o genoma humano • Muitos métodos diferentes de análise com resultados divergentes

RNA-sequencing – análise dos dados

- **Solução → combinar mais de uma estratégia**
  - **Ex1:** Alinhar no genoma de referência e depois alinhar no transcriptoma de referência os reads que não foram mapeados
  - **Ex2:** Alinhar no transcriptoma de referência e depois fazer a montagem (*assembly*) dos reads que não foram mapeados
  - **Ex3:** Fazer primeiro a montagem e depois usar esse "transcriptoma montado" como referência para realinhar os reads
  - **Etc...**

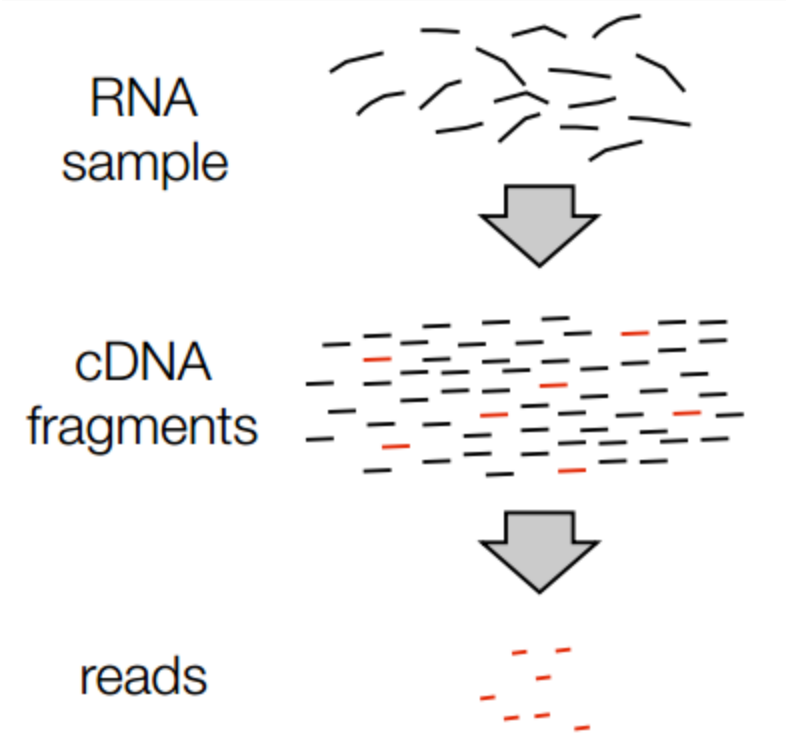

RNA-Seq
• Reference genome
• Reference transcriptome
Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Reference Genome
Read mapping
fasta
SAM/BAM
Reference Transcriptome
GFF/GTF
Differential Expression Analysis
RNA-Seq
• Reference genome
• No reference transcriptome
Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Reference Genome
Read mapping
fasta
SAM/BAM
Transcriptome assembly
GFF/GTF
Differential Expression Analysis
RNA-Seq
• No reference genome
• No reference transcriptome
Experimental Design
RNA
Sequencing
fastq
Data Quality Control
fastq
Transcriptome assembly
fasta
Read mapping
SAM/BAM
Differential Expression Analysis

RNA-sequencing – análise dos dados

- Perguntas:
  - Onde alinhar? Genoma de referência / Transcriptoma de referência / montagem (*assembly*) dos reads
  - **O que é a expressão gênica?** Contagem dos reads / Cobertura dos éxons
  - **Normalização?** Como garantir que amostras são comparáveis e que diferenças de reads são os únicos fatores relacionados à expressão?
  - Qual software usar?
  - Desenho experimental?

RNA-sequencing – análise dos dados

- **O que é a expressão gênica mensurada no RNA-seq?**



- O que quantificamos é uma amostra aleatória de fragmentos de cDNA derivados do RNA em nossa amostra
- Probabilisticamente, um gene com maior expressão será representado por mais reads

RNA-sequencing – análise dos dados

- Contagem dos reads
  - Número de reads que foram alinhados em um determinado éxon (genoma de referência) ou transcrito (transcriptoma de referência)

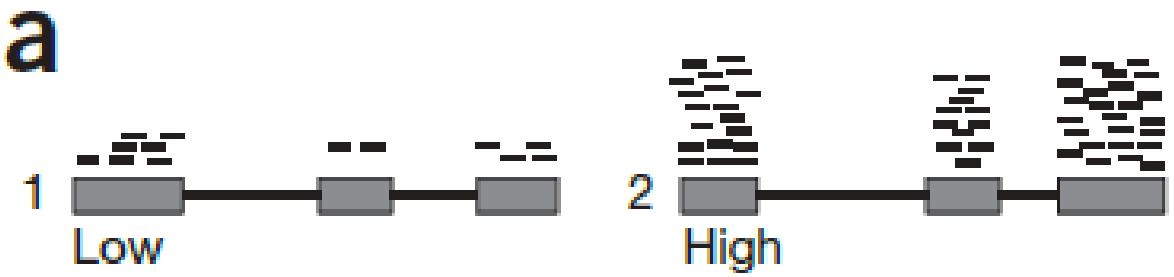
gene	wt1	wt2	mut1	mut2
24	3203	3215	2304	2220
12	23	30	14	5
5	2	3	0	5
2	1	5	0	6
34	0	3	0	2
21	13	14	14	0
56	54	59	32	31
3	12	155	12	16

RNA-sequencing – análise dos dados

- Contagem dos reads
  - **Podemos usar a contagem bruta dos reads para comparar a expressão gênica?**

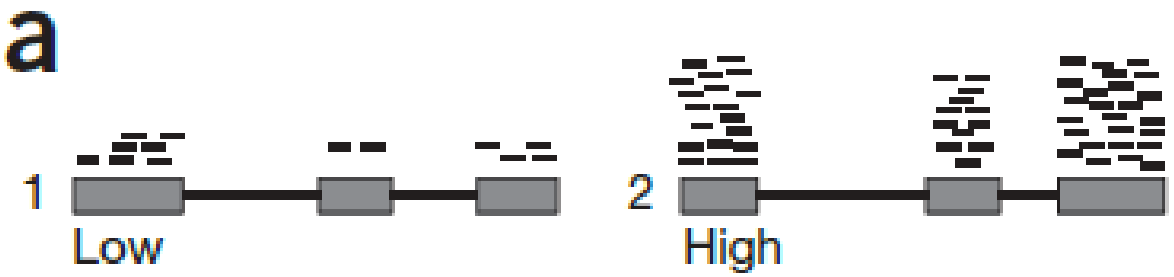
RNA-sequencing – análise dos dados

- Contagem dos reads
  - **Podemos usar a contagem bruta dos reads para comparar a expressão gênica?**



RNA-sequencing – análise dos dados

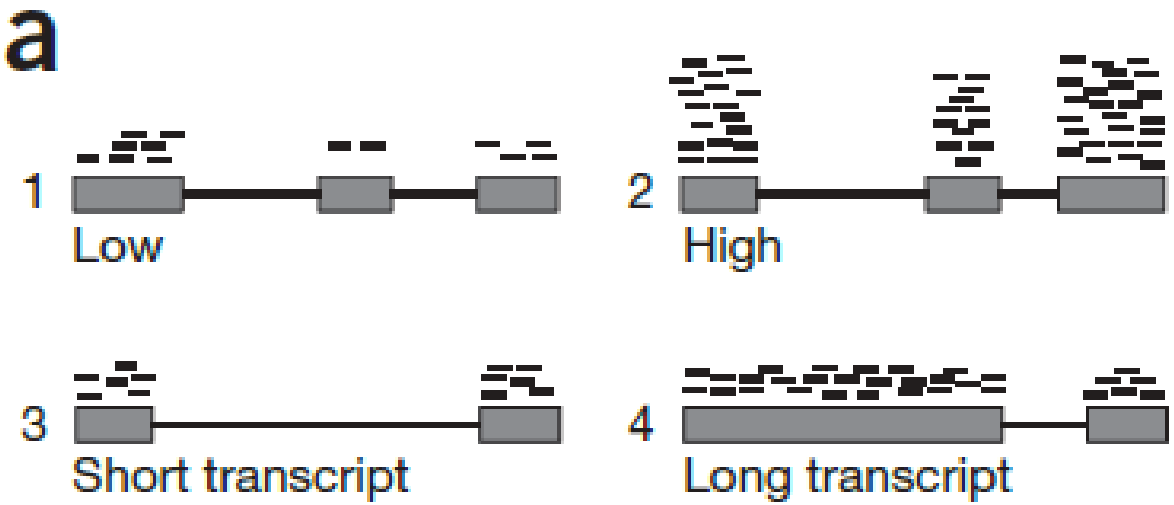
- Contagem dos reads
  - **Podemos usar a contagem bruta dos reads para comparar a expressão gênica?**



E se na amostra 2 eu simplesmente tinha uma cobertura maior?

RNA-sequencing – análise dos dados

- Contagem dos reads
  - **Podemos usar a contagem bruta dos reads para comparar a expressão gênica?**



RNA-sequencing – análise dos dados

- Contagem dos reads
  - Podemos usar a contagem bruta dos reads para comparar a expressão gênica?
    - **Não!** Uma *normalização* é necessária

RNA-sequencing – análise dos dados

- O que um método de normalização de dados de RNA-seq deve levar em consideração?
  - **Quantidade total de reads gerados**
  - **Tamanho do gene**
  - **Presença de genes muito expressos**

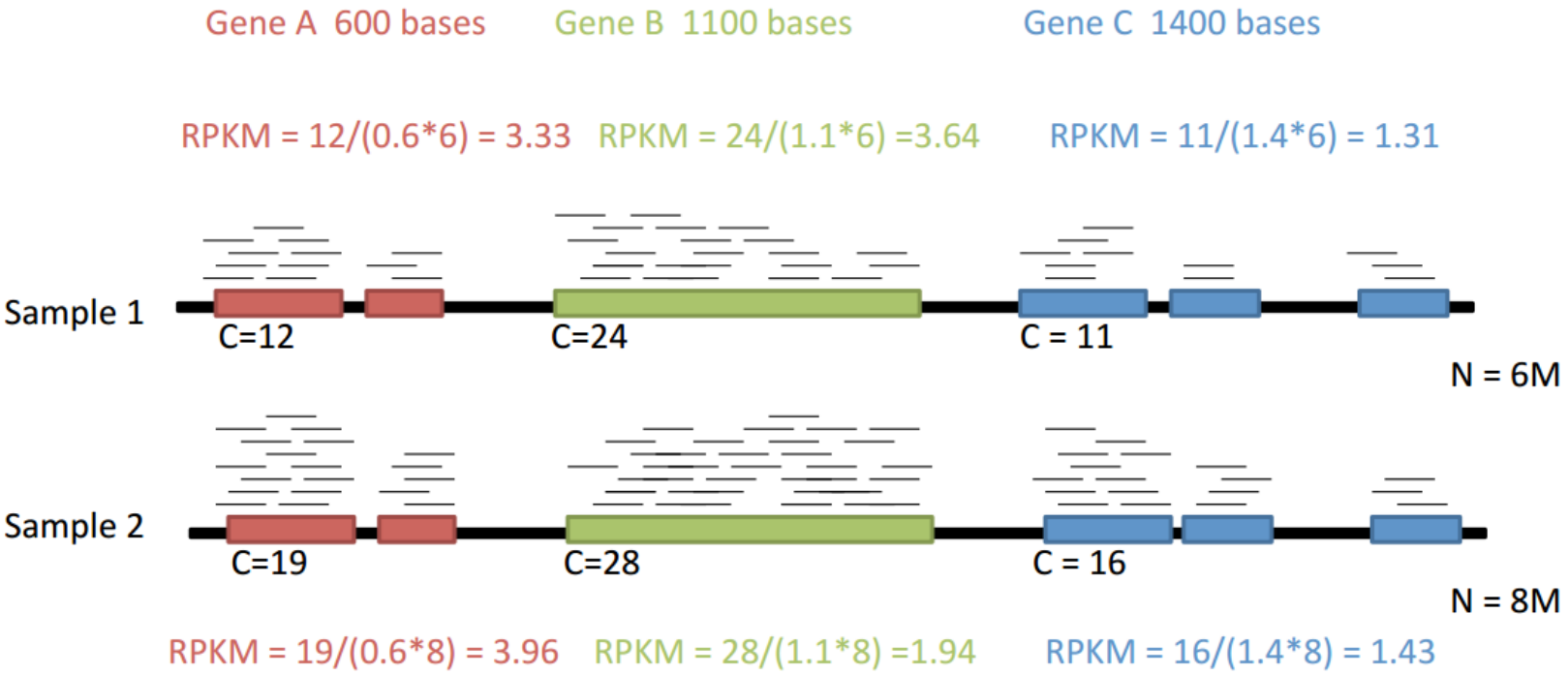
RNA-sequencing – análise dos dados

- Quantidade de reads e tamanho do gene
  - RPKM – *reads per kilobase million*

$$\text{RPKM} = \frac{\text{total mapped to gene}}{\text{total mapped to lane (in millions) x gene length (in kilobases)}}$$

RNA-sequencing – análise dos dados

- Quantidade de reads e tamanho do gene
  - RPKM – *reads per kilobase million*



RNA-sequencing – análise dos dados

- Quantidade de reads e tamanho do gene
  - RPKM – *reads per kilobase million*
  - **Desvantagem:** genes muito expressos podem dominar a capacidade de sequenciamento e afetar de maneira desproporcional o valor de expressão de genes com baixa expressão.

RNA-sequencing – análise dos dados

- Presença de genes muito expressos

Gene	Sample 1 absolute abundance	Sample 1 relative abundance	Sample 2 absolute abundance	Sample 2 relative abundance
1	20	10%	20	5%
2	20	10%	20	5%
3	20	10%	20	5%
4	20	10%	20	5%
5	20	10%	20	5%
6	100	50%	300	75%

RNA-sequencing – análise dos dados

- Outras estratégias de normalização:
  - **Total count**
  - **Upper quartile**
  - **Median**
  - **DESeq**
  - **Trimmed Mean of M-values**
  - **Quantile**
  - **RPKM**

RNA-sequencing – análise dos dados

- Outras estratégias de normalização:

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

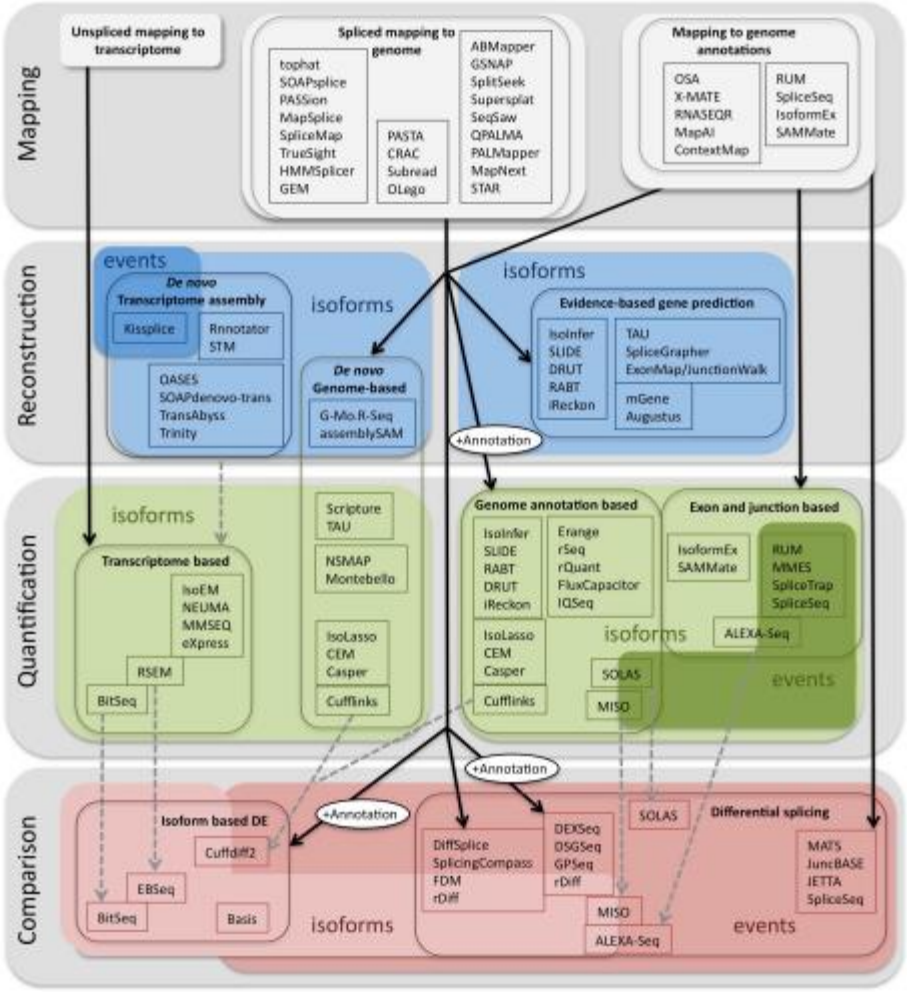
RNA-sequencing – análise dos dados

- Perguntas:
  - Onde alinhar? Genoma de referência / Transcriptoma de referência / montagem (*assembly*) dos reads
  - O que é a expressão gênica? Contagem dos reads / Cobertura dos éxons
  - Normalização? Como garantir que amostras são comparáveis e que diferenças de reads são os únicos fatores relacionados à expressão?
  - **Qual software usar?**
  - Desenho experimental?

RNA-sequencing – análise dos dados

- Qual software usar para cada uma das etapas?
  - Depende da estratégia adotada (alinhamento no genoma, transcriptoma e montagem)
  - Depende da estratégia de normalização
- Problemas → **muitas opções!**

RNA-sequencing – análise dos dados



RNA-sequencing – análise dos dados

- Qual software usar para cada uma das etapas?

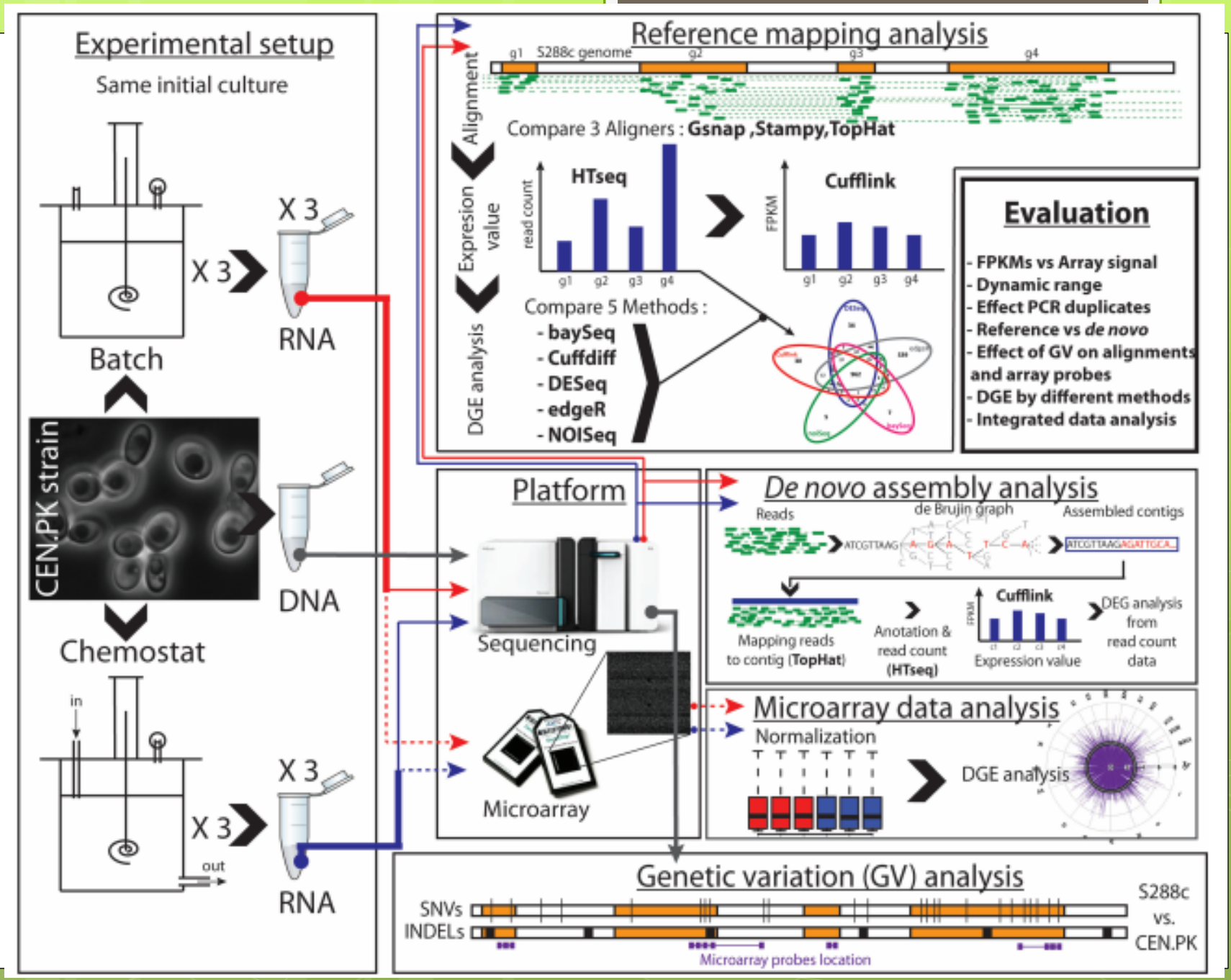
10084–10097 Nucleic Acids Research, 2012, Vol. 40, No. 20 *Published online 10 September 2012*
doi:10.1093/nar/gks804

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*

Intawat Nookaew[1], Marta Papini[1], Natapol Pornputtapong[1], Gionata Scalcinati[1], Linn Fagerberg[2], Matthias Uhlén[2,3] and Jens Nielsen[1,3,*]

[1]Novo Nordisk Foundation Center for Biosustainability, Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-41296, Gothenburg, Sweden, [2]Novo Nordisk Foundation Center for Biosustainability, Department of Biotechnology, Royal Institute of Technology, SE-10691, Stockholm, Sweden and [3]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2970 Hørsholm, Denmark

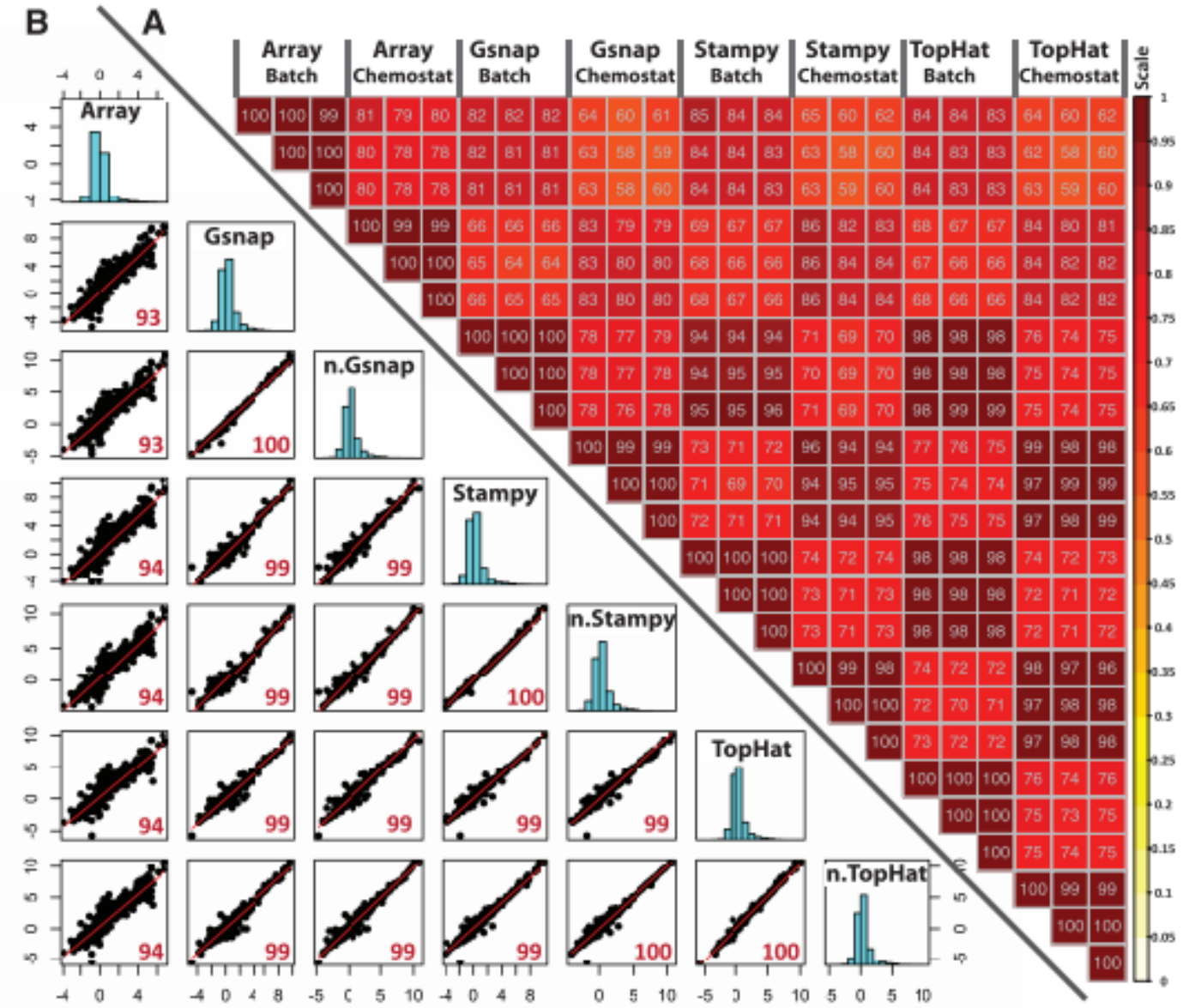
Received May 9, 2012; Revised and Accepted July 31, 2012


Experimental setup
Same initial culture
X 3
X 3
RNA
Batch
CEN.PK strain
DNA
Chemostat
in
out
X 3
X 3
RNA
Reference mapping analysis
g1 S288c genome g2 g3 g4
Alignment
Compare 3 Aligners : Gsnap ,Stampy,TopHat
Expresion value
HTseq
read count
g1 g2 g3 g4
Cufflink
FPKM
g1 g2 g3 g4
DGE analysis
Compare 5 Methods :
- baySeq
- Cuffdiff
- DESeq
- edgeR
- NOISeq
Evaluation
- FPKMs vs Array signal
- Dynamic range
- Effect PCR duplicates
- Reference vs de novo
- Effect of GV on alignments and array probes
- DGE by different methods
- Integrated data analysis
Platform
Sequencing
Microarray
De novo assembly analysis
Reads
de Brujin graph
Assembled contigs
ATCGTTAAG
ATCGTTAAGAGATTGCA
Mapping reads to contig (TopHat)
Anotation & read count (HTseq)
Cufflink
Expression value
DEG analysis from read count data
Microarray data analysis
Normalization
DGE analysis
Genetic variation (GV) analysis
SNVs
INDELs
Microarray probes location
S288c vs. CEN.PK

RNA-sequencing – análise dos dados

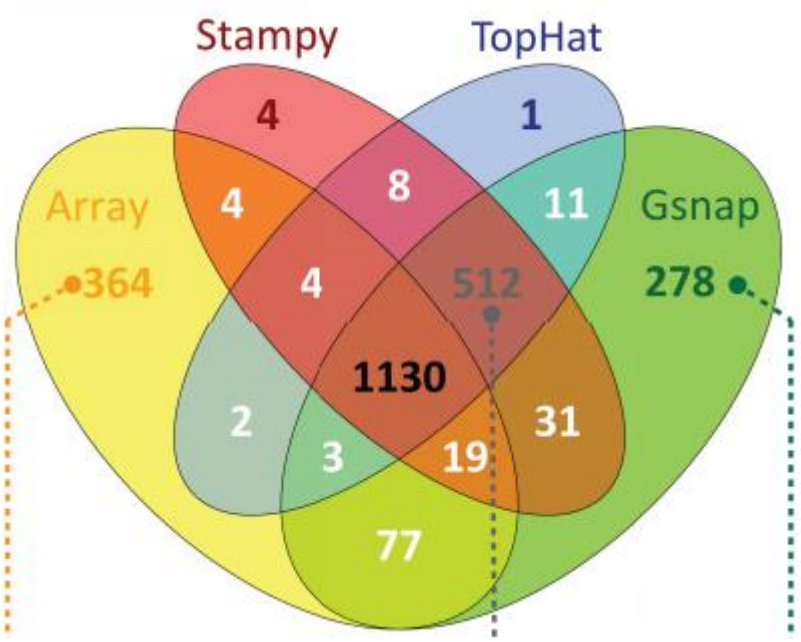
- Qual software usar para cada uma das etapas?
  - Alinhamento:
    - Gsnap (rápido)
    - Stampy (sensível)
    - TopHat (mais usado)
  - Identificação de genes diferencialmente expressos
    - Cuffdiff
    - baySeq
    - DESeq
    - edgeR
    - NOISeq

RNA-Seq

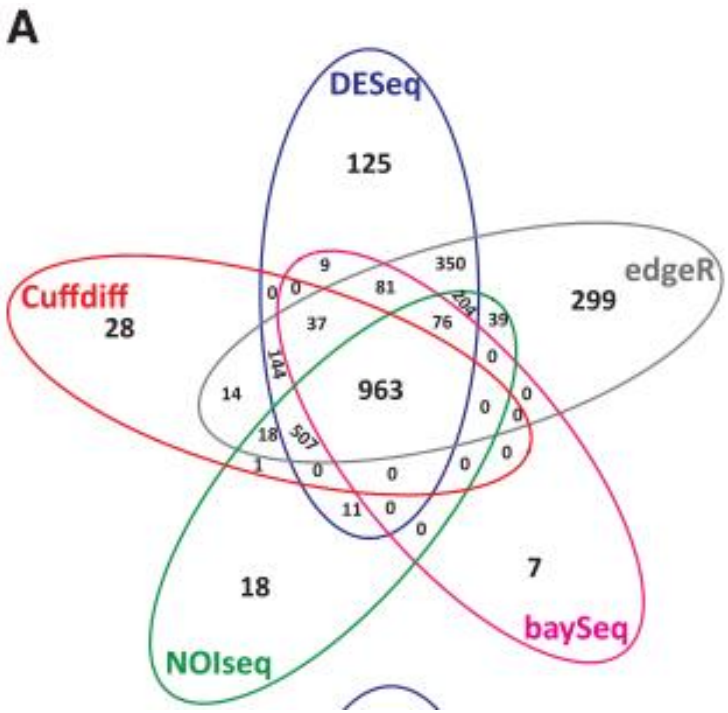

B
A
Array
Gsnap
n.Gsnap
Stampy
n.Stampy
TopHat
n.TopHat
Array Batch
Array Chemostat
Gsnap Batch
Gsnap Chemostat
Stampy Batch
Stampy Chemostat
TopHat Batch
TopHat Chemostat
Scale

RNA-sequencing – análise dos dados

- Genes diferencialmente expressos



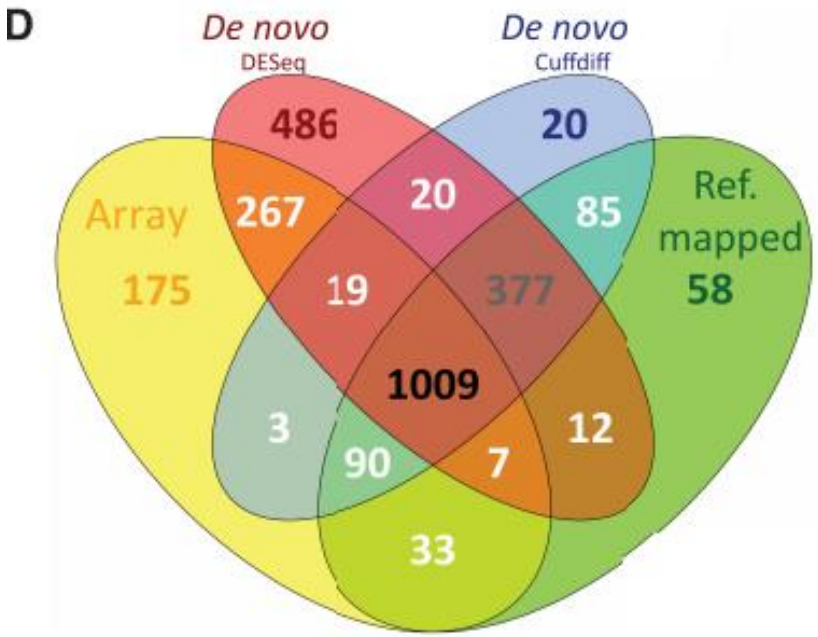
Diferentes alinhadores



Diferentes métodos de análise de genes dif. expressos

RNA-sequencing – análise dos dados

- Genes diferencialmente expressos



Alinhamento no transcriptoma versus montagem *de novo*

RNA-sequencing – análise dos dados

- Conclusões

In order to address the impact of different statistical methods on the identification of DGE, we found that Cuffdiff, baySeq, DESeq, edgeR and NOISeq generated consistent results. Additionally, the results obtained based on RNA-seq data were in good agreement with microarray data. Interestingly, edgeR identified more DGE than the other methods at the same cut-off, which might infer less control of type 1 error with this method. Using results derived from different statistical methods of RNA-seq gave similar biological interpretations as is shown in GO enrichment analysis. This result strongly supports the robustness and reliability of different processing and analysis of RNA-seq data. Furthermore, we identified high consistency between microarray and RNA-seq platforms, thus encouraging the continual use of microarray as a versatile tool for differential gene expression analysis. In conclusion, our study provides a comprehensive comparison of different methods for analyses of *S. cerevisiae* transcriptome based on RNA-seq data using Illumina platform, elucidating the contribution of the different steps involved in analysis of RNA-seq data.

RNA-sequencing – análise dos dados

- Perguntas:
  - Onde alinhar? Genoma de referência / Transcriptoma de referência / montagem (*assembly*) dos reads
  - O que é a expressão gênica? Contagem dos reads / Cobertura dos éxons
  - Normalização? Como garantir que amostras são comparáveis e que diferenças de reads são os únicos fatores relacionados à expressão?
  - Qual software usar?
  - **Desenho experimental?**

RNA-sequencing – análise dos dados

Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge[1]

Department of Statistics, Purdue University, West Lafayette, Indiana 47907

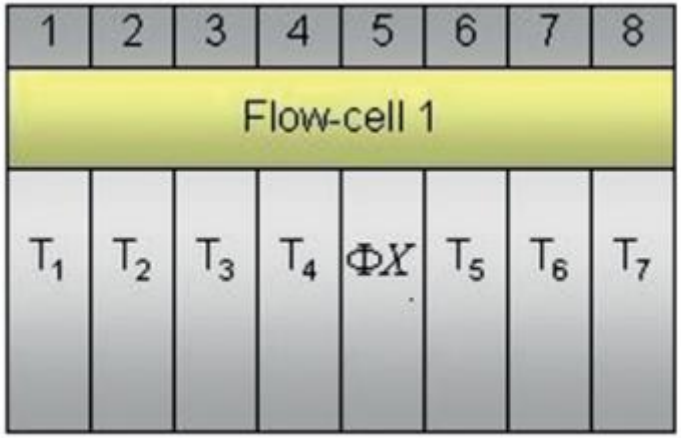
Manuscript received January 31, 2010
Accepted for publication March 15, 2010

ABSTRACT

Next-generation sequencing technologies are quickly becoming the preferred approach for characterizing and quantifying entire genomes. Even though data produced from these technologies are proving to be the most informative of any thus far, very little attention has been paid to fundamental design aspects of data collection and analysis, namely sampling, randomization, replication, and blocking. We discuss these concepts in an RNA sequencing framework. Using simulations we demonstrate the benefits of collecting replicated RNA sequencing data according to well known statistical designs that partition the sources of biological and technical variation. Examples of these designs and their corresponding models are presented with the goal of testing differential expression.

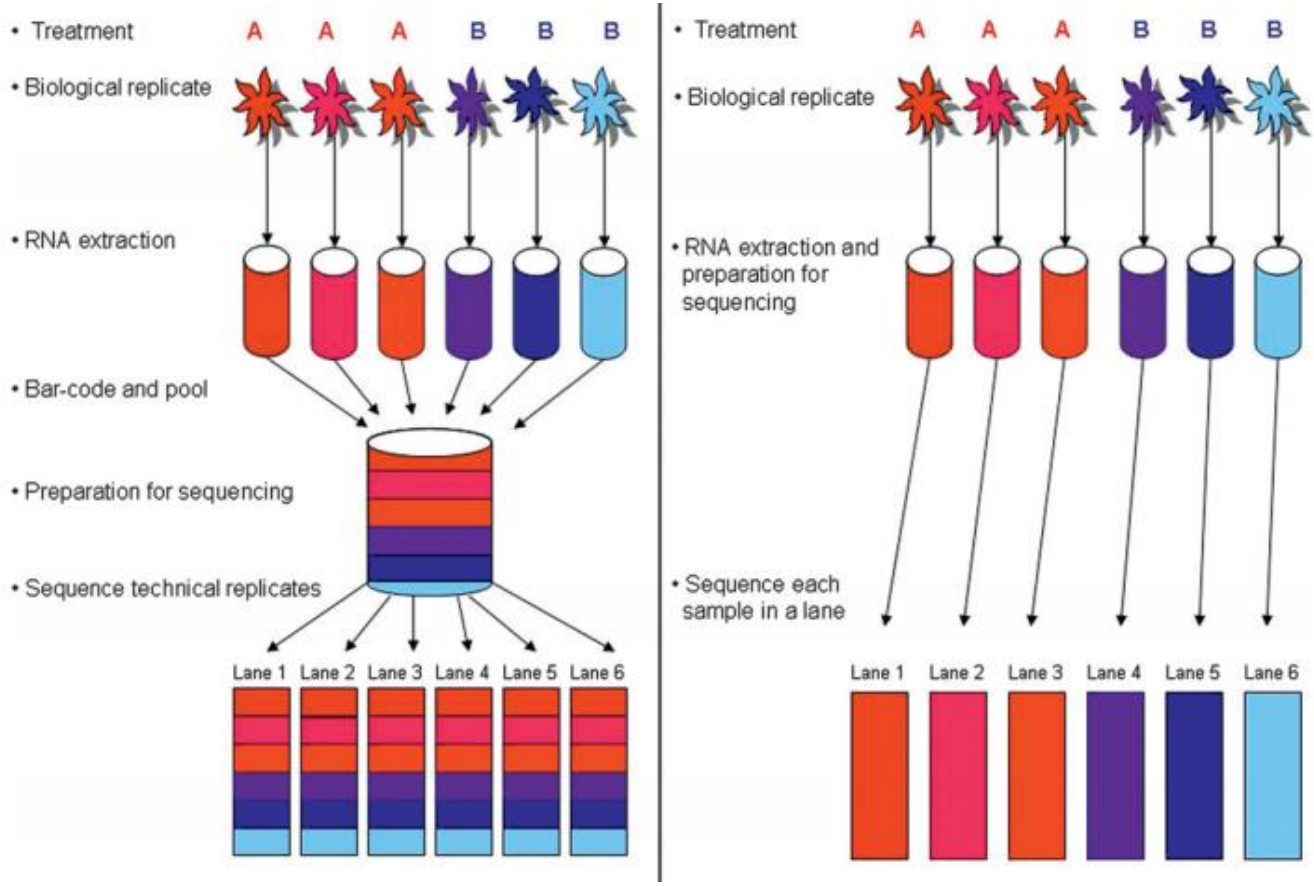
RNA-sequencing – análise dos dados

- Avaliação de desenhos experimentais mais favoráveis para análise de RNA-seq
- Situação de sequenciamento de 8 lanes (Ex: Illumina Genome Analyzer)



RNA-sequencing – análise dos dados

- Qual desenho experimental mais favorável e por quê?



RNA-sequencing – análise dos dados

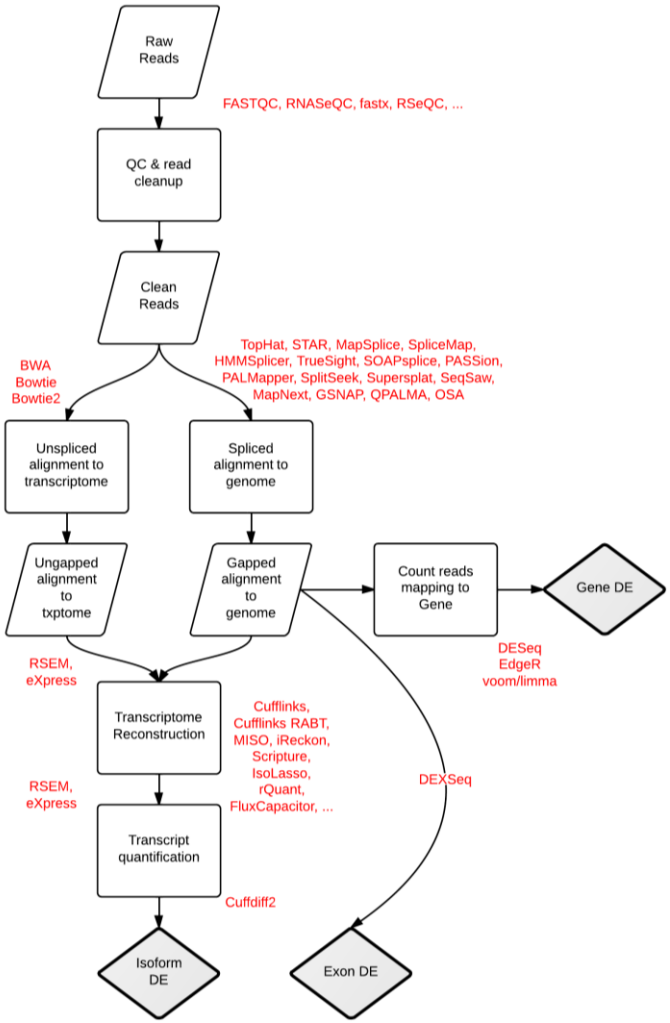
- Parâmetros definidos com antecedência:
  - Quais/ quantos grupos serão comparados?
  - Dados Paired-end ou single-end?
  - Tamanho do read
  - Cobertura
  - Replicatas biológicas (não há necessidade de replicata técnica)
  - Pool de amostras
  - Amostragem
  - *Blocking*

RNA-sequencing – análise dos dados

- *ENCODE RNA-seq standards*

- *http://encodeproject.org/ENCODE/protocols/dataStandards/RNA_standards_v1_2011_May.pdf*

RNA-seq pipelines



RNA-seq Tuxedo pipeline

Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package
Cufflinks assembles transcripts
Cuffdiff identifies differential expression of genes/ transcripts/promoters

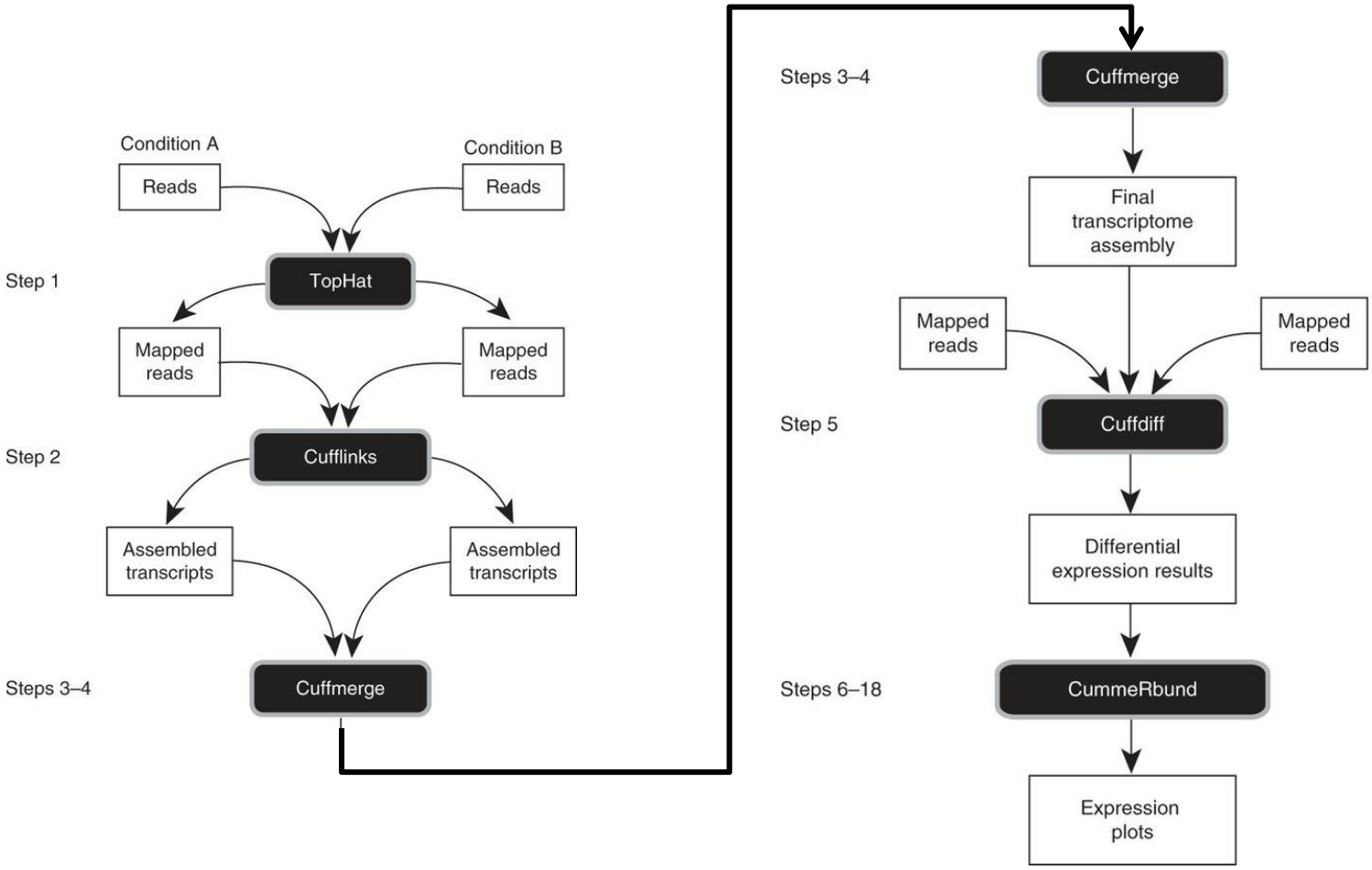
CummeRbund
Plots abundance and differential expression results from Cuffdiff

RNA-seq Tuxedo pipeline

- TopHat → alinhamento dos reads em um genoma de referência, usando o Bowtie de maneira *splice aware*
- Cufflinks → usa resultados do alinhamento para montar diferentes transcritos (cufflinks), compara resultados com transcriptoma de referência (cuffcompare), identificando assim diferentes isoformas de um mesmo gene (cuffmerge) e faz o teste para verificar transcritos diferencialmente expressos (cuffdiff)
- CummeRbund (pacote do R) → cria gráficos com os resultados do Cufflinks

- **Requerimento** → existência de um genoma de referência bem anotado

RNA-seq Tuxedo pipeline



RNA-seq Tuxedo pipeline

PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Computer Science, University of California, Berkeley, California, USA. [4]Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [7]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [8]Department of Mathematics, University of California, Berkeley, California, USA. [9]Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. Correspondence should be addressed to C.T. (cole@broadinstitute.org).

RNA-seq Aula prática

- Reads *single-end* do cromossomo 20 de uma amostra de Fígado e Rim humanos.
- Alinhamento no transcriptoma de referência (chr20)
- Processamento dos reads
- Contagem dos reads por transcrito
- Análise de expressão gênica diferencial (apenas transcritos anotados)