INF-BIOx121 2017

RNA-seq

differential expression analysis

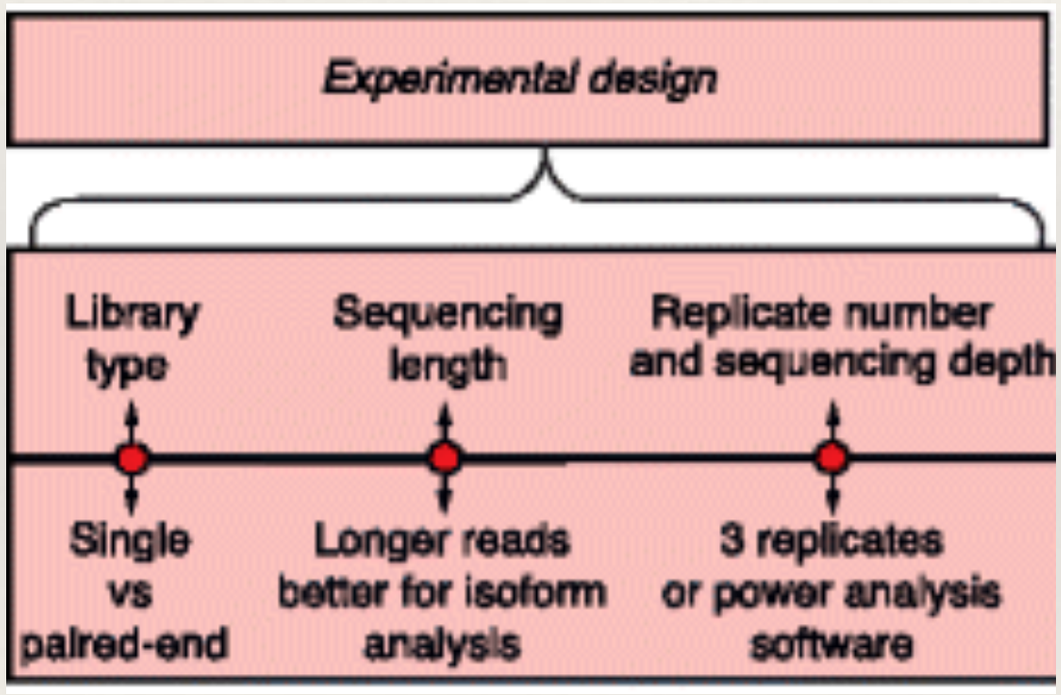
Arvind Sundaram

Sep 18-20, 2017

Design, library prep, sequencing and analysis

Arvind Sundaram
Sep 19, 2017

Pipeline(s) - too many



Design of the experiment and sequencing plan are very important!!

A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

Experimental design

- Biological question
- Species-specific information
  - Is there a genome sequence available??
  - Is it well annotated??
- Sample variation
- Replicates
- Platform choice
  - Technology-specific variation
  - Technical bias
- Library prep
- Sequencing depth
- Data analysis

Replicates and Depth

- Sound experimental design
  - Number of replicates
    - Biological variation
    - Technical replicates - not so important
  - Sequencing depth

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

Replicates vs Depth

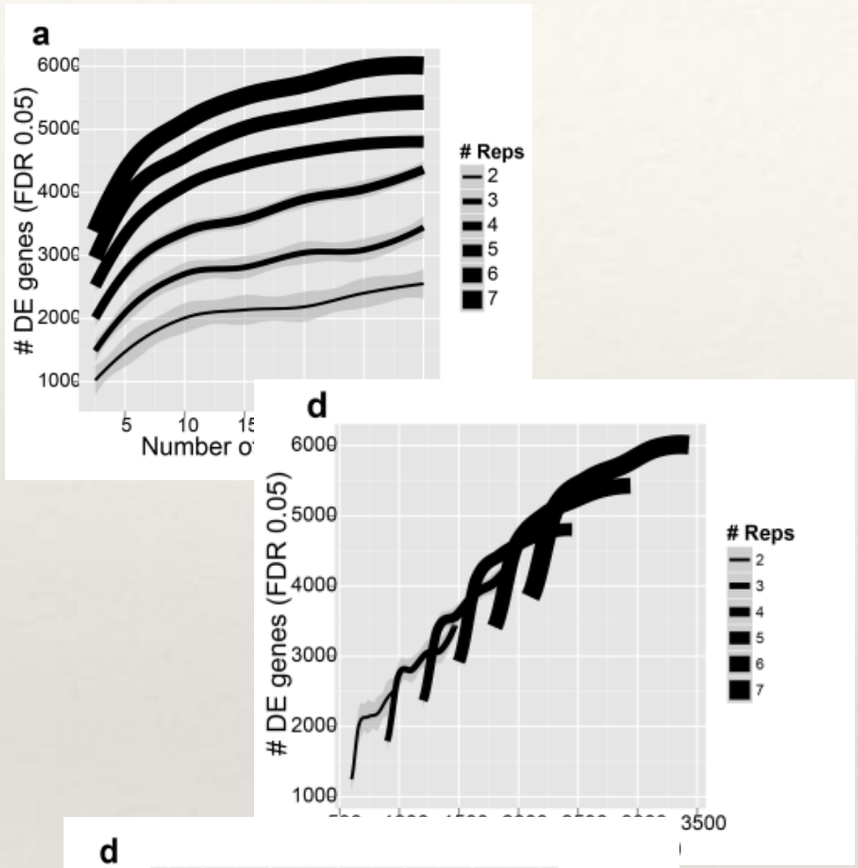


Table 1. Cost efficiency for power to detect DE genes (cost per 1% power given each experimental design where the variables are). Assumptions made during calculations are described in Methods. * indicates lowest cost per 1% power in each replication level. Units are in dollars.

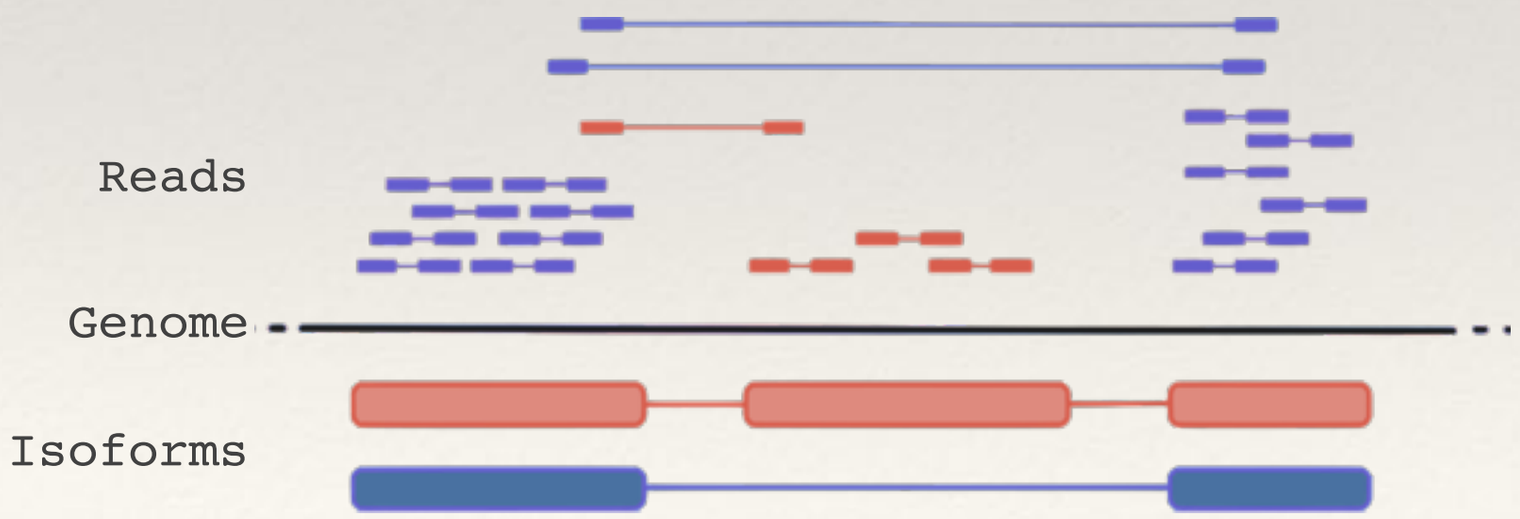
Relative Cost	2.5M	5M	10M	15M	20M	25M	30M
2 replicates	24.2	17.2	14.4*	15.8	16.7	17.0	17.8
3 replicates	23.4	17.2	15.3*	16.3	17.1	18.5	19.4
4 replicates	23.1	17.7	16.5*	17.5	18.6	19.8	21.2
5 replicates	23.8	19.0	18.1*	19.4	21.0	22.8	24.9
6 replicates	25.0	20.7	20.6*	22.4	24.6	27.0	29.4
7 replicates	26.8	23.0*	23.5	26.0	28.7	31.5	34.3

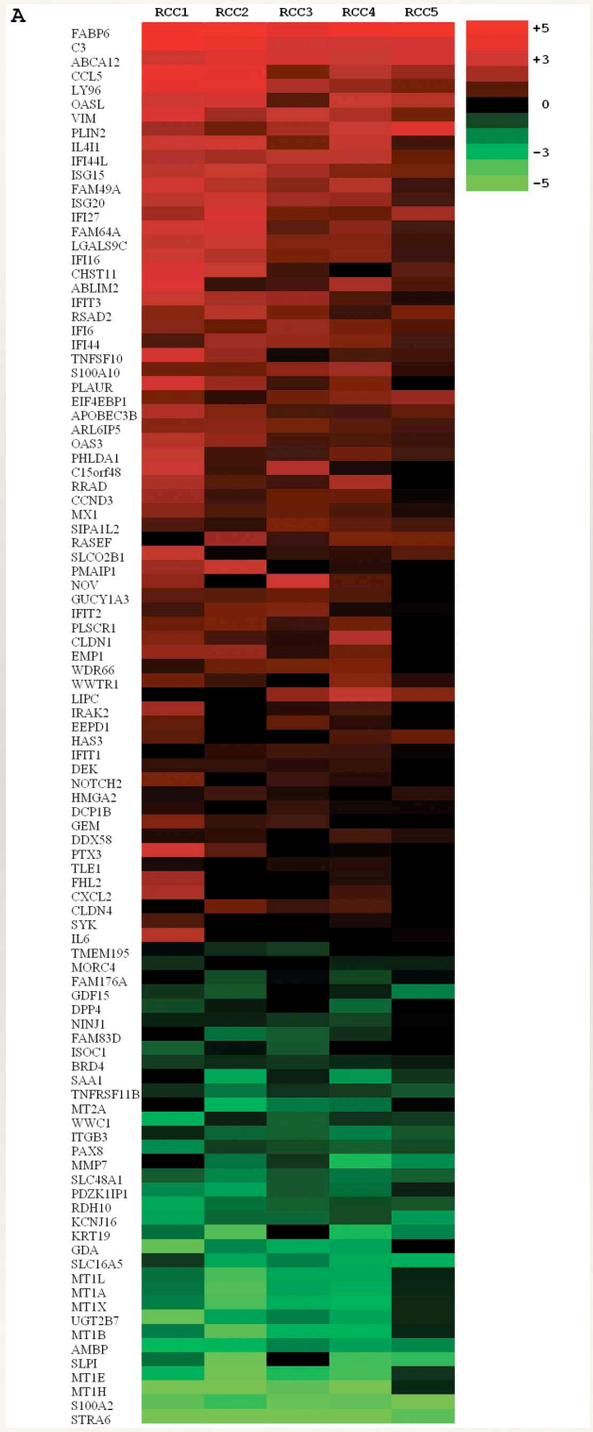
http://www.doi.org/10.1093/bioinformatics/btt688

Depth

- RNA sequencing
  - Highly expressed known transcripts
  - Novel isoforms
  - Low expressed / rare transcripts

More depth





Sequencing technology

Short-read (Illumina) or Long-read (PacBio)??

Deep sequencing?

Model or non-model species

Are you interested in gene-level, transcript-level expression?

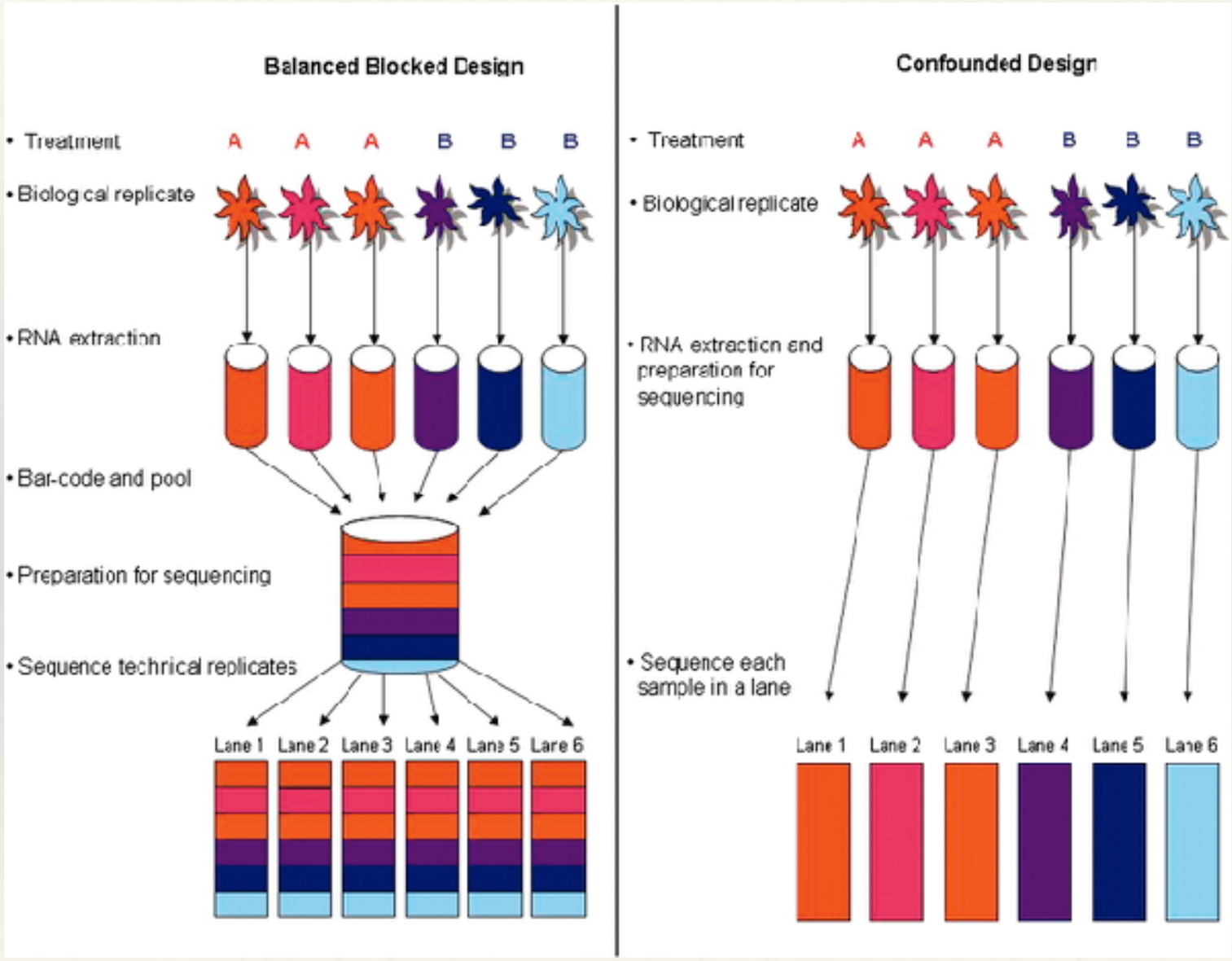
What type and amount of genome resource is available?

Interested in finding new genes, novel transcripts?

Read length, paired end?

Technical bias

- Lane / flowcell bias
- Index / barcode bias
- Batch effect
- Randomisation is key

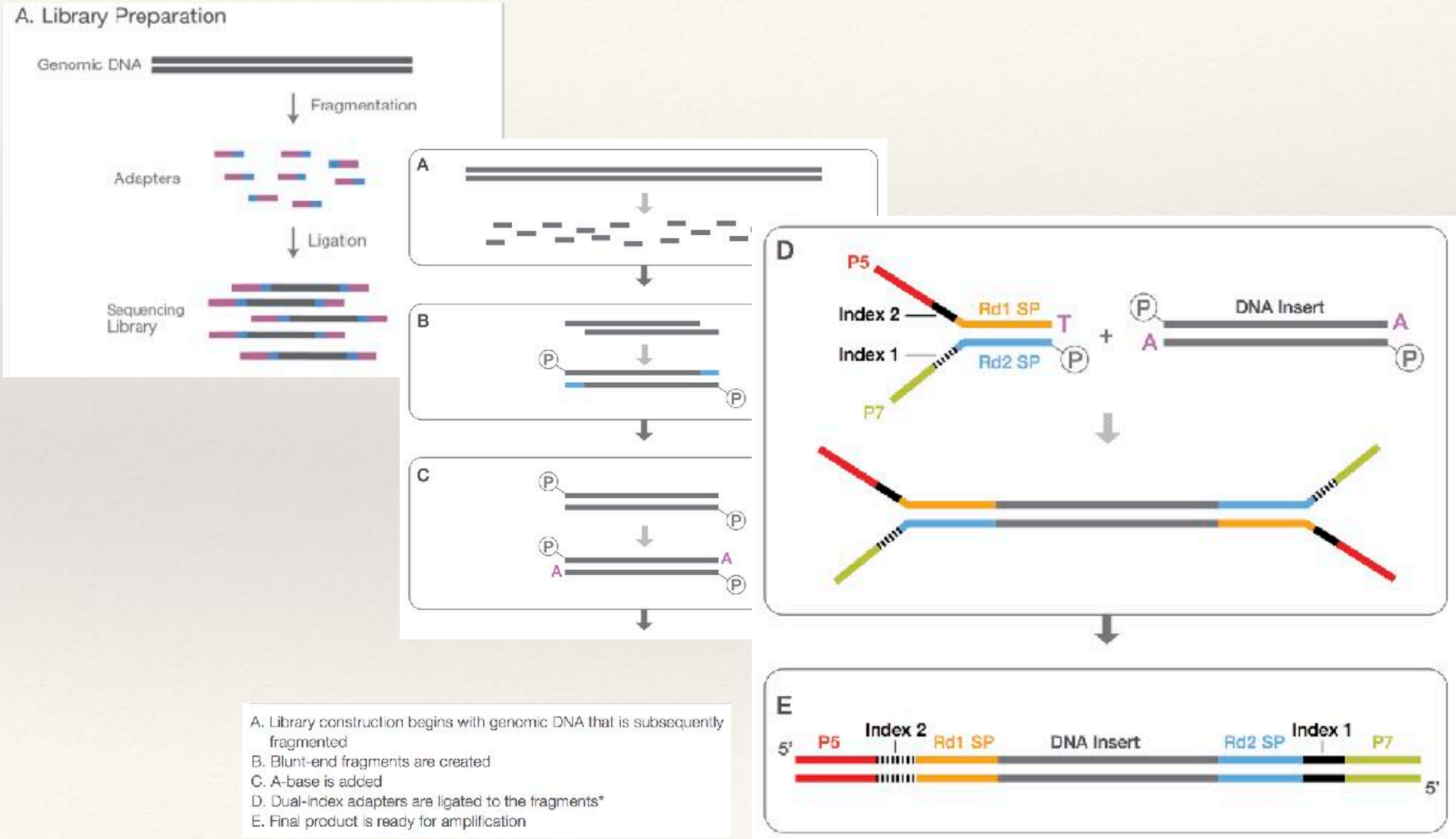


http://www.genetics.org/content/185/2/405.long

Design prior to sequencing

- Sources of variation
  - Dynamic range - Not all samples get sequenced the same way (Normalisation)
  - Technical variation - Bias inherent to the technology
  - Biological variation
- Controlling for variation
  - Randomisation
  - Blocking
    - Pool and sequence across several lanes
  - Replication

Library prep (Illumina)



A. Library construction begins with genomic DNA that is subsequently fragmented
B. Blunt-end fragments are created
C. A-base is added
D. Dual-index adapters are ligated to the fragments*
E. Final product is ready for amplification

Library prep (Illumina)

- RNA sequencing
  - Total RNA
    - TruSeq **Stranded** Total RNA kit
  - mRNA
    - TruSeq **Stranded** mRNA kit
  - small RNA
    - TruSeq small RNA kit
  - Ribosome profiling

    - High quality and quantity of RNA
    - Do you want to sequence rRNA??

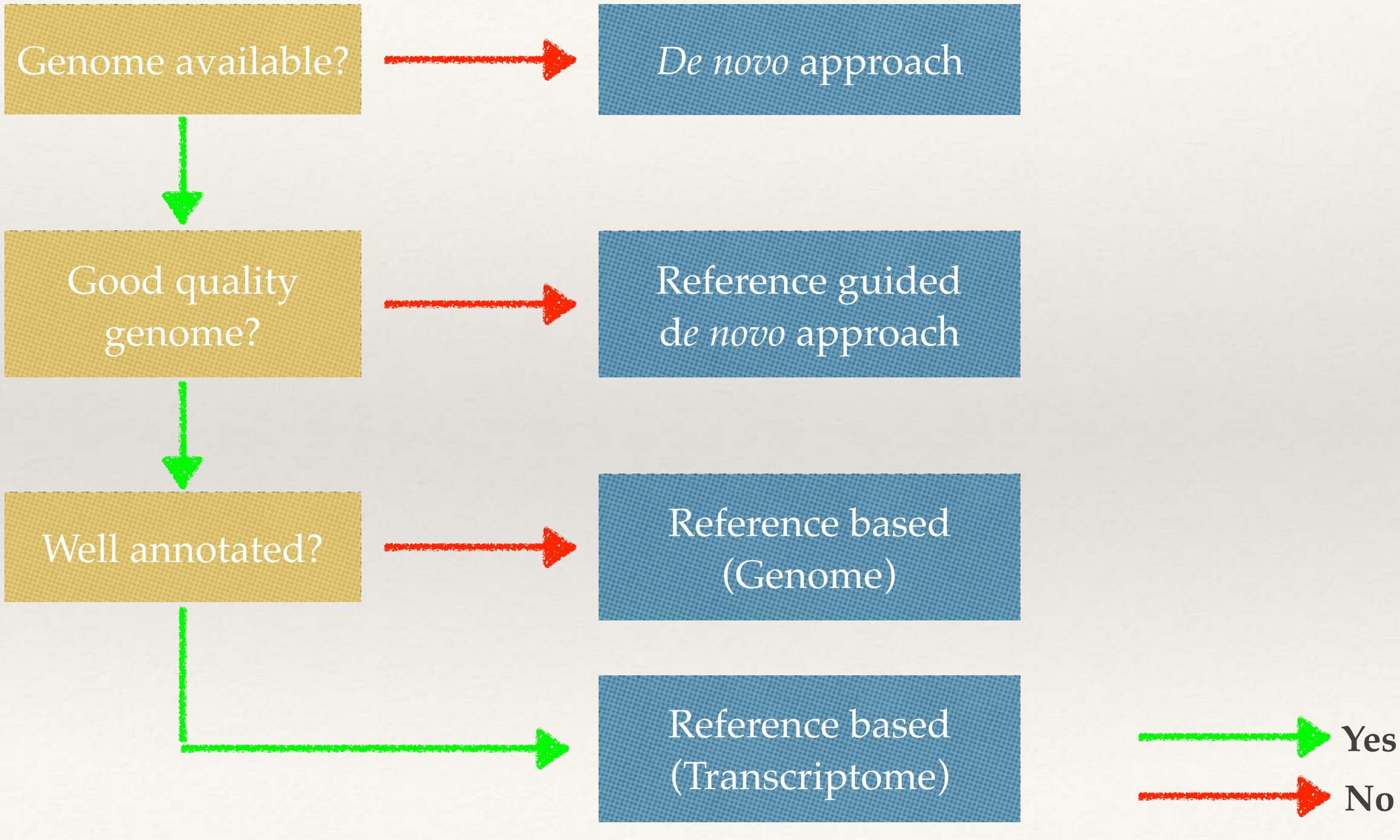
Sequence data analysis

- Is genome available?
- Well annotated?

- *De novo* approach
- Reference based approach
- Transcriptome
- Genome+Transcriptome
- Mixed approach??

- Short reads (Illumina) + Long reads (PacBio)

Pipeline choice



Reference choice

Well annotated?

Known transcripts? → Reference based (Transcriptome)

Novel isoforms? → Reference based (Genome + Transcriptome)

Novel genes? → Mixed approach (Reference based + *de novo* approach)

→ **Yes**

→ **No**

Pre-processing

- Remove sequencing adapters
- Trim/remove low quality reads
- Remove sequencing spike-ins (PhiX for Illumina), if any

➡ Make sure paired end data is always paired and in correct order!

Simple truth

To consult the statistician after an experiment is finished is often merely to ask him(her) to conduct a post mortem examination. He(she) can perhaps say what the experiment died of.

- Ronald Fischer