RNA-seq experimental considerations

Mikhail Dozmorov

Spring 2018

Library preparation

Library preparation steps

- **RNA isolation and QC**, to extract RNA relevant to the experimental question
- **Fragmentation**, to recover short reads across full length of long genes
- **Size selection**, suitable for RNA sequencing. 300-500bp - mRNA, 20-150bp - small/miRNA
- **Amplification**, typically by PCR. Up to $0.5 - 10ng$ of RNA
- **Library normalization/Exome capture**
- **Barcoding** and **multiplexing**
- Optionally, add **External RNA Control Consortium (ERCC) spike-in controls**
- **Single** or **paired end** sequencing. The latter is preferrable for the *de novo* transcript discovery or isoform expression analysis

Sample prepatation and library construction strategies:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s005

RNA isolation

- **Ribosomal RNA (rRNA) depletion**
  - $0.1 - 1\mu g$ original total RNA (One cell contains ~10 picogram of total RNA)
  - rRNAs constitute over 90 % of total RNA in the cell, leaving the 1–2 % comprising messenger RNA (mRNA) that we are normally interested in (One cell contains ~0.1 picogram mRNA)
  - Enriches for mRNA + long noncoding RNA.
  - Hybridization to bead-bound rRNA probes

RNA isolation

- **Poly(A) selection (for eukaryotes only)**
  - Enrich for mRNA.
  - Hybridization to oligo-dT beads
- **Small RNA extraction**
  - Specific kits required to retain small RNAs
  - Optionally, size-selection by gel

Description of RNA-seq library enrichment strategies:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s006

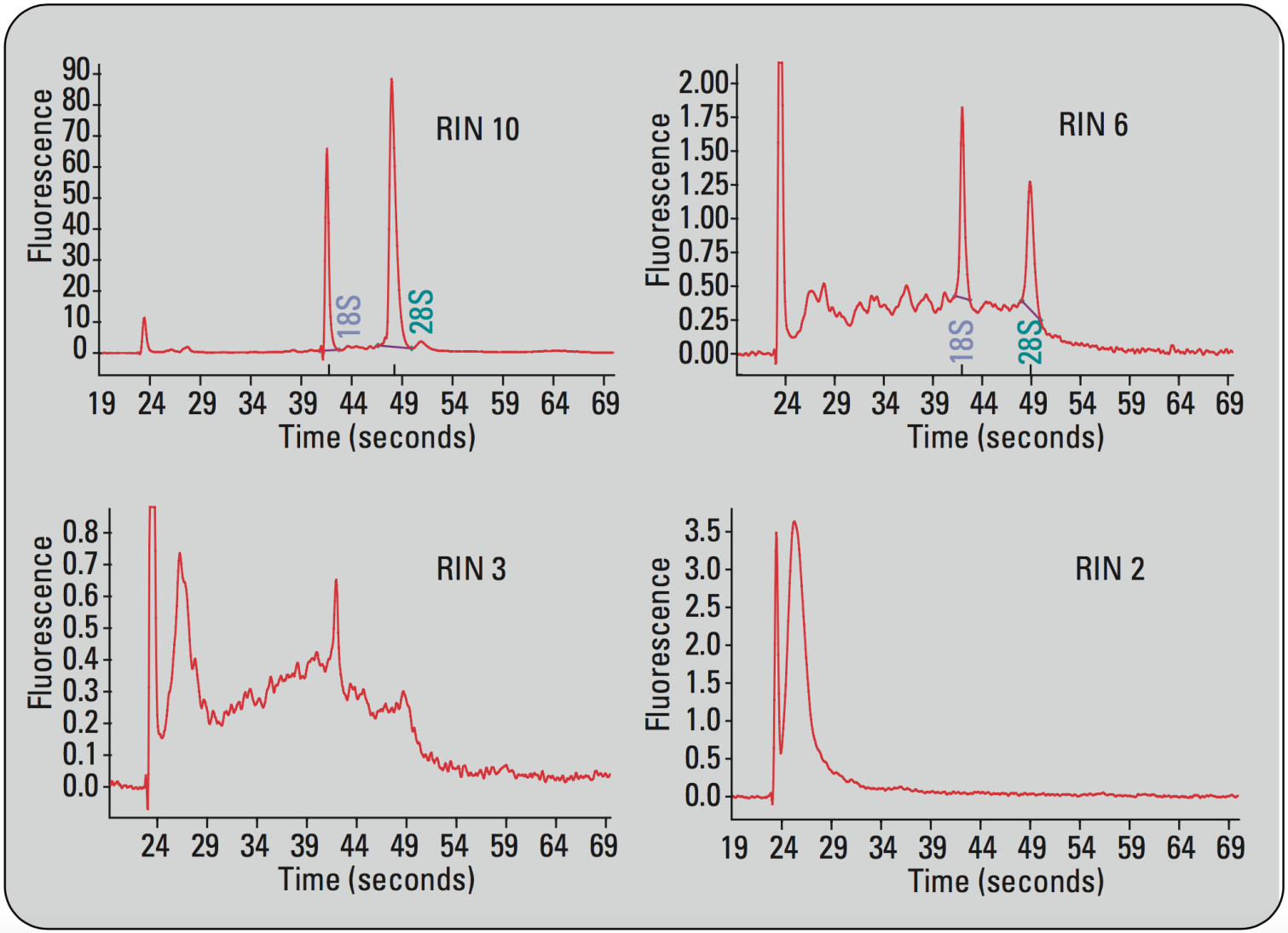
Poly-A selection or ribosome depletion protocol?

- Poly-A excels at gene quantification for classification/prediction purposes, better represents total RNA content
- Ribosomal depletion - more noncoding RNAs, better alignment of reads, more gene fusion events
- Overall, comparable performance

Detailed comparison of RNA-seq library construction protocols:
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4039-1

RNA quality

Agilent 2100 bioanalyzer. RIN - RNA integrity number (should be >7)

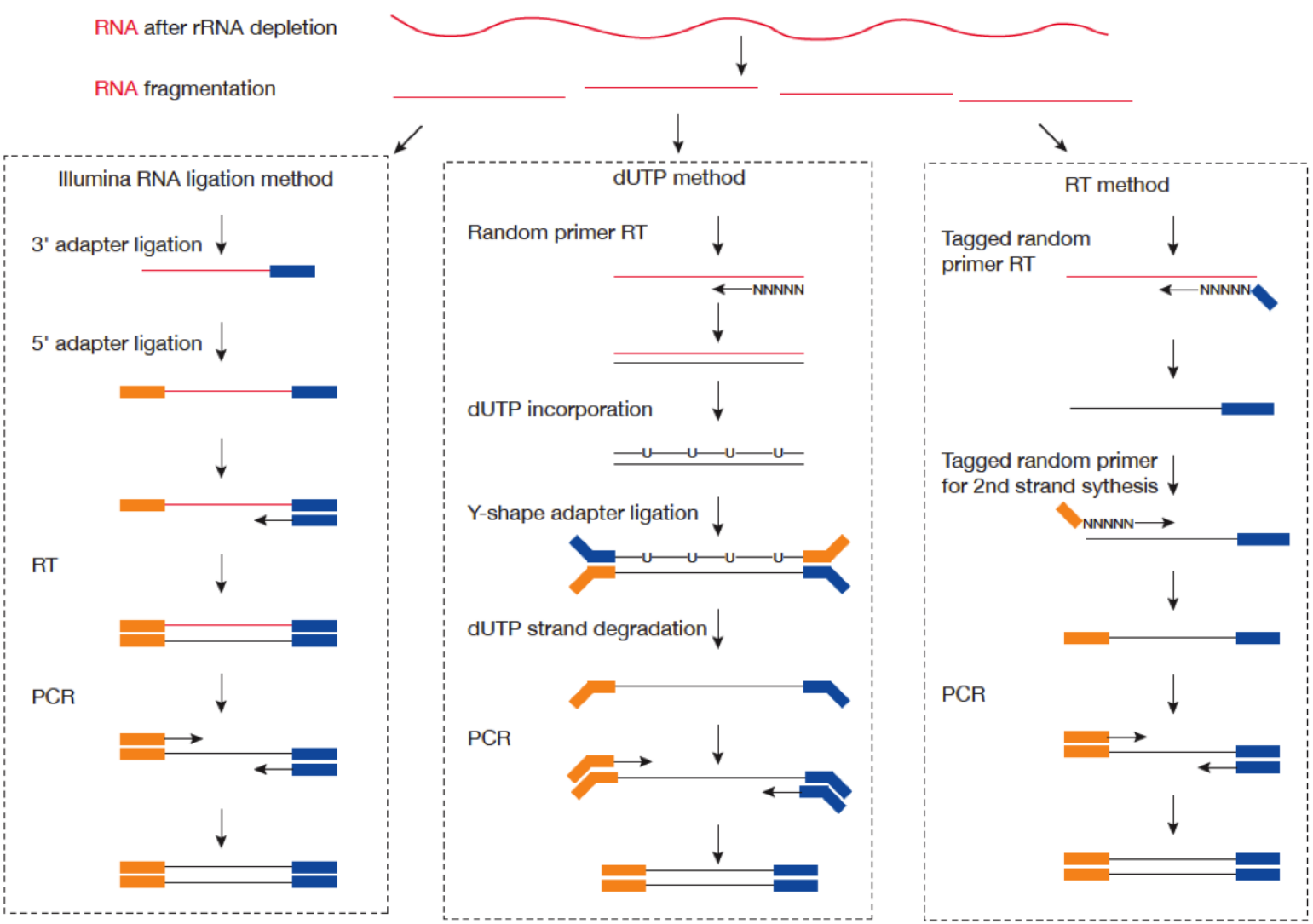


Unstranded vs. Strand-specific library

- **Unstranded**: Random hexamer priming to reverse-transcribe mRNA
- **Stranded**: dUTP method - incorporating UTP nucleotides during the second cDNA synthesis, followed by digestion of the strand containing dUTP

Strand-related settings for RNA-seq tools:
http://journals.plos.org/ploscompbiol/article/file?type=supplementary&id=info:doi/10.1371/journal.pcbi.1004393.s007}

Unstranded vs. Strand-specific library



Experimental design

Sources of variability in RNA-seq measures

In RNA-seq, we have multiple levels of randomness:

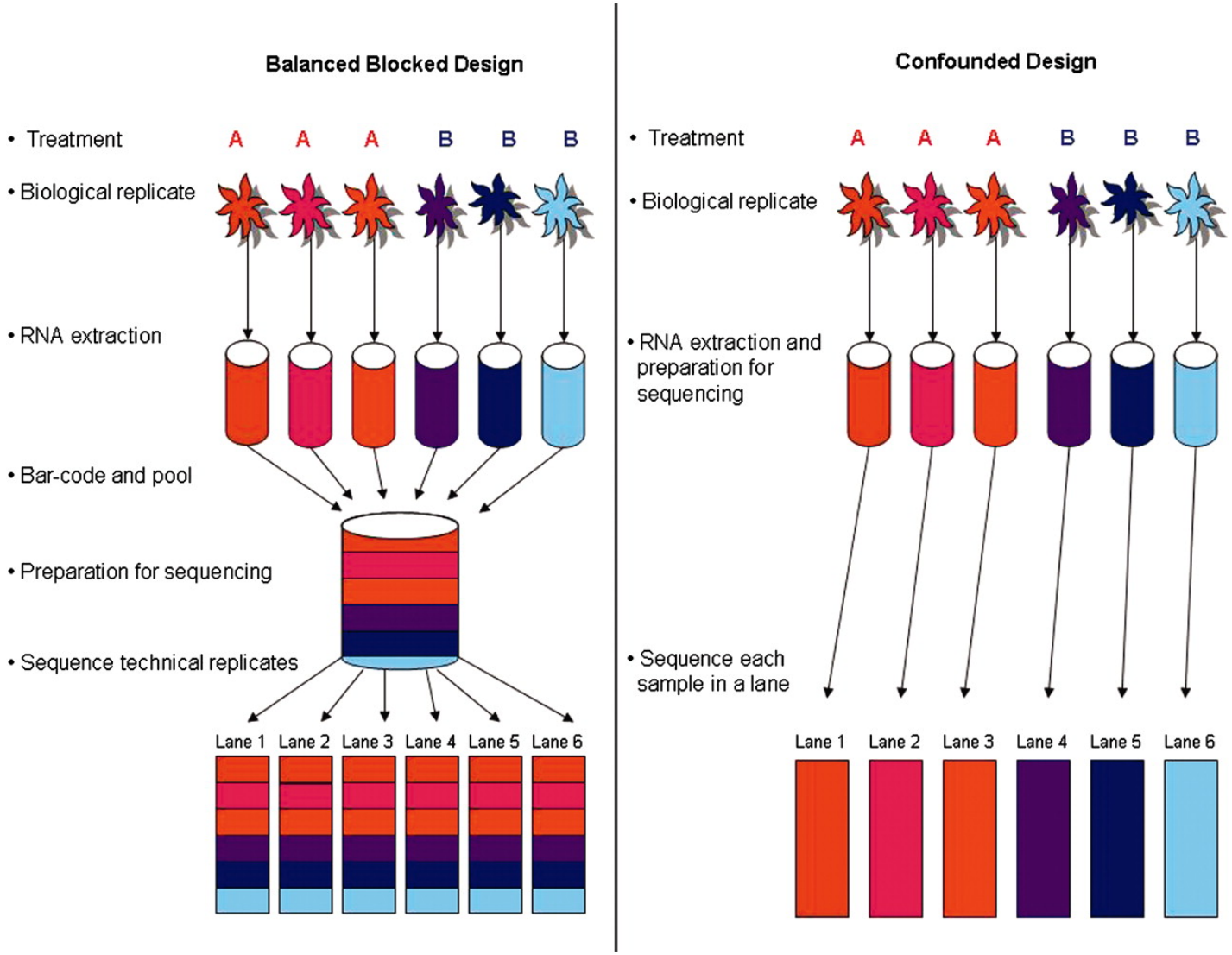
- Biological variability in samples
- Stochasticity of RNA content
- Randomness of fragments being sequenced
- Technical variability

Auer, P.,RW Doerge. "Statistical Design and Analysis of RNA Sequencing Data." Genetics, 2010
http://www.genetics.org/content/185/2/405.long

RNA-seq consideraations

- **Replication**. It allows the experimenter to obtain an estimate of the experimental error
- **Randomization**. It requires the experimenter to use a random choice of every factor that is not of interest but might influence the outcome of the experiment. Such factors are called nuisance factors
- **Blocking**. Creating homogeneous blocks of data in which a nuisance factor is kept constant while the factor of interest is allowed to vary. Used to increase the accuracy with which the influence of the various factors is assessed in a given experiment
- **Block what you can, randomize what you cannot**

Experimental design: Multiplexing balances technical variability



Sequencing length/depth

- Longer reads improve mappability and transcript quantification
- More transcripts will be detected and their quantification will be more precise as the sample is sequenced to a deeper level
- Up to 100 million reads is needed to precisely quantify low expressed transcripts.
- In reality, 20-30 million reads is OK for human genome.

Power calculations

- **Scotty** - Power Analysis for RNA Seq Experiments
- **powerSampleSizeCalculator** - R scripts for power analysis and sample size estimation for RNA-Seq differential expression
- **RnaSeqSampleSize** - R package and a Shiny app for RNA sequencing data sample size estimation
- **RNASeqPower** - R package for RNA-seq sample size analysis

http://scotty.genetics.utah.edu/, Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression". *Bioinformatics* 2013 https://www.ncbi.nlm.nih.gov/pubmed/23314327

http://www2.hawaii.edu/~lgarmire/RNASeqPowerCalculator.htm, Travers C. et.al. "Power analysis and sample size estimation for RNA-Seq differential expression" *RNA* 2014 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4201821/

https://cqs.mc.vanderbilt.edu/shiny/RNAseqPS/, Guo et.al. "RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment" *Cancer Informatics* 2014 http://insights.sagepub.com/rnaseqps-a-web-tool-for-estimating-sample-size-and-power-for-rnaseq-ex-article-a4433

https://bioconductor.org/packages/release/bioc/html/RNASeqPower.html, Svensson, V. et.al. "Power Analysis of Single-Cell RNA-Sequencing Experiments." *Nature Methods* 2017 http://www.nature.com/nmeth/journal/v14/n4/pdf/nmeth.4220.pdf