



CCAGGCGAGTTTCCCCAAAGGG GGCATTATTGGCCAATCGAATC GATCCAGCCTTCAAACGGGTTC TGCCACTGGAGGCCCAATACC

> OBERT R. SPRAGUE FOUNDATION HALL



# Commonly Used Analytical Tools for Bulk RNA-seq

Fabio Macciardi MD PhD

Laboratory of Molecular Psychiatry

**UC** Irvine

# Outline

### **Pipeline(s) for RNA-Seq data analysis**

- Digital RNA-sequencing = measures gene expression in true genome-wide fashion (all the c/nc RNA)
- Also it enables detection of mutations (SNPs), alternative splicing, allele specific expression, and fusion genes
- Can also be used to detect non-coding / regulatory RNA "genes" (e.g., miRNA, siRNA, IncRNAs, TSS, enhancers, ...)
- Hence, methods and tools to perform RNA-seq analyses must be chosen based on the given goal(s) of the research

### **RNA SEQUENCING: A QUICK RECAP**

### **RNA – seq quick recap**

### From the genomic region (DNA) to the RNA ... and back



### **Step 1: prepare RNA-Seq Libraries**



### **Step 2: select the good-sized RNA molecules**



From the "n" RNA fragments select those fragments that range from ~150 bp to ~ 400 bp

Step 3: For each "fragment" of 150 to 300 bp long you then sequence the first 100bp FWD (Read1) and the last 100bp REV (Read2)





Step 4: The random fragmentation process generates <u>millions</u> of RNA (or cDNA) fragments, in excess of the total length of the transcriptome. By chance, fragments can partially or completely overlap to each other, "covering" the whole transcriptome



#### Step 5: the read are aligned against a reference genome



# Short reads mRNA Sequencing (Illumina)



#### QC of RNA reads: an example using CTRL # 58

#### RAW data

#### Filtered data



Position in read (bp)

Position in read (bp)



### RNA SEQUENCING: DIFFERENT ANALYTICAL PIPELINES FOR RNA-SEQ

### A survey of best practices for RNA-seq data analysis

There is no optimal pipeline for the variety of different applications and analysis scenarios in which RNA-seq can be used.

Every RNA-seq – aka, **digital RNA sequencing** - experimental scenario could potentially have different optimal methods for

- transcript mapping (= read alignment against a known genome, "new" transcript discovery, ...)
- transcript quantification
- *normalization*, and ultimately
- differential expression analysis

Conesa, Ana, Madrigal, Pedro, Tarazona, Sonia, et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17, 13.

### Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction

### = 278 RNA-SEQ PIPELINES .....

#### **Good-Performing RNA-seq Pipelines for Various Applications**

| <b>RNA-seq Application</b>  | Metric  | RNA-seq Pipelines  |  |  |  |  |  |  |
|---|---|--|--|--|--|--|--|--|
| Accurate estimation of<br>relative gene expression with<br>benefit for differentially<br>expressed gene detection                   | <b>Accuracy</b><br>(Deviation from<br>qPCR)   | <ul> <li>Bowtie + RSEM + Median</li> <li>Bowtie 2 Single-Hit +<br/>[Count-Based/Cufflinks/RSEM] + Median</li> <li>Bowtie 2 Multi-Hit + RSEM + Median</li> <li>Bowtie 2 Multi-Hit + RSEM + Median</li> <li>BWA + [Count-Based/RSEM] + Median</li> <li>GSNAP Spliced [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks] + Median</li> <li>GSNAP Un-spliced Multi-Hit +<br/>RSEM + Median</li> <li>MAGIC [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks] + Median</li> <li>MAGIC [Single-/Multi-Hit] +<br/>Count-Based + Median</li> <li>MAGIC [Single-/Multi-Hit] +<br/>Count-Based + Median</li> <li>STAR + Count-Based + Median</li> <li>TopHat [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks] + Median</li> <li>WHAM Single-Hit + Count-Based + Median</li> <li>WHAM Multi-Hit + RSEM + Median</li> </ul> |  |  |  |  |  |  |
| Small variation in gene<br>expression <u>across</u> all<br>replicate libraries for a<br>single sample                               | <b>Precision</b><br>(Coefficient of<br>variation across<br>replicate libraries)         | <ul> <li>Bowtie2 Multi-Hit + Count-Based + [FPM/FPKM/Upper Quartile]</li> <li>Bowtie2 Multi-Hit + Cufflinks + RLE</li> <li>GSNAP Un-Spliced [Single-/Multi-Hit] + [Count-Based/Cufflinks] + [FPM/FPKM/Upper Quartile/RLE]</li> <li>GSNAP Un-Spliced [Single-/Multi-Hit] + Cufflinks + TMM</li> </ul>   |  |  |  |  |  |  |
| Small within-sample<br>variation in gene expression<br>across all replicate libraries<br>compared with between-<br>sample variation | <b>Reliability</b><br>(Intraclass<br>[intra-sample]<br>correlation for<br>grouped data) | <ul> <li>Bowtie2 [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks/RSEM] + Median</li> <li>BWA + [Count-Based/Cufflinks/RSEM] + Median</li> <li>GSNAP Spliced [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks] + Median</li> <li>MAGIC [Single-/Multi-Hit] +<br/>Count-Based + Median</li> <li>MapSplice + [Count-Based/Cufflinks] + Median</li> <li>MapSplice + [Count-Based/Cufflinks] + Median</li> <li>MapSplice + [Count-Based/Cufflinks] + Median</li> </ul>   |  |  |  |  |  |  |

Tong, L., Wu, P. Y., Phan, J. H., et al. 2020. Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction. *Sci Rep*, 10, 17925.

# **RNA-seq informatics**

- Filter out rRNA, tRNA, mitoRNA
- Align to genome
- Find splice junction fragments (join exon boundaries)
- Differential expression
- Alternatively spliced transcripts
- Novel genes/exons
- Sequence variants (SNPs, indels, translocations)
- Allele-specific expression

### RNA-sequencing pipeline = the original TUXEDO workflow ....



### **Real data generally support existing annotation**



# **RNA-seq informatics**

- Filter out rRNA, tRNA, mitoRNA
- Align to genome
- Find splice junction fragments (join exon boundaries)
- Alternatively spliced transcripts
- Differential expression
- Novel genes/exons
- Sequence variants (SNPs, indels, translocations)
- Allele-specific expression



### **Alternative Splicing**



# Map reads to exons & junctions = mapping <u>transcripts</u>, NOT genes



In RNA sequencing analyses we DO NOT usually use the absolute # of reads (read counts), but we use a normalized measure known as FPKM (Fragments Per Kilobase of exon per Million fragments mapped) to allow for undistorted comparisons across subjetcs





Pertea et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33, 290-295.



#### Step 5: the read are aligned against a reference genome



### IGV visualization of a RNA-seq analysis: coverage and reads display along the MEF2C gene



# IGV visualization of a RNA-seq analysis: coverage and reads display along the MEF2C gene





(a) The **coverage plot** shows the sum of mapped reads at each position as grey peaks. In the middle, each read is displayed where it maps. The blue lines indicate the junction events (or splice sites)

(b) **Sashimi plot** showing the coverage in red with the arcs representing the splice junctions. The numbers refer to the number of reads spanning the junctions. On the bottom, the different groups of linked boxes represent the different transcripts from the genes at this location that are present in the GTF file

### A survey of best practices for RNA-seq data analysis



Conesa, Ana, Madrigal, Pedro, Tarazona, Sonia, et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17, 13.

### Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction

### = 278 RNA-SEQ PIPELINES .....

#### **Good-Performing RNA-seq Pipelines for Various Applications**

| <b>RNA-seq Application</b>  | Metric  | RNA-seq Pipelines  |  |  |  |  |  |  |
|---|---|--|--|--|--|--|--|--|
| Accurate estimation of<br>relative gene expression with<br>benefit for differentially<br>expressed gene detection                   | <b>Accuracy</b><br>(Deviation from<br>qPCR)   | <ul> <li>Bowtie + RSEM + Median</li> <li>Bowtie 2 Single-Hit +<br/>[Count-Based/Cufflinks/RSEM] + Median</li> <li>Bowtie 2 Multi-Hit + RSEM + Median</li> <li>Bowtie 2 Multi-Hit + RSEM + Median</li> <li>BWA + [Count-Based/RSEM] + Median</li> <li>GSNAP Spliced [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks] + Median</li> <li>GSNAP Un-spliced Multi-Hit +<br/>RSEM + Median</li> <li>MAGIC [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks] + Median</li> <li>MAGIC [Single-/Multi-Hit] +<br/>Count-Based + Median</li> <li>MAGIC [Single-/Multi-Hit] +<br/>Count-Based + Median</li> <li>STAR + Count-Based + Median</li> <li>TopHat [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks] + Median</li> <li>WHAM Single-Hit + Count-Based + Median</li> <li>WHAM Multi-Hit + RSEM + Median</li> </ul> |  |  |  |  |  |  |
| Small variation in gene<br>expression <u>across</u> all<br>replicate libraries for a<br>single sample                               | <b>Precision</b><br>(Coefficient of<br>variation across<br>replicate libraries)         | <ul> <li>Bowtie2 Multi-Hit + Count-Based + [FPM/FPKM/Upper Quartile]</li> <li>Bowtie2 Multi-Hit + Cufflinks + RLE</li> <li>GSNAP Un-Spliced [Single-/Multi-Hit] + [Count-Based/Cufflinks] + [FPM/FPKM/Upper Quartile/RLE]</li> <li>GSNAP Un-Spliced [Single-/Multi-Hit] + Cufflinks + TMM</li> </ul>   |  |  |  |  |  |  |
| Small within-sample<br>variation in gene expression<br>across all replicate libraries<br>compared with between-<br>sample variation | <b>Reliability</b><br>(Intraclass<br>[intra-sample]<br>correlation for<br>grouped data) | <ul> <li>Bowtie2 [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks/RSEM] + Median</li> <li>BWA + [Count-Based/Cufflinks/RSEM] + Median</li> <li>GSNAP Spliced [Single-/Multi-Hit] +<br/>[Count-Based/Cufflinks] + Median</li> <li>MAGIC [Single-/Multi-Hit] +<br/>Count-Based + Median</li> <li>MapSplice + [Count-Based/Cufflinks] + Median</li> <li>MapSplice + [Count-Based/Cufflinks] + Median</li> <li>MapSplice + [Count-Based/Cufflinks] + Median</li> </ul>   |  |  |  |  |  |  |

Tong, L., Wu, P. Y., Phan, J. H., et al. 2020. Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction. *Sci Rep*, 10, 17925.

Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction



Tong, L., Wu, P. Y., Phan, J. H., et al. 2020. Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction. *Sci Rep*, 10, 17925.

Analysis workflow of publicly available RNA sequencing datasets



Step 1

Step 2

datasets

**Dataset selection** 

Æ

Browse public repositories for eligible

Download clinical and sequencing data

Data analysis

2 days

Delineate your eligibility criteria

Sanchis et al., STAR Protocols 2, 100478 2021 https://doi.org/10.1016/j.xpro.2021.100478

Sanchis, P., Lavignolle, R., Abbate, M., et al. 2021. Analysis workflow of publicly available RNA-sequencing datasets. *STAR Protoc*, 2, 100478.

Before you begin

2 hours

and packages

Bioconductor

Download and install: R. Rstudio

Counts and clinical

### RNA SEQUENCING: THE CHOICE OF A SPECIFIC ANALYTICAL PIPELINE DEPENDS ON THE COMPLEXITY OF THE TRANSCRIPTOME

### Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis



Corchete, L. A., Rojas, E. A., Alonso-Lopez, D., et al. 2020. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci Rep*, 10, 19737

# Transcripts complexity ALOX12, K=6

|   | Cluster   | Transcript                   | pval        | qval       |                  |
|---|-----------|------------------------------|-------------|------------|------------------|
| 1 | cluster 1 | 118180                       | 0.005537547 | 0.03322528 | 118166           |
| 2 | cluster 2 | 118182,<br>118165            | 0.454418905 | 0.55115251 | 118163<br>118181 |
| 3 | cluster 3 | 118177                       | 0.414217008 | 0.55115251 | 118169           |
| 4 | cluster 4 | 118181                       | 0.602722552 | 0.60272255 | 118164           |
| 5 | cluster 5 | 118166,<br>118163            | 0.459293760 | 0.55115251 | 118180           |
| 6 | cluster 6 | 118164,<br>118167,<br>118169 | 0.269000193 | 0.55115251 | 118182<br>118165 |

MSTRG.48499: transcripts clustered with kmeans, k=6



genomic position

## Transcripts complexity ALOX12, K=6



### Transcripts complexity MEF2C = 15 transcripts .... & counting

|   |          |          |   |           |        |             |            |            |             |                             |    |      |           | FIOLEIII COUIIIg - LI FA7 (5524 |
|---|----------|----------|---|-----------|--------|-------------|------------|------------|-------------|-----------------------------|----|------|-----------|---------------------------------|
|   |          |          | - |           |        |             |            |            |             |                             |    |      |           | bp) only human and it is in a   |
| 5 | 88722489 | 88883181 |   | MEF2C     | 236486 | mediumpurp  | 0.57469151 | 0.00019921 | MSTRG.98135 | transcript:ENST00000508569  | 10 | 1882 | MEF2C     | highly transcribed region       |
| 5 | 89418807 | 89466398 | + | MEF2C-AS1 | 236544 | salmon      | -0.5461187 | 0.00047143 | MSTRG.98141 | transcript:ENST00000508742  | 3  | 679  | MEF2C-AS1 |                                 |
| 5 | 88722837 | 88729519 | - | MEF2C     | 236491 | saddlebrown | -0.5288615 | 0.00076495 | MSTRG.98135 | transcript:ENST00000510980  | 3  | 527  | MEF2C     | Retained intron - shortest      |
| 5 | 88880844 | 88881264 | - | MEF2C     | 236511 | green       | 0.50830235 | 0.00131805 | MSTRG.98135 | transcript:ENST00000629847  | 1  | 421  | MEF2C     | Processed transcript            |
| 5 | 88720829 | 88883184 | - | MEF2C     | 236481 | mediumpurp  | 0.47950018 | 0.00267435 | MSTRG.98135 | transcript:ENST00000510942  | 10 | 3446 | MEF2C     |                                 |
| 5 | 88804526 | 88883174 | - | MEF2C     | 236503 | lightcyan   | -0.214636  | 0.20206241 | MSTRG.98135 | transcript:ENST00000511086  | 3  | 689  | MEF2C     |                                 |
| 5 | 88722498 | 88749115 | - | MEF2C     | 236488 | darkgrey    | 0.2100348  | 0.21213528 | MSTRG.98135 | transcript:ENST00000627717  | 5  | 815  | MEF2C     |                                 |
| 5 | 88804770 | 88824275 | - | MEF2C     | 236504 | lightgreen  | 0.14291944 | 0.39875254 | MSTRG.98135 | transcript: ENST00000509373 | 2  | 573  | MEF2C     |                                 |
| 5 | 88731437 | 88749124 | - | MEF2C     | 236492 | tan         | -0.1333192 | 0.4314948  | MSTRG.98135 | transcript: ENST00000515715 | 2  | 520  | MEF2C     |                                 |
| 5 | 88761023 | 88827158 | - | MEF2C     | 236499 | white       | 0.12439238 | 0.46323188 | MSTRG.98135 | transcript: ENST00000507984 | 4  | 623  | MEF2C     |                                 |
| 5 | 88731852 | 88824340 | - | MEF2C     | 236494 | greenyellow | 0.09637388 | 0.57042721 | MSTRG.98135 | transcript: ENST00000506716 | 7  | 854  | MEF2C     |                                 |
| 5 | 88965875 | 89168668 | + | MEF2C-AS1 | 236531 | red         | 0.09637388 | 0.57042721 | MSTRG.98141 | transcript:ENST00000514571  | 4  | 594  | MEF2C-AS1 |                                 |
| 5 | 88731805 | 88883147 | - | MEF2C     | 236493 | white       | 0.08383415 | 0.62179311 | MSTRG.98135 | transcript:ENST00000513252  | 7  | 1063 | MEF2C     |                                 |
| 5 | 88823861 | 88883224 | - | MEF2C     | 236507 | darkgreen   | NA         | NA         | MSTRG.98135 | transcript: ENST00000509349 | 3  | 582  | MEF2C     |                                 |
| 5 | 88889308 | 88943343 | + | MEF2C-AS1 | 236517 | sienna3     | NA         | NA         | MSTRG.98141 | transcript:ENST00000513704  | 3  | 563  | MEF2C-AS1 |                                 |
|   |          |          |   |           |        |             |            |            |             |                             |    |      |           |                                 |



Ductoin coding 11 DAZ (FF24



Differential expression of MEF2C

T3 = retained intron - regulator

T1 = protein coding

#### MEF2C



MEF2C



#### **BOLA2**





#### combined



### C4A and an intronic LTR in Schizophrenia



#### C4A and an intronic LTR in Schizophrenia

