SRP Bioinformatics and Statistics Workshop

SUSAN TILTON

Oregon State University
Corvallis, OR
September 13-14, 2017

Instructors

Susan Tilton
Assistant Professor
Environmental & Molecular Toxicology Dept, OSU
susan.tilton@oregonstate.edu

Ryan McClure
Scientist
Pacific Northwest National Laboratory
Ryan.Mcclure@pnnl.gov

Lisa Bramer
Scientist
Pacific Northwest National Laboratory
Lisa.Bramer@pnnl.gov

Mark Jankowski
Toxicologist
EPA Region 10
jankowski.mark@epa.gov

Workshop Overview

Day 1

SRP BIOINFORMATICS AND STATISTICS WORKSHOP

JUNE 27-28, 2019

ALS 4000

DAY 1: THURSDAY JUNE 27TH

8:00 – 8:30 AM: ARRIVAL, SETUP

RNASEQ I: OVERVIEW OF CONCEPTS AND TOOLS

8:30 – 9:30 AM: RNASEQ OVERVIEW (LECTURE)
EXPERIMENTAL DESIGN FOR OMICS STUDIES
RNA ISOLATION/LIBRARY PREP- CONSIDERATIONS AND CONCERNS

9:30-10:00 AM: POST-SEQUENCING QC (LECTURE)
FASTQC
TRIMMING

10:00-10:30 AM: PRIMARY PROCESSING AND QC (LECTURE)
RNASEQ PROCESSING STEPS AND TOOLS (OVERVIEW)

10:30-10:45: BREAK

10:45-12:45 PM: INTRO TO R / R STUDIO (HANDS-ON)
REFRESHER AND INTRODUCTION TO R STUDIO
CONSTRUCTING PLOTS IN R, SUBSETTING DATA

12:45-1:45 PM: LUNCH (ON YOUR OWN)

1:45 - 2:45 PM: POWER ANALYSIS IN R & STUDY DESIGN (LECTURE AND HANDS-ON)
COMPONENTS OF POWER ANALYSIS, WHAT TO CONSIDER
EXAMPLE LINES OF CODE

RNASEQ II: RNASEQ DATA ANALYSIS IN R

2:45 - 3:45 PM: DESEQ (HANDS-ON)
NORMALIZATION
STATISTICAL ANALYSIS
Scripts: DESeq_Workshop

Day 2

3:45-4:00 PM: BREAK

4:00 – 5:00 PM: SEMINAR: MARK JANKOWSKI

BEING AN ECOTOXICOLOGIST IN EPA REGION 10: HARNESSING DATA TO SUPPORT DECISION-MAKING NOW (MORE EMPIRICAL DATA) AND INTO THE FUTURE (MORE MECHANISTIC DATA)

DAY 2: FRIDAY JUNE 28TH

RNASEQ II: RNASEQ DATA ANALYSIS IN R (CONTINUED)

8:30-10:00 AM: QC AND DATA VISUALIZATION (HANDS-ON)

QC AND OUTLIERS

DATA ANALYSIS AND VISUALIZATION IN R

Scripts: Transcriptomic_Outlier_Finder_Workshop

Make_Colored_PCA_Plot_Workshop

makeVolcanoPlot_Workshop

Make_Heat_Map_Workshop

makeVenn_Workshop

10:00-10:15 AM: BREAK

RNASEQ III: BIOINFORMATICS TOOLS

10:15-11:30 AM: OMICS DATA INTEGRATION (LECTURE AND HANDS-ON)

OVERVIEW OF CONCEPTS

INTRODUCTION TO BRM AND OTHER WEB TOOLS

CASE STUDIES WITH ZEBRAFISH DATA:

CROSS-SPECIES DATA INTEGRATION

MIRNA TARGET PREDICTION AND DATA INTEGRATION

11:30-12:30 PM: LUNCH (ON YOUR OWN)

12:30 -2:00 PM: NETWORK ANALYSIS AND VISUALIZATION

CYTOSCAPE, R

Scripts: MINET_Workshop

Determine_edge_cutoff_Workshop

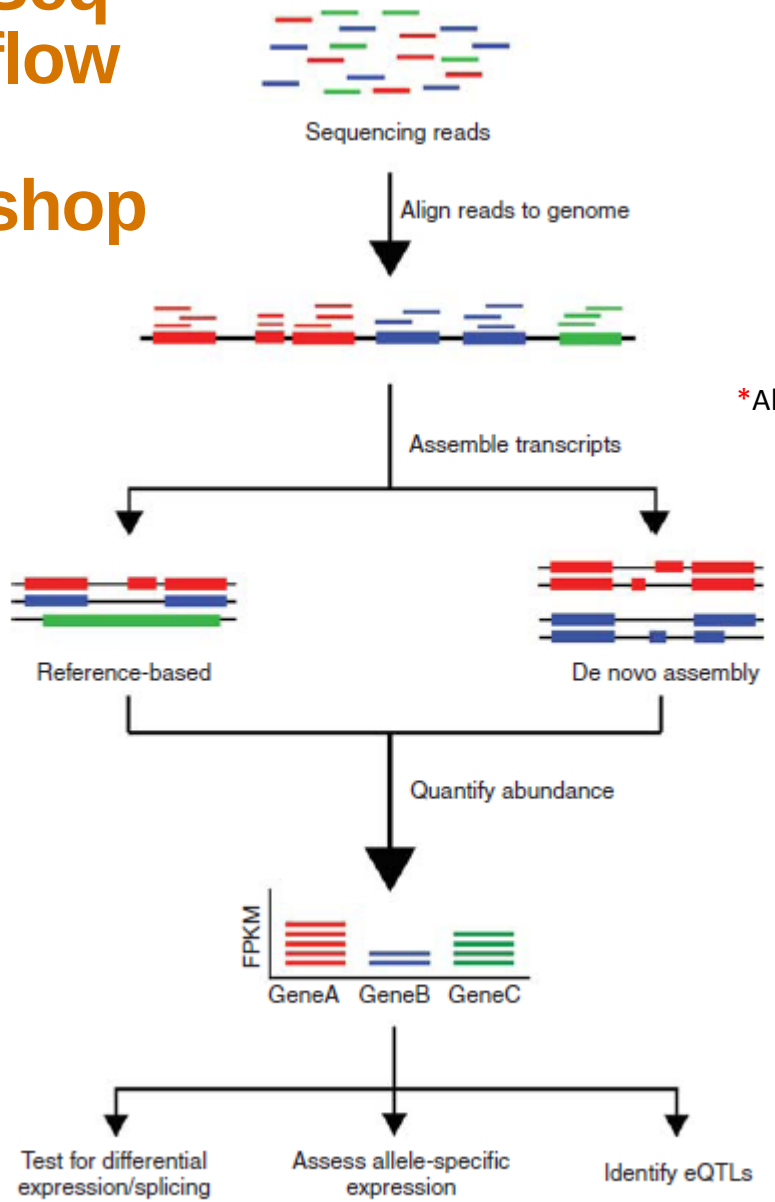
Filter_and_Create_Network_NEW_Workshop

Find_Modules_Workshop

2:00-3:00 PM: DATA VISUALIZATION WITH TRELLISCOPE (HANDS-ON)

RNAseq Overview

RNA-Seq Workflow 2017 Workshop



QC/Trim Adaptors:
FastQC/MultiQC
Trimmomatic/Cutadapt

Aligning reads to genome*:
Hisat2/Tophat2/Bowtie (miRNA)

*Alignment: Allows for detection of novel transcripts, variants, SNPs

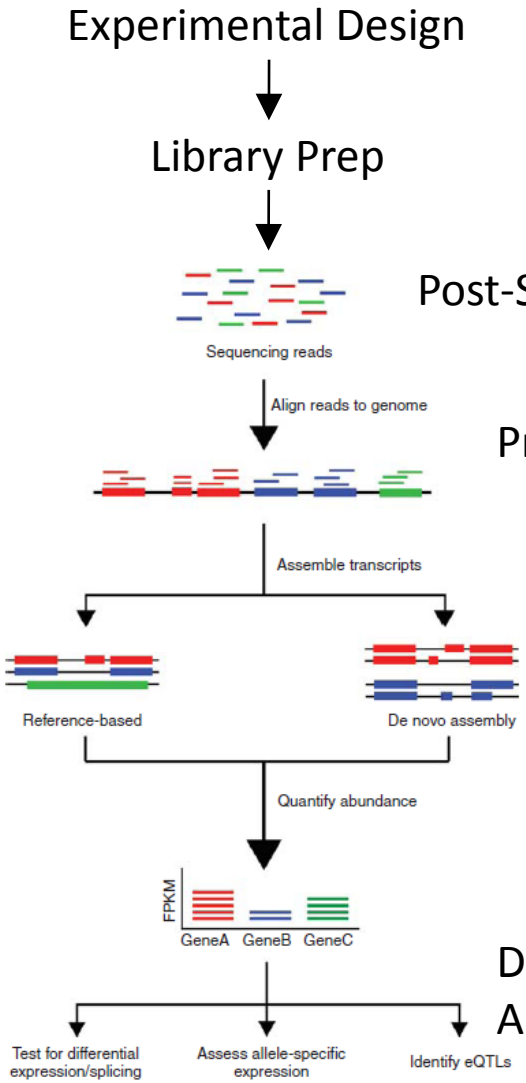
Transcript assembly/ quantification:
StringTie/Cufflinks (FPKM)

Transcript quantification from quasi-mapped reads (no alignment):
Salmon (counts)

Statistical/DE Analysis:
DESeq2
EdgeR
Ballgown/Cuffdiff

Analysis of DE genes:
biomaRt (ID conversion from Ensembl BioMart)
TopGO (functional enrichment with Gene Ontology)

RNA-Seq Workflow 2019 Workshop



Post-Sequencing QC

Primary Processing

Day 1 AM

Intro to R/R Studio
Power Analysis

DESeq, Normalization, Statistical Analysis
Applications of RNAseq Data

Day 1 PM

QC
Data Visualization/Analysis

Day 2

Experimental Design for Omics Studies

Formulate Questions/Hypothesis for Study

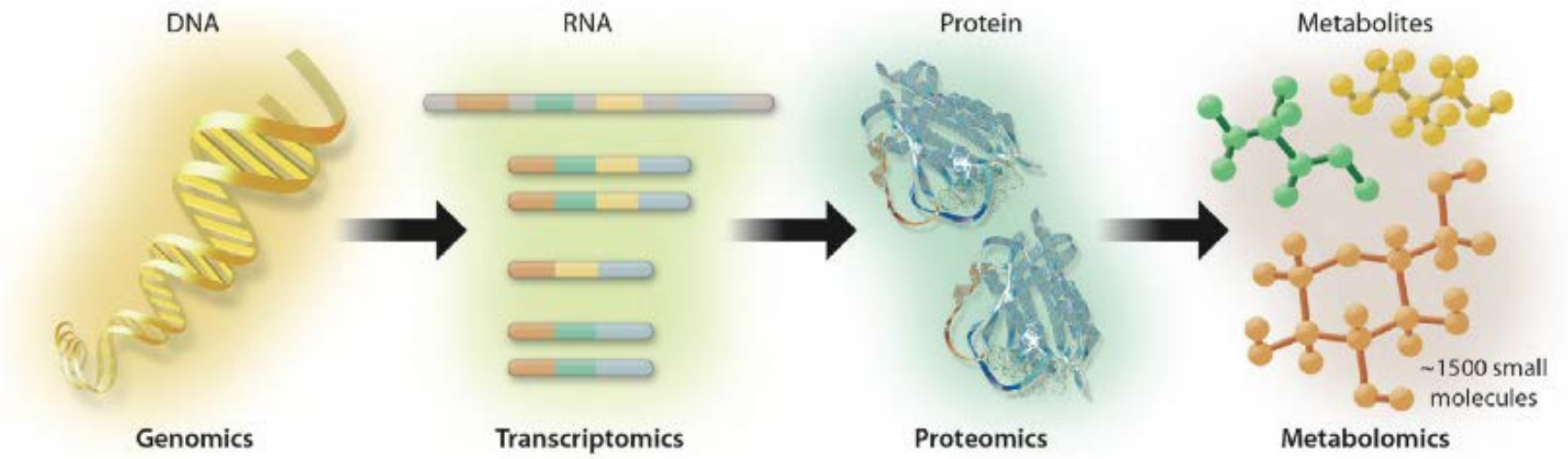
Myth: Omics experiments are entirely ‘hypothesis generating’; therefore, we don’t need a biological question in the experiment

- Always need to have a biological question
- Could be nebulous (e.g. What happens to the transcriptome in response to chemical A) or more focused (e.g. What is the mechanism for chemical A in causing phenotype B)
- The purpose of the question drives the experimental design
- Make sure the samples and experimental design answer the question

Omics Experimental Design Considerations

- Choose which Omics technology to measure based on the question to be addressed
- Choice of appropriate controls
  - Time-matched, negative control, positive control
- If multiple OMICS analyses will be compared, collecting samples in the same experiment is the preferred approach
  - This will allow for better data integration downstream
- Collect as much phenotypic data as you can to allow correlation of system response to specific phenotypic outputs
- An added advantage is that phenotypic data can be used to help eliminate outlier samples

Choose which Omics technology to measure based on the question to be addressed



DNA sequencing
Chip-Seq
NGS/SNP
Methylation

Microarrays
RNA sequencing
Q-PCR arrays

SDS-PAGE/MS
LC-MS/MS
ELISA Protein microarrays
Post translational modifications

LC-MS
GC-MS
CE-MS
NMR

What question are you asking?

Highlighted are the technologies included in these workflows

Omics Experimental Design Considerations

- Choose which Omics technology to measure based on the question to be addressed
- Choice of appropriate controls
  - Time-matched, negative control, positive control
- If multiple OMICS analyses will be compared, collecting samples in the same experiment is the preferred approach
  - This will allow for better data integration downstream
- Collect as much phenotypic data as you can to allow correlation of system response to specific phenotypic outputs
- An added advantage is that phenotypic data can be used to help eliminate outlier samples

Choice of appropriate controls

Time Points:

Multiple sample collection time points enable study of a system over time

Dose-Response:

Sample collection over a range of doses/concentrations allow evaluation of phenotype across dose

Preliminary experiments should be performed to establish attributes of response, so that appropriate experimental conditions (dose/time) can be selected.

Appropriate controls for each experimental factor should be included based on study questions. (e.g. treatment, time, genotype)

Omics Experimental Design Considerations

- Choose which Omics technology to measure based on the question to be addressed
- Choice of appropriate controls
  - Time-matched, negative control, positive control
- If multiple OMICS analyses will be compared, collecting samples in the same experiment is the preferred approach
  - This will allow for better data integration downstream
- Collect as much phenotypic data as you can to allow correlation of system response to specific phenotypic outputs
- An added advantage is that phenotypic data can be used to help eliminate outlier samples

Other Experimental Design Considerations

- Biological Systems
- Replication
- Randomization/Blocking

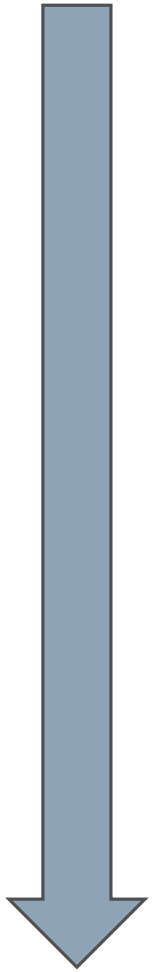
Experimental Design Considerations: Biological Systems

Immortalized Cell Lines

Primary Cells

Organotypic cultures (ex vivo)

Tissues from whole animals
(embryos)



Decreased biological relevance and cost
Decreased variability
Easy to perturb

Increased biological relevance
More difficult to perturb

Reflects complex biological signaling
Increased variability
Perturbation methods may be available (KO/KD)

Experimental Design Considerations: Replication

- Replication is repeating a phenomenon so that variability associated with the phenomenon can be estimated.

- What to replicate?

  - Biological replicates
    - Replicates at the experimental level, provides sense of population response
    - Typically more important (depends on purpose of experiment)
    - Often more effective in increasing power for detecting differential genes/metabolites

  - Technical replicates
    - Replication below experimental level - provides estimate of variance for a single biological sample
    - Tends to account for variance in technology
    - Useful when technical variability is large and technical replicates are inexpensive

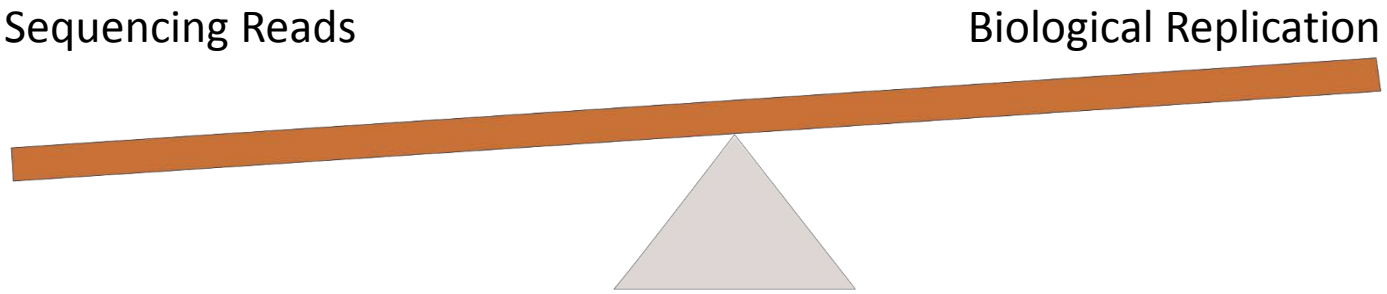
Experimental Design Considerations: Replication

- For RNAseq:
  - Technical reproducibility is very good
  - Biological variation is much greater!

- How many replicates?
  - Biological variance
  - Read count
    - Compare to mRNA samples for differential expression (5-30M reads per sample)
    - Discover novel elements, perform more precise quantification, especially of low expressed transcripts (50-200M reads per sample)
  - What resources do you have?
    - Well assembled / annotated genomes – single ends, shorter reads
      - SE is more cost effective, sufficient for studies of gene expression levels in well annotated genomes
    - De novo – longer reads, paired ends
      - PE will have more accurate transcript assembly and provide more accurate isoform expression levels
  - Resource Allocation: Balance sequencing reads and replication

Experimental Design Considerations: Replication

- Resource Allocation: Balance sequencing reads and replication



- Greater sequencing depth correlates with better genomic coverage and more robust differential gene expression analysis
- Greater biological replication improves power to detect differences among treatments/samples

Study type	Reads Needed
Expression profiling	5-30 M
Alternative splicing, SNPs	50-100 M
De novo transcriptome assembly	100+ M
HiSEQ 3000	300-400 reads/lane

Wang, Y, N Ghaffari, CD Johnson, UM Braga-Neto, H Wang, R Chen, H Zhou. (2011). Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics* 12 Suppl 10:S5

BIOINFORMATICS **DISCOVERY NOTE** *Vol. 30 no. 3 2014, pages 301–304*
doi:10.1093/bioinformatics/btt688

Gene expression Advance Access publication December 6, 2013

RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

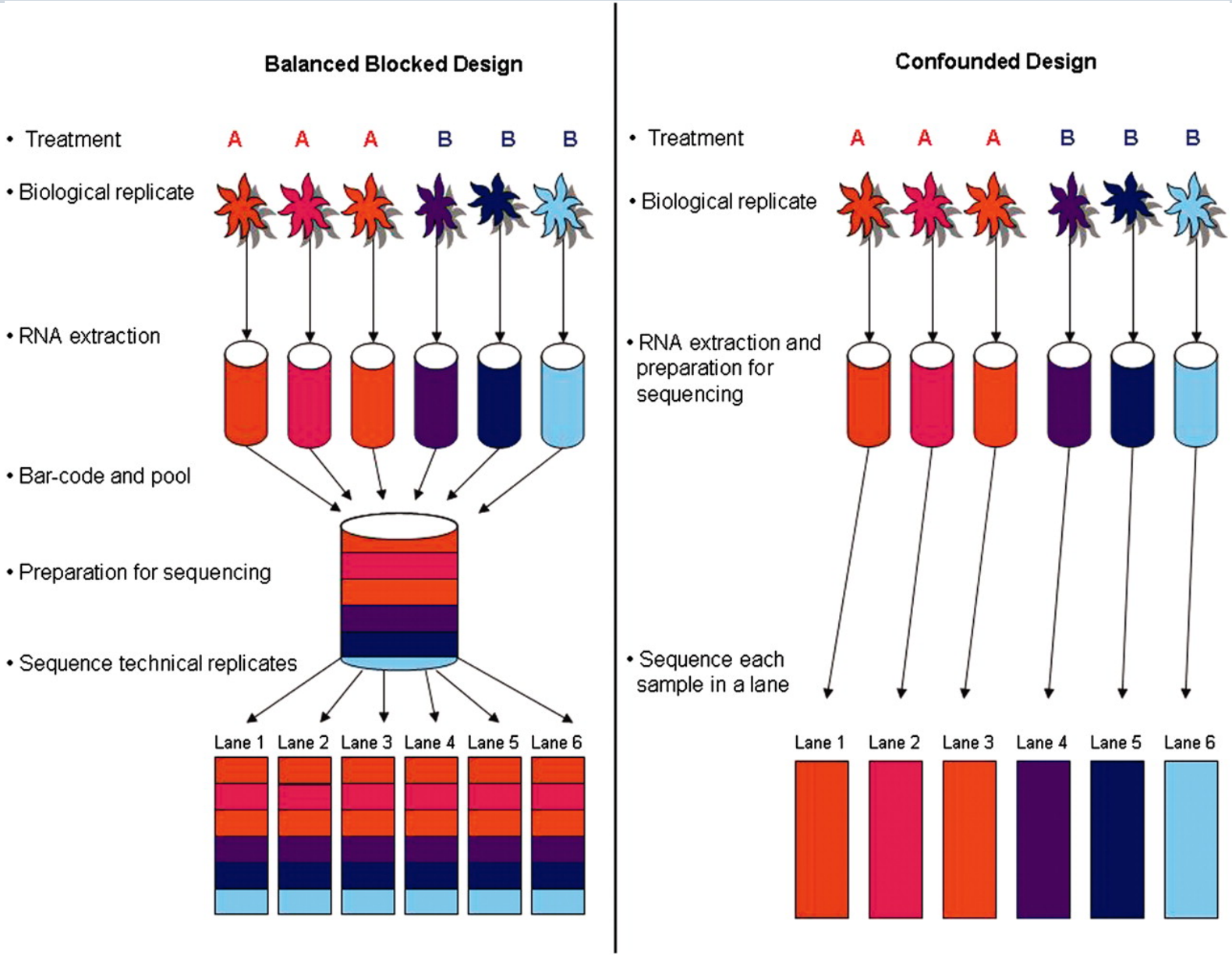
[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Experimental Design Considerations: Randomization

- Experimental treatments/conditions are assigned in a random fashion.

- Why randomize samples?
  - Helps to eliminate effect of uncontrolled factors unrelated to biological conditions/questions under study (batch effect)
- When to randomize samples?
  - Consider randomization at all stages of experimentation (treatments, order of sample handling, run order, sample process order)
  - Use appropriate randomization method depending on size/type of experiment (random generator/blocking)

Experimental Design Considerations: Randomization



Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge

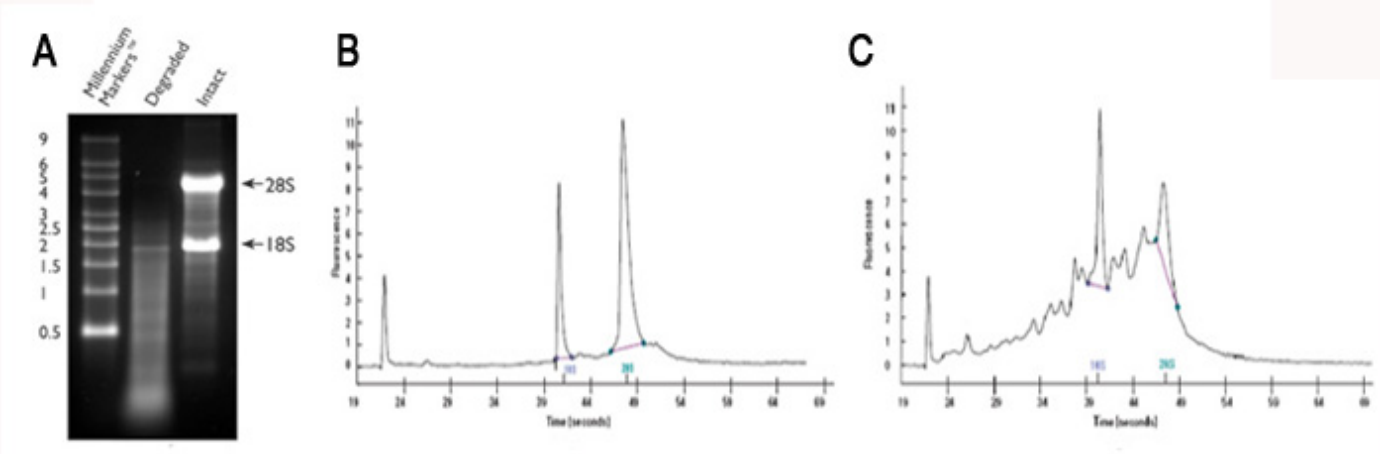
GENETICS *June 1, 2010 vol. 185 no. 2 405-416; https://doi.org/10.1534/genetics.110.114983*

RNA Isolation and Library Prep

RNA Isolation and Library Preparation

Success of RNAseq experiments is highly dependent on recovering pure and intact RNA.

- Multiple type of RNA isolation procedures (organic extraction vs. solid-phase extraction)
- Some protocols for isolating mRNAs are not optimized for isolating small RNA, so check first
- Quality assessment (Bioanalyzer trace)

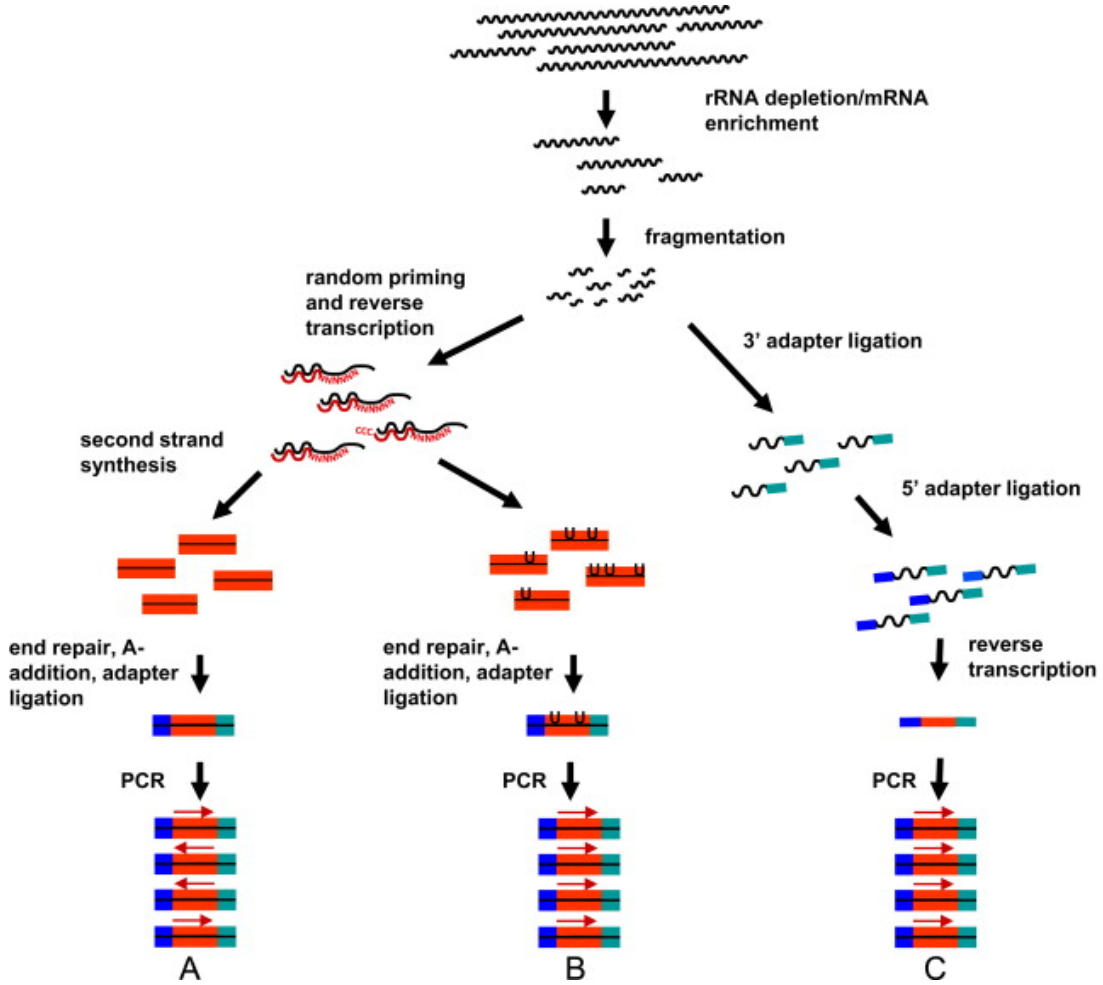


RNA Integrity Number (RIN)
0-10

RNA Isolation and Library Preparation

Library Preparation

- DNAse treatment
- RNA fragmentation / size selection
- cDNA synthesis
- Add sequencing adaptors



EXPERIMENTAL CELL RESEARCH 322 (2014) 12–20

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/yexcr

ELSEVIER

Review Article

Library preparation methods for next-generation sequencing: Tone down the bias

Erwin L. van Dijk[a,*], *Yan Jaszczyszyn*[b], *Claude Thermes*[a]

[a]Centre de Génétique Moléculaire – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France
[b]Plateforme Intégrée IMAGIF – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France