

RNA-Seq Module

BD2K

JAX

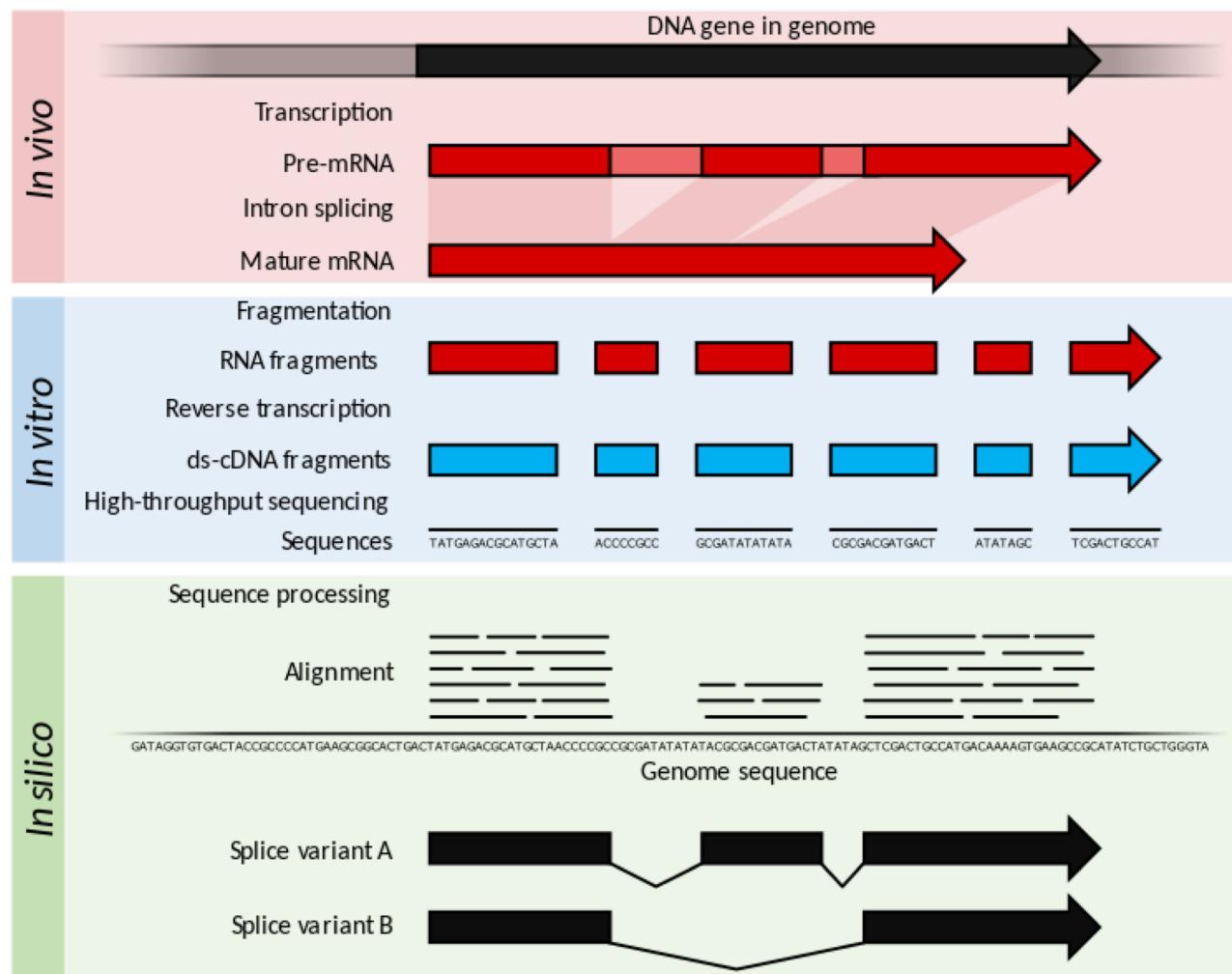
2019



Introduction

RNA-Seq statement	True	False
Next-generation sequencing		
Quantify RNA in a biological sample at a given moment		
To analyze transcriptome		
3 rd -generation sequencing		
Sequencing performs on cDNA not on RNA		

RNA-Seq



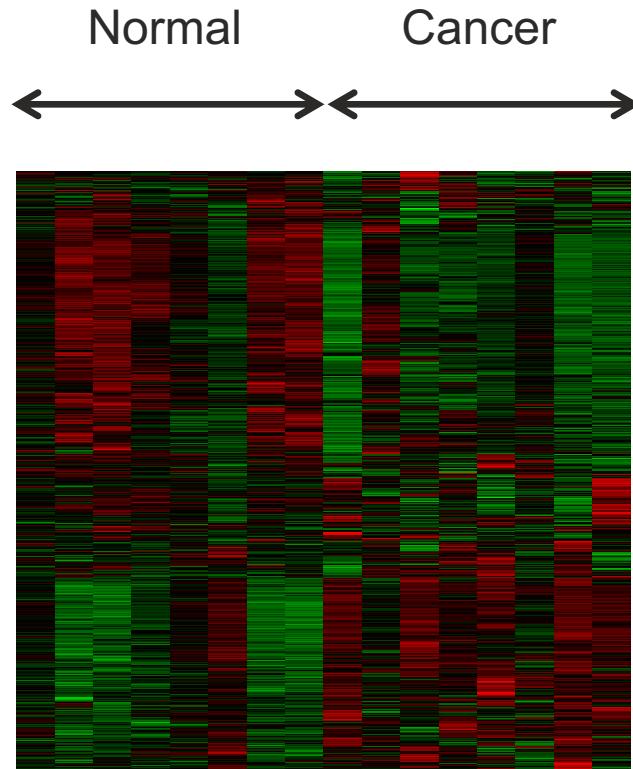
Wikimedia Commons

THE JACKSON LABORATORY

What information you can learn from RNA-Seq?



Applications of RNA-seq



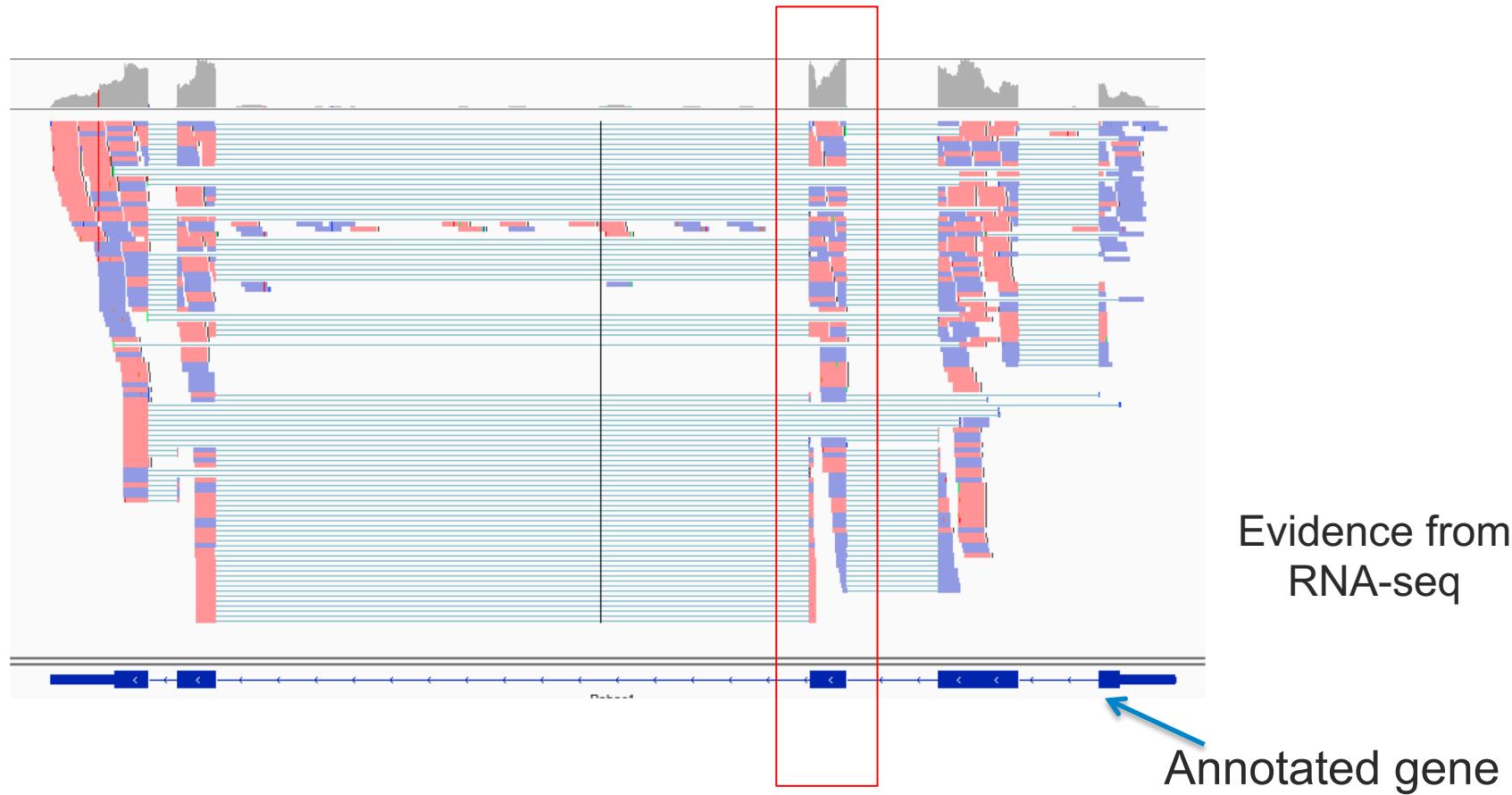
Differential Gene Expression analysis



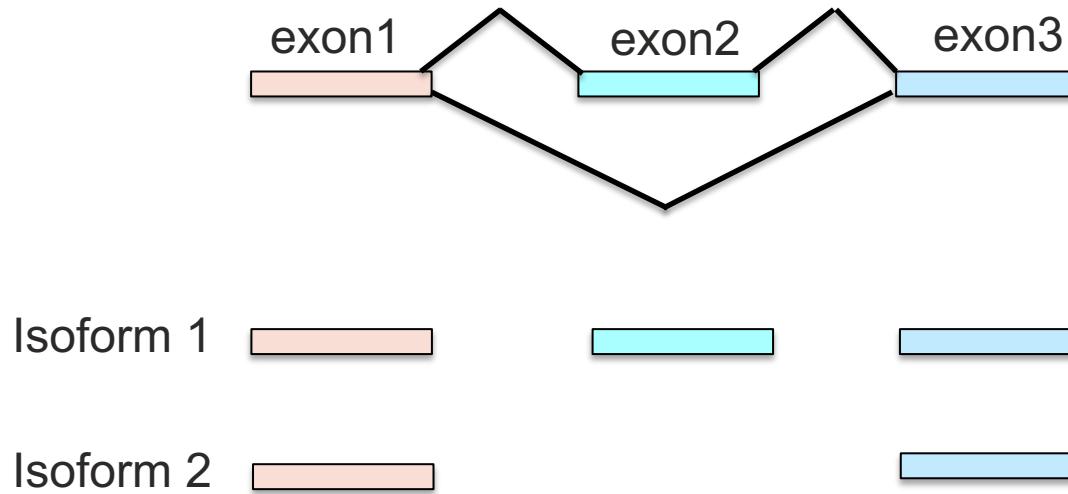
Steve Munger, 2017

THE JACKSON LABORATORY

Applications of RNA-seq

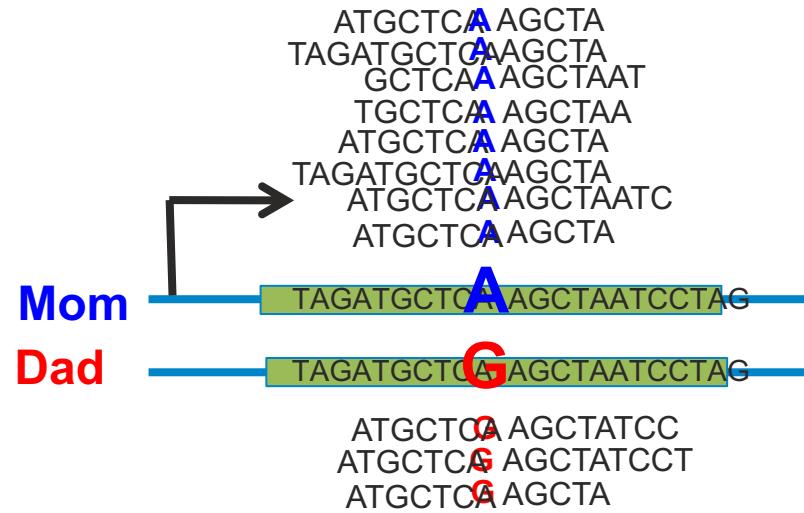


Applications of RNA-seq



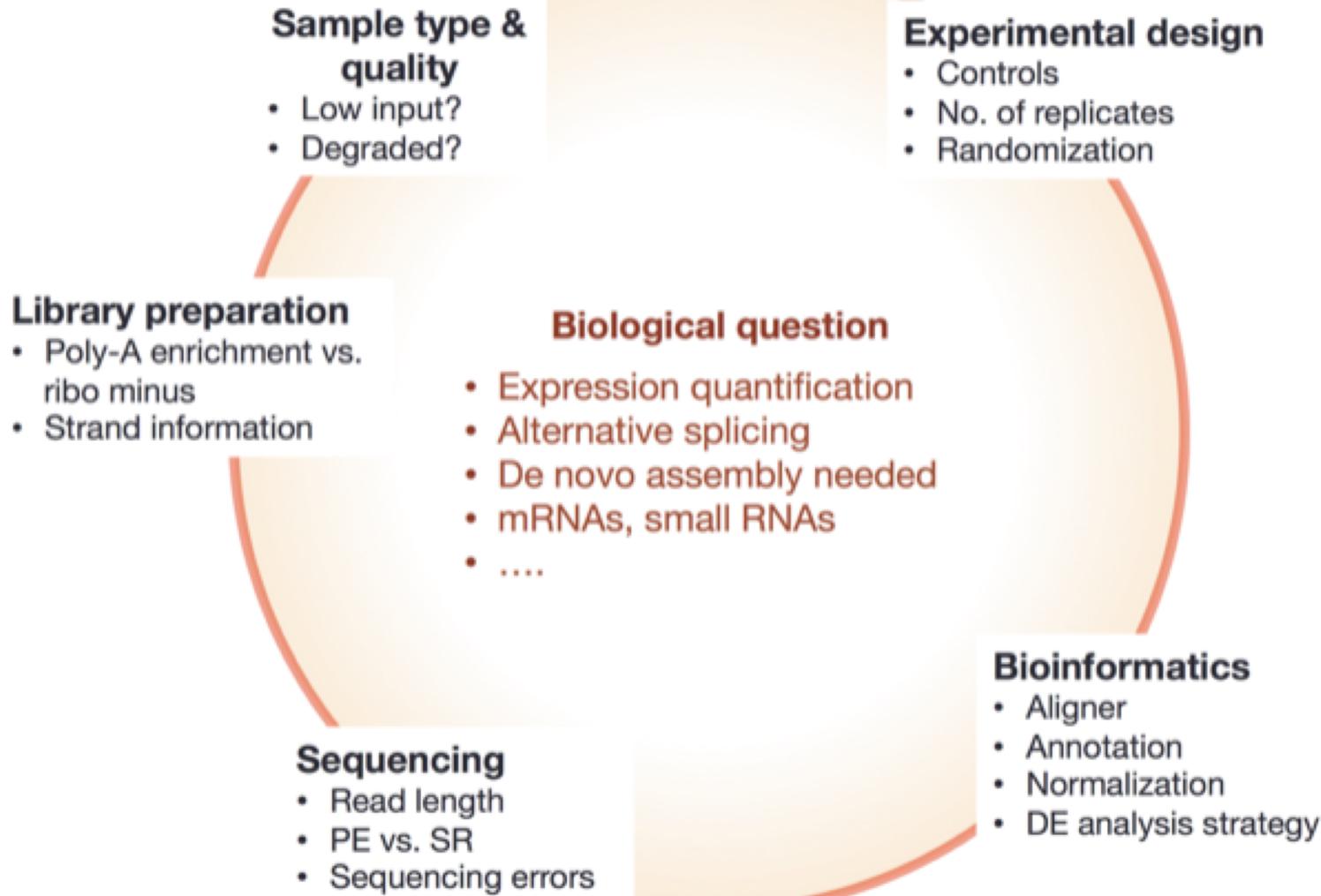
Alternative splicing

Applications of RNA-seq



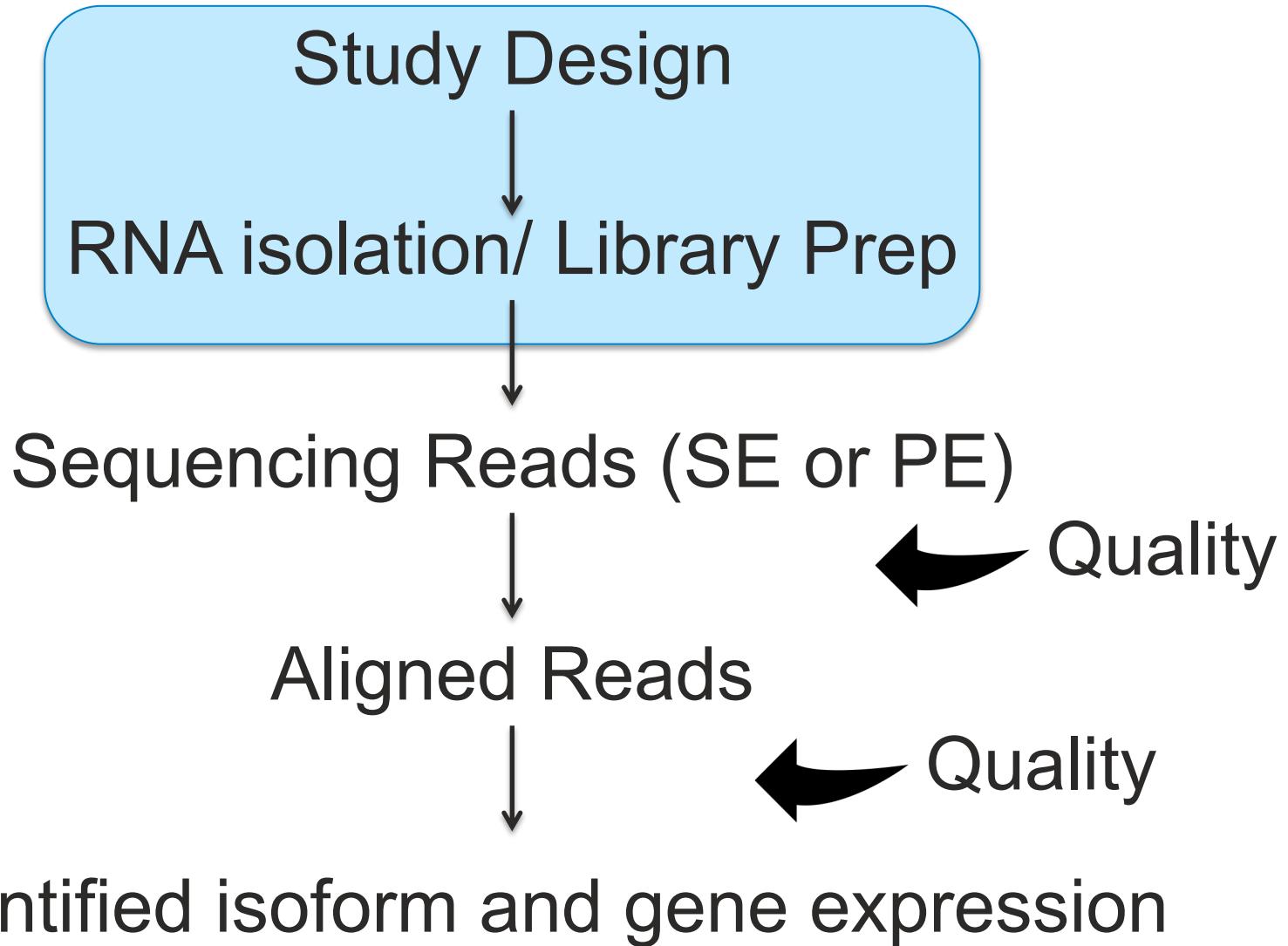
Allele-Specific gene Expression (ASE)

Preferential expression of one allele over the other.

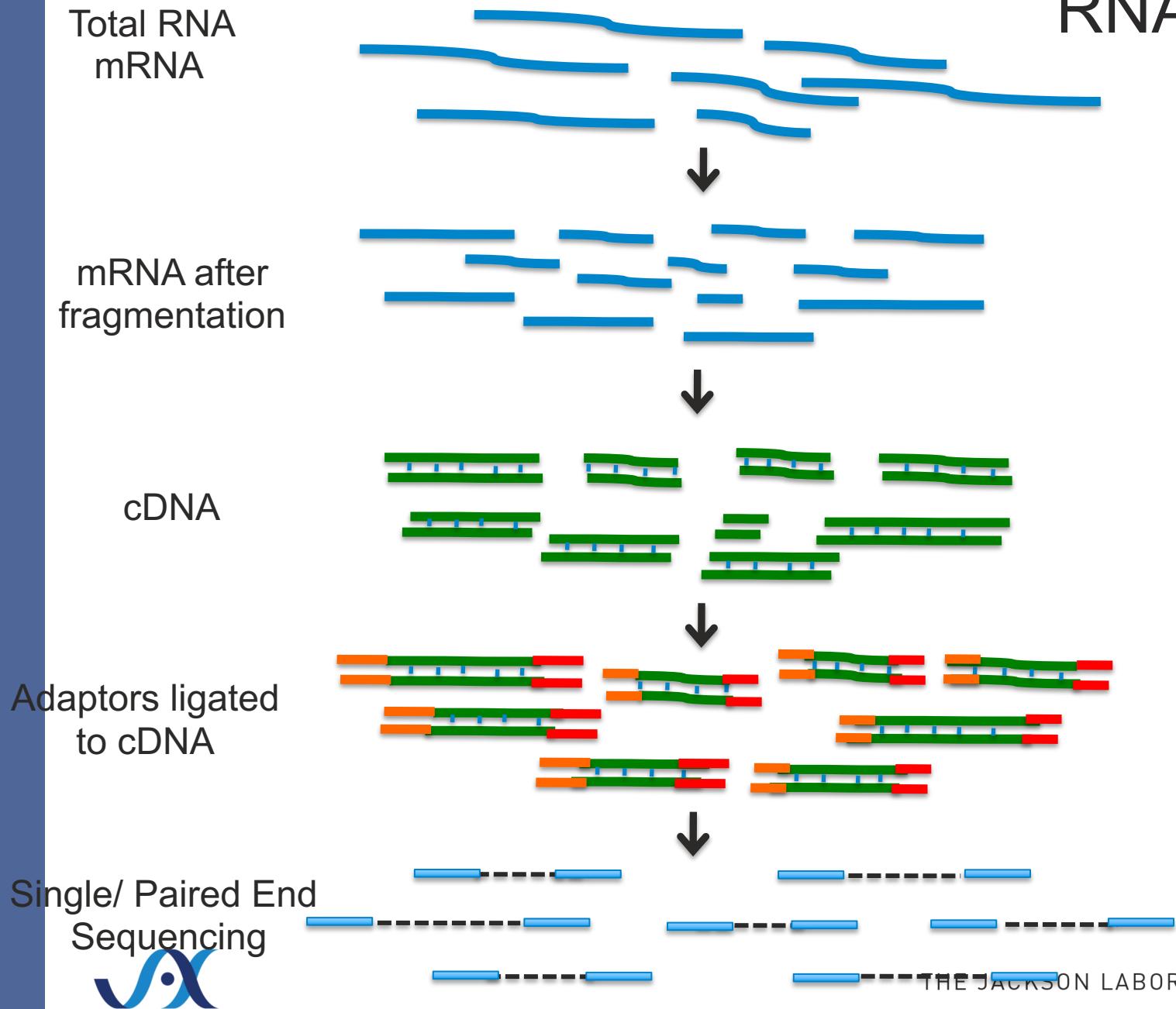


Please order the terms on the circle?

RNA-seq Work Flow



RNA-Seq

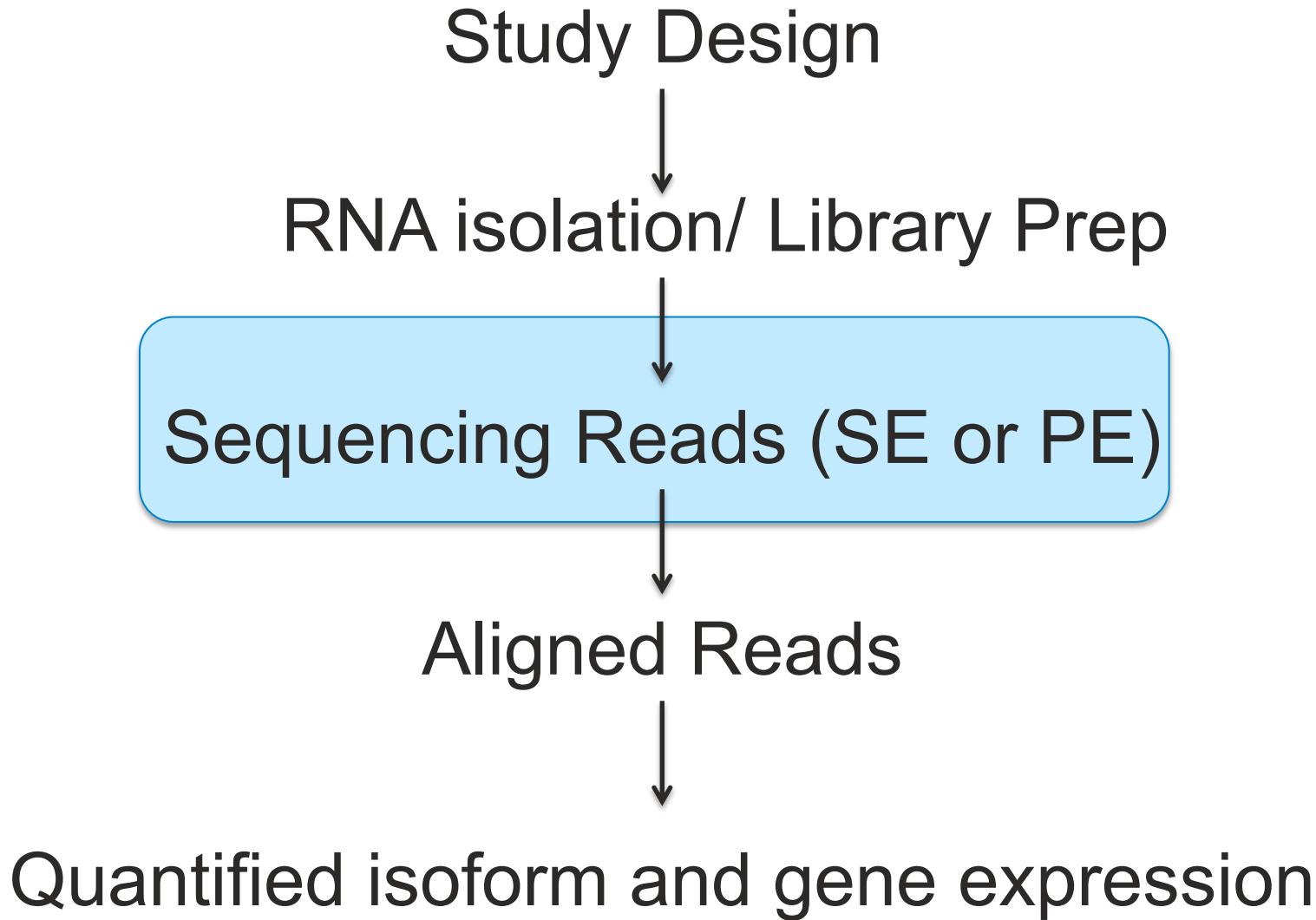


Know your experiments

- 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](#))



RNA-seq Work Flow



Millions and millions of reads...

Fastq format

```
@HISEQ2000_0074:8:1101:7544:2225#TAGCTT/1
TCACCCGTAAGGTAACAAACCGAAAGTATCAAAGCTAAAAGAAGTGGACGACGTGCTTGGTG
GAGCAGCTGCATG
+
CCCFHHHHDDHHJJJJJJJJJJ?FGIIJJJJJJJJFHIJJJIJHHHFFFFD>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



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

TCACCCGTAAGGTAACAAACCGAAAGTATCCAAAGCTAAAAGAAGTGGAGAGCAGCTGCATG

+

CCCFFFFFFHHHHDHHJJJJJJJJJJ?FGIIIIJJJJJJJJFHIJJJIJHHHFFFFD>AC?B??C?ACCAC>BB<<<>C@CCCACCCDCCIJ

Sequenc

Quality value

Quality value can be converted to Phred score

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,

Note: different sequencing platforms have different ways of recording quality score



Steve Munger, 2017

THE JACKSON LABORATORY

Quality Control: How to tell if your data is clean

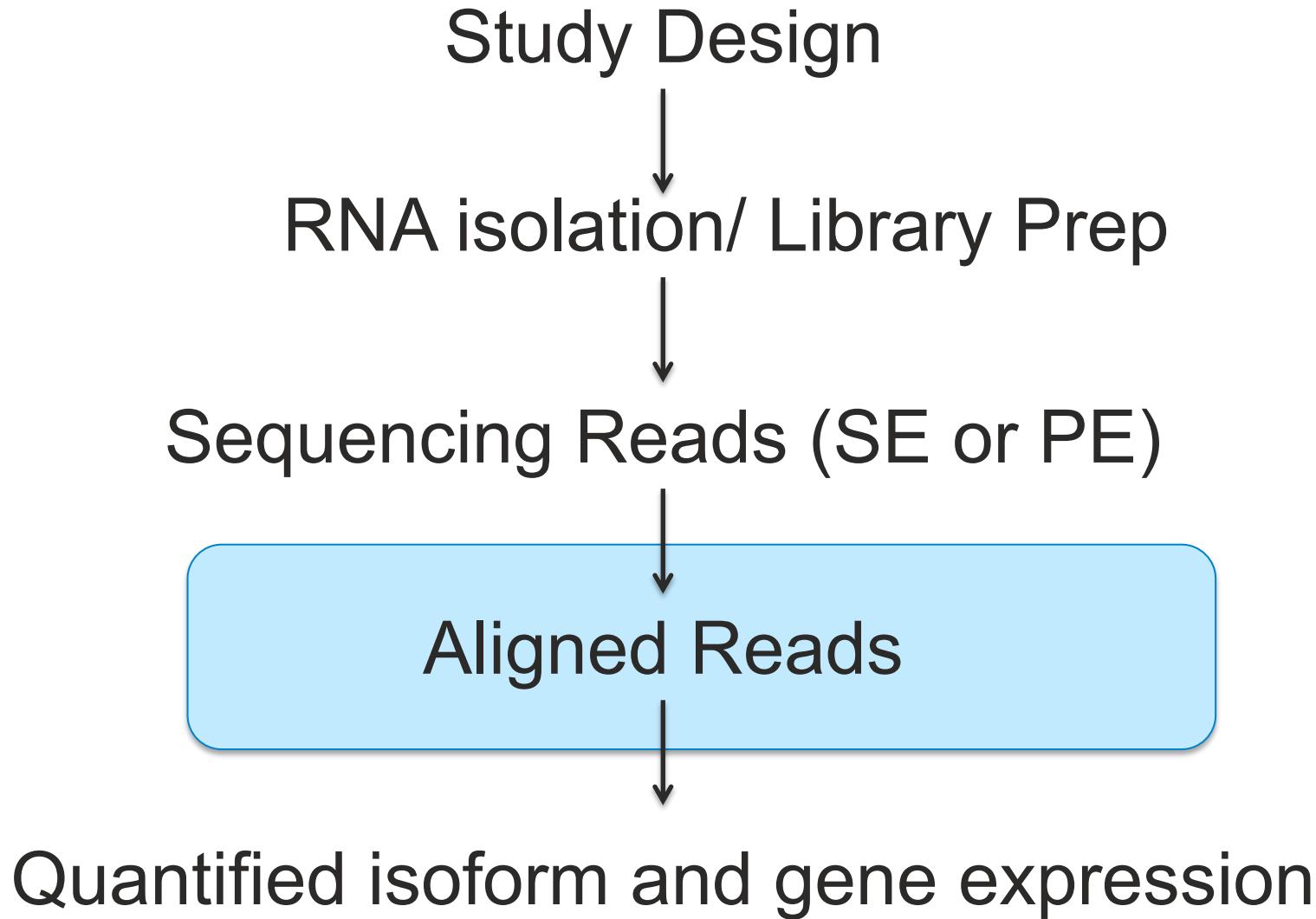
- FASTX-Toolkit
 - http://hannonlab.cshl.edu/fastx_toolkit/
- FastQC
 - <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>



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

100bp Read

ACATGCTGCGGA

Reference sequence

Chr 3



Chr 2



Chr 1



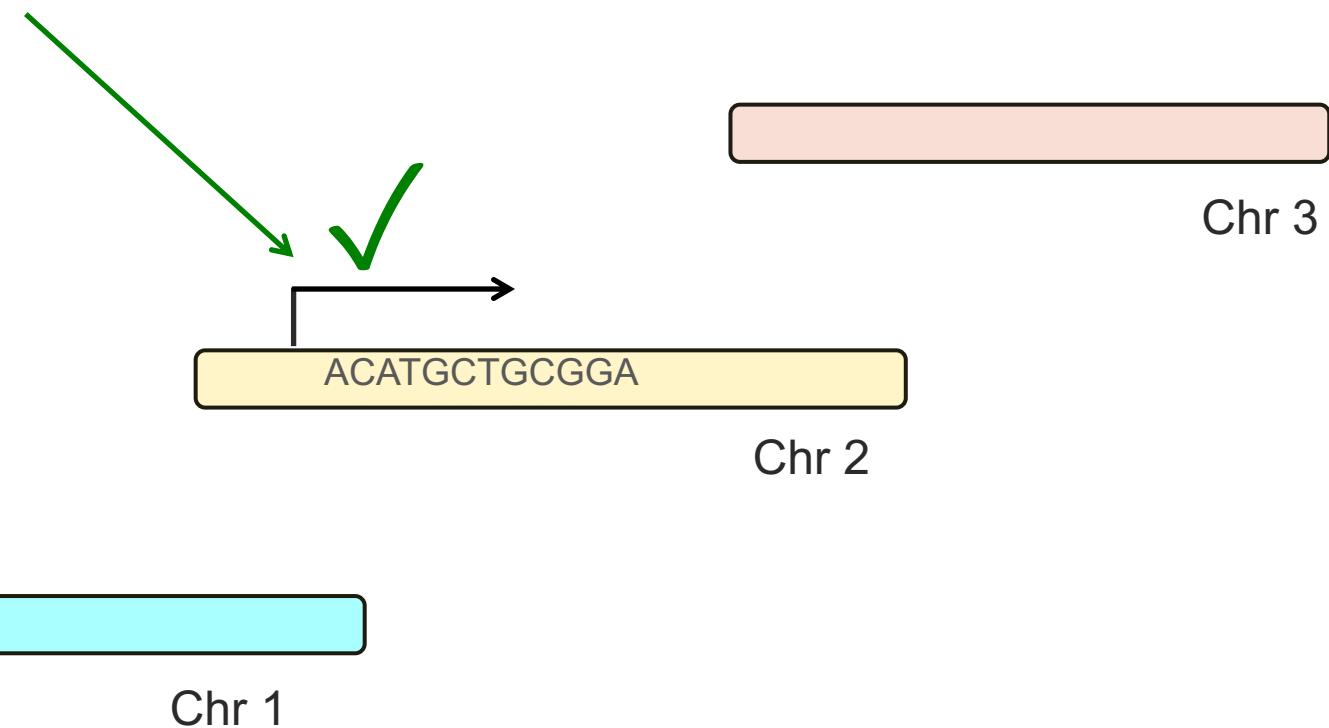
Steve Munger, 2017

THE JACKSON LABORATORY

The perfect read: 1 read = 1 unique alignment.

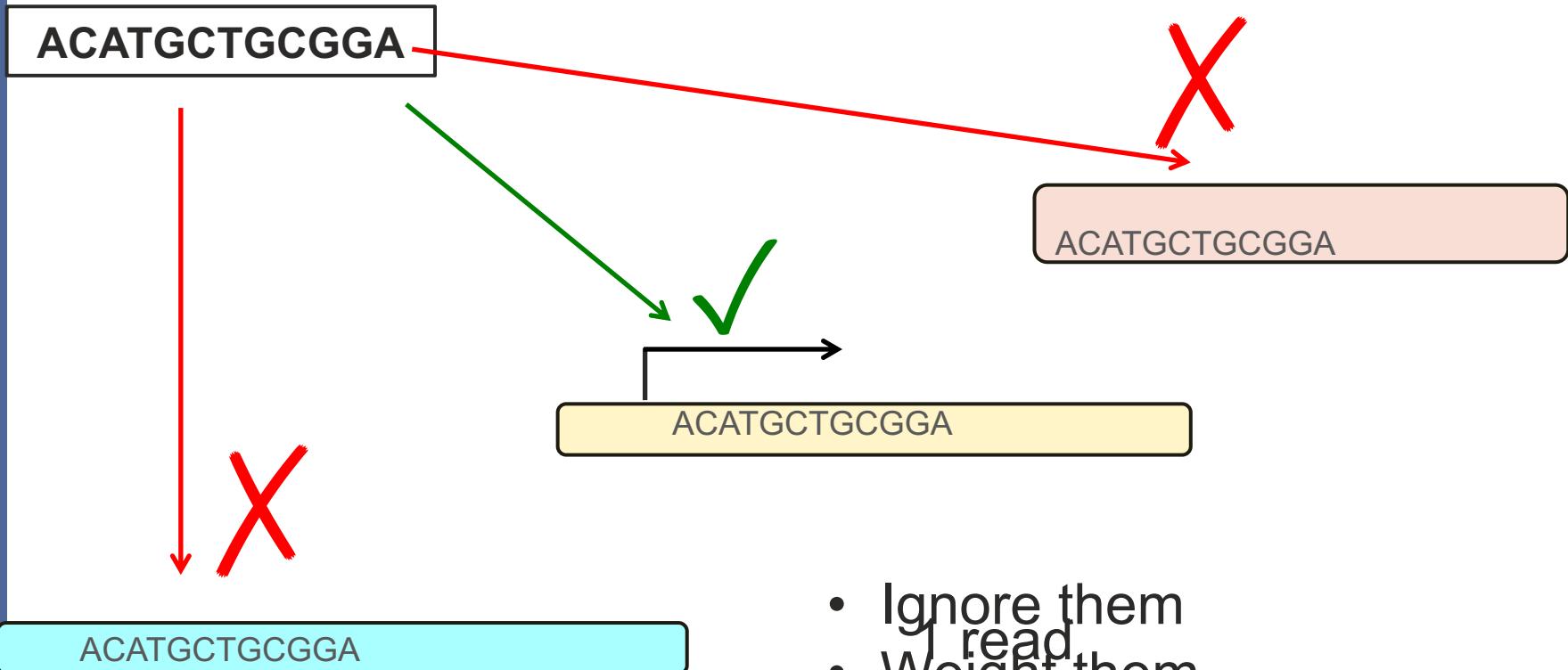
100bp Read

ACATGCTGCGGA



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

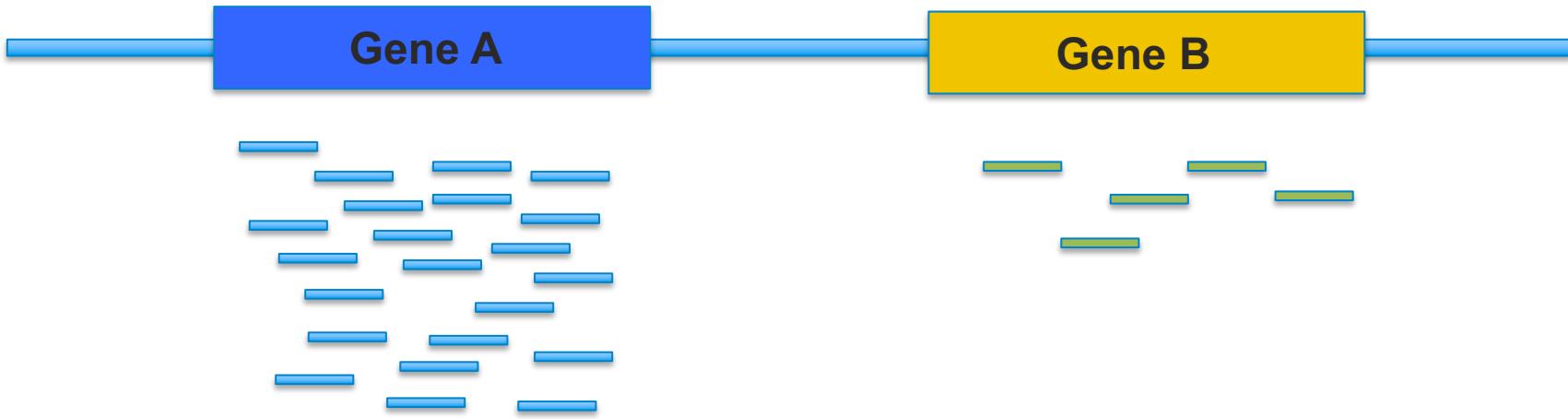
100bp Read



- Ignore them
1 read
- Weight them
3 valid alignments
- Tools: mmquant, MMR,
Only 1 alignment is correct
seqcluster, etc.



Aligning Millions of Short Sequence Reads



Aligners: STAR, HISAT2, TopHat2

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

Tools to handle Bam/Sam files: samtools, bamUtil



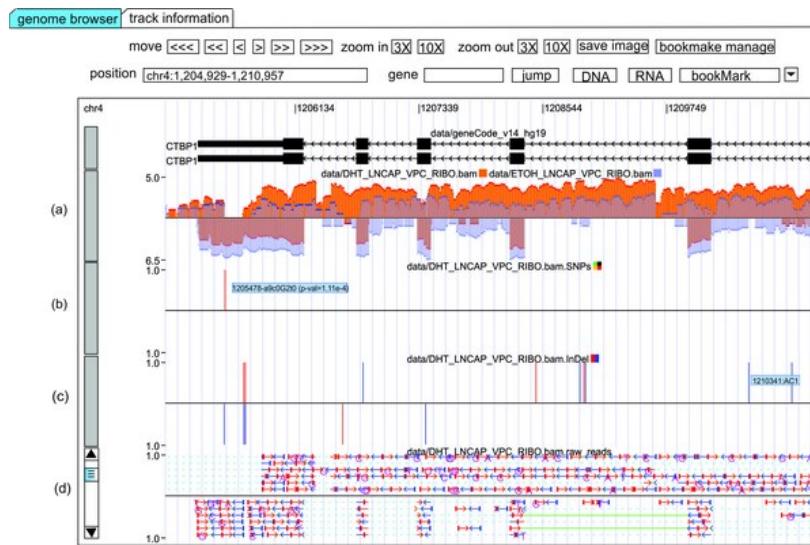
Quality control post alignment

- RNA-SeQC
- <https://github.com/broadinstitute/rnaseqc>
- Output metrics: Mapping Rate, unique rate of mapped, duplicate rate, expression profiling efficiency, coverage, ...



Visualization of alignment data (BAM/SAM)

Genome browsers – IGV and RNASeqBrowser



Integrative Genome Viewer (IGV)

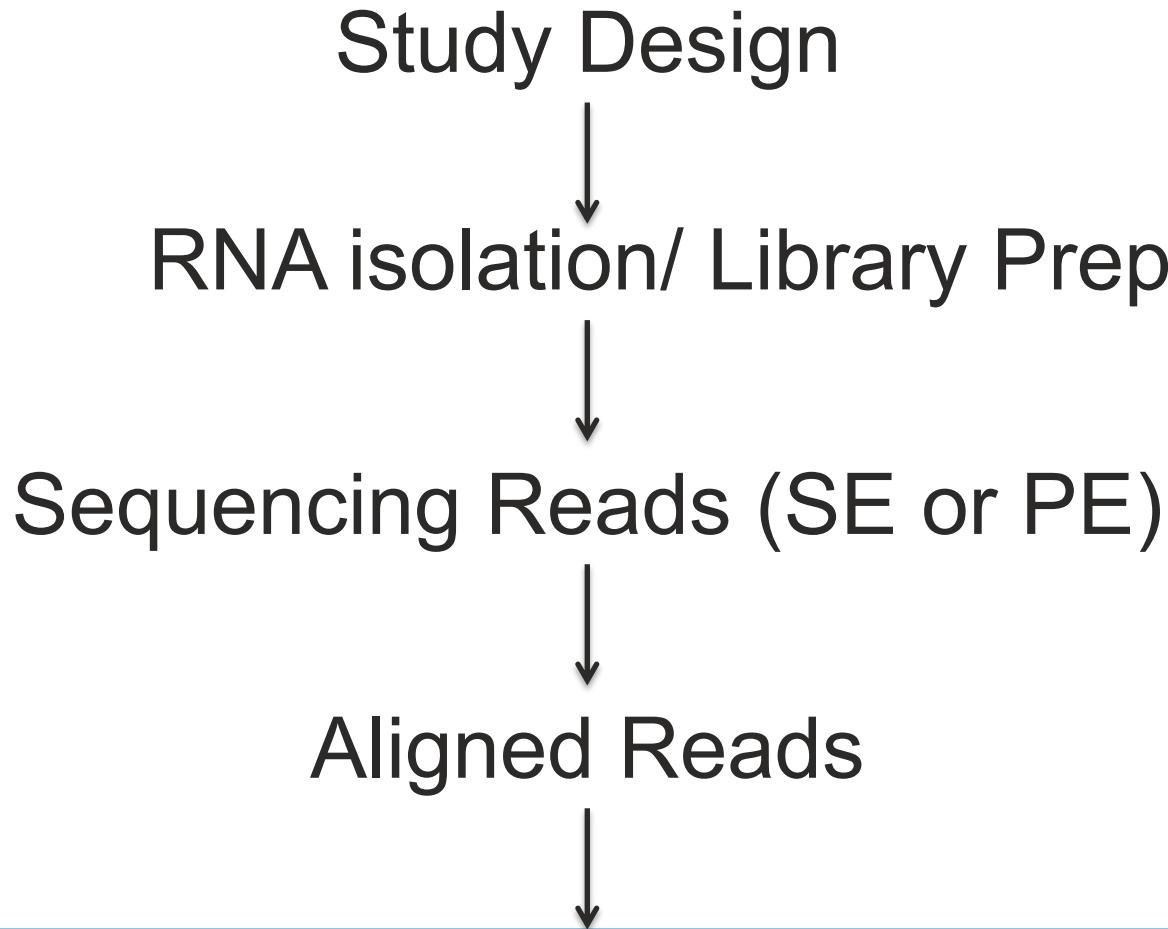
<http://software.broadinstitute.org/software/igv/download>

RNASeqBrowser

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

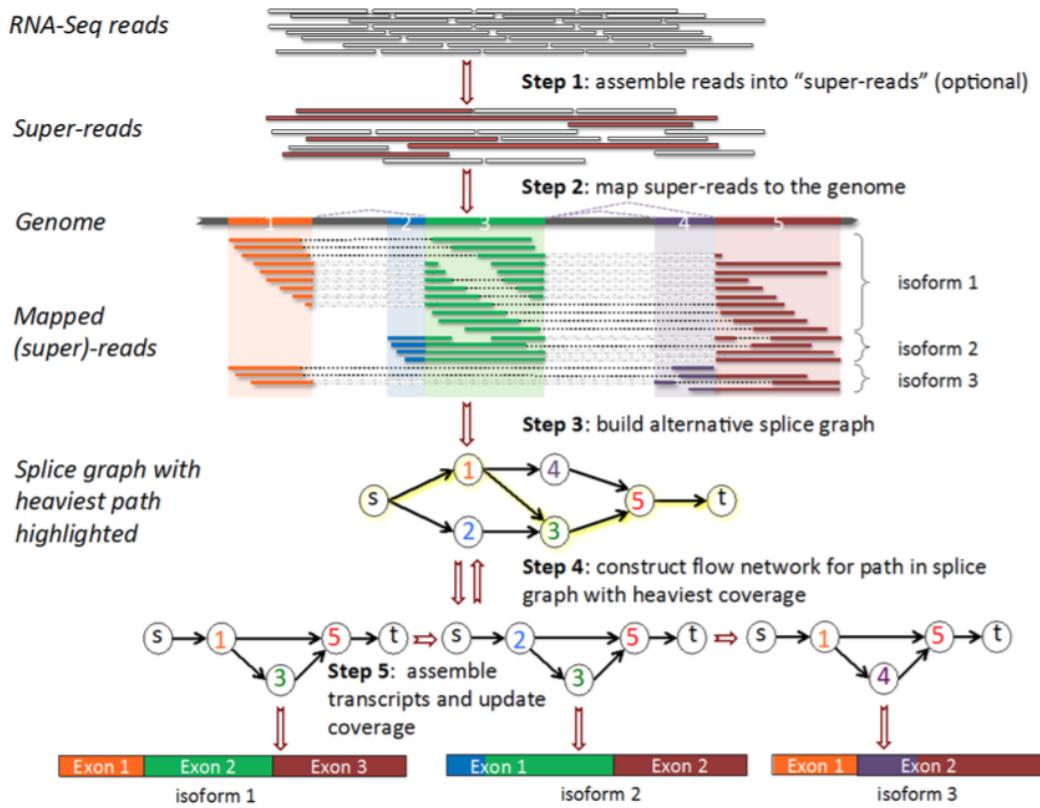


RNA-seq Work Flow



Quantified isoform and gene expression

Transcript quantification

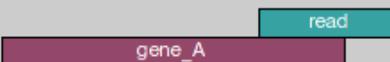
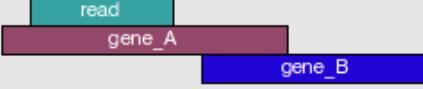
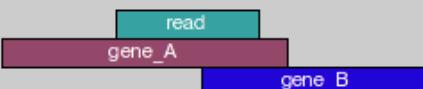


StringTie reconstruct transcripts from spliced read alignments generated by previously mentioned aligners

* 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.

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/>)



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)	<ol style="list-style-type: none">For each gene, count the number of reads mapping to it.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. $RPKM_i = \frac{\text{read count of gene } i}{\left(\frac{\text{length of gene } i}{10^3}\right)\left(\frac{\text{library size}}{10^6}\right)}$	<ul style="list-style-type: none">introduces a bias in the per-gene variances, in particular for lowly expressed genes (Oshlack and Wakefield, 2009)implemented in edgeR's <code>rpk()</code> function
FPKM (fragments per kilobase...)	<ol style="list-style-type: none">Same as RPKM, but for paired-end reads:The number of fragments (defined by two reads each) is used.	<ul style="list-style-type: none">implemented in DESeq2's <code>fpkm()</code> function <p>Good for comparison within one sample but not for cross sample comparison</p>
TPM	<p>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.</p> <ol style="list-style-type: none">For each gene, count the number of reads mapping to it and divide by its length in base pairs (= counts per base).Multiply that value by 1 divided by the sum of all counts per base of every gene.Multiply that number by 10^6.	<ul style="list-style-type: none">details in Wagner et al. (2012)



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

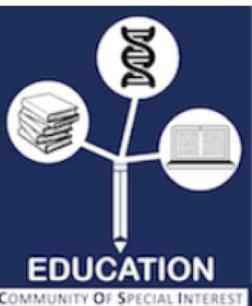
Keep updated

“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/>
- Pipeline tool: Bcbio-nextgen <https://bcbio-nextgen.readthedocs.io/en/latest/>





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 Education 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 which 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 to get information about bioinformatics education, to start discussions on 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 provision of 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 infrastructure 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^{§¶¶11,2}, 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



THE JACKSON LABORATORY

Background

Pax6

MGI

Keywords, Symbols, or IDs

About Help FAQ

Home Genes Phenotypes Human Disease Expression Recombinases Function Strains / SNPs Homology P

Search Download More Resources Submit Data Find Mice (IMSR) Analysis Tools Contact Us Browsers

Pax6 Gene Detail

?

Summary	Symbol Pax6 Name paired box 6 Synonyms 1500038E17Rik, AEY11, Dey, Dickie's small eye, Gsfaey11, Pax-6	Feature Type protein coding gene IDs MGI:97490 NCBI Gene: 18508 Gene Overview MyGene.info: PAX6 Alliance gene page														
Location & Maps	more ▶ Sequence Map Chr2:105668900-105697364 bp, + strand	Genetic Map Chromosome 2, 55.31 cM														
Homology	more ▶ Human Ortholog PAX6, paired box 6	Vertebrate Orthologs 9														
Human Diseases	less ▽ Diseases 4 with Pax6 mouse models; 4 with human PAX6 associations <table border="1"><tr><th>Human Disease</th><th>Mouse Models</th></tr><tr><td>aniridia</td><td>IDs View 2 models</td></tr><tr><td>Peters anomaly</td><td>IDs View 5 models</td></tr><tr><td>cataract</td><td>IDs View 1 model</td></tr><tr><td>juvenile glaucoma</td><td>IDs View 1 model</td></tr><tr><td>coloboma of optic nerve</td><td>IDs View 1 "NOT" model</td></tr><tr><td>WAGR syndrome</td><td>IDs View 1 "NOT" model</td></tr></table> Click on a disease name to see all genes associated with that disease.	Human Disease	Mouse Models	aniridia	IDs View 2 models	Peters anomaly	IDs View 5 models	cataract	IDs View 1 model	juvenile glaucoma	IDs View 1 model	coloboma of optic nerve	IDs View 1 "NOT" model	WAGR syndrome	IDs View 1 "NOT" model	References 6 with disease annotations
Human Disease	Mouse Models															
aniridia	IDs View 2 models															
Peters anomaly	IDs View 5 models															
cataract	IDs View 1 model															
juvenile glaucoma	IDs View 1 model															
coloboma of optic nerve	IDs View 1 "NOT" model															
WAGR syndrome	IDs View 1 "NOT" model															
Mutations, Alleles, and Phenotypes	less ▽ Phenotype Summary 179 phenotypes from 42 alleles in 46 genetic backgrounds 44 phenotypes from multigenic genotypes 10 images 291 phenotype references Phenotype Overview ? Click cells to view annotations;	All Mutations and Alleles 59 Chemically and radiation induced 3 Chemically induced (ENU) 23 Chemically induced (other) 2 Gene trapped 7 Radiation induced 4 Spontaneous 4 Targeted 12 Transgenic 4 Genomic Mutations 6 involving Pax6 Incidental Mutations Mutagenetix , APF Find Mice (IMSR) 67 strains or lines available Comparison Matrix Gene Expression + Phenotype Recombinase Activity 2														



Data fact

[Data](#)

 ArrayExpress

Search Examples: E-MEXP-31, cancer, p53, Geuvadis [advanced search](#)

Contact Us | [Login](#)

Home | Browse | Submit | Help | About ArrayExpress

ARRAYEXPRESS / BROWSE / E-MTAB-5708

E-MTAB-5708 - RNA-Seq of pancreatic islets from beta cell-specific Pax6 knockout mice

Status	<i>Submitted on 8 February 2017, last updated on 26 April 2017, released on 26 April 2017</i>
Organism	Mus musculus
Samples (9)	Click for detailed sample information and links to data
Protocols (5)	Click for detailed protocol information
Description	RNA-Seq to investigate significant transcriptional differences underlying the defective glucose-stimulated insulin secretion of Pax6 knockout mice in comparison to floxed littermate controls.
Experiment types	RNA-seq of coding RNA, genetic modification design
Contacts	✉ Timothy Pullen <t.pullen@imperial.ac.uk> , ✉ Guy Rutter <g.rutter@imperial.ac.uk>
Citation	The transcription factor Pax6 is required for pancreatic β cell identity, glucose-regulated ATP synthesis and Ca ²⁺ dynamics in adult mice. Mitchell RK, Nguyen-Tu MS, Chabosseau P, Callingham RM, Pullen TJ, Cheung R, Leclerc I, Hodson DJ, Rutter GA. , PMID:28377501
MINSEQE	    
	Exp. design Protocols Variables Processed Seq. reads
Files	Investigation description  E-MTAB-5708.idf.txt Sample and data relationship  E-MTAB-5708.sdrf.txt Click to browse all available files
Links	Expression Atlas - E-MTAB-5708 ENA - ERP022747 Send E-MTAB-5708 data to  GENOME SPACE

- Poly-A selected
- Reverse stranded

THE JACKSON LABORATORY



Open up galaxy ...