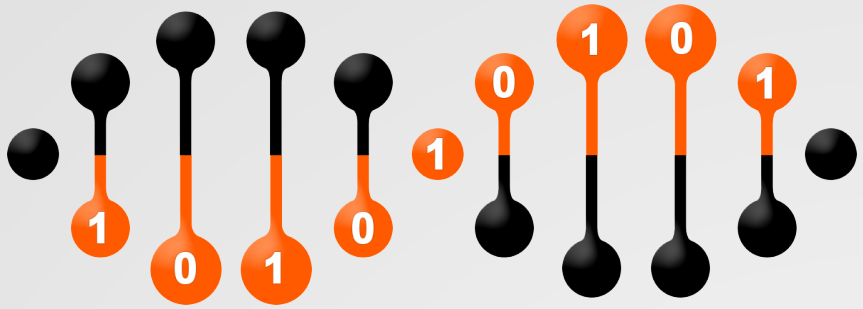
CBC Data Therapy

RNA-Seq Discussion



Computational Biology Core

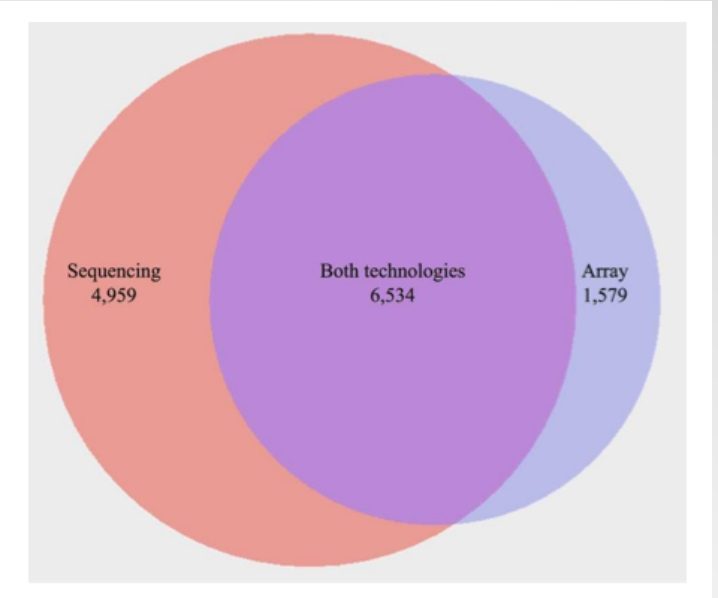
UCONN
UNIVERSITY OF CONNECTICUT

RNA-Seq versus Microarrays

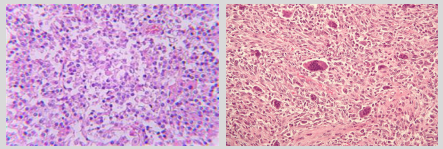
- Correlation of fold change between arrays and RNAseq is similar to correlation between array platforms
- Technical replicates have less variation
- Extra analysis: prediction of alternative splicing, SNPs
- Low- and high-expressed genes do not match

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

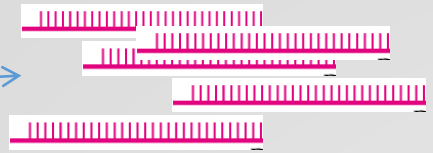


RNA-Seq workflow

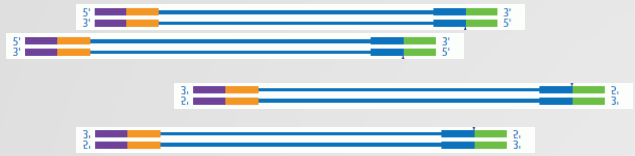


Samples from two conditions

Isolate RNA

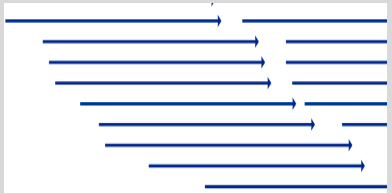


Generate cDNA



Create sequencing library by fragmenting, size selection and adding adaptors

Run sequencer



Generate short reads

Identify differentially expressed genes

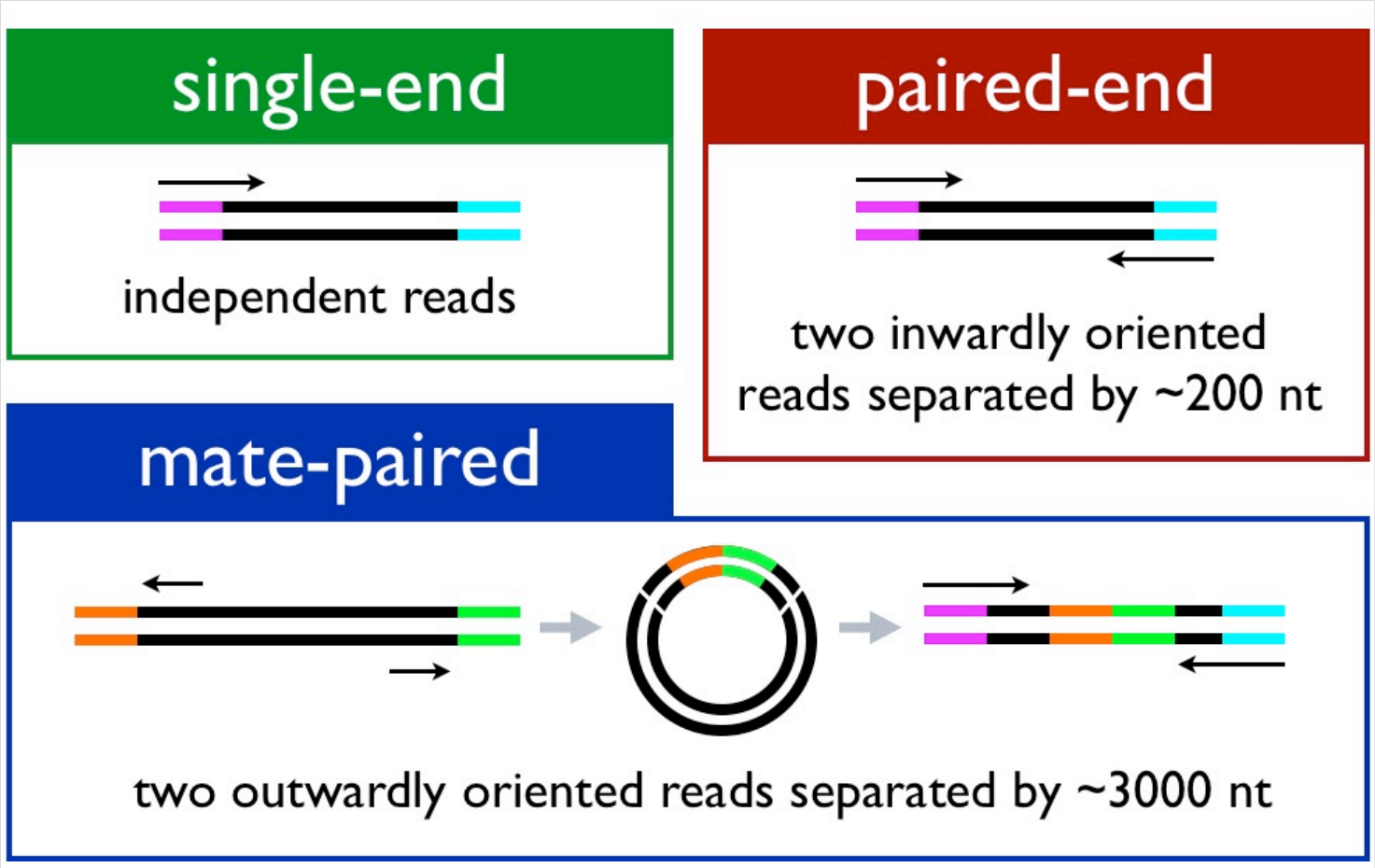
Designing the Study

- Final Goals?
  - Transcriptome assembly (characterization of the gene space)
  - Differential gene expression?
  - Identify rare isoforms?
  - Identify variants?
- Characteristics of the System?
  - Genome?
    - Quality of the reference
    - Availability of an annotation?
    - Introns?
  - Close relative? How close?
  - Other transcriptomic resources?

Designing the Study

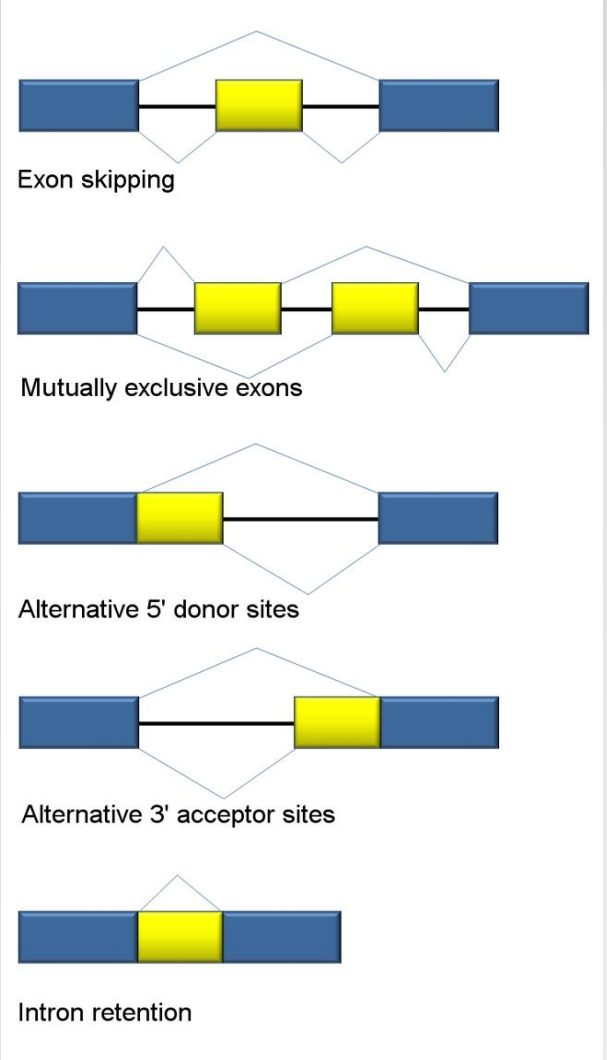
- Experimental design
  - Biological replicates
  - Technical replicates
  - Minimize lane effects
- Appropriate Sequencing Technologies
  - HiSeq 3000/4000/TenX
- Read depth
- Barcoding (multiplex -> how much?)
- Read length
- Paired vs. single-end

Paired versus Single-End

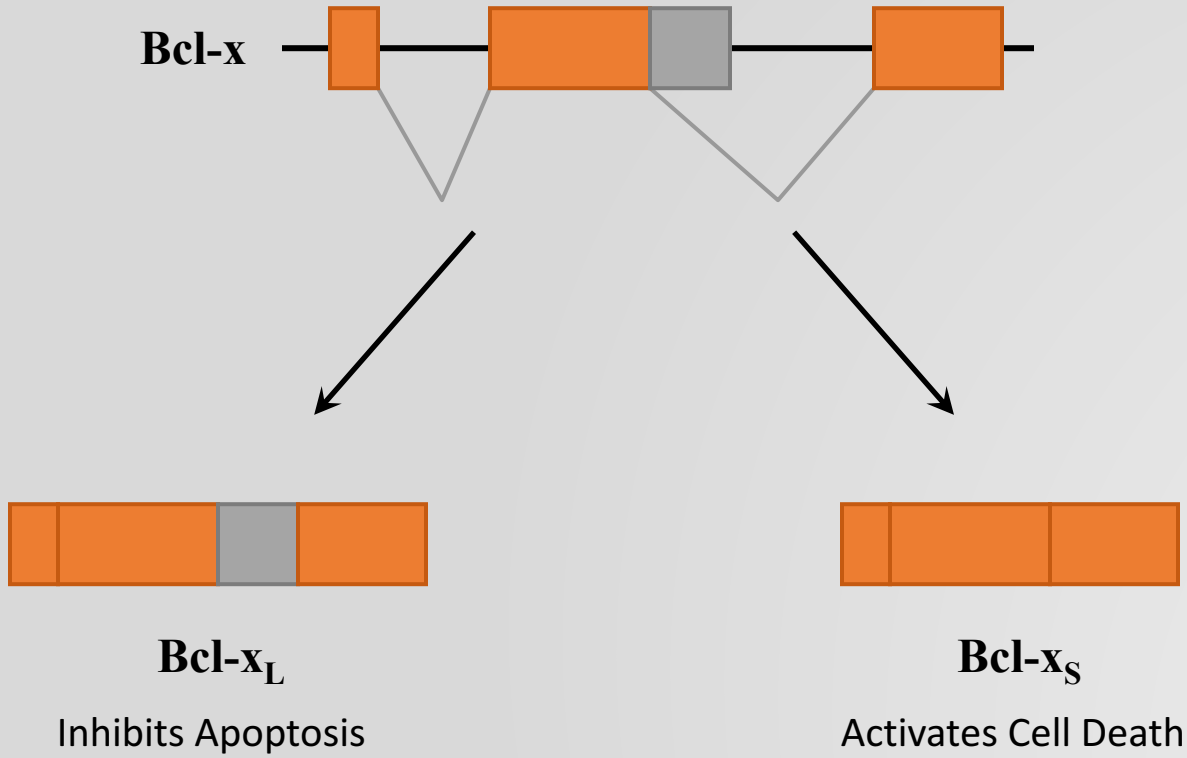


Analysis Challenges

- Biases may mean what we are seeing is not reflective of true state of the transcriptome.
- Alternative splicing!
- Gene level, exon level?
- Multimapping, partial mapping, not mapping
- Normalization issues
  - Size (depth) of datasets
  - Gene length differences



One Gene: Two Isoforms: Opposite Functions

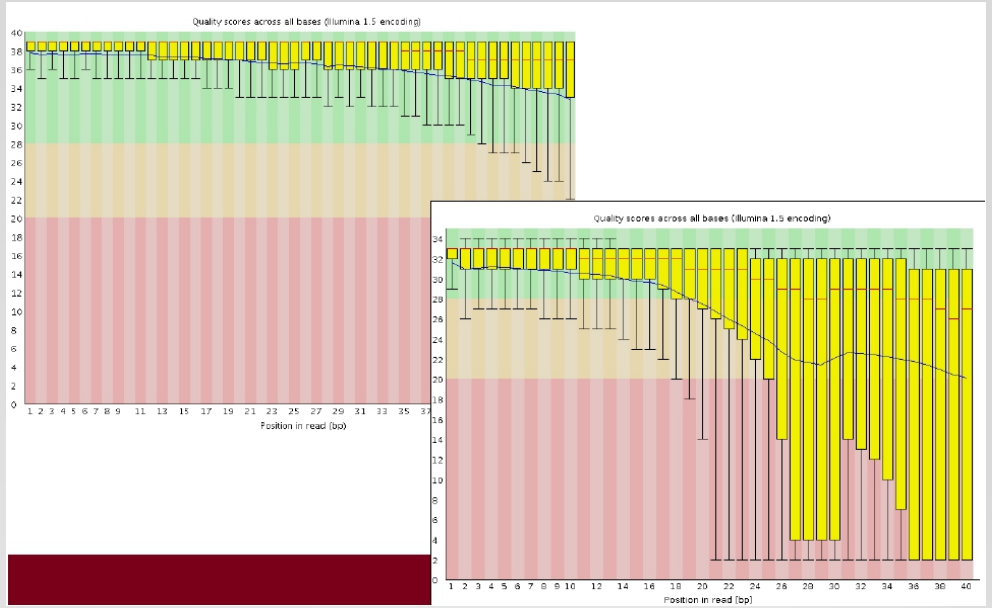


Example in Apoptosis:

For complete biological understanding, you need to know which isoforms are expressed

Quality Control

- Quality Control:
  - FASTQ Files
    - FASTQC (Preview)
      - Barcodes and adaptors should be removed by Illumina Casava/BaseSpace
      - Examine lane effects, quality issues, library depth considerations
    - Trimmomatic or Sickle
      - Trim poor quaity bases
      - Remove short reads
      - Identify proper pairs



Read Mapping or Assembly

- Read Mapping
  - Genome reference available?
    - Annotation available?
      - Bowtie2/TopHAT
      - HiSAT2
      - STAR
      - StringTie
    - No Genome Available
      - De novo transcriptome assembly
        - Reference Guided (StringTie)
        - No Reference or very fragmented (Trinity)

Read Mapping or Assembly

- Mapping to genome vs transcriptome?

- Is your reference the right version?
- Does your annotation match your reference?

Generating Raw Reads

- Generate an alignment file
  - SAM or BAM file format
    - Detailed information on how each read aligns to the genome
  - Interrogate file to convert to raw reads
    - Read counts across each gene (not normalized)
      - HTSeq
    - FPKM/RPKM -> TopHat/Cufflinks -> Full Solution for Differential Expression
  - From de novo assembly
    - Non-splice aware aligner (Bowtie2 or BWA)
      - Convert to raw reads via Express

Differential Expression

- TopHat/Cufflinks
- Alternatives mostly exist in R
  - TopHat/Cufflinks also interfaces with Cummerbund for visualization
  - DESeq2 – more conservative – ideal for proper replication
  - EdgeR – slightly more permissive – similar normalization
  - G-FOLD - poor replication – preliminary view of DE candidates