RNA-Seq Module

BD2K

JAX

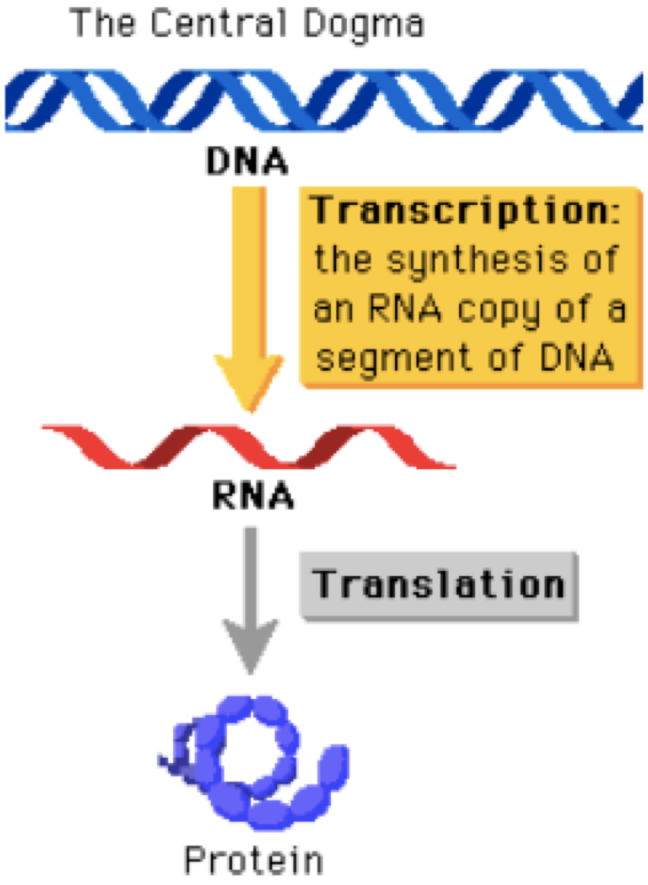
2018

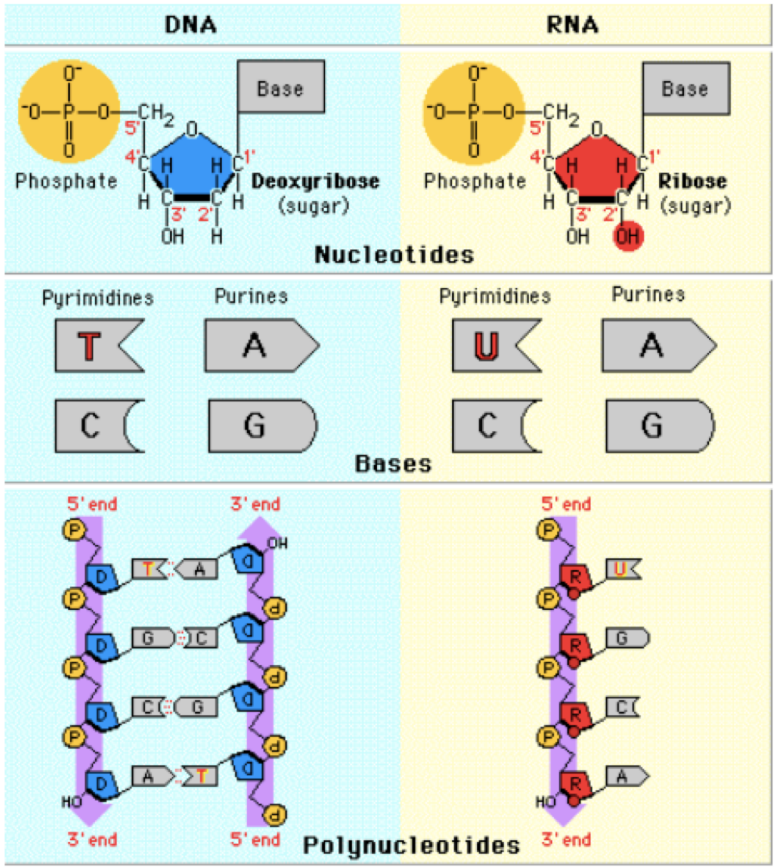
The Jackson Laboratory

Leading the search for tomorrow's cures

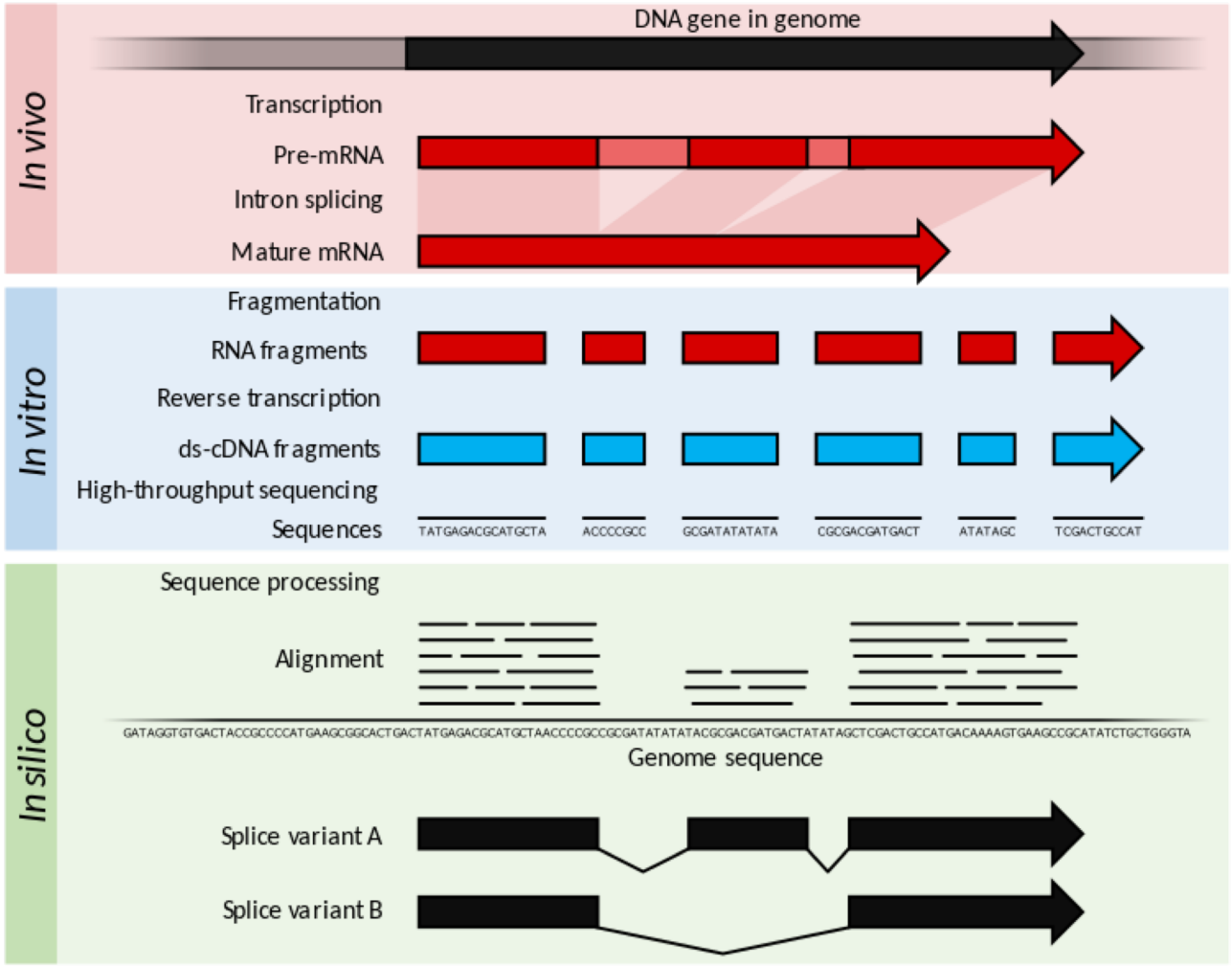
RNA-Seq Introduction

RNA

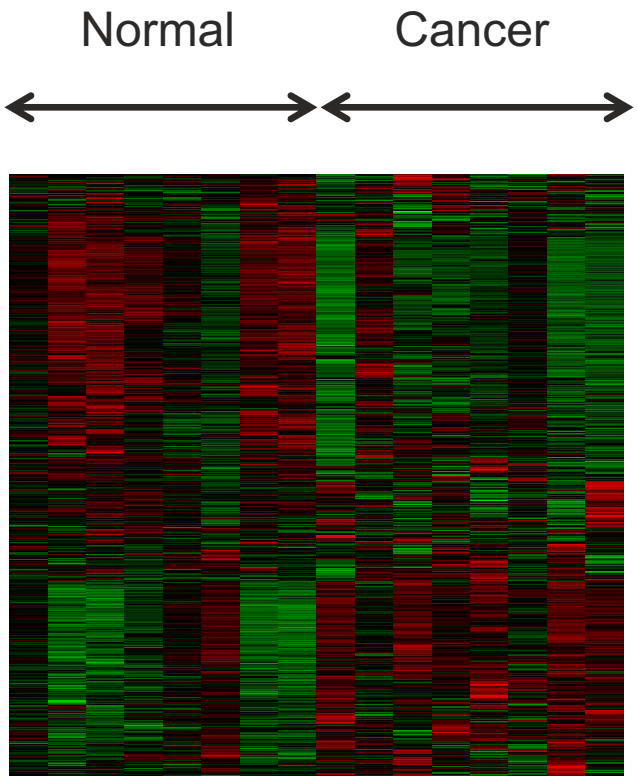

The Central Dogma
DNA
Transcription: the synthesis of an RNA copy of a segment of DNA
RNA
Translation
Protein


DNA
RNA
Base
Phosphate
Deoxyribose (sugar)
Ribose (sugar)
Nucleotides
Pyrimidines
Purines
T
A
C
G
U
Bases
5' end
3' end
Polynucleotides

RNA Sequencing (RNA-Seq)

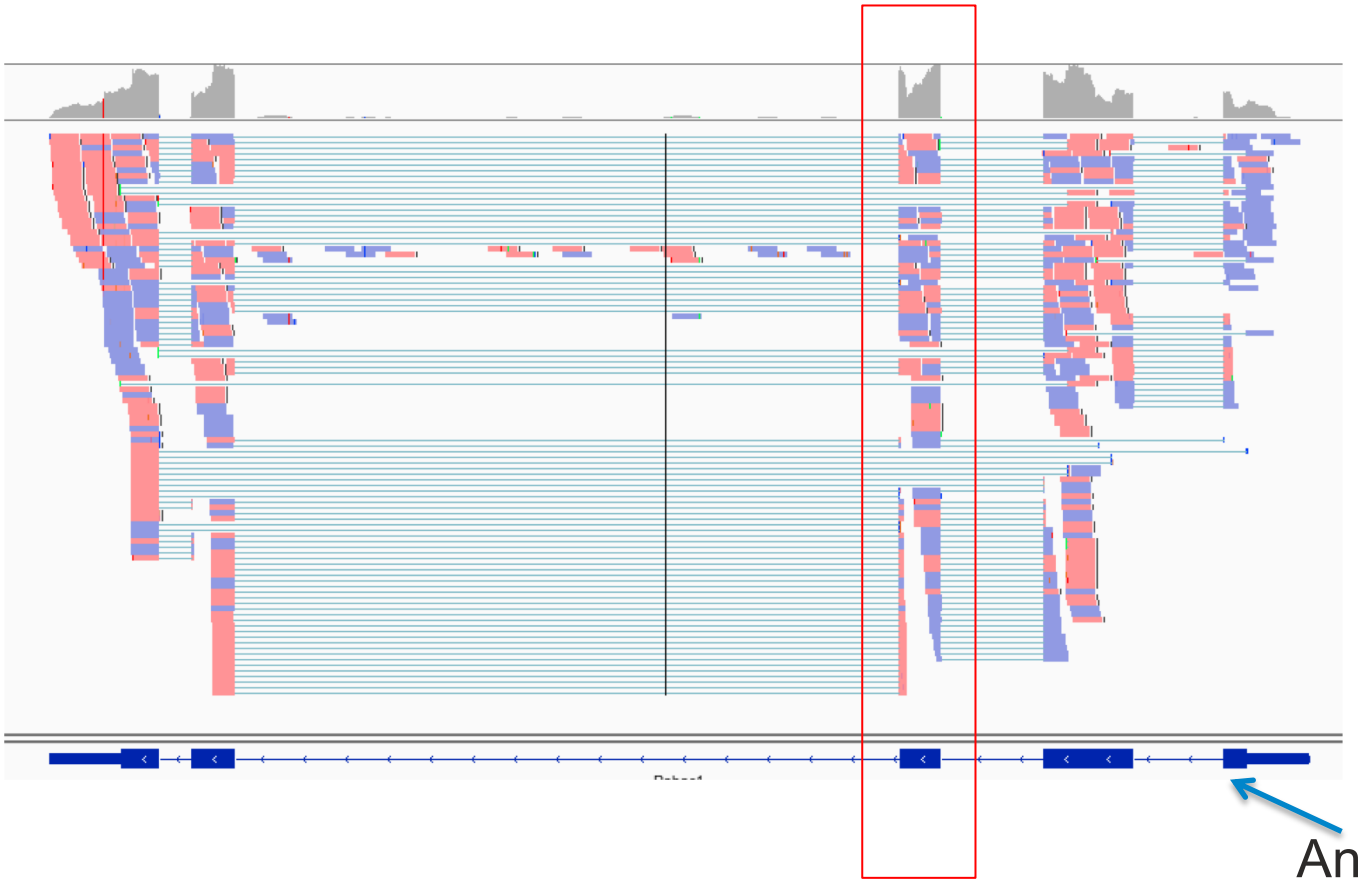


Applications of RNA-seq



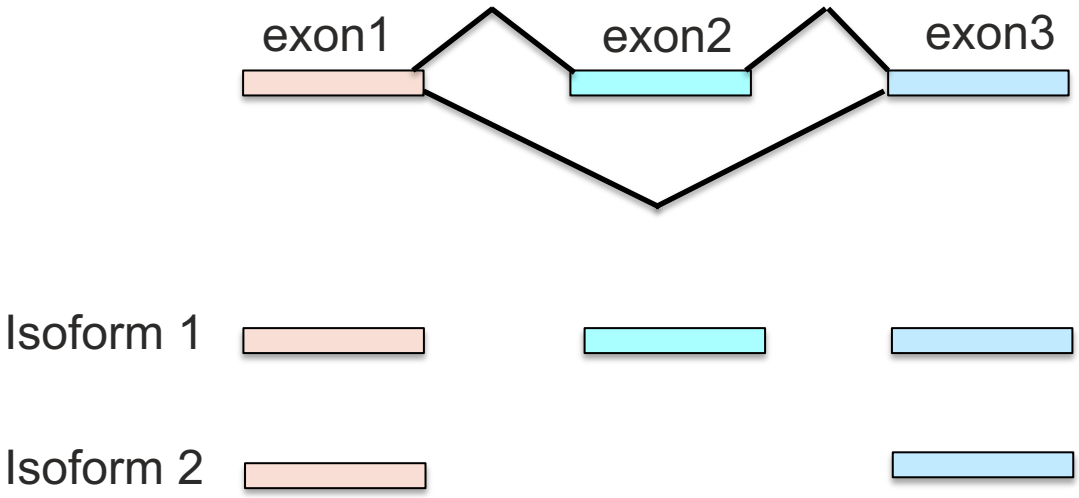
Differential Gene Expression analysis

Applications of RNA-seq



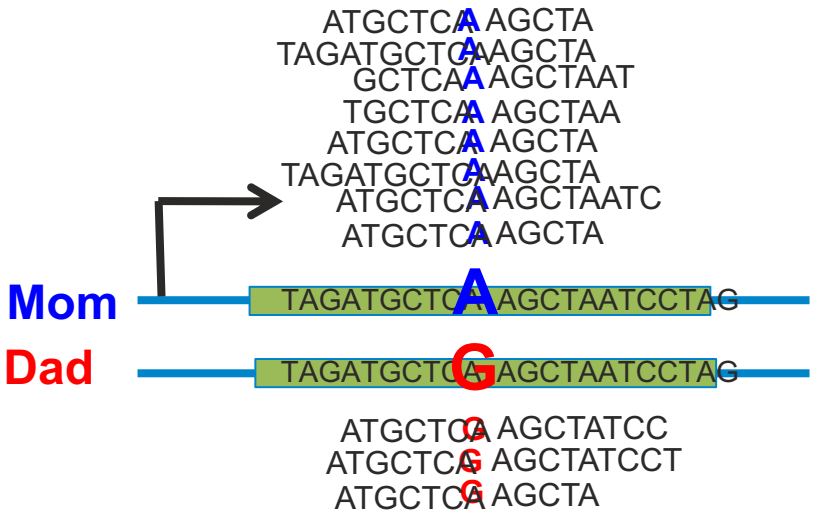
Novel exon discovery

Applications of RNA-seq



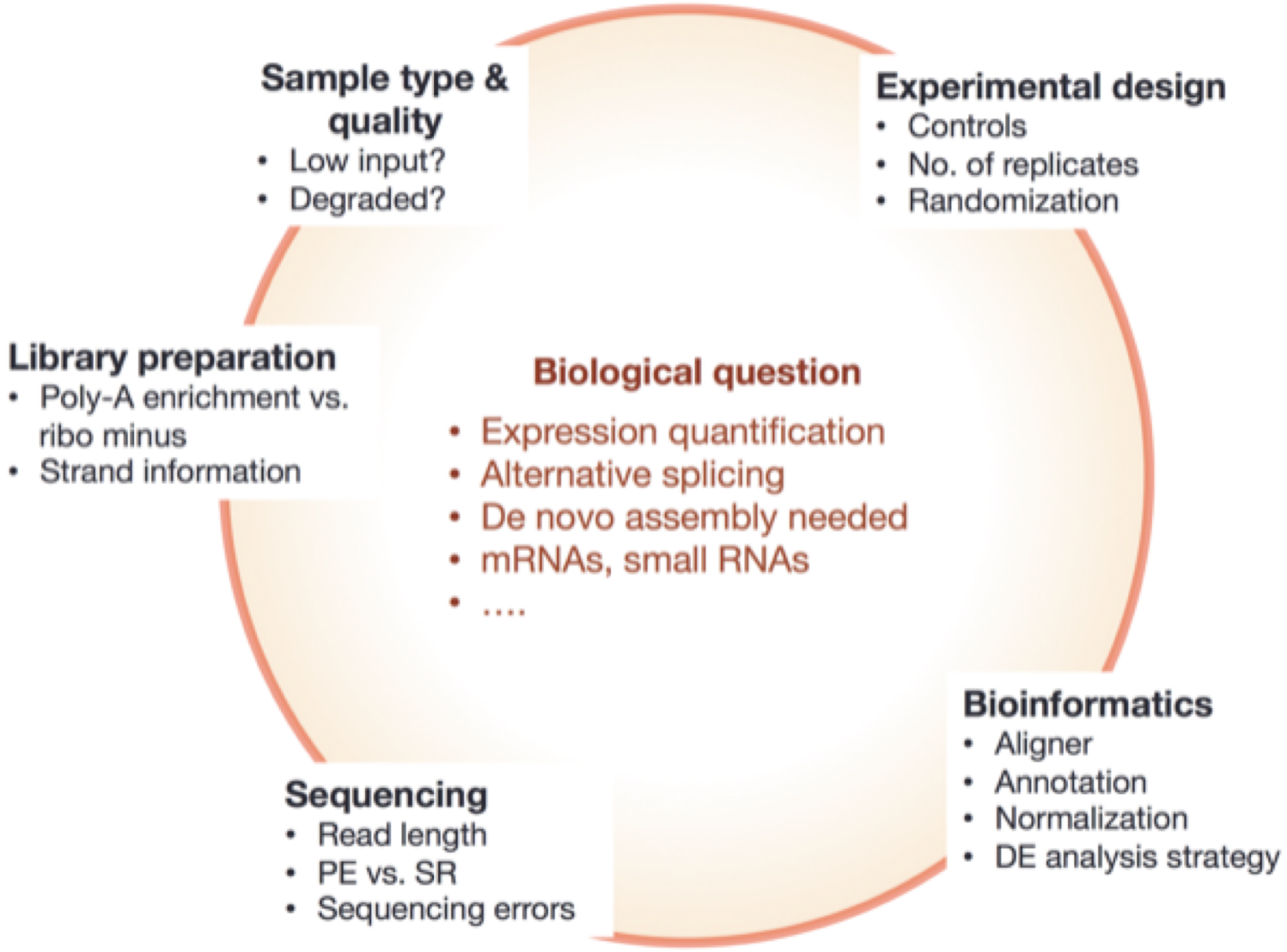
Alternative splicing

Applications of RNA-seq



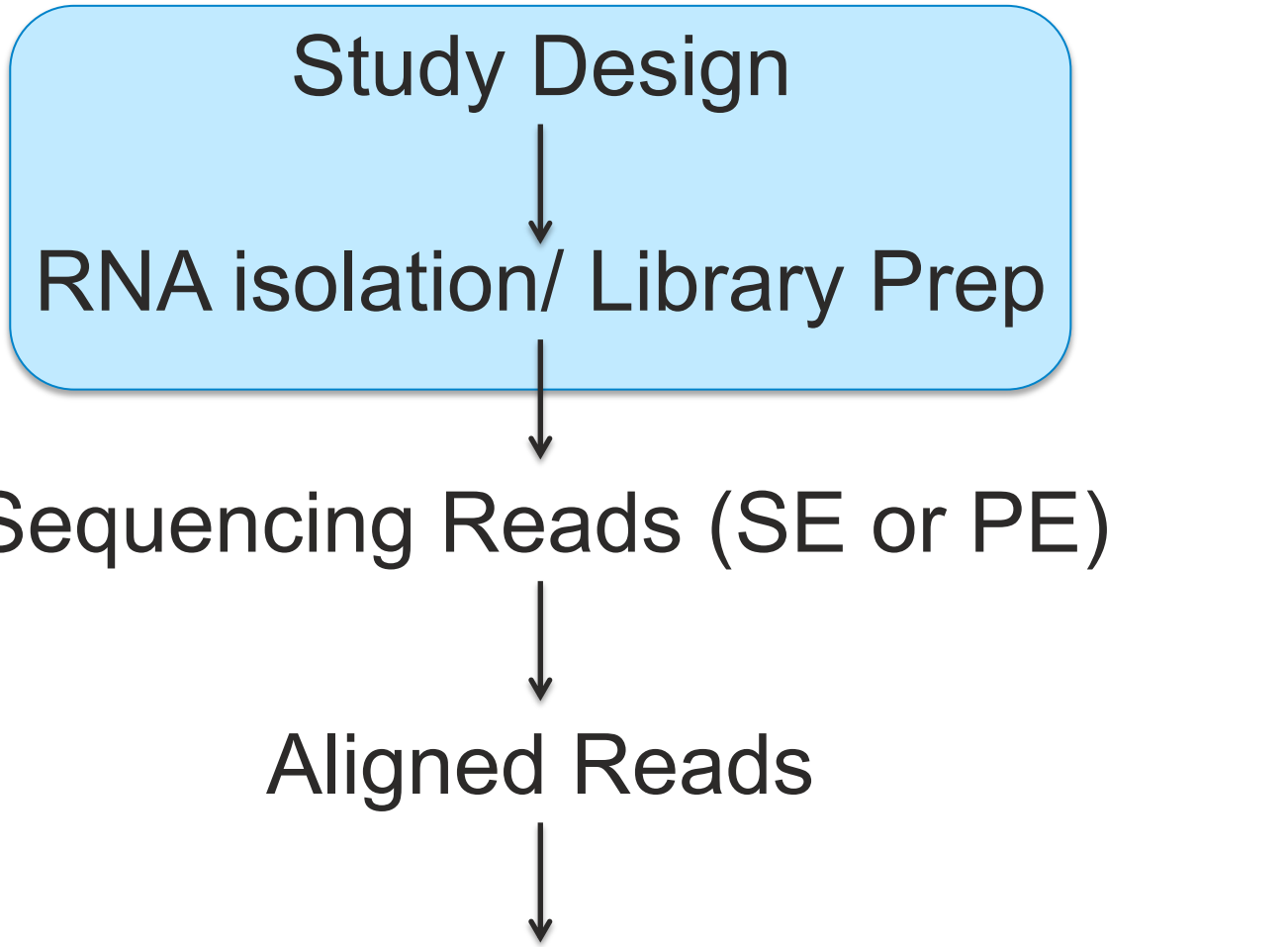
Allele-Specific gene Expression (ASE)

Preferential expression of one allele over the other.

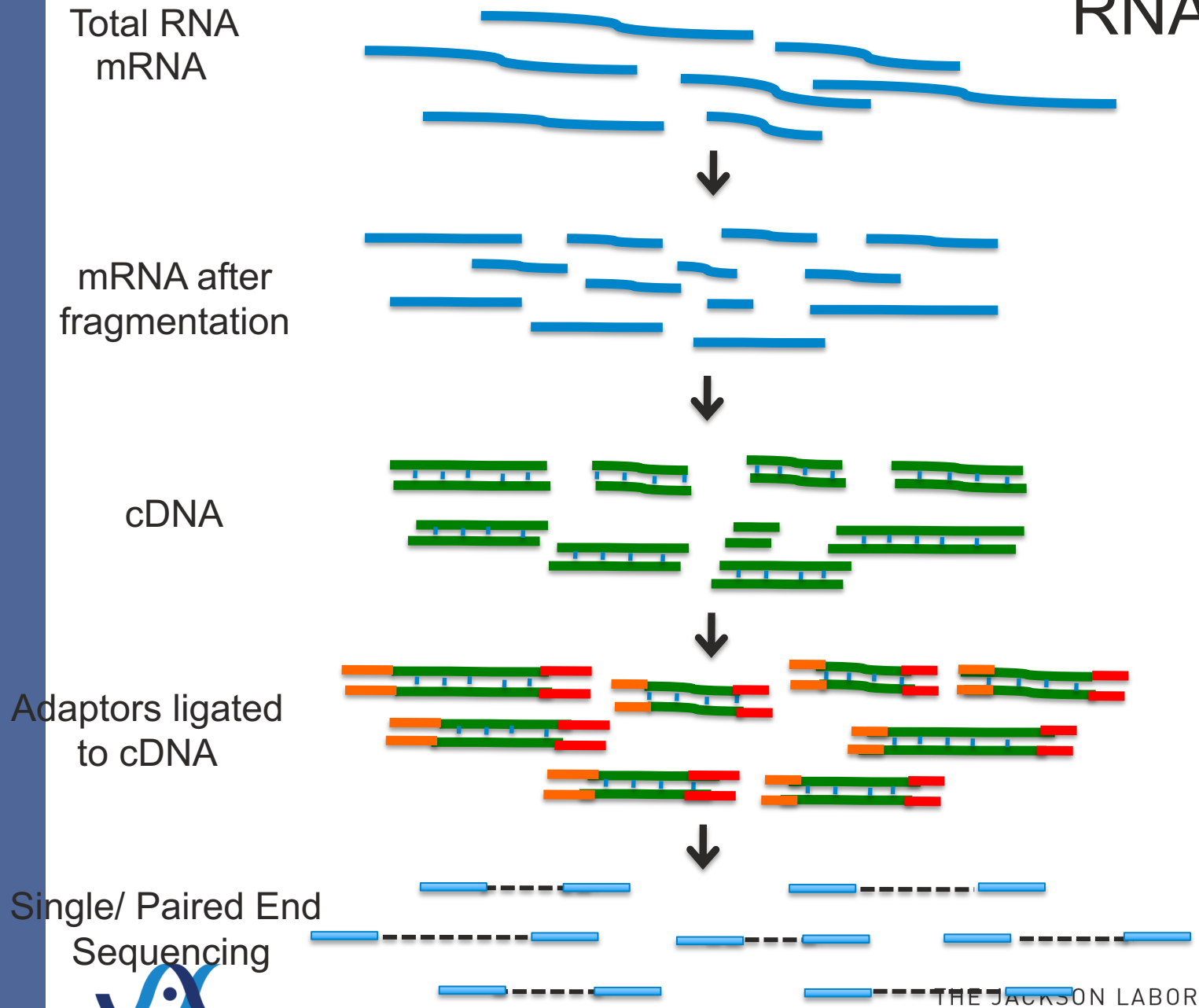


Dündar et al. (2015)

RNA-seq Work Flow



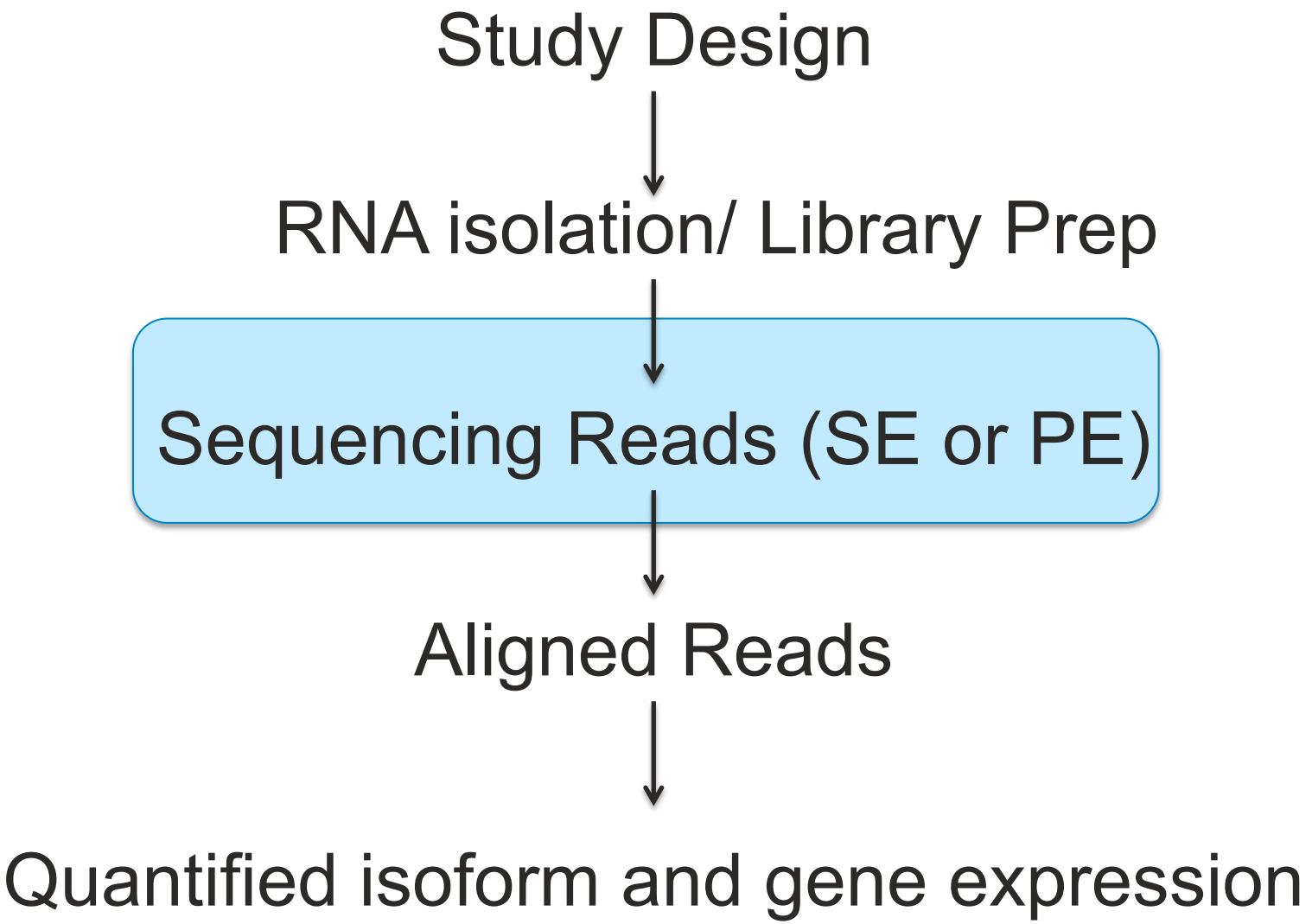
RNA-Seq


Total RNA
mRNA
mRNA after fragmentation
cDNA
Adaptors ligated to cDNA
Single/ Paired End Sequencing

Know your experiments

- How was the library constructed? (total RNA, polyA selection, rRNA depletion or RNA capture) – libraries produced by different methods are not comparable
- Single or pair end – mapping differently
- Unstranded or forward/reversed – critical for quantification
- Number of replicates – At least 2
- Read depth and read length – depend on the experimental goal (ENCODE_guideline https://www.encodeproject.org/documents/cede0cbe-d324-4ce7-ace4-f0c3eddf5972/@@download/attachment/ENCODE Best Practices for RNA_v2.pdf)

RNA-seq Work Flow



Millions and millions of reads…

```
@HISEQ2000_0074:8:1101:7544:2225#TAGCTT/1
TCACCCGTAAGGTAACAAACCGAAAGTATCCAAAGCTAAAAGAAGTGGACGACGTGCTTGGTG
GAGCAGCTGCATG
+

CCCFFFFFHHHHDHHJJJJJJJJIJJ?FGIIIJJJJJJIJJJJJJFHIJJJIJHHHFFFFD>AC?B??C?ACCAC>
BB<<<>C@CCCACCCDCCIJ
```

@HISEQ2000_0074:8:1101:7544:2225#TAGCTT/1

- Instrument: run/flowcell id
- Flowcell lane and tile number
- X-Y Coordinate in flowcell
- The member of a pair
- Index Sequence

Phred Score: $Q = -10 \log_{10} P$

10 indicates 1 in 10 chance of error
20 indicates 1 in 100,
30 indicates 1 in 1000,

Steve Munger, 2017

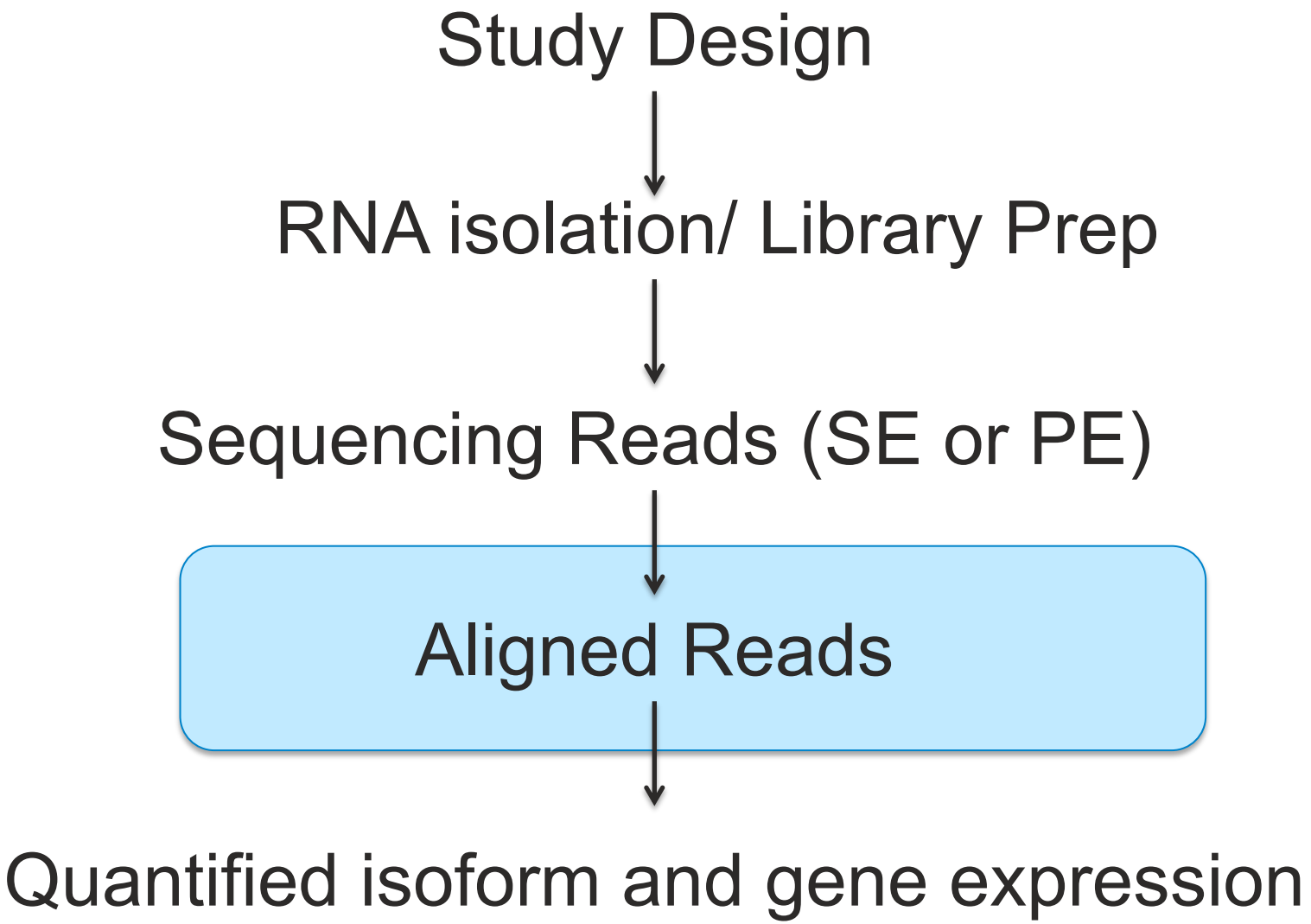
Quality Control: How to tell if your data is clean

- FASTX-Toolkit
  - http://hannonlab.cshl.edu/fastx_toolkit/
- FastQC
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

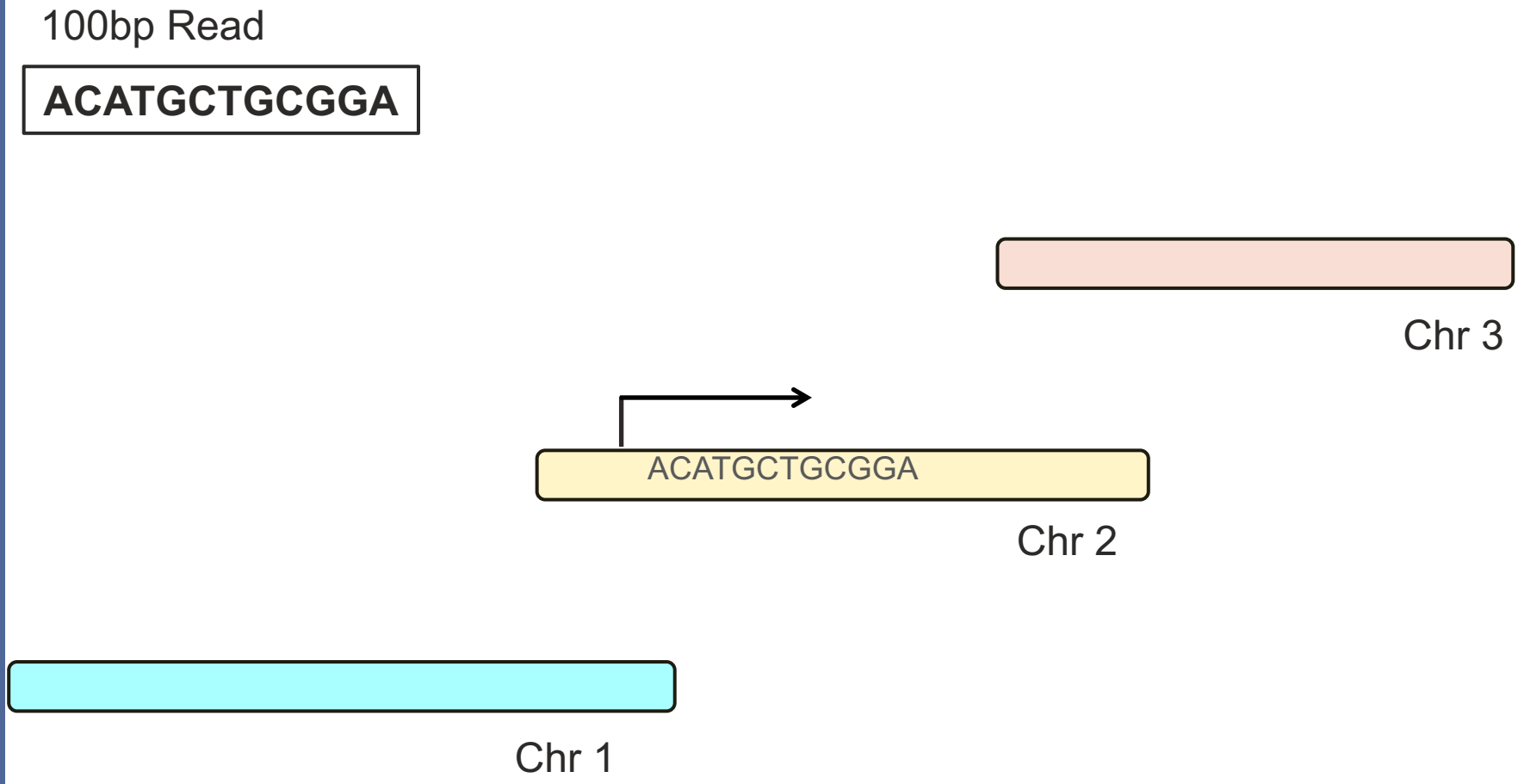
Trim or not trim?

- Signal/noise -> Preprocessing can remove low-quality “noise”, and adapters but the cost is information loss.
  - Some uniformly low-quality reads map uniquely to the genome.
  - Trimming reads to remove lower quality ends can adversely affect alignment, especially if aligning to the genome and the read spans a splice site.
  - **Most aligners can take quality scores into consideration.**
  - Currently, we do not recommend preprocessing reads aside from removing uniformly low quality samples.
  - Debate: http://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary

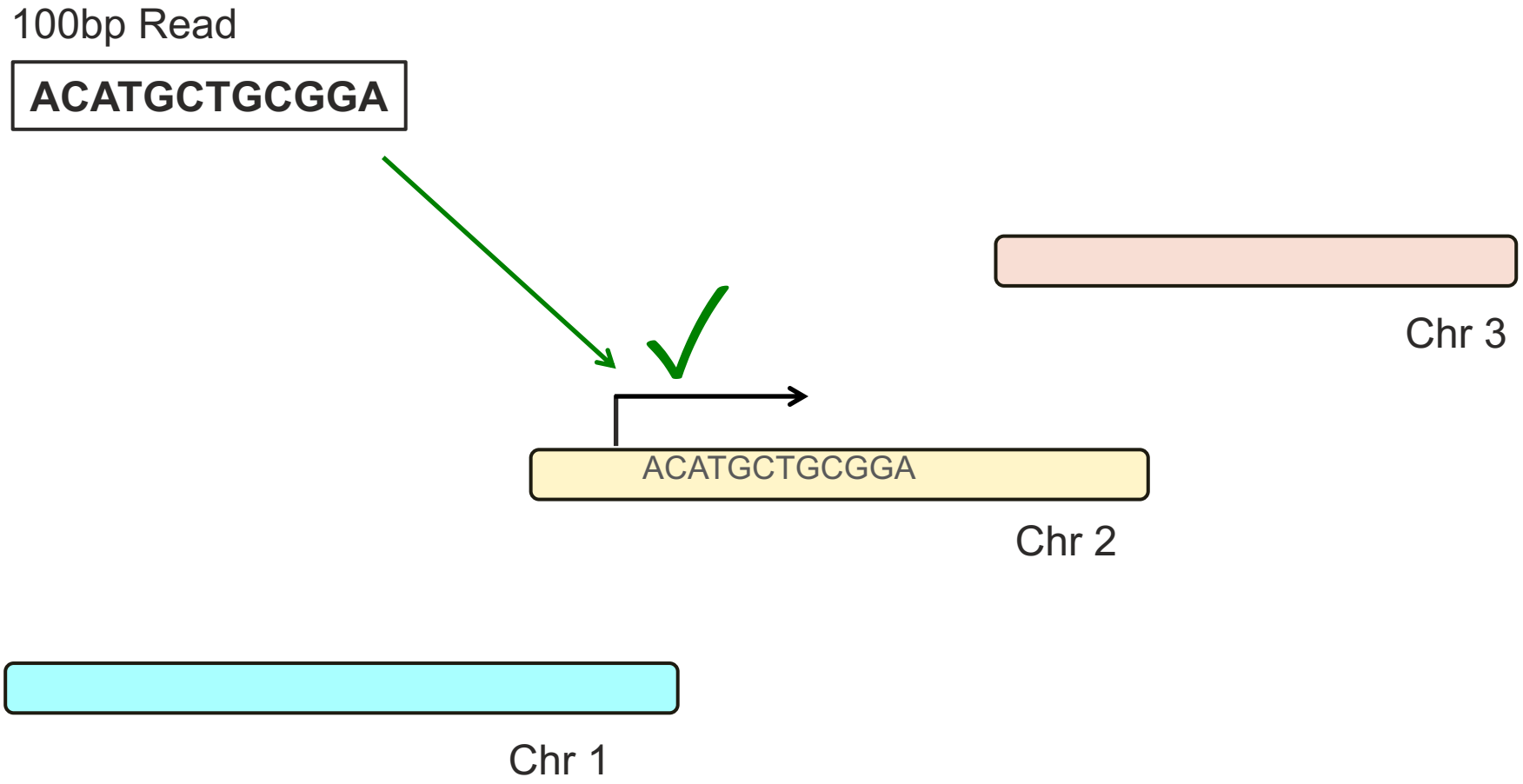
RNA-seq Work Flow



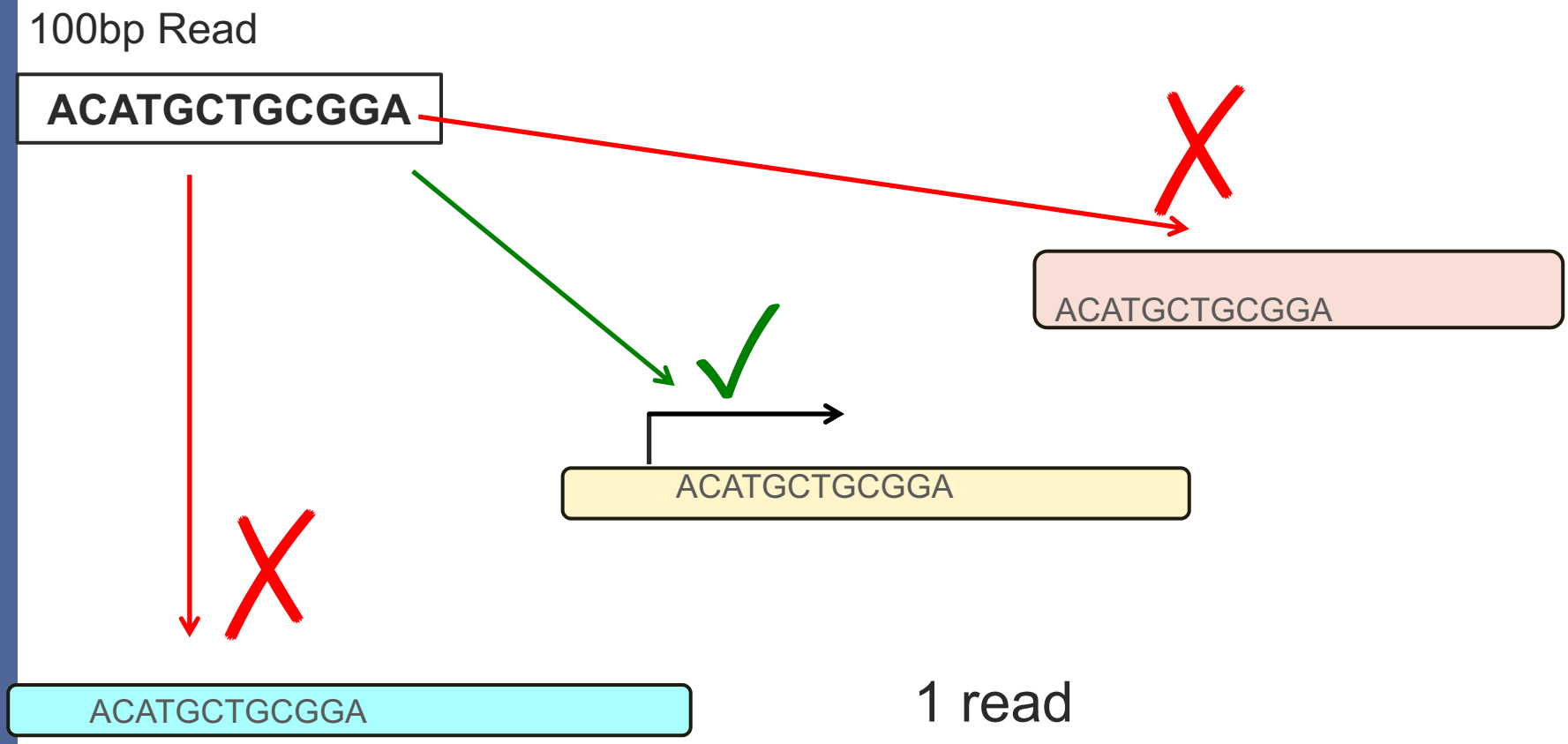
Alignment 101



The perfect read: 1 read = 1 unique alignment.

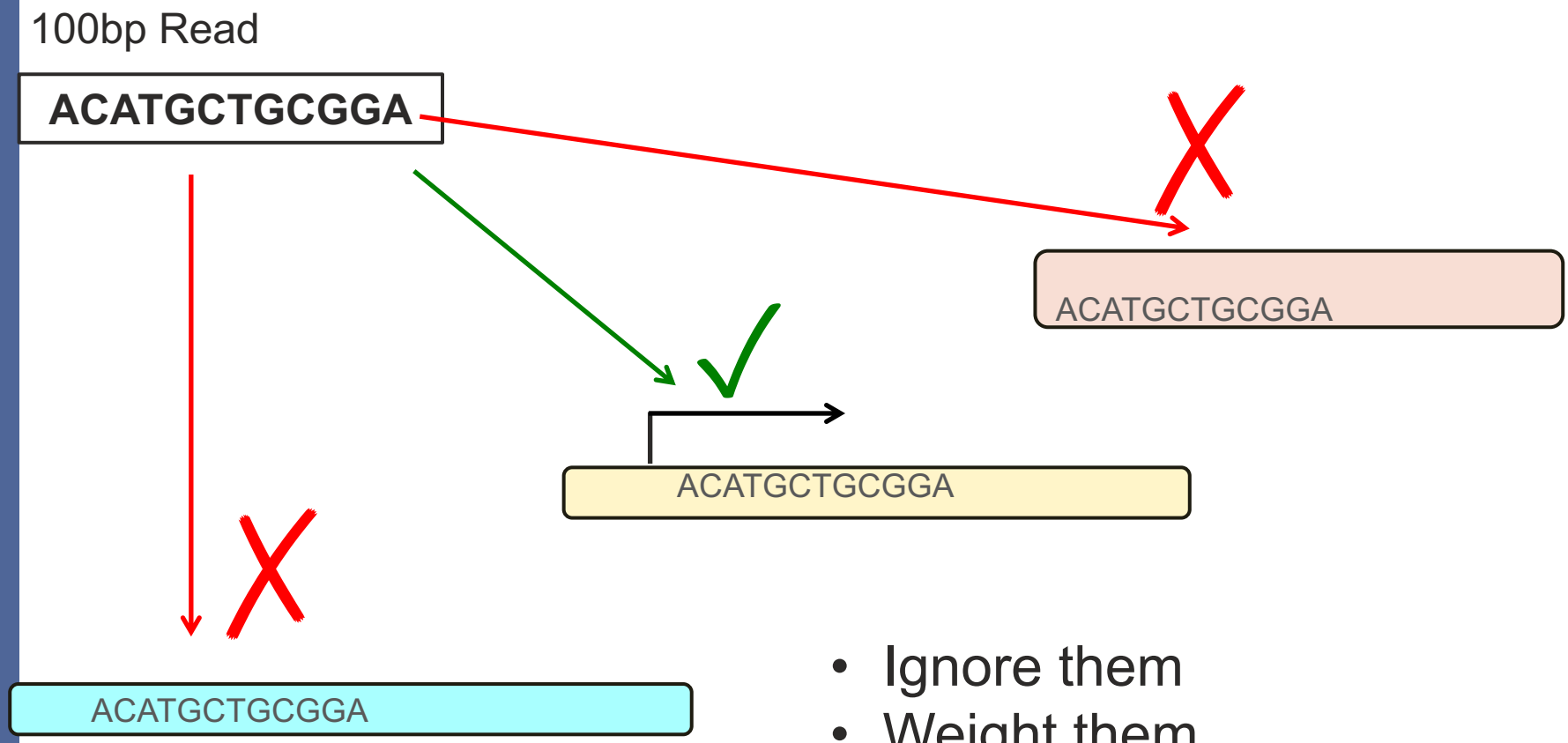


Some reads will align equally well to multiple locations. “Multireads”



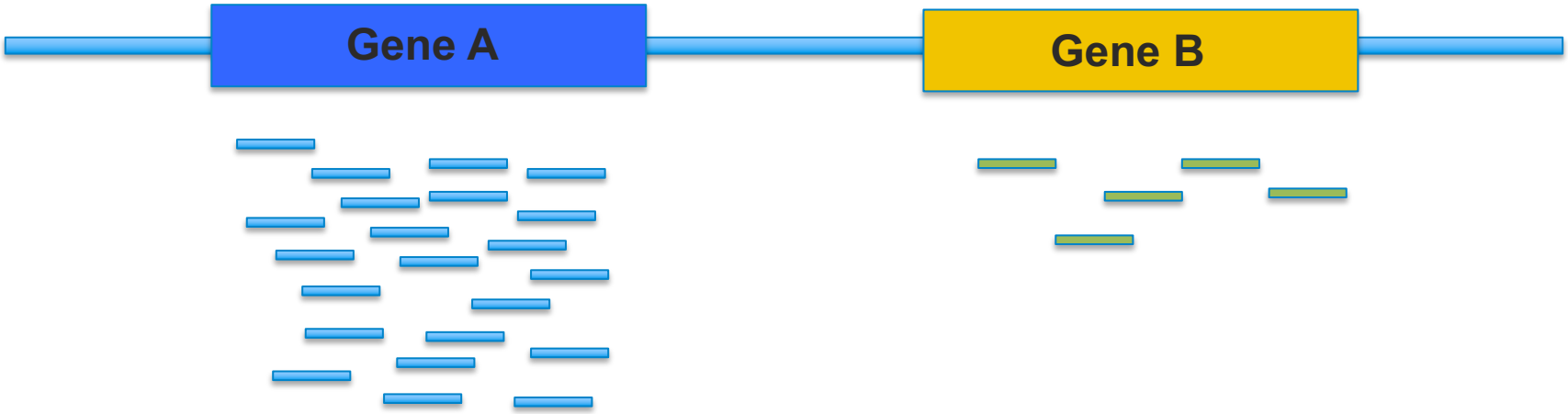
1 read
3 valid alignments
Only 1 alignment is correct

Some reads will align equally well to multiple locations. “Multireads”



- Ignore them
- Weight them
- Tools: mmquant, MMR, seqcluster, etc.

Aligning Millions of Short Sequence Reads



Aligners: STAR, HISAT2, TopHat2

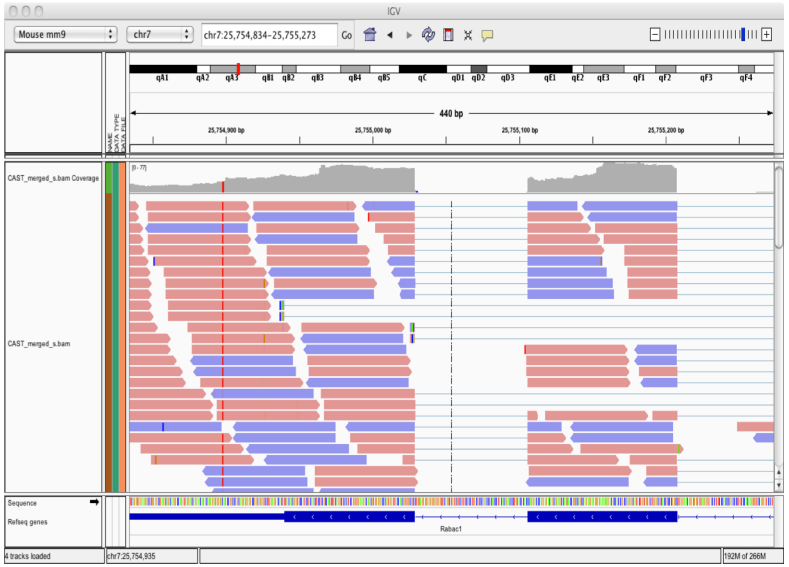
Output of most aligners: Bam/Sam file of reads and genome positions

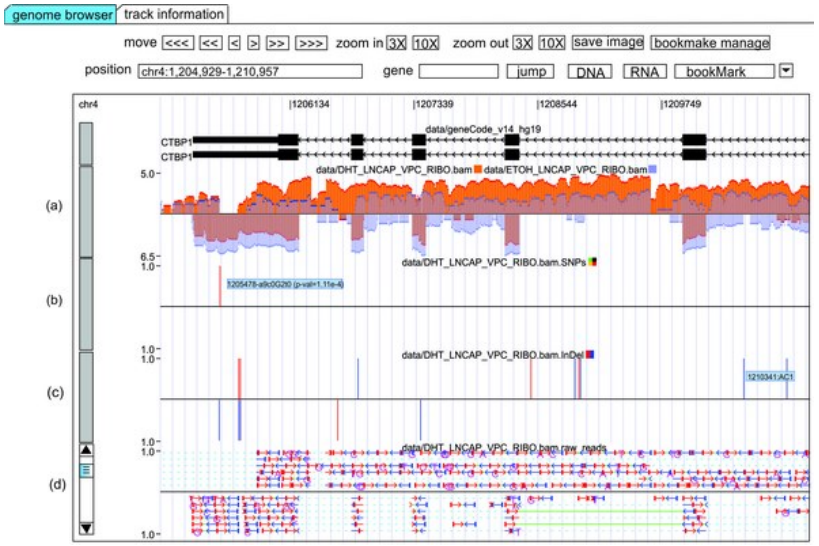
```
HISEQ2000:113:D0636ACXX:1:2308:15958:82100      409     1       3035058 0       100M    *       0       0
      AACCTGGGTATGCCTCGTAGTTAAAACATTCCTGGGAACATCTTGACCATAAGATAAAGGGGACTGTGAAGACATAGCAGGGCTATATTATCTAAGTCAA
    5:?<DC?C<CBADDDECEA=IGEHA@JJJIHEIGGIIHJIHEJJJJIIIIIIJJIIJJJIIGFJJIJJIJIHJIIJJIJIJJJJIGJHHHFHFFFFFCCC
NM:i:0  NH:i:5  CC:Z:10 CP:i:4143337    HI:i:0
HISEQ2000:113:D0636ACXX:1:2206:6975:110266      163     1       3035206 3       100M    =       3035297 191
     GTAAAAGTCACACATCAACTGGTTGCTATGTGAACAAAGATAAGCCCCCAGCCCACAGGAACAAAGTCCTGATGCACTGTGTTCTTTCTGTTAATGTTTG
    @BCFFFFDHHHHHJJIJJJJJJIJJJJJJJIJJJJIIJJJJJJJIJJJIJJJJJJJJJJIJIJJHHHHHHFFFFFEEEEEECDDEDDDDDEDDDEDEEE<
NM:i:0  NH:i:2  CC:Z:4  CP:i:118529266  HI:i:0
HISEQ2000:113:D0636ACXX:1:2206:6975:110266      83      1       3035297 3       100M    =       3035206 -191
    TAATGTTTGAATAAGCCAATAGTGTGTTGCTATGCTGAATTCCACACCCCTAAGCCCCGTACCCCATAAAAGCCCCTGGCTTTCGAGCCTCGTGGCCGGC
CCC?:DEDEEDDDDDDDEDDCDD@?@DDDDDDDDDC@DDDCB?55,DDDDCDDDDFFHHB;JJJJIGJJJJIJJJJIJGIIGIJJJJHHHHHFFFFFCCC    NM:i
:2  NH:i:2  CC:Z:4  CP:i:118529175  HI:i:0
HISEQ2000:113:D0636ACXX:1:1204:3972:146753      329     1       3044627 1       100M    *       0       0
       ATTCATGGCCCATGCCGACTTTGTTTCTAGAGGACAAACAGTTTCAAGGGCTCCTGGATACCGGGGCAGATGTGACAGTAATTTCCTCAACACATTGGCC
    BCCFFFDFHHGHHJIJJFIIJJFIIIGIIGGGGGIIGIGBG0BFGIG?DHGCGHGHIIIBEHEEDDDBBBCDDDC>@C3>>BDDDECDD<?@??CC@:>:
NM:i:1  NH:i:4  CC:Z:15 CP:i:9154226    HI:i:0
HISEQ2000:113:D0636ACXX:1:2201:2762:178840      355     1       3045593 0       100M    =       3045612 119
     ATAAATTAAAAGTTTTAGATGCTTGCCAGAAACTGTTAGAAAATTTTGGATTTTAATCTTGGTTTGACAAGCTACCTCTTCTTACAAGCAGGAAAGGAAA
    CC@FFFFFHHGHDHHJJIJJJJJIJIIIJJIGIIJIIIGIIJIHIIJIIDCGHIGIJIIJJDIIIGGIIIGIIJJJEFEHFFEFFFDD?ED?7=CDA?CD
NM:i:0  NH:i:7  CC:Z:=  CP:i:16822458   HI:i:0
```

Tools to handle Bam/Sam files: samtools, bamUtil

Visualization of alignment data (BAM/SAM)

Genome browsers – IGV and RNASeqBrowser





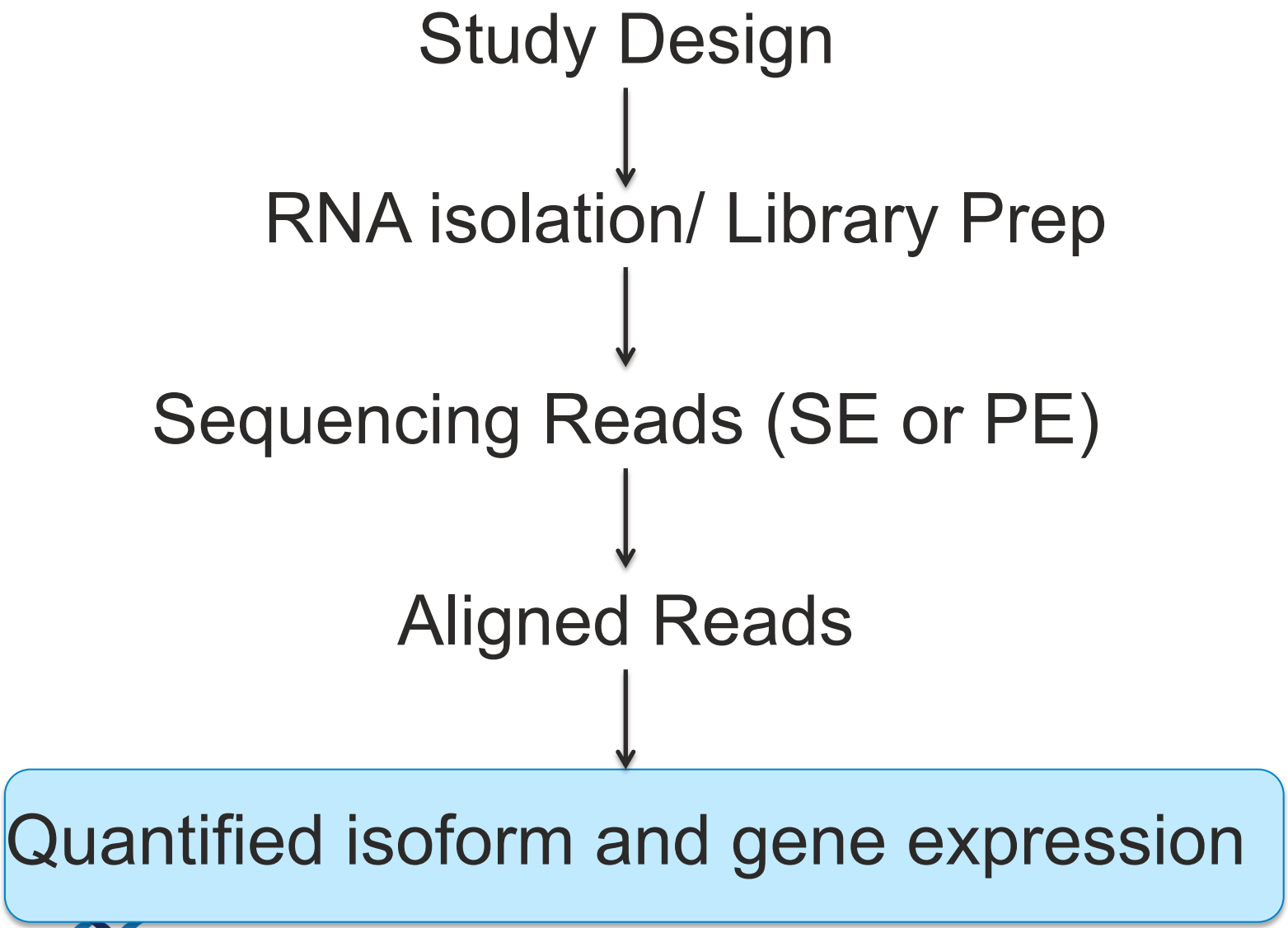
Integrative Genome Viewer (IGV)

http://software.broadinstitute.org/software/igv/download

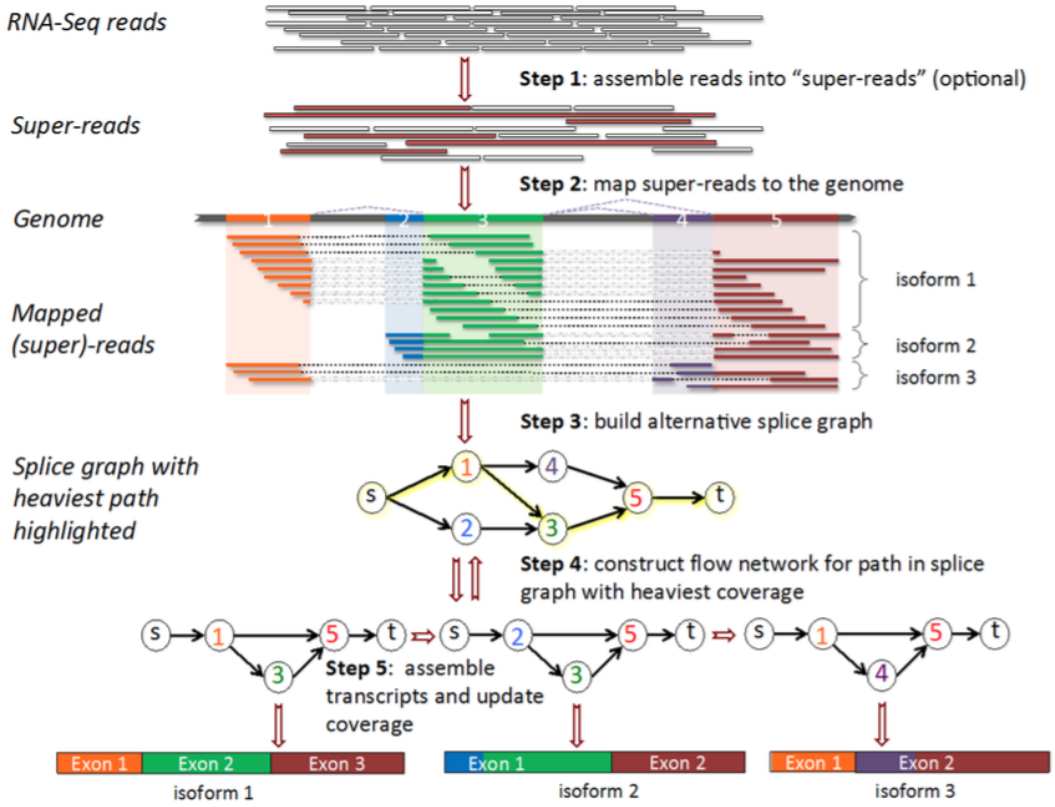
RNASeqBrowser

http://www.australianprostatecentre.org/research/software/rnaseqbrowser

RNA-seq Work Flow



Transcript quantification



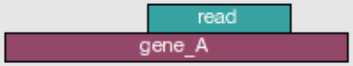
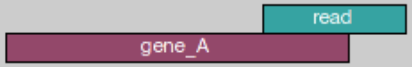
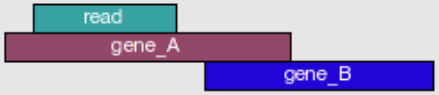
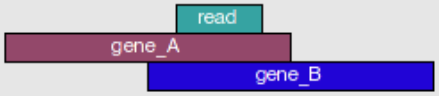
StringTie reconstruct transcripts from spliced read alignments generated by previously mentioned aligners

* Sailfish, Kallisto and Salmon align reads to annotated transcriptome sequences

Supplementary Figure 12. The StringTie algorithm: RNA-seq reads are assembled into super-reads (Step 1) and then super-reads plus un-assembled reads are mapped to the genome (Step 2). In Step 3, mapped reads and super-reads are used to build an alternative splice graph. We use the path from source (s) to sink (t) with the heaviest coverage to build a flow network corresponding to the transcript represented by that path (Step 4). The maximum flow in this network represents the coverage of one assembled transcript, which is removed from the splice graph (Step 5). Steps 4 and 5 are repeated until no more transcripts can be assembled.

Pertea, 2015

Gene expression

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

HTseq-count read/feature overlap modes

* Another popular tool: featureCounts (http://bioinf.wehi.edu.au/featureCounts/)

HTseq-count documentation

Expression Abundance: Counts, RPKM/FPKM, TPM

* Raw counts are required input for differential analysis by DESeq2 and EdgeR

Table 11: Normalization methods for the comparison of gene read counts within the same sample.

Name	Details	Comment
RPKM (reads per kilobase of exons per million mapped reads)	1. For each gene, count the number of reads mapping to it. 2. Divide that count by: the length of the gene in base pairs divided by 1,000 multiplied by the total number of mapped reads divided by 10^6. $RPKM_i = \frac{read\ count\ of\ gene\ i}{(\frac{length\ of\ gene\ i}{10^3})(\frac{library\ size}{10^6})}$	• introduces a bias in the per-gene variances, in particular for lowly expressed genes (Oshlack and Wakefield, 2009) • implemented in `edgeR`'s `rpkm()` function
FPKM (fragments per kilobase...)	1. Same as RPKM, but for paired-end reads: 2. The number of fragments (defined by two reads each) is used.	• implemented in `DESeq2`'s `fpkm()` function
TPM	Instead of normalizing to the total library size, TPM represents the abundance of an individual gene i in relation to the abundances of the other transcripts (e.g., j) in the sample. 1. For each gene, count the number of reads mapping to it and divide by its length in base pairs (= counts per base). 2. Multiply that value by 1 divided by the sum of all counts per base of every gene. 3. Multiply that number by 10^6. $TPM_i = \frac{X_i}{l_i} * \frac{1}{\sum_j \frac{X_j}{l_k}}$	• details in Wagner et al. (2012)

Good for comparison within one sample but not for cross sample comparison

Dündar et al. 2015

More downstream analysis

- Differential expression analysis – DESeq2, EdgeR, and limma-voom (Schurch et al. 2015 for reviews of DE tools)
- Gene set enrichment analysis i.e. Gene Ontology (GO) – GOrilla, DAVID, g:profiler
- Network-based - GeneMania

Interactive web-based tools

- Galaxy (https://usegalaxy.org)
- GenomeSpace (http://www.genomespace.org)
- Degust: Perform RNA-seq analysis and visualization (http://degust.erc.monash.edu/degust-old/index.html)

Where to get the data?

- GenBank (http://www.ncbi.nlm.nih.gov/genbank/)
- ENA (https://www.ebi.ac.uk/ena/)
- DDBJ (http://www.ddbj.nig.ac.jp/intro-e.html)
- ENCODE (https://www.encodeproject.org)

RNA-Seq is still evolving

- Single cell
- Longer reads
- Nascent RNA-Seq

Keep updated

"RNA-Seq is not a mature technology. It is undergoing rapid evolution of biochemistry of sample preparation; of sequencing platforms; of computational pipelines; and of subsequent analysis methods that include statistical treatments and transcript model building." From ENCODE RNA-Seq analysis guidelines

Resources

- https://www.ebi.ac.uk/gxa/home
- https://www.ncbi.nlm.nih.gov/gds
- https://portal.gdc.cancer.gov
- https://www.ebi.ac.uk/arrayexpress/

page | discussion | view source | history

http://cosi.iscb.org/wiki/Education:Home

log in / create

NAL SOCIETY FOR
ONAL BIOLOGY
OF SPECIAL INTEREST

age
changes
n page

Search

nks here
changes
pages
le version
nent link
formation

EDUCATION
COMMUNITY OF SPECIAL INTEREST

The life sciences are increasingly reliant on computational and mathematical approaches for biological data storage, analysis, visualisation and interpretation. But bioinformatics and computational biology, and the technologies that underpin them, are swift-moving disciplines and it can be difficult to keep pace. The ISCB Edu COSI focuses on education and training in this fast-moving arena. A major goal of this Community of Special Interest is to foster a collaborative community in whic bioscientists can share bioinformatics education and training resources and experiences, and facilitate the development of education programs, courses, curricula teaching tools and methods.

The Education COSI was established in 2014. We hope that this space becomes a place for you get information about bioinformatics education, to start discussio relevant topics and get connected to other like-minded people.

Getting involved

- We have an active ISCB Education Committee and welcome your participation.
- We regularly organise Workshops on Education in Bioinformatics (WEB) at ISMB and ECCB conferences, offering opportunities and strategies for the provisio bioinformatics training to engaged audiences - contact us with suggestions for future workshops.
- We are members of the Global Organisation for Bioinformatics Learning, Education & Training (GOBLET), an umbrella organisation that brings together major international and national bioinformatics and computational biology societies and networks, aiming to provide a global, sustainable support and networking infr for bioinformatics trainers and trainees. GOBLET's main meetings are held annually, with interim events throughout the year - join us.
- We have an active Curriculum & Competencies task-force - get involved.
- At ISMB/ECCB in Berlin, we established a poster track dedicated to educational activities. This activity recurs each year..
- During ISMB 2018 in Chicago, we will host our first full day Education COSI.

Important dates

- January 29, 2018 - Deadline for Proceedings Submission for ISMB 2018 [Closed]
- March 15, 2018 - Meeting Registration Opens for ISMB 2018
- April 5, 2018 - Deadline for Abstract Submission for Talks and Posters for ISMB 2018
- July 8, 2018 - One day **Education COSI at ISMB 2018**

Steering committee

- Fran Lewitter, Whitehead (Chair)
- Lonnie Welch, Ohio University
- Terri Attwood, The University of Manchester

RNA-Seq hands-on

Outline

- Prepare data
- Alignment
- Count feature
- DE analysis
- Gene set enrichment analysis

Background

JBC ARTICLE

Author's Choice

The transcription factor *Pax6* is required for pancreatic β cell identity, glucose-regulated ATP synthesis, and Ca^{2+} dynamics in adult mice

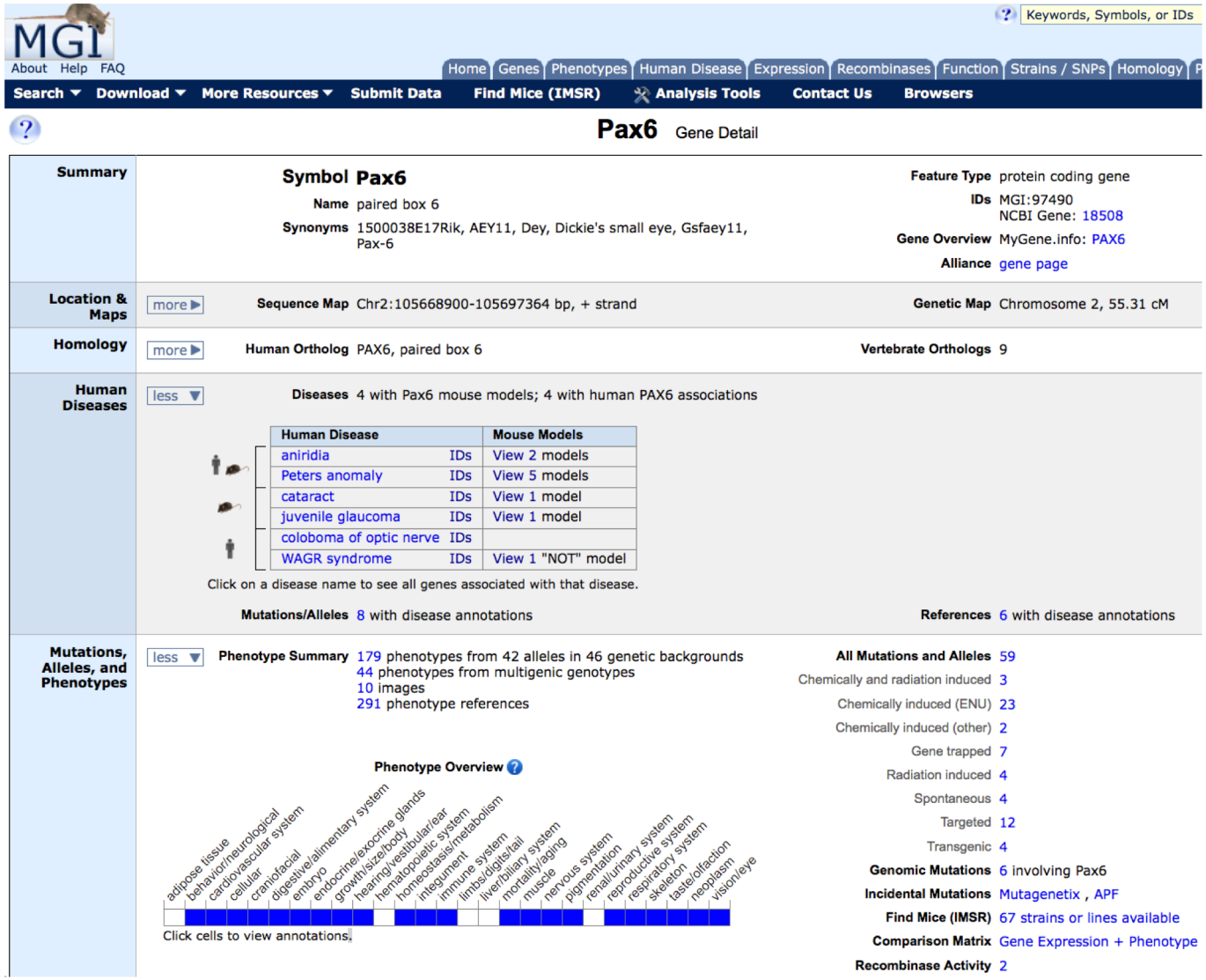
Received for publication, March 6, 2017, and in revised form, April 3, 2017 Published, Papers in Press, April 4, 2017, DOI 10.1074/jbc.M117.784629

Ryan K. Mitchell[‡], Marie-Sophie Nguyen-Tu[‡], Pauline Chabosseau[‡], Rebecca M. Callingham[‡], Timothy J. Pullen[‡], Rebecca Cheung[‡], Isabelle Leclerc[‡], David J. Hodson[‡§¶1,2], and Guy A. Rutter[‡2,3]

From the [‡]Section of Cell Biology and Functional Genomics, Division of Diabetes, Endocrinology, and Metabolism, Imperial College London, Du Cane Road, London W12 0NN, United Kingdom, the [§]Institute of Metabolism and Systems Research and Centre of Membrane Proteins and Receptors, University of Birmingham, Edgbaston B15 2TT, United Kingdom, and the [¶]Centre for Endocrinology, Diabetes, and Metabolism, Birmingham Health Partners, Birmingham B15 2TH, United Kingdom

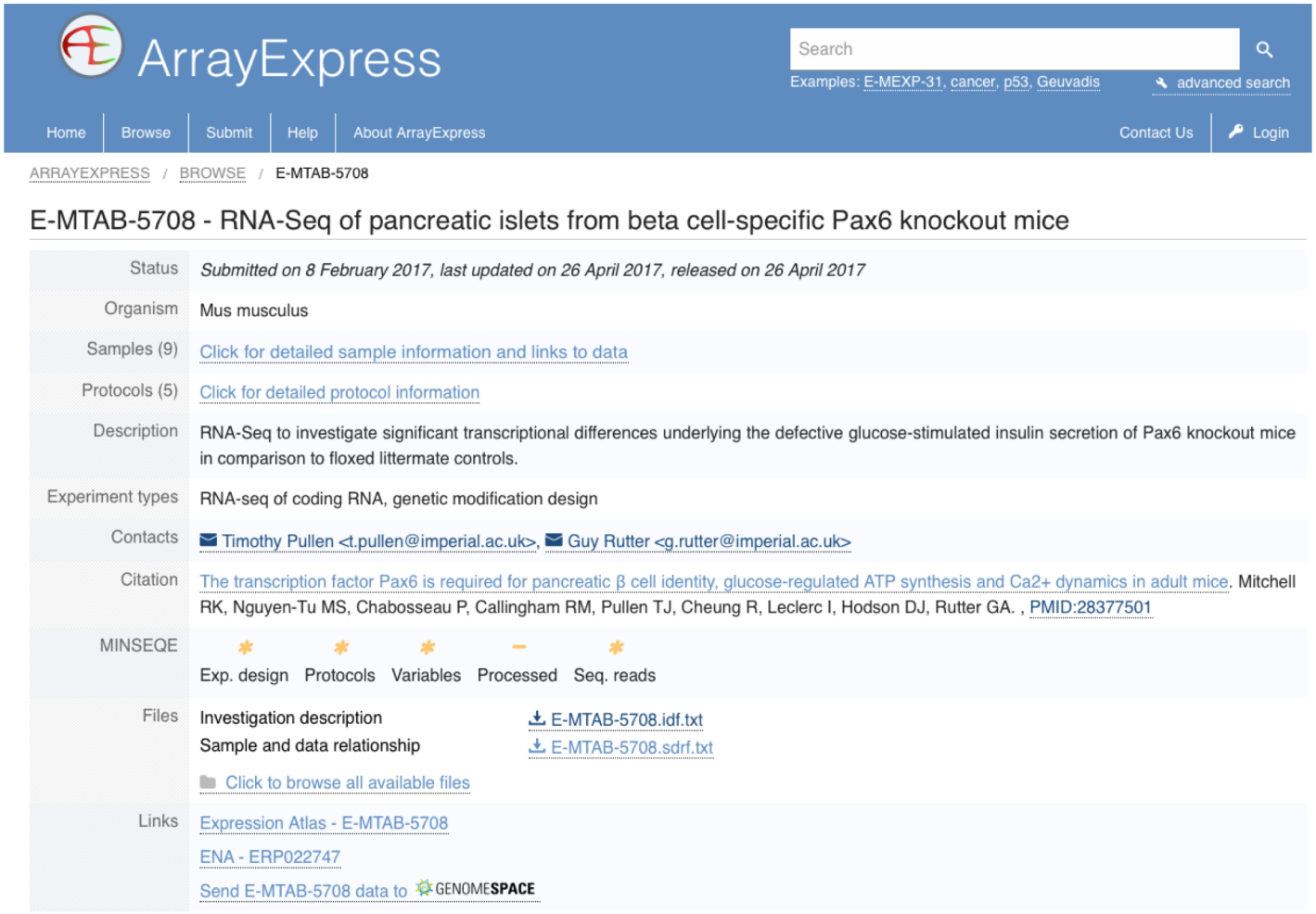
Background

Pax6



Data fact

Data



- Poly-A selected
- Reverse stranded

Open up galaxy …