RNA-Sequencing

BOHTA 2019

Kristoffer Vitting-Seerup

ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGACGTACGTACGTACGTACGTACGTACGTACGTCGTAC

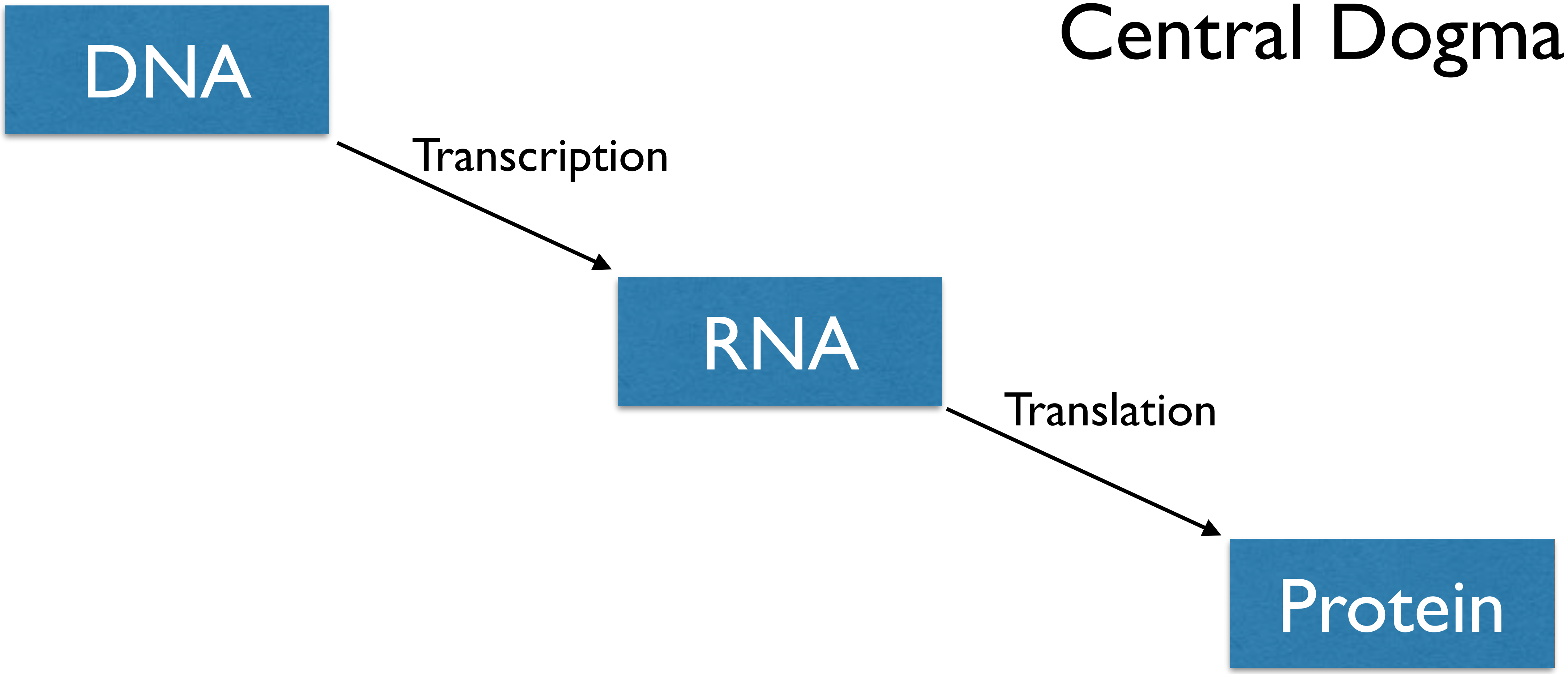
Agenda

1. Introduction to RNA-seq
2. RNA-seq workflow
    1. Do-it-yourself exercise
3. Isoform Switch Analysis
    1. Do-it-yourself exercise
4. Perspective

Agenda

1. **Introduction to RNA-seq**
2. RNA-seq workflow
    1. Do-it-yourself exercise
3. Isoform Switch Analysis
    1. Do-it-yourself exercise
4. Perspective

What is RNA



Central Dogma

DNA

Transcription

RNA

Translation

Protein

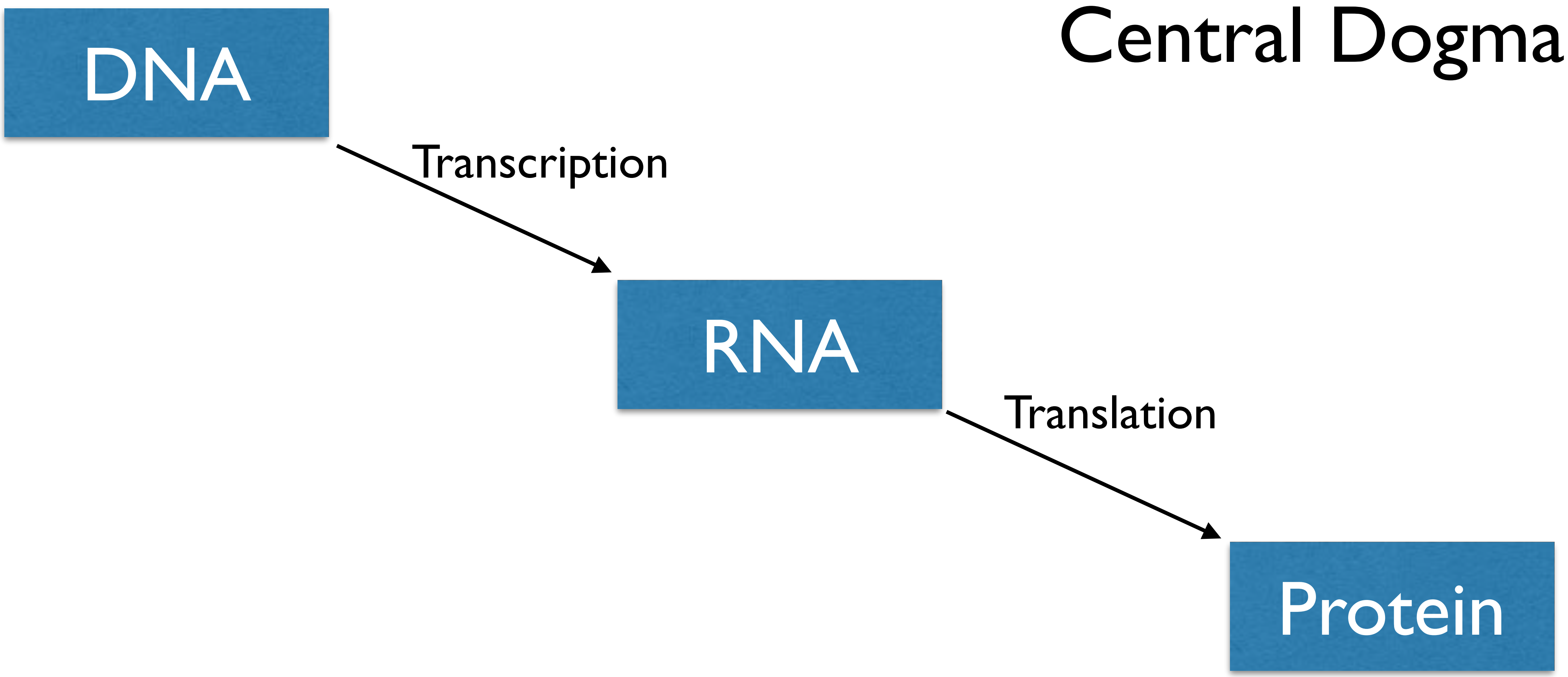
All cells in an
organism have the same DNA

Exercise:

- 3 minutes with neighbour:

  Why sequence RNA at all?

What is RNA



Central Dogma

DNA

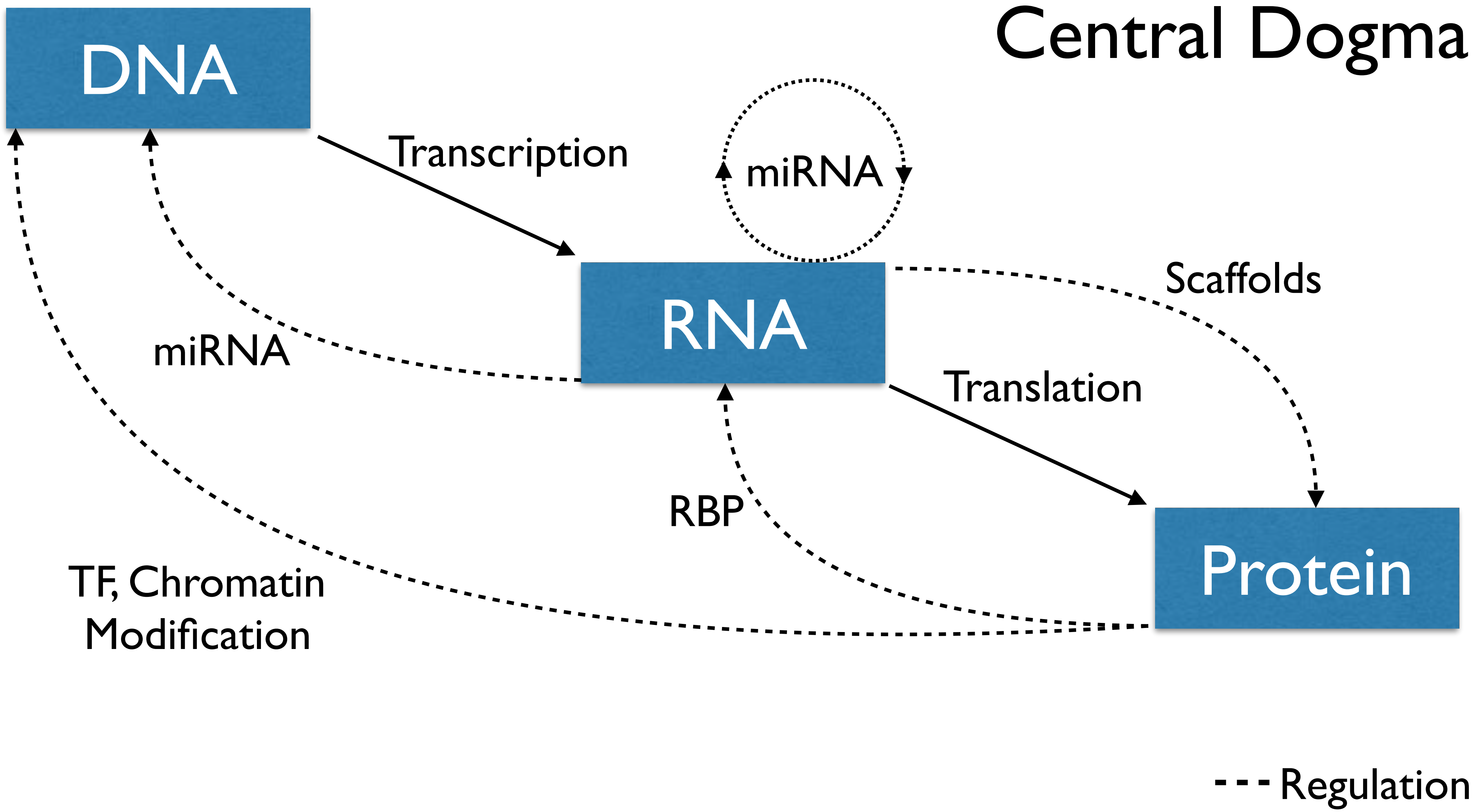
Transcription

RNA

Translation

Protein

What is RNA

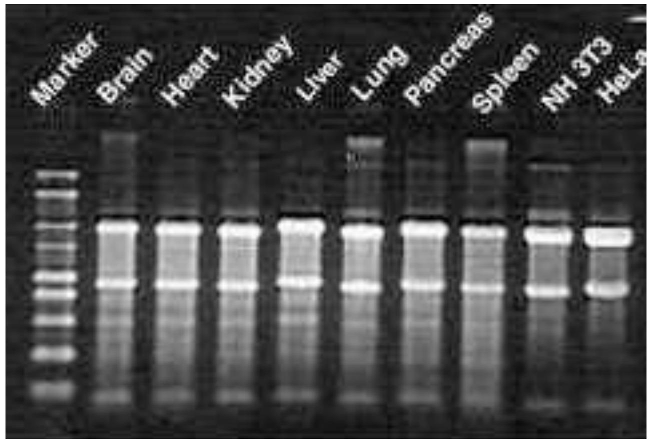

Central Dogma
DNA
Transcription
miRNA
RNA
Scaffolds
miRNA
Translation
RBP
Protein
TF, Chromatin Modification
- - - Regulation

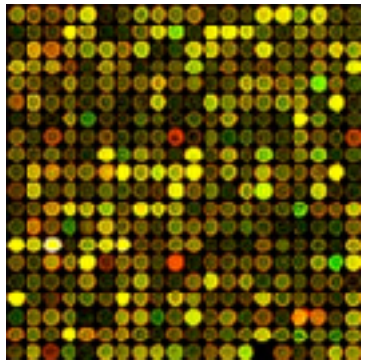
RNA-sequencing

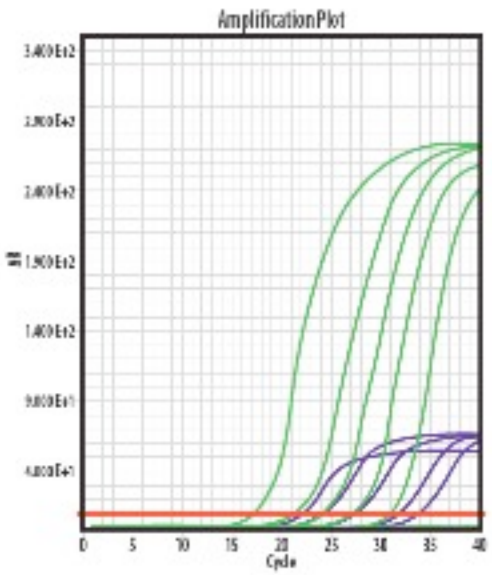
- Sequencing of purified RNA
- Called RNA-seq, whole cell sequencing, Next generation sequencing etc.
- A method to characterise, qualitatively and quantitatively, a RNA population in a sample
- More importantly samples can be compared!
- Furthermore these RNA-populations can be quite specific

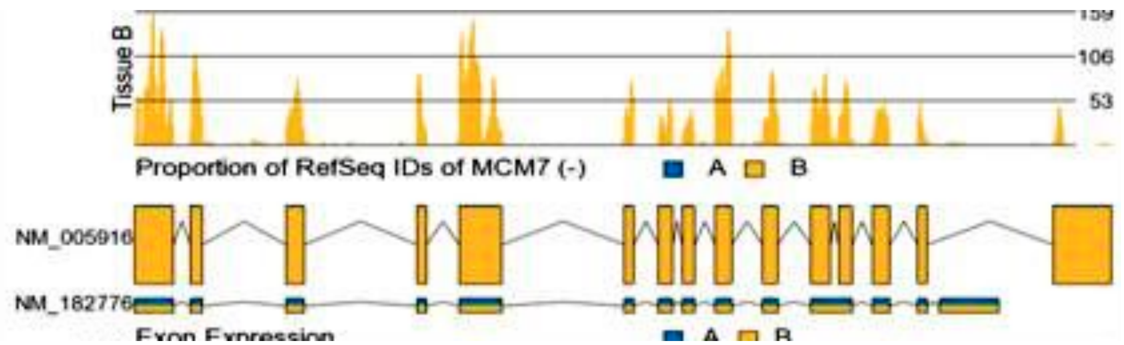
History of RNA-analysis

- 1977: Northern Blot (low sensitivity, low throughput, hard to quantify)
- 1977 Sanger sequencing highly accurate - low throughput - not quantitative - expensive
- 1987: Microarray (high-throuput, low cost, low dynamic range, low specificity)
- 1997: qPCR (high dynamic range, low throughput)
- 2005: 5’ RNA-seq (high spec., high dynamic range, high specificity)
- 2009: Paired-end sequencing (high spec., high dynamic range, high specificity)
- 2013: Single cell RNA-seq
- 2014: Long read RNA-seq







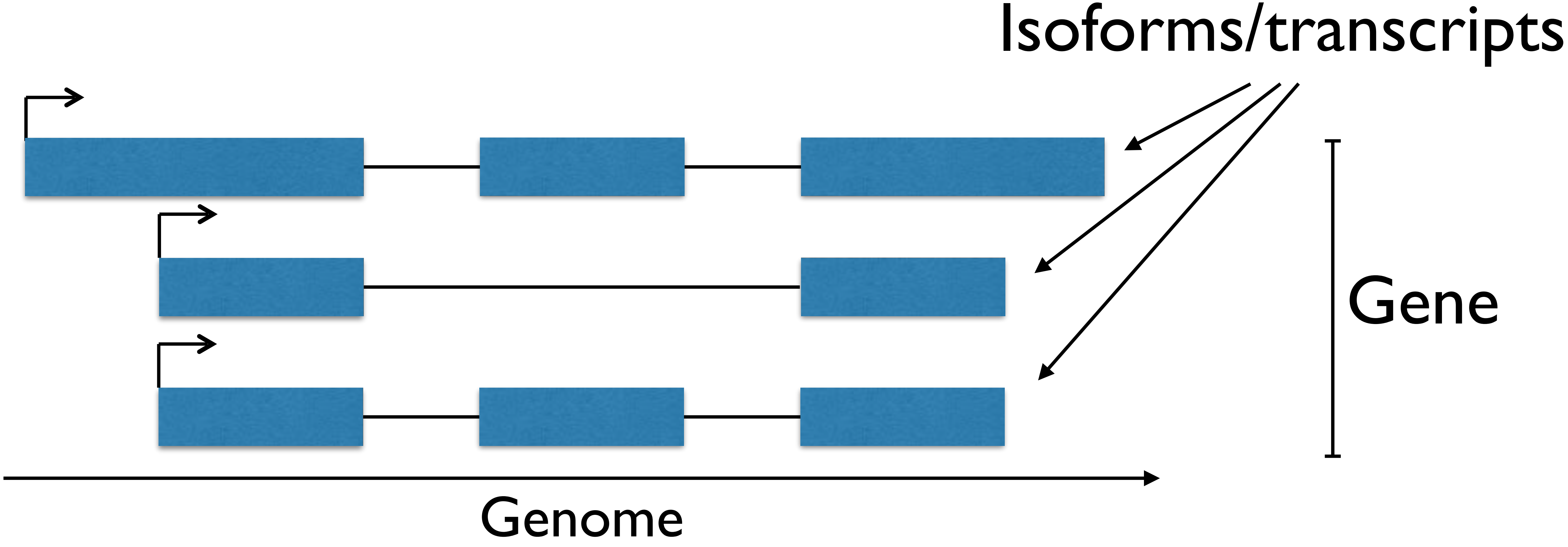


Gene vs Isoform

- It is quite hard to define a gene because you can always find biological exceptions to rules
- One suggestion, that will be used here, is that a gene is a loci from which one or more transcripts originate (strand specific). Furthermore these transcripts should share some exon information

Gene vs Isoform

The terms “transcript” and “isoform” is here used interchangeably



Exercise:

- 5 minutes with neighbour:

  What do you gain by profiling the transcriptome with isoform resolution (compared to gene resolution)?

Agenda

1. Introduction to RNA-seq
2. **RNA-seq workflow**
    1. Do-it-yourself exercise
3. Isoform Switch Analysis
    1. Do-it-yourself exercise
4. Perspective

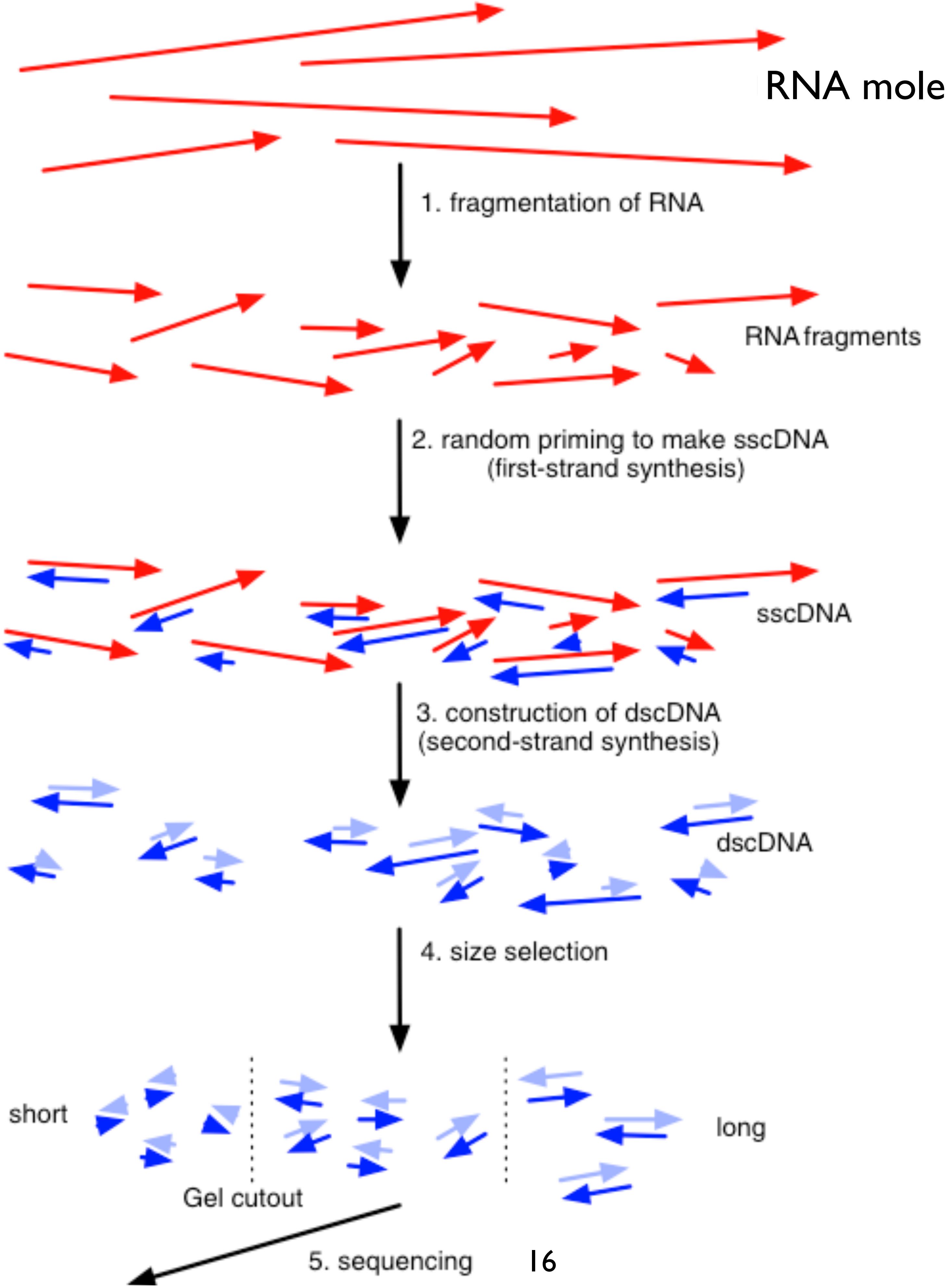
Conceptual Overview

1. Experiment -> RNA-Seq libraries (lab-work)
2. Sequencing (company)
3. Data analysis (you)
    - A. QC and Trimming
    - B. Mapping
    - C. Quantification
    - D. Post analysis

Conceptual Overview

1. **Experiment -> RNA-Seq libraries (lab-work)**
2. Sequencing (company)
3. Data analysis (you)
    - A. QC and Trimming
    - B. Mapping
    - C. Quantification
    - D. Post analysis

Experiment Design:

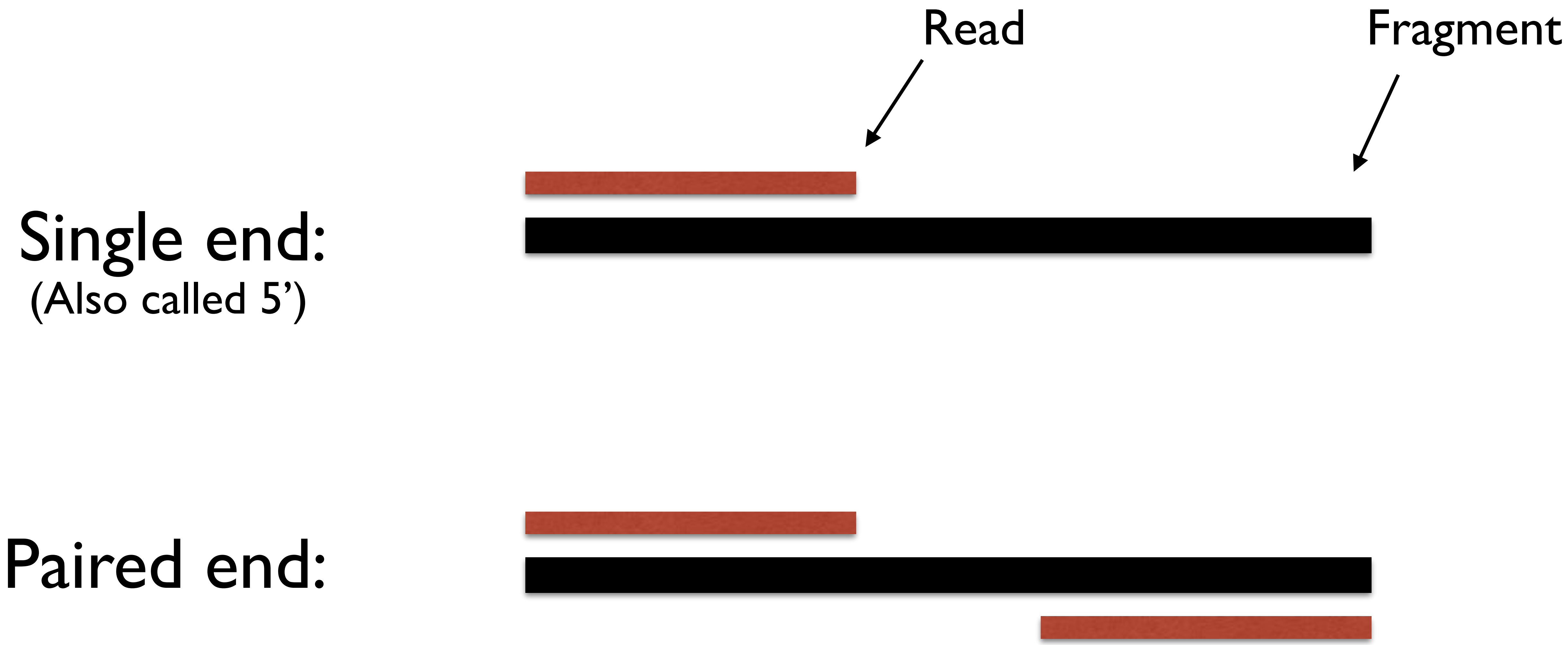


Note:
Random primers - they do not cause biases

Conceptual Overview

1. Experiment -> RNA-Seq libraries (lab-work)
2. **Sequencing (company or institution)**
3. Data analysis (you)
    A. QC and Trimming
    B. Mapping
    C. Quantification
    D. Post analysis

Sequencing:
Single vs paired end



Read

Fragment

Single end:

(Also called 5')

Paired end:

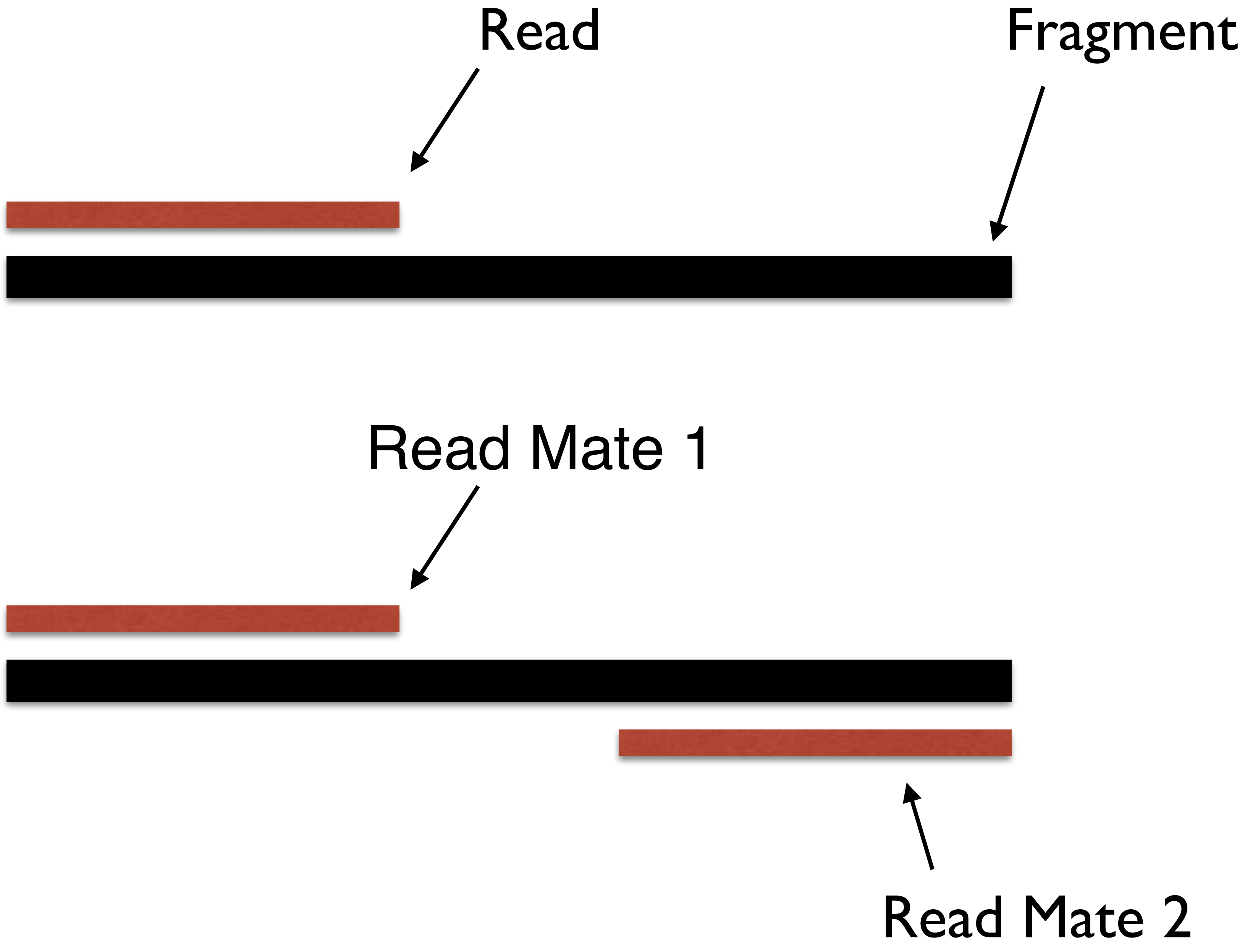
Exercise

- 3 minutes with neighbour:

  Why are paired en sequencing (mostly) preferred to single end sequencing?

  Hint 1: Does reads map uniquely?
  Hint 2: Think about the transcript structure

Some Terminology



Single end:
(Also called 5’)

Read

Fragment

Paired end:

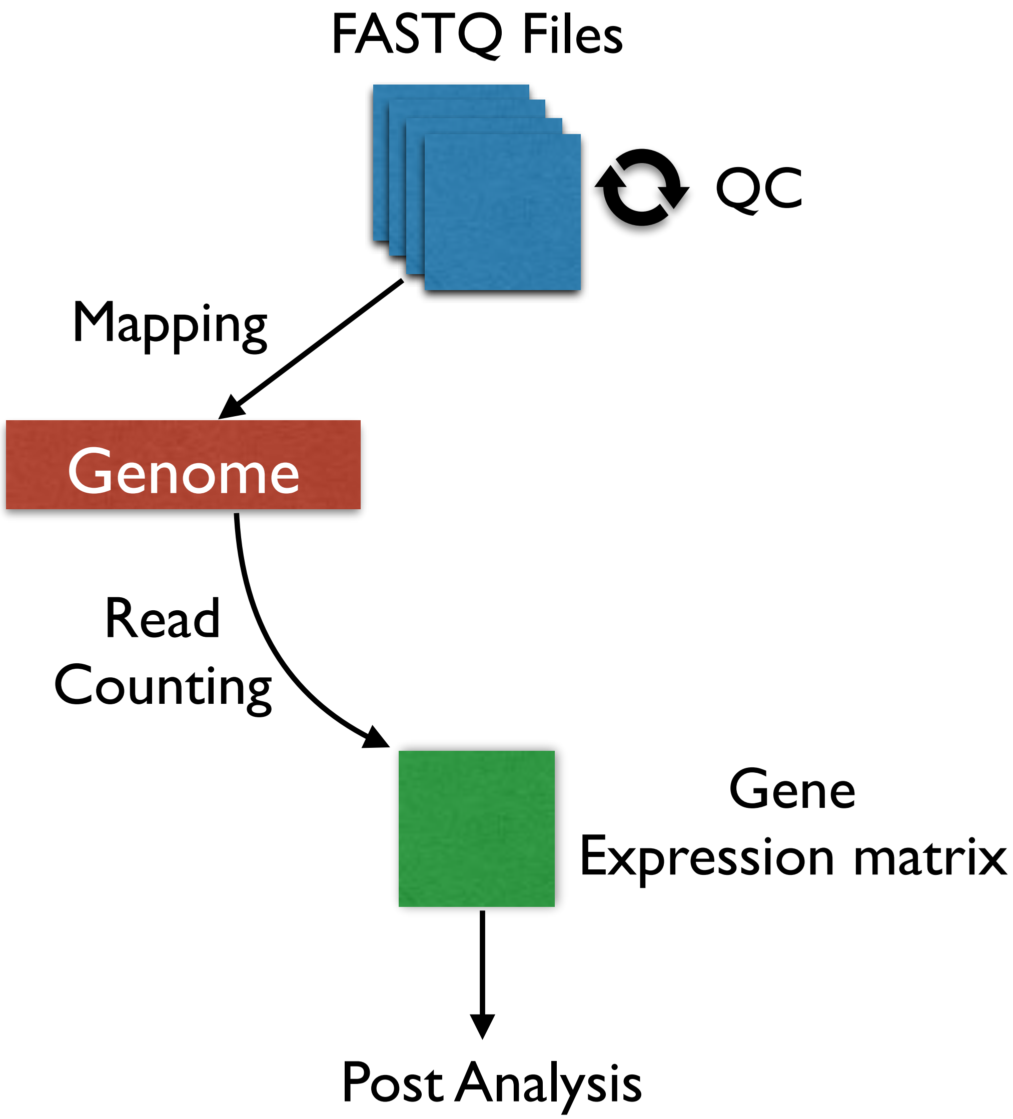
Read Mate 1

Read Mate 2

Goal: Quantify number of Fragments

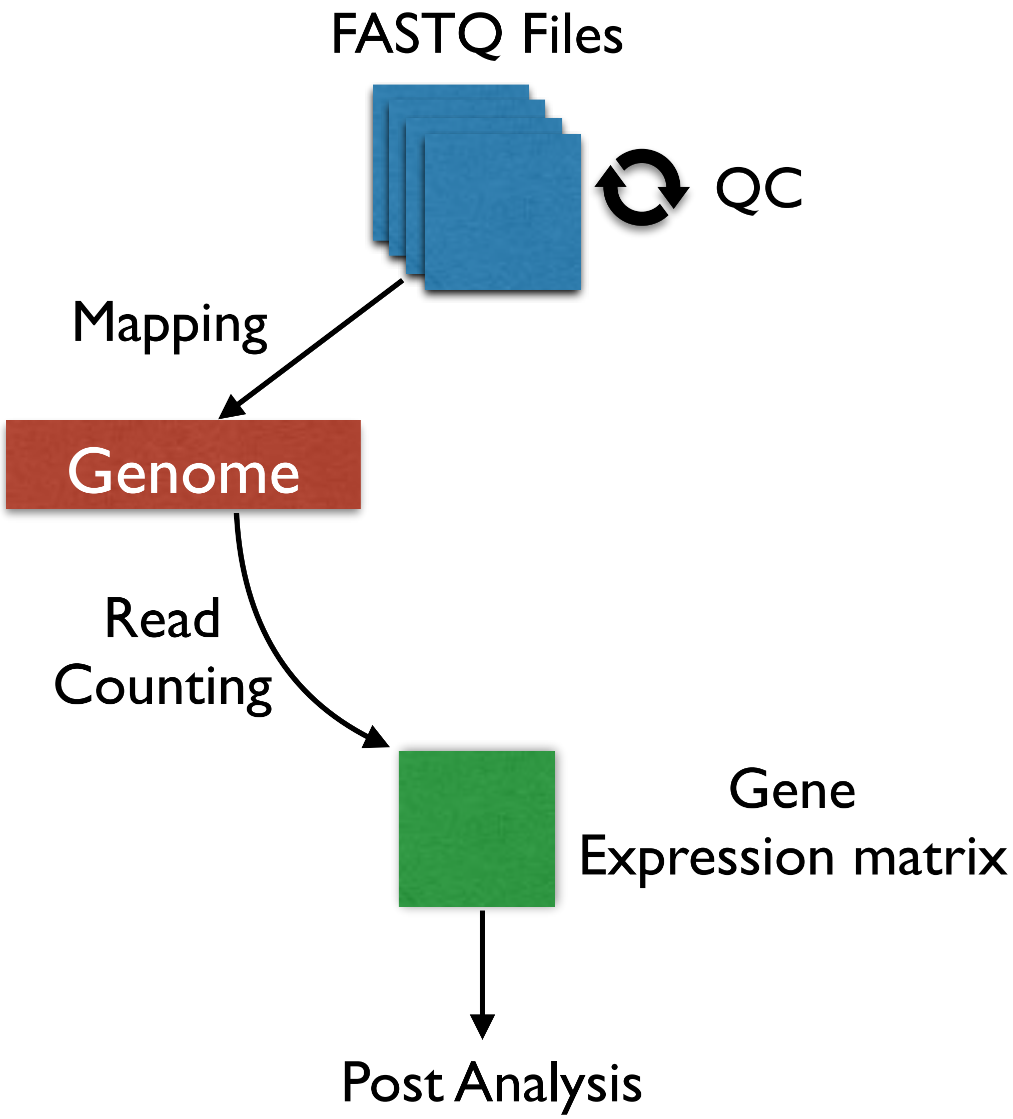
Conceptual Overview

1. Experiment -> RNA-Seq libraries (lab-work)
2. Sequencing (company or institution)
3. **Data analysis (you)**
    A. QC and Trimming
    B. Mapping
    C. Quantification
    D. Post analysis



Conceptual Overview

1. Experiment -> RNA-Seq libraries (lab-work)
2. Sequencing (company)
3. Data analysis (you)
    A. **QC and Trimming**
    B. Mapping
    C. Quantification
    D. Post analysis



Tool: FastQC

- Fast and comprehensive quality control of FASTQ files
- The one we already told you about
  (might accidentally have been called FastX QC)
- Links:
  - tool and examples of good and poor quality :

    https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

  - Manual:

    https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

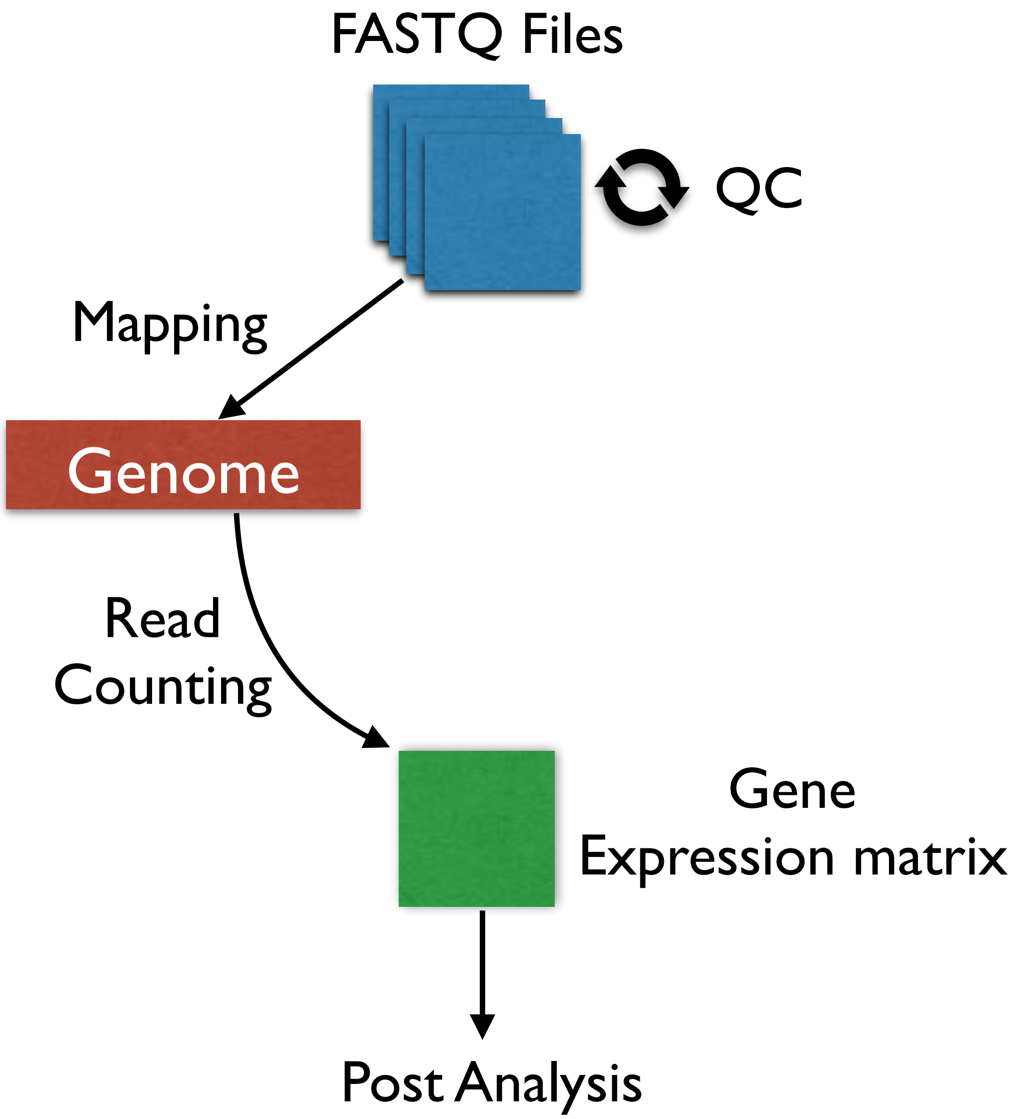
Recap

- 3 min with neighbour:

  Why can it a good idea to perform quality trimming before mapping RNA-seq reads to the genome?

Conceptual Overview

1. Experiment -> RNA-Seq libraries (lab-work)
2. Sequencing (company)
3. Data analysis (you)
    - A. QC and Trimming
    - **B. Mapping**
    - C. Quantification
    - D. Post analysis



Mapping

TCGGCGATTCAGTCTCAGAATCGA

Read

TTCAGTCTCAGAATCGAGATACGATTACGATATCGAGATACGATCGGCGATTCAGTCTCAGAATCGAGATACAGAGCGA

Genome

Mapping

Read

TCGGCGATTCAGTCTCAGAATCGA

TTCAGTCTCAGAATCGAGATACGATTACGATATCGAGATACGATCGGCGATTCAGTCTCAGAATCGAGATACAGAGCGA

Genome

Individual basepair matching

Mapping

Read

TCGGCGATTCAGTCTCAGAATCGA

TTCAGTCTCAGAATCGAGATACGATTACGATATCGAGATACGATCGGCGATTCAGTCTCAGAATCGAGATACAGAGCGA

Genome

Individual basepair matching

Mappers

Naturally modern algorithms are a lot smarter than that:

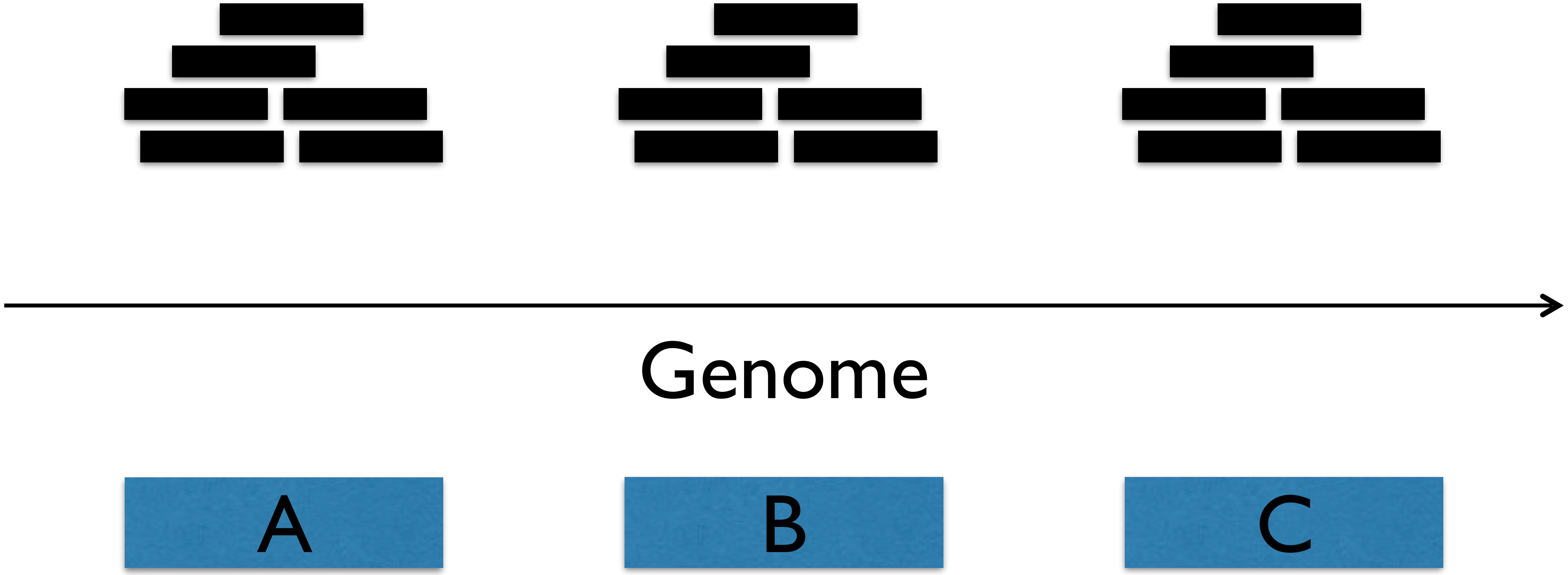
- Clever genome indexing
- Allows for mismatches
- Consider quality score
- Consider position in read
- Considers read pairs
- Etc

Aligned Reads

- Reads can be divided into 4:

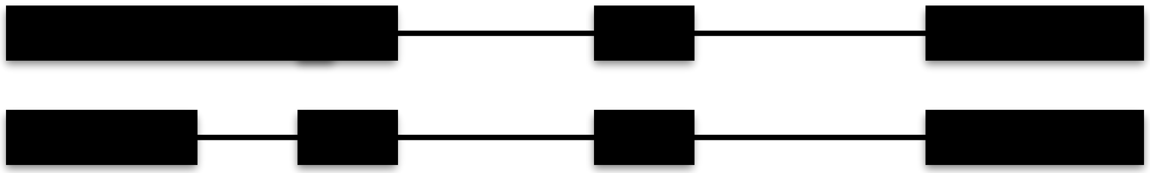
  1. Reads not mapping
  2. Reads mapping uniquely
  3. Multi-mappers

Uniquely Mapped Reads

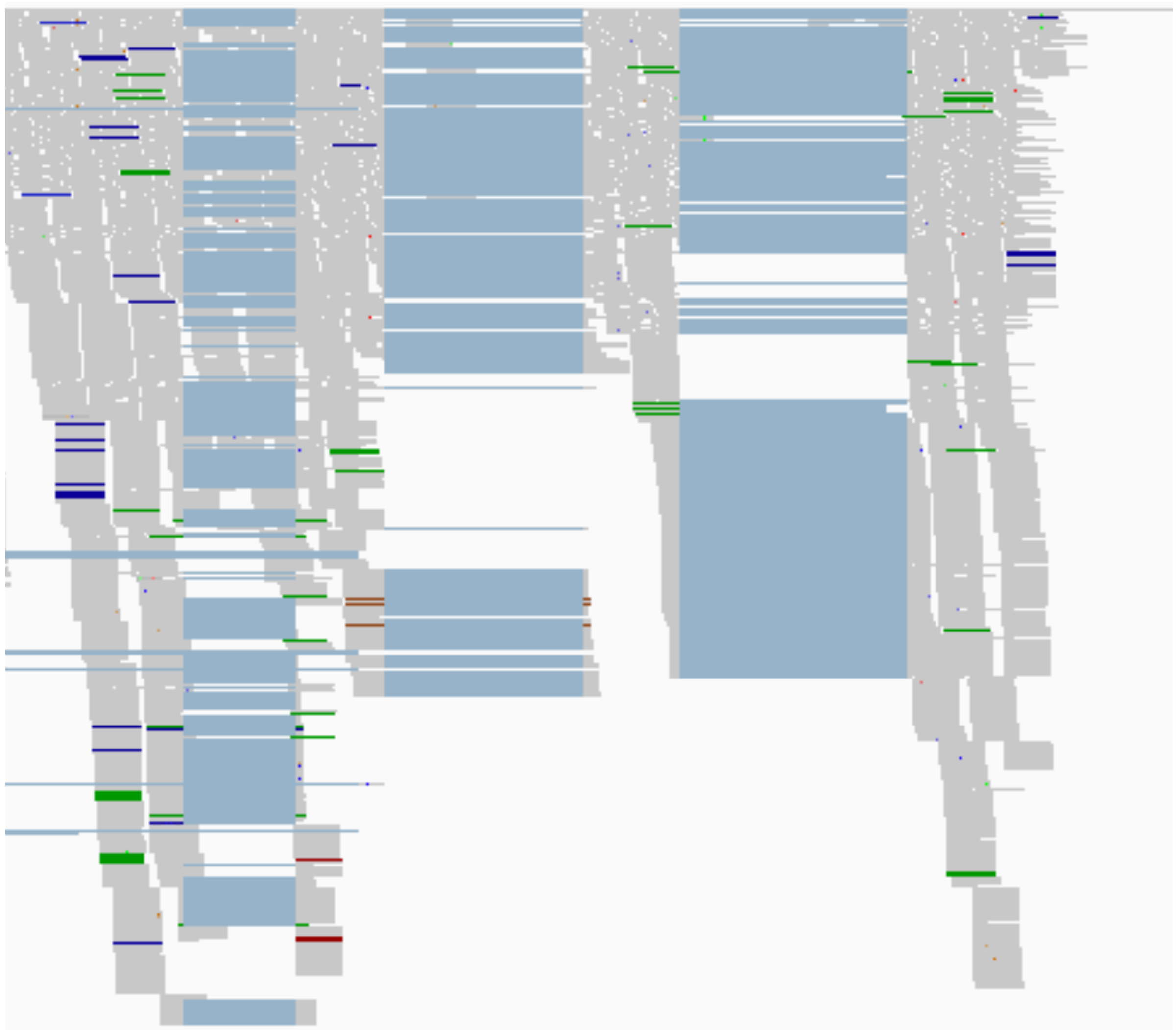


Side Note: Real Data

Transcripts:



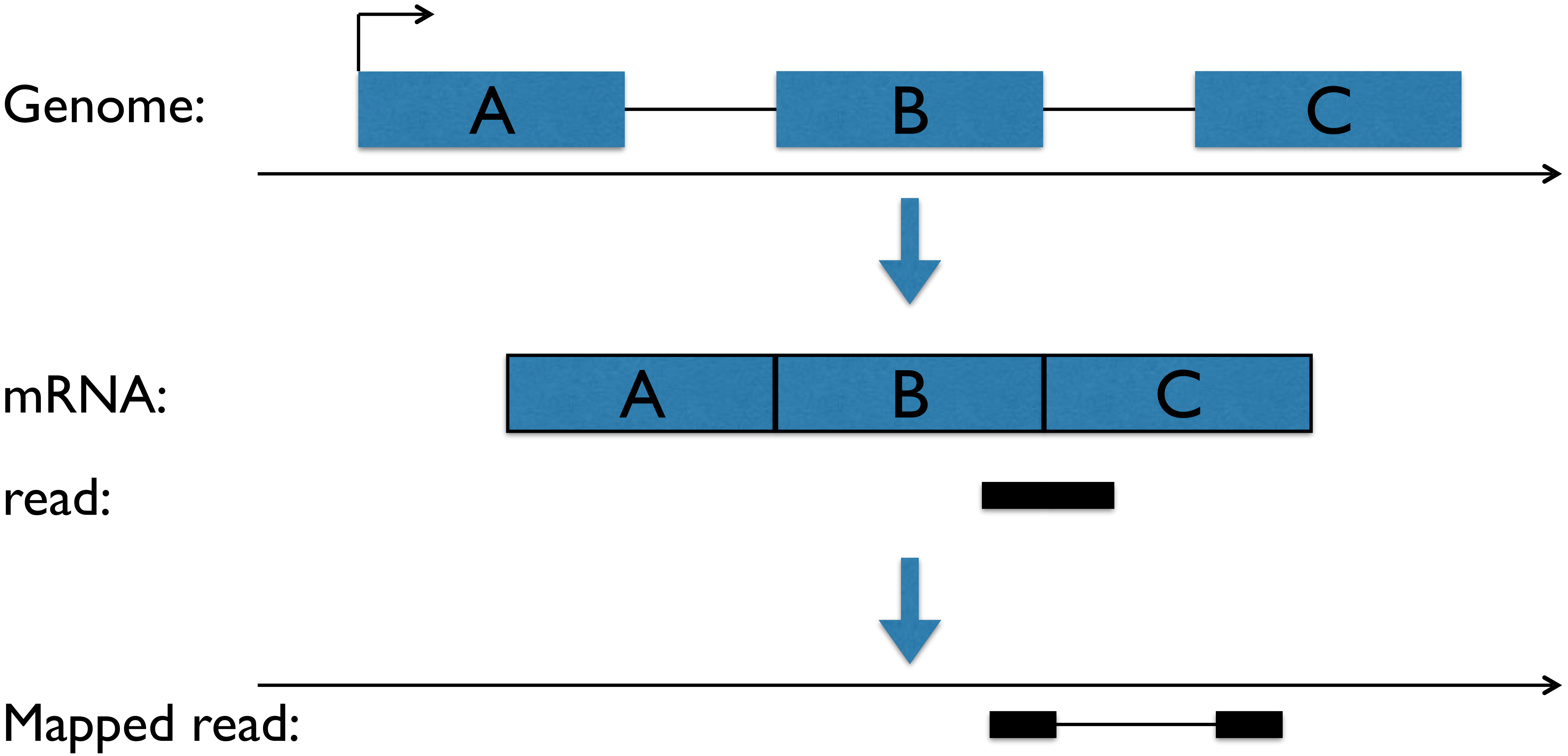
Mapped Reads:



Mapping of Reads

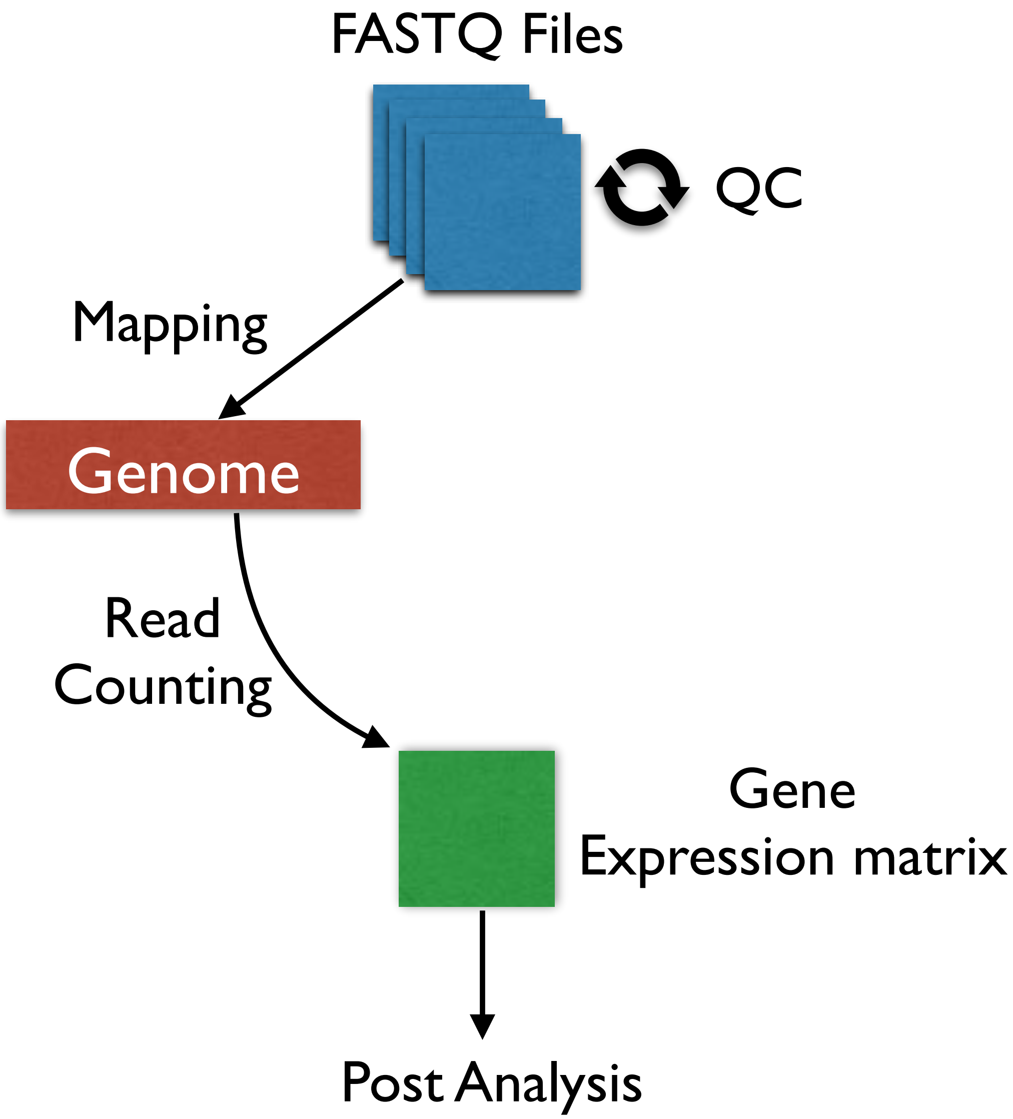
- Reads can be divided into 4:
  1. Reads not mapping
  2. Reads mapping perfectly
  3. Multi-mappers
  4. Reads that maps to two (or more) exons (junction spanning reads)

Mapping:
Junction-Spanning Reads



Conceptual Overview

1. Experiment -> RNA-Seq libraries (lab-work)
2. Sequencing (company)
3. Data analysis (you)
    A. QC and Trimming
    B. Mapping
    C. Quantification
    D. Post analysis

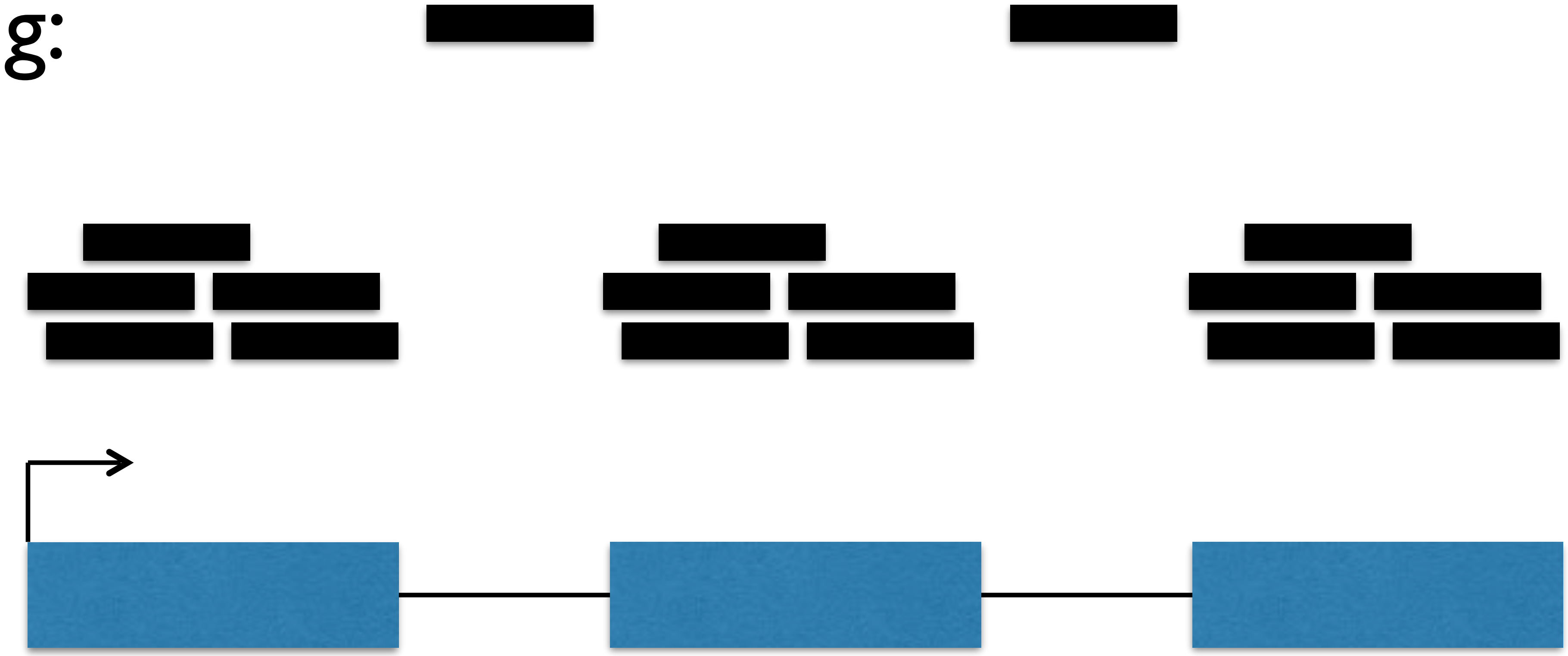


Expression estimation

Non-Matching:

Matching Fragments:

Gene:



Read count: 5 5 5

Total read count: 15

NB: Only uniquely mapping!

Expression estimation exercise

- 5 minutes with neighbour:
- You are analysing 2 genes (gene A and B) in two conditions (condition 1 and 2) on the basis of an single end RNA-seq experiment that resulted the following number of reads (= fragments):

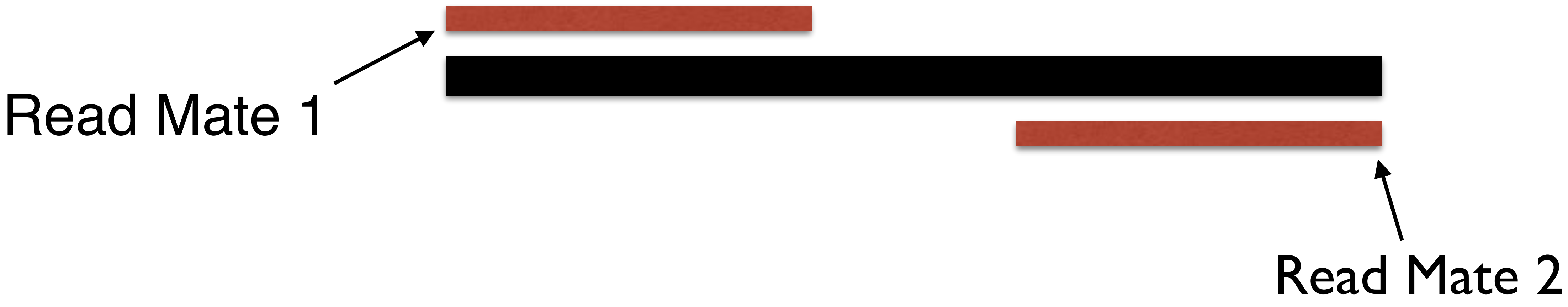
	Condition 1	Condition 2
Gene A	1000	3000
Gene B	2000	4000

- Question: Is the following statement correct?

  Both gene A and B are more expressed in condition 2. Explain why/why not.

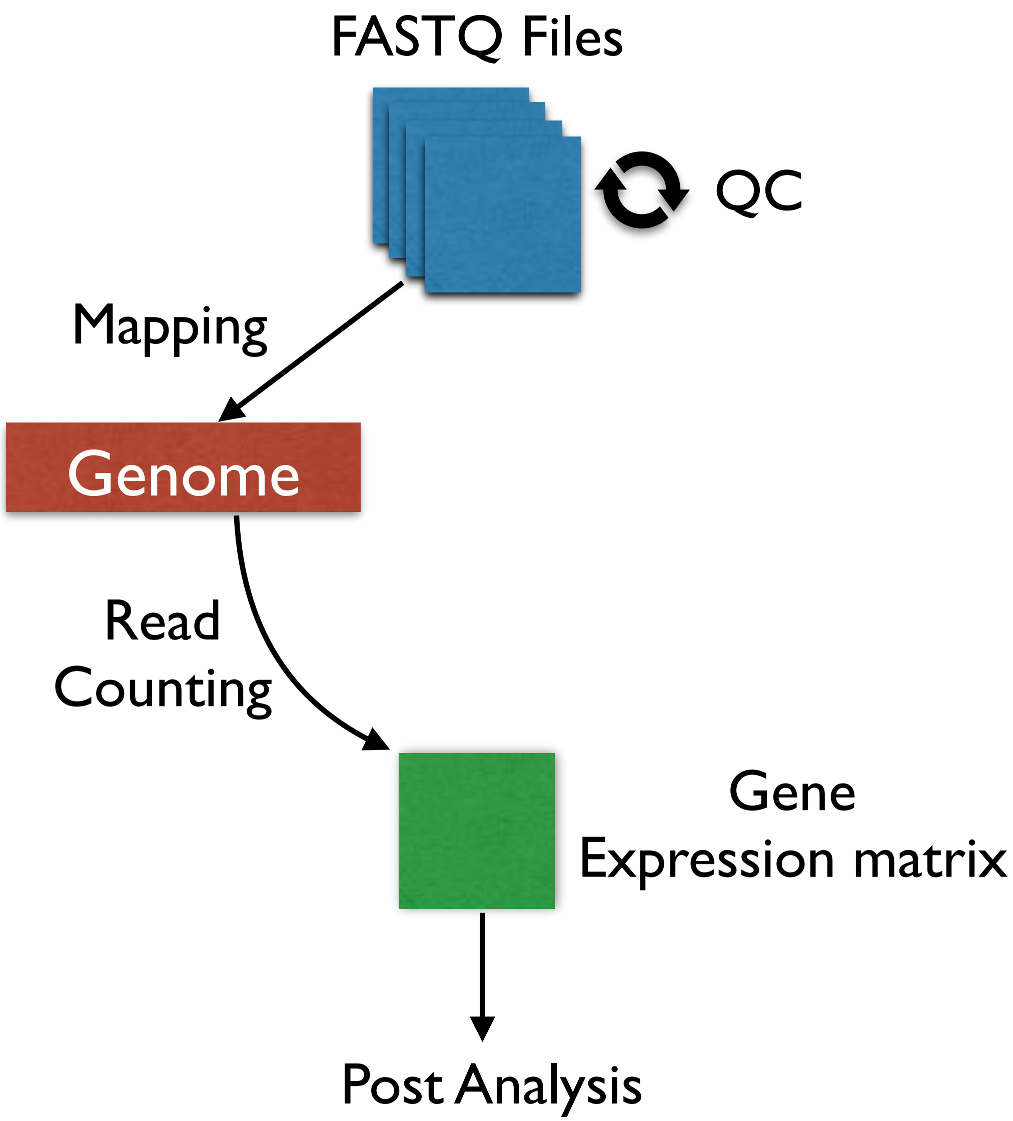
FPKM

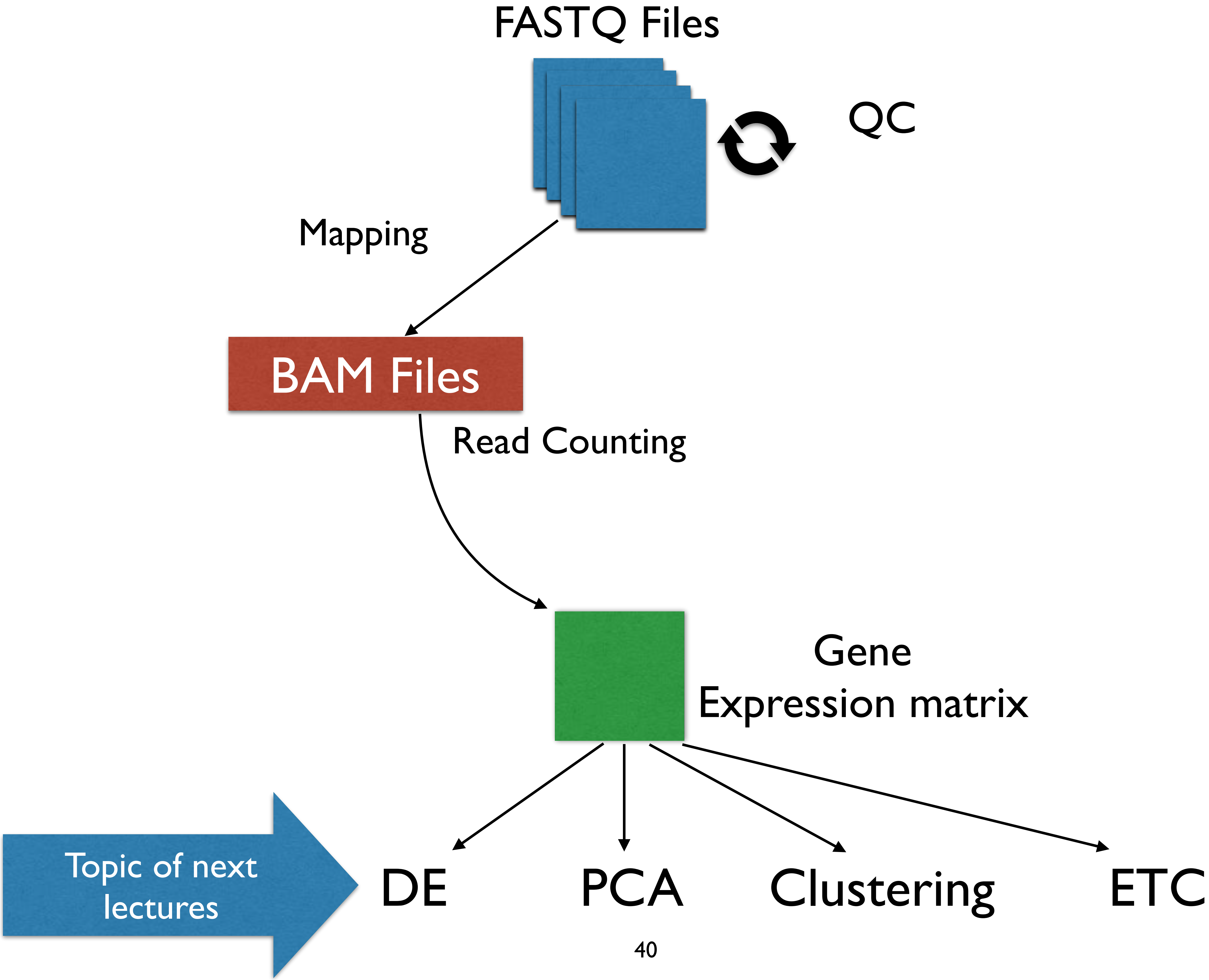
- A common measure of expression in RNA-seq:
  - FPKM - Fragments Per Kilobase transcript per Million mapped reads
  - Analogous to RPKM, just adjusted to multiple reads originating from same fragment (paired end sequencing)
  - Allows comparison of different genes and between samples



Conceptual Overview

1. Experiment -> RNA-Seq libraries (lab-work)
2. Sequencing (company)
3. Data analysis (you)
    A. QC and Trimming
    B. Mapping
    C. Quantification
    D. **Post analysis**





FASTQ Files

QC

Mapping

BAM Files

Read Counting

Gene Expression matrix

Topic of next lectures

DE

PCA

Clustering

ETC

But what about isoforms?

Rember: Solution

What do you gain by profiling the transcriptome with isoform resolution (compared to gene resolution)?

- Greater details
- Alternative splicing
- Isoform switching
- Sequence analysis (e.g. protein domains (Pfam))
- Improved gene-level analysis

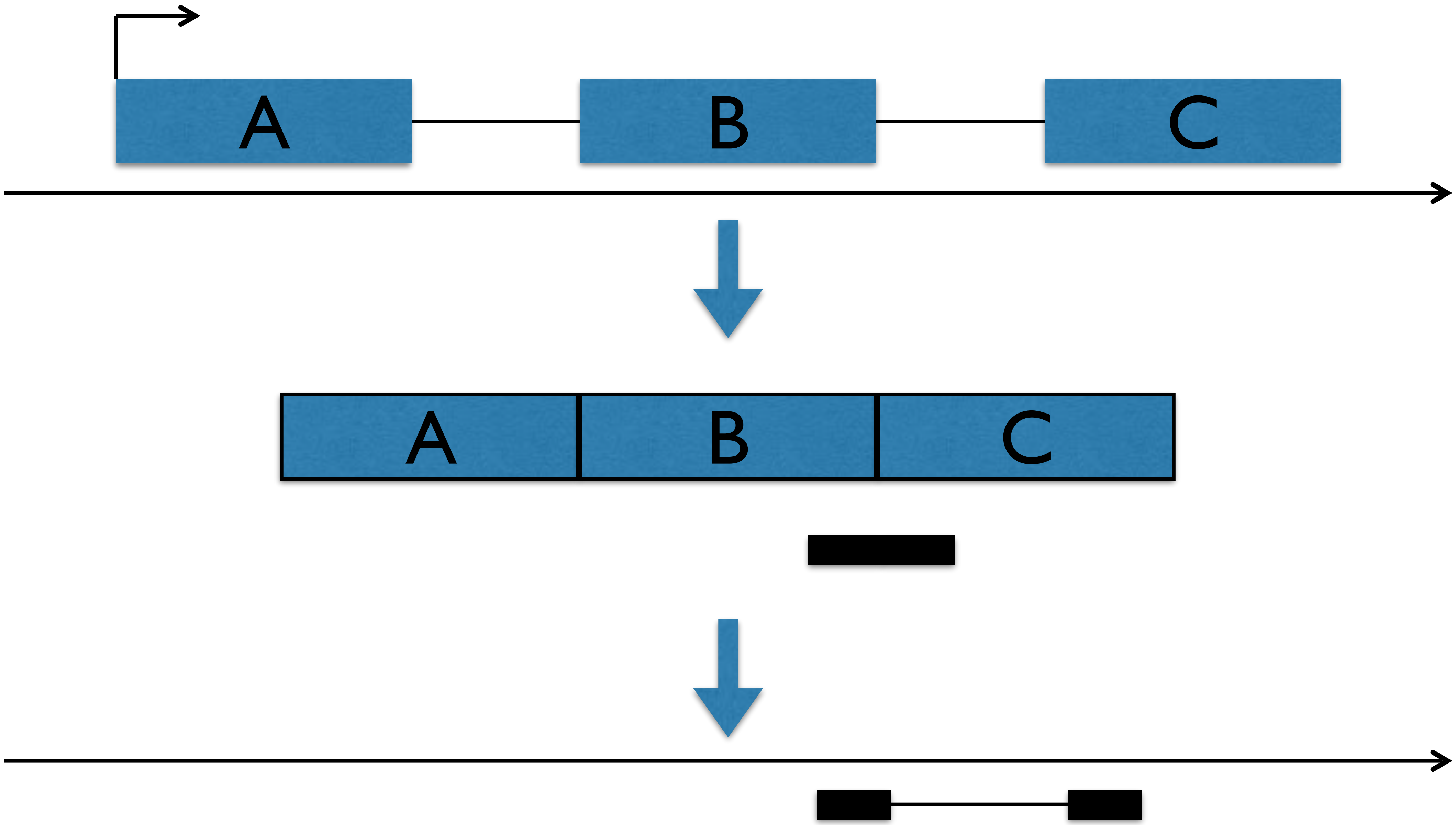
Types of Isoform Analysis

1. Predict new isoforms (reconstruct)
2. Quantify annotated (aka known) isoforms

Types of Isoform Analysis

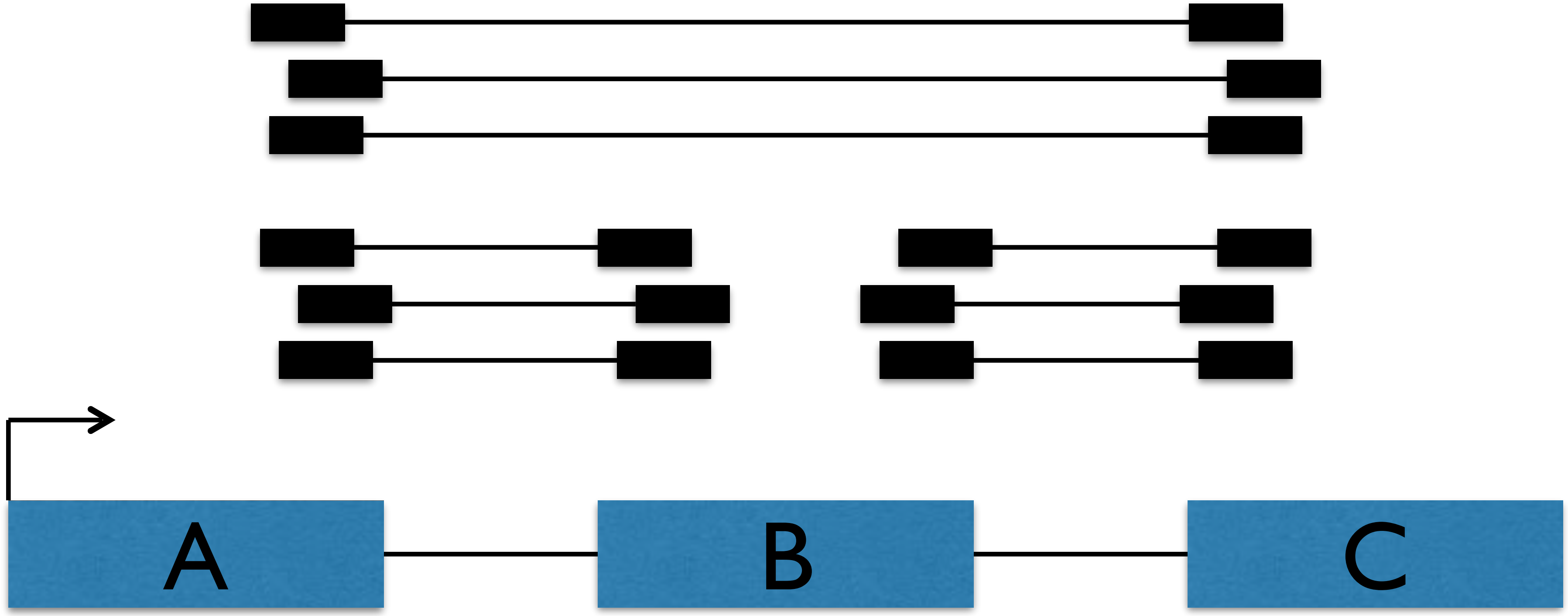
1. **Predict new isoforms (reconstruct)**
2. Quantify annotated (aka known) isoforms

Remember: Junction-Spanning Reads

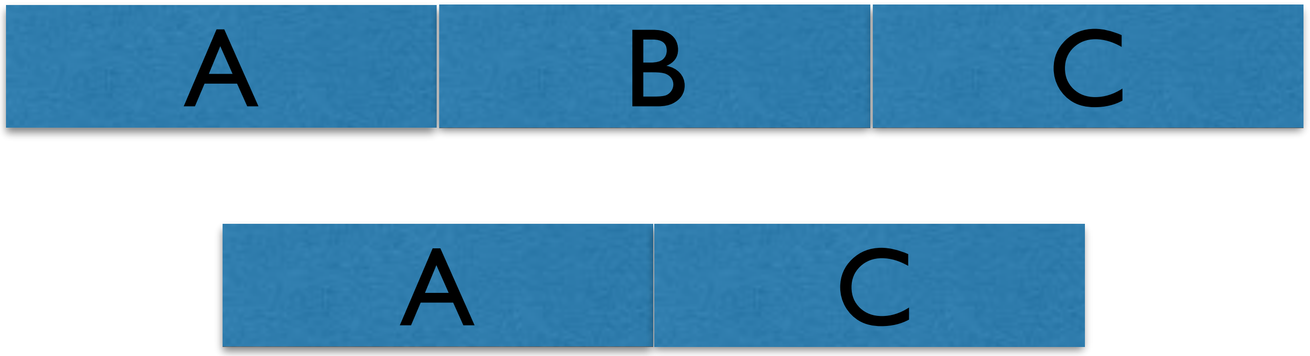


Isoform Reconstruction - Concept

Mapping:



Isoform deconvolution:

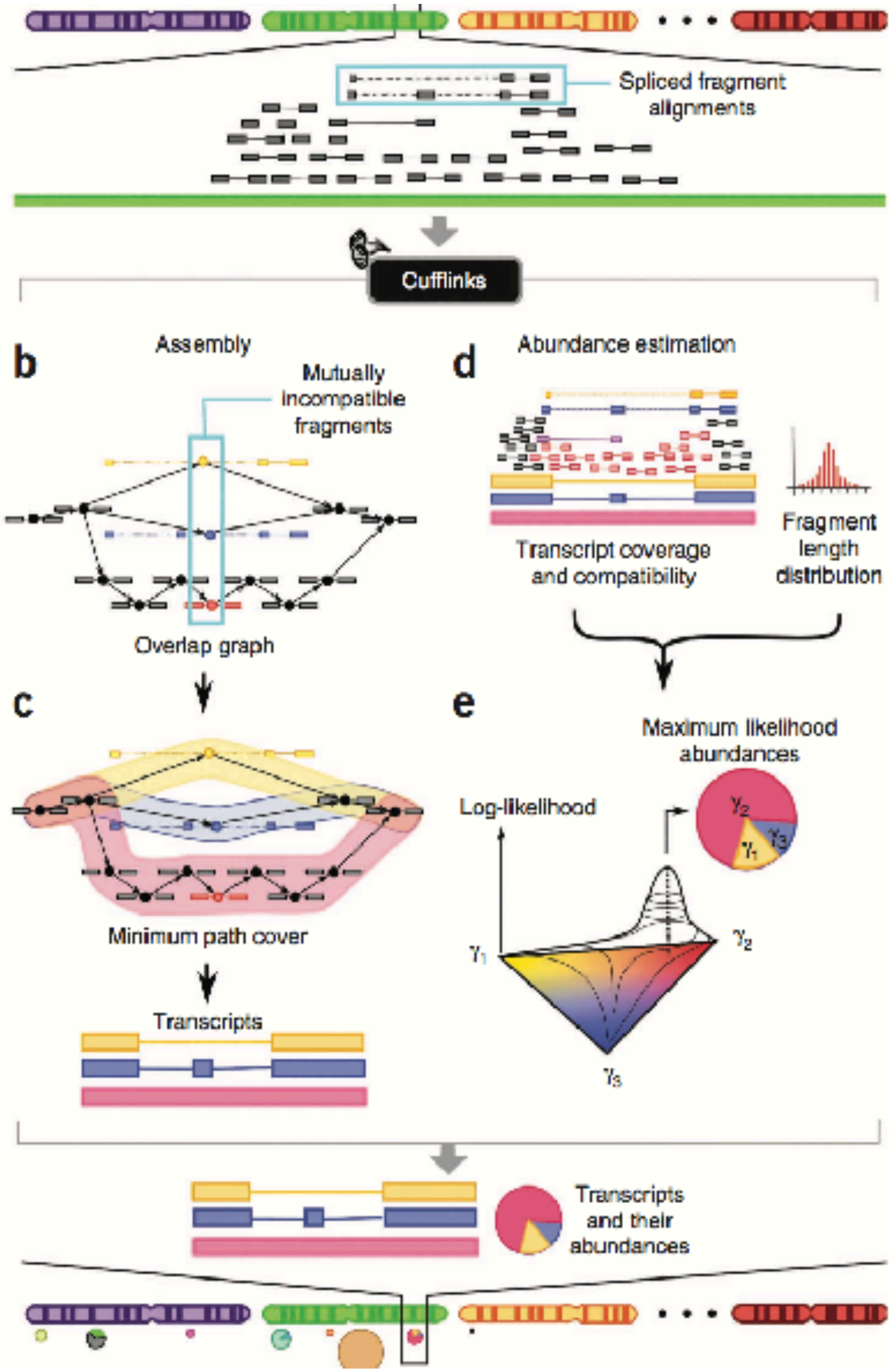


Which isoform is expressed more?

Isoform Reconstruction - Reality

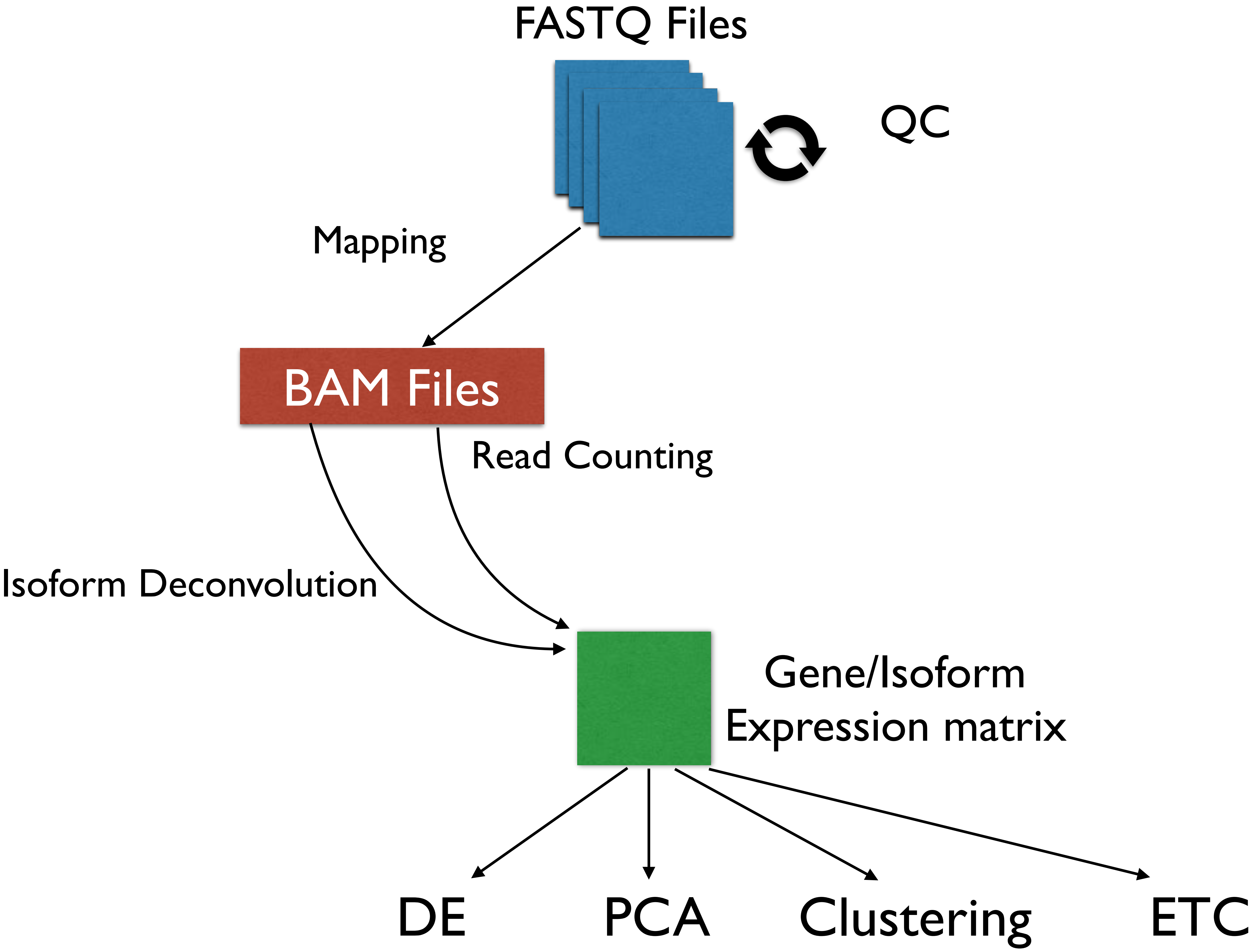
Mapping:

Isoform deconvolution


Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ_1
γ_2
γ_3
Transcripts and their abundances

(Not curriculum)

http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1621.html

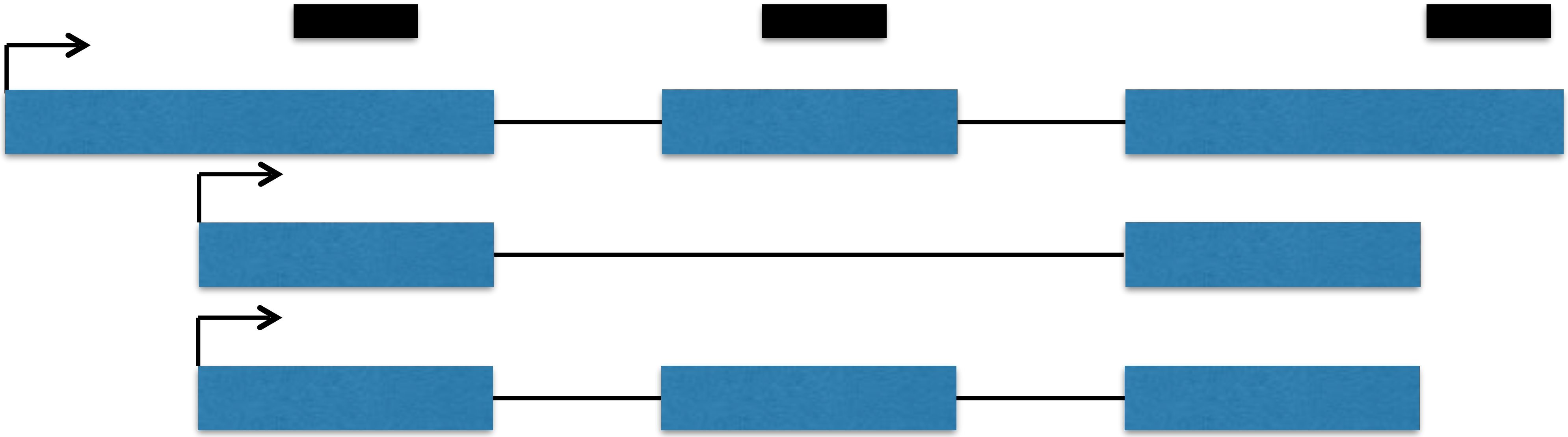

FASTQ Files
QC
Mapping
BAM Files
Read Counting
Isoform Deconvolution
Gene/Isoform
Expression matrix
DE
PCA
Clustering
ETC

Types of Isoform Analysis

1. Predict new isoforms (reconstruct)
2. **Quantify annotated (aka known) isoforms**

Quantify Annotated Isoforms

A tough problem

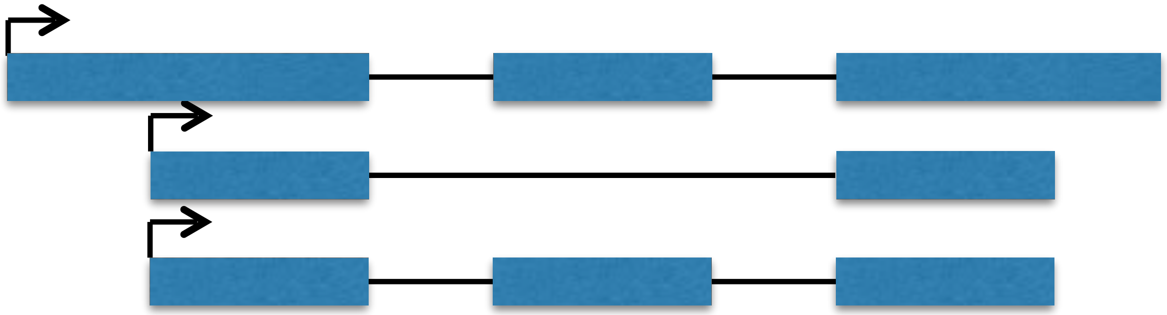


Quantify Annotated Isoforms

A tough problem

Solution: Pseudo-allignment

Pseudo Allignment



Extract sequence

TTCAGTCTCAGAATCGA GATACGATTACG ATATCGAGATACGATCGGCG

AGAATCGA ATATCGAGAT

AGAATCGA GATACGATTACG ATATCGAGAT

Concatenate sequence

TTCAGTCTCAGAATCGAGATACGATTACGATATCGAGATACGATCGGCG

AGAATCGAATATCGAGAT

AGAATCGAGATACGATTACGATATCGAGAT

Pseudo Allignment

TACGAT

Read

TTCAGTCTCAGAATCGAGATACGATTACGATATCGAGATACGATCGGCG

AGAATCGAATATCGAGAT

Reference
Transcriptome

AGAATCGAGATACGATTACGATATCGAGAT

Pseudo Allignment

TACGAT
TTCAGTCTCAGAATCGAGATACGATTACGATATCGAGATACGATCGGCG

Match

TACGAT
AGAATCGAATATCGAGAT

No match

Reference
Transcriptome

TACGAT
AGAATCGAGATACGATTACGATATCGAGAT

Match

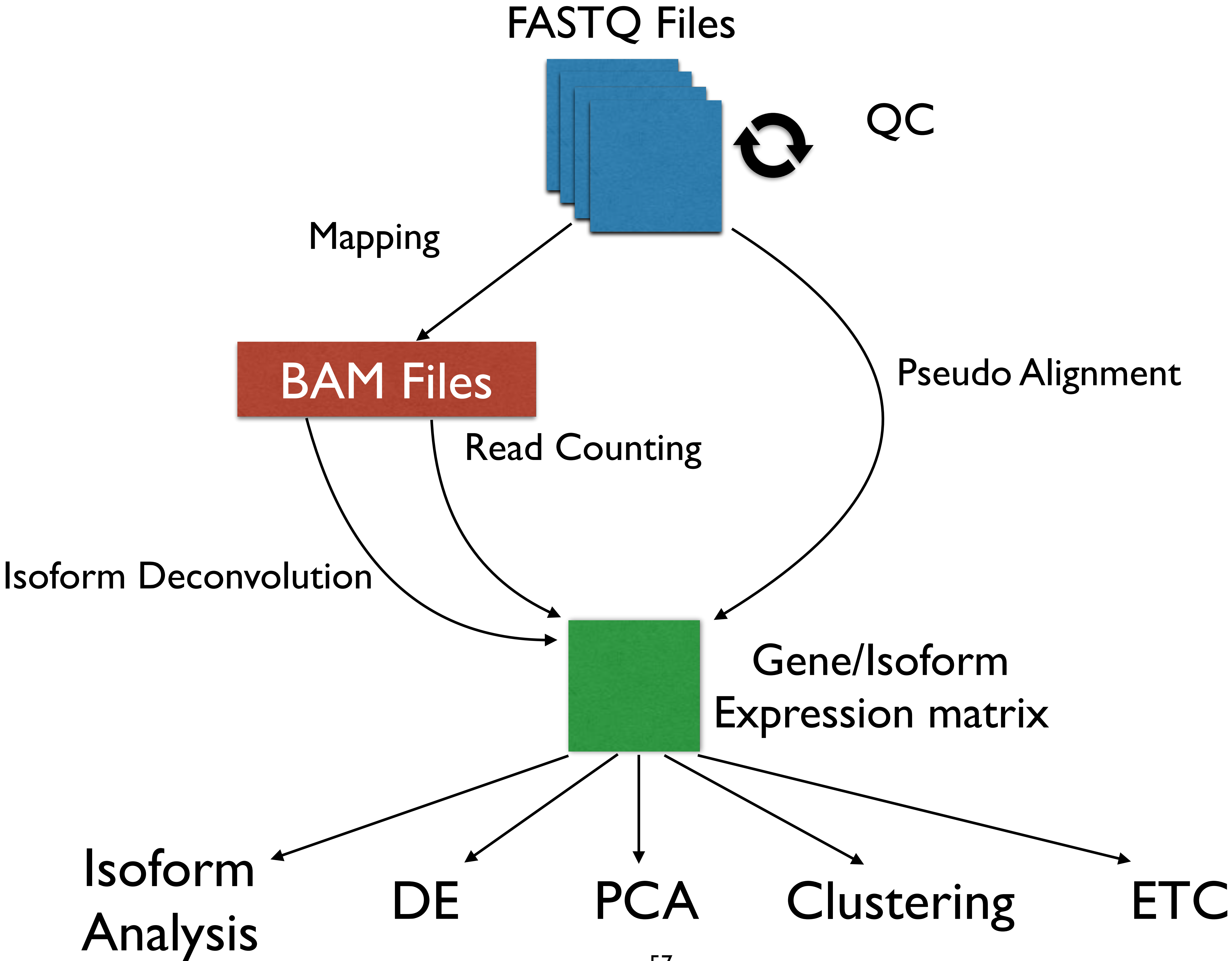
Mapping

Naturally modern algorithms are a lot smarter than that:

- Clever transciptome indexing
- Advanced read matching which considers read pairs
- Advanced quantification algorithm
- Bias corrections
- Etc

TPM / TxPM

- Currently the best measure of expression in RNA-seq:
  - TPM - Transcript Per Million
  - Not the same as sometimes used for CAGE!!!
  - Analogous to FPKM except also normalised for other features biasing the FPKM measure



FASTQ Files

QC

Mapping

BAM Files

Pseudo Alignment

Read Counting

Isoform Deconvolution

Gene/Isoform Expression matrix

Isoform Analysis

DE

PCA

Clustering

ETC

Isoform vs Gene quantification

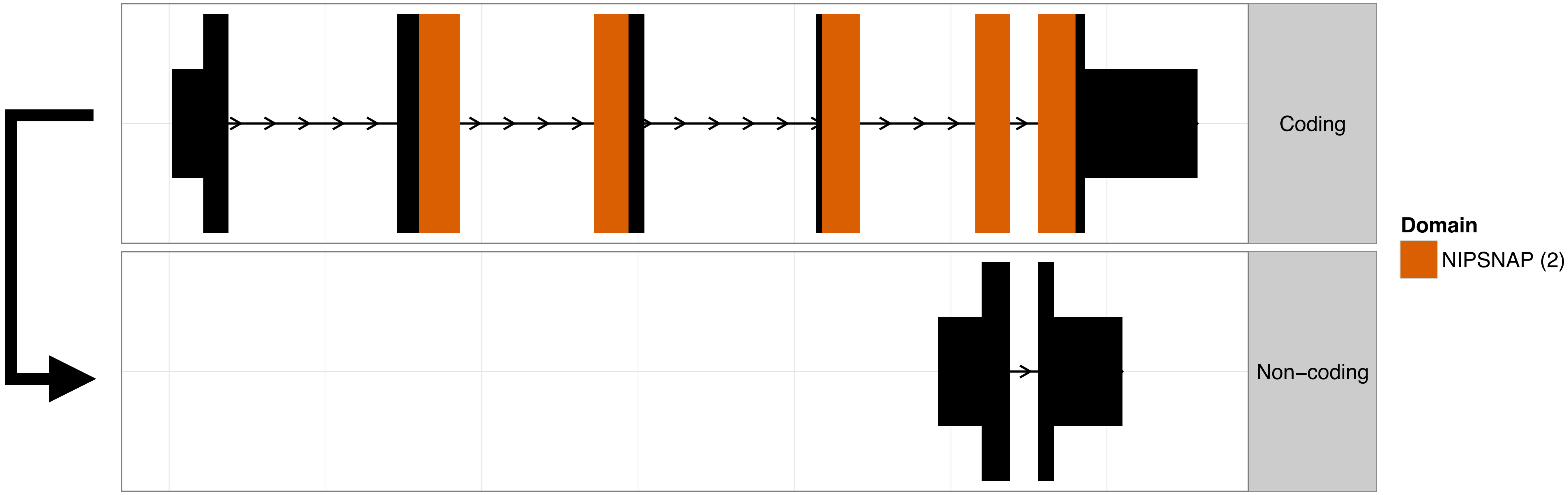
Improved gene level analysis

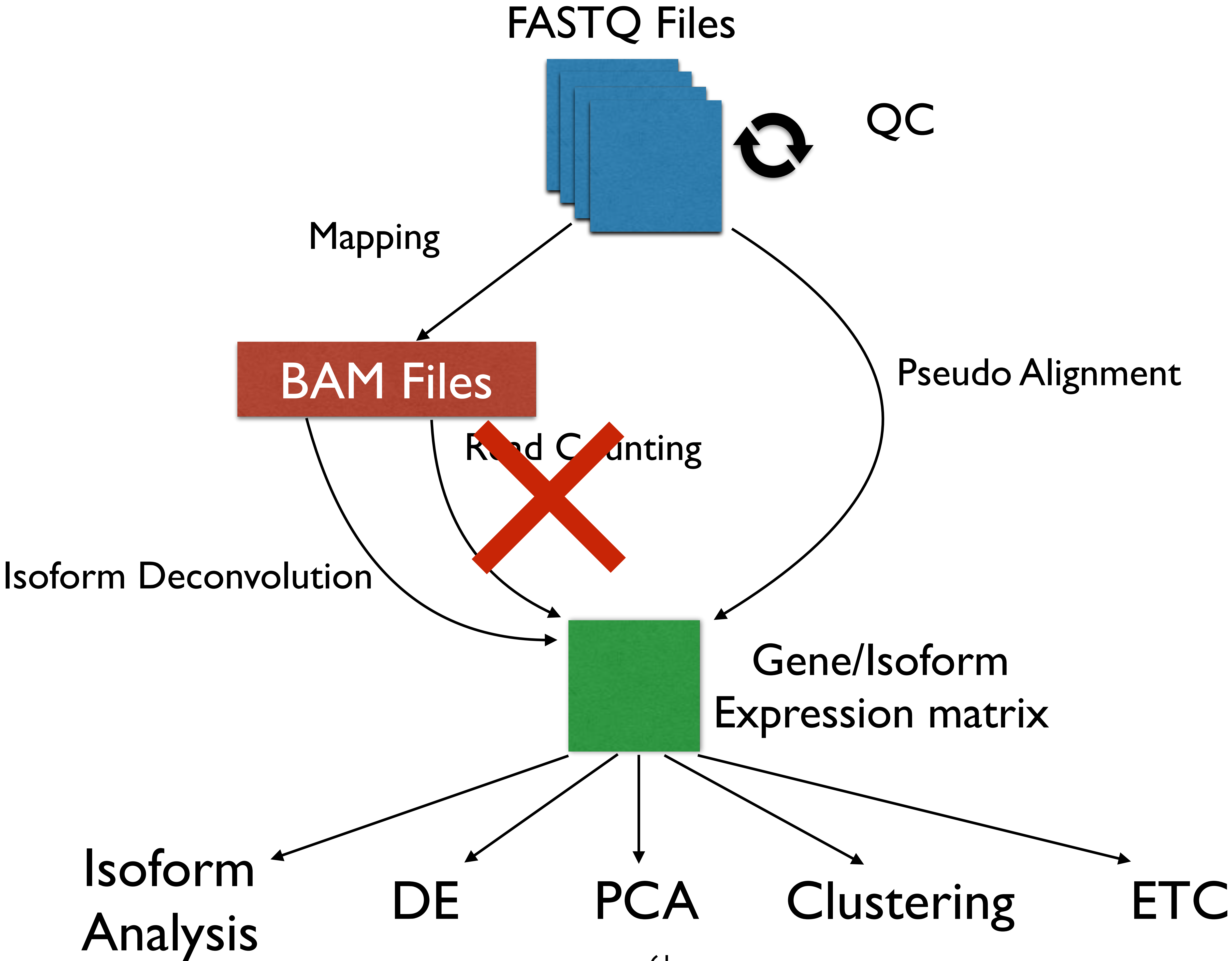
1. Multi-mapping reads can be counted
2. Counting uniquely mapping reads is problematic as genes differ in terms of how large a fraction of
   the gene is uniquely mappable
3. Isoform switches are a problem

Exercise:

5 minuts with your neighbour:

When counting uniquely mapping reads and normalising via (RPKM/FPKM) what problem(s) would the isoform switch illustrated below cause for quantification in the two conditions?





FASTQ Files

QC

Mapping

BAM Files

Read Counting

Pseudo Alignment

Isoform Deconvolution

Gene/Isoform Expression matrix

Isoform Analysis

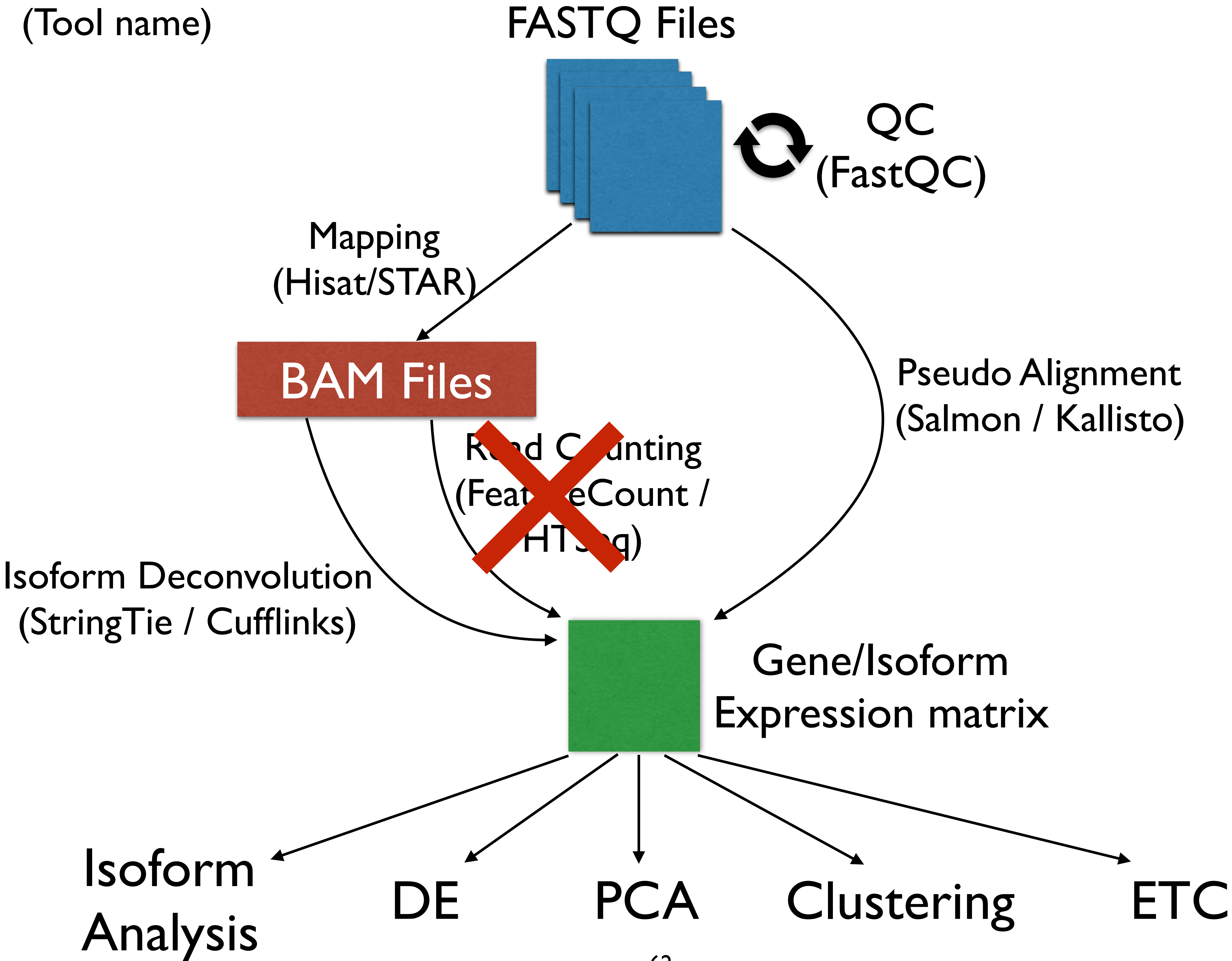
DE

PCA

Clustering

ETC

(Tool name)



FASTQ Files

QC (FastQC)

Mapping (Hisat/STAR)

BAM Files

~~Read Counting (FeatureCount / HTSeq)~~

Pseudo Alignment (Salmon / Kallisto)

Isoform Deconvolution (StringTie / Cufflinks)

Gene/Isoform Expression matrix

Isoform Analysis

DE

PCA

Clustering

ETC

Summary

- Quality control of FASTQ files is always needed
- Gene/isoform quantification should almost always be done with pseudo aligners
- To get gene/isoform expression a lot of normalisation is needed (library size, feature length etc)
- There are good tools for doing all of this

Agenda

1. Introduction to RNA-seq
2. RNA-seq workflow
    1. **Do-it-yourself exercise**
3. Isoform Switch Analysis
    1. Do-it-yourself exercise
4. Perspective

Conceptual Overview

1. Experiment -> RNA-Seq libraries (lab-work)
2. Sequencing (company)
3. Data analysis (you)

    A. QC and Trimming

    B. Mapping

    C. Quantification

    D. Post analysis
    - Isoform analysis
    - PCA
    - Clustering
    - Differential expression analysis

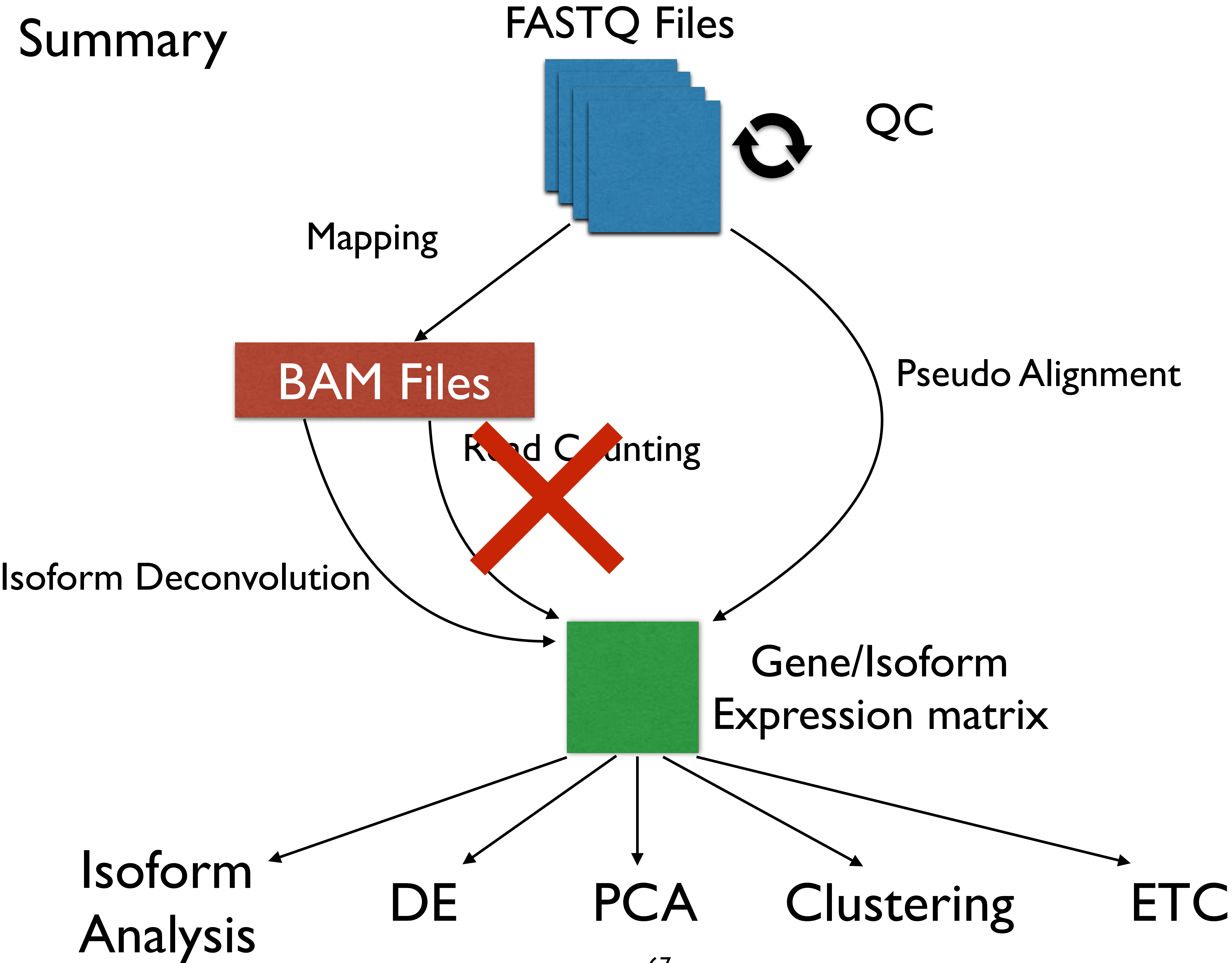
Focus for today
(pseudo-allignment)

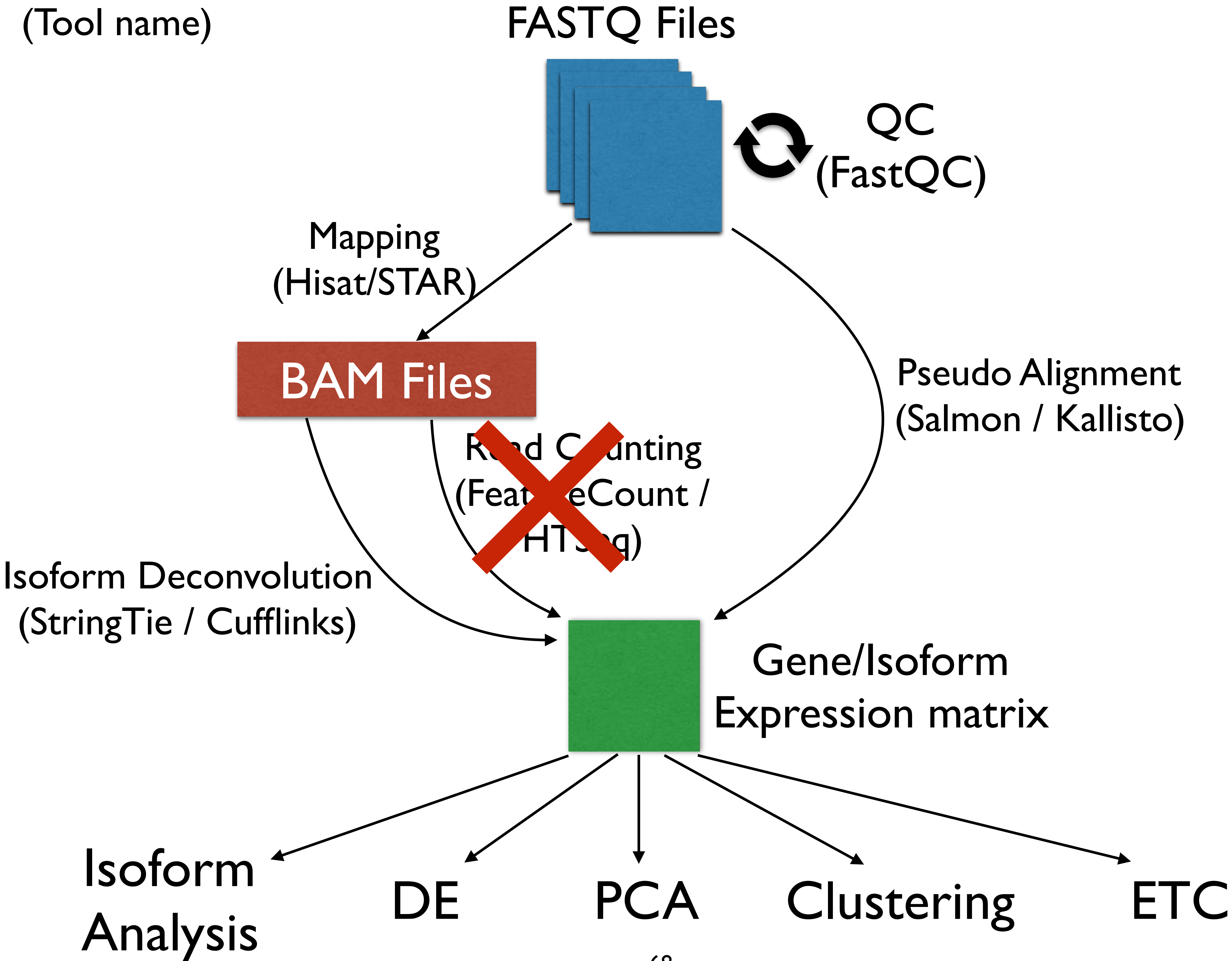
RNA-seq exercise

Find in the document
“rnaseq_quantification_exercise_wo_solutions.docx”

On Absalon and do the exercise

Summary


FASTQ Files
QC
Mapping
BAM Files
Pseudo Alignment
Read Counting
Isoform Deconvolution
Gene/Isoform
Expression matrix
Isoform
Analysis
DE
PCA
Clustering
ETC



(Tool name)

FASTQ Files

QC
(FastQC)

Mapping
(Hisat/STAR)

BAM Files

~~Read Counting
(FeatureCount /
HTSeq)~~

Pseudo Alignment
(Salmon / Kallisto)

Isoform Deconvolution
(StringTie / Cufflinks)

Gene/Isoform
Expression matrix

Isoform
Analysis

DE

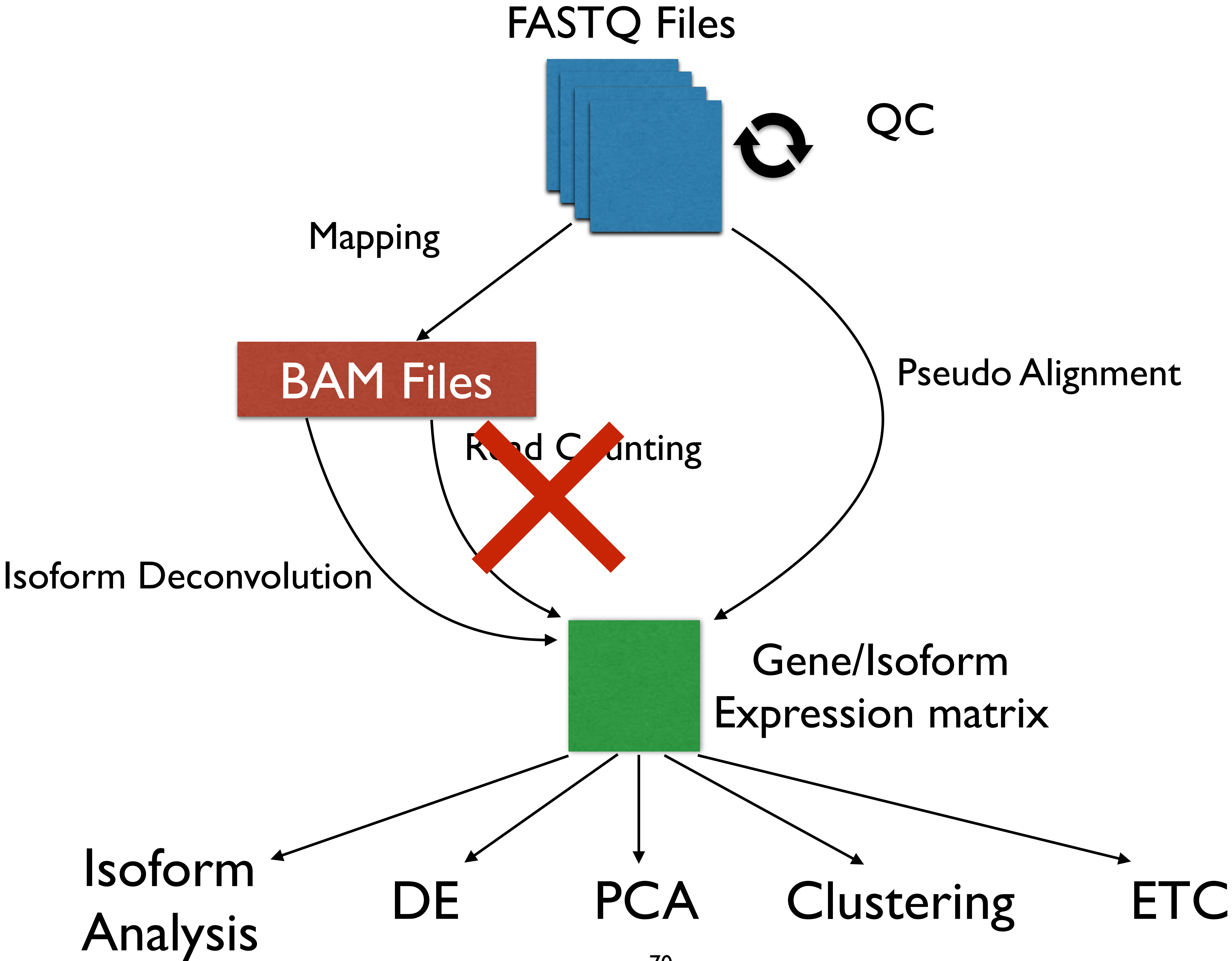
PCA

Clustering

ETC

Agenda

1. Introduction to RNA-seq
2. RNA-seq workflow
    1. Do-it-yourself exercise
3. **Isoform Switch Analysis**
    1. Do-it-yourself exercise
4. Perspective



FASTQ Files

QC

Mapping

BAM Files

Pseudo Alignment

Read Counting

Isoform Deconvolution

Gene/Isoform
Expression matrix

Isoform
Analysis

DE

PCA

Clustering

ETC

Isoform Switching

Isoform Fraction (IF values)

IF= isoform_exp / gene_exp

Expression	TxPM	IF
Isoform 1	10	0.1
Isoform 2	90	0.9
Gene (total)	100	1

Extra important with accurate abundance estimats!

Isoform Switching

Isoform Fraction (IF values)

IF= isoform_exp / gene_exp

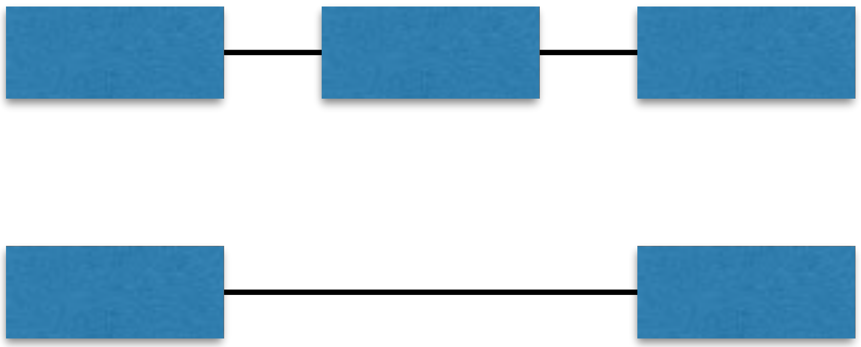
dIF= IF2 - IF1

	IF1 (Condition 1)	IF2 (Condition 2)	dIF (IF2 - IF1)
Isoform 1	0.2	0.8	+0.6
Isoform 2	0.8	0.2	-0.6
Gene (total)	1	1	0

Isoform Switching

- A IF value measures how much an isoform contribute to the parent gene expression
- A dIF values measures the change, between conditions, in how much an isoform contribute to the parent gene expression!
- Both values can be interpreted as the (change in) the relative importance of an isoform

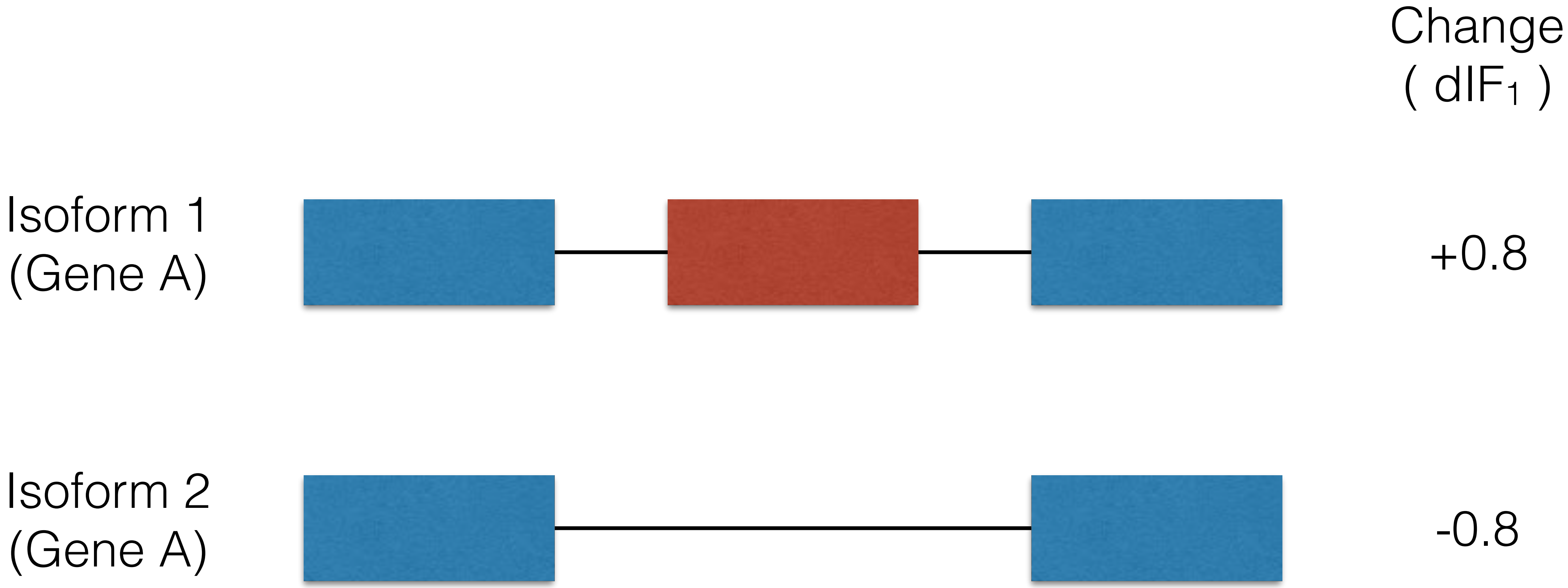
Isoform Switching



	Condition A (IF_{A1}, …, IF_{An})		Condition B (IF_{B1}, …, IF_{Bn})	Change (avg dIF)
Isoform 1 (Gene A)	0.1, … , 0.2	→	0.9, … , 0.8	+0.7
Isoform 2 (Gene A)	0.9, … , 0.8	→	0.1, … , 0.2	-0.7

Remember the difference between p-values and effect size

Protein Domains



Change
(dIF_1)

Isoform 1
(Gene A)

+0.8

Isoform 2
(Gene A)

-0.8

PFAM

- Database of protein domains
- Tool for finding protein domains in amino acid sequence

Only ~11% of scientific articles from the start of 2016 analysing RNA-seq data does so at isoform resolution

Systematic High throughput Analysis of Isoform Switches

- there is an R package for that

IsoformSwitchAnalyzeR

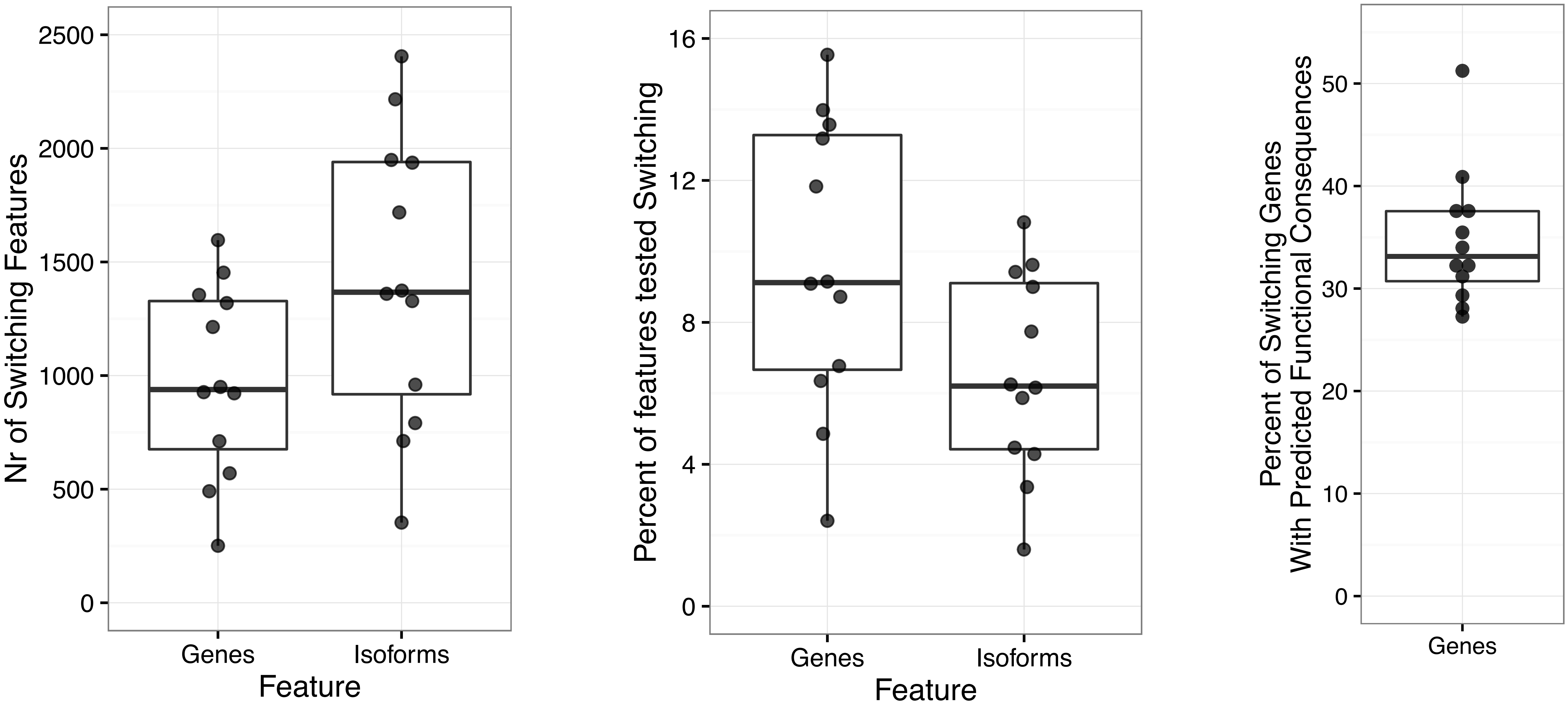
An R package which from full-length isoform quantifications:

1. Identify isoform switches
2. Combine multiple sources of annotations
3. Prediction functional consequences

Showcase: Data

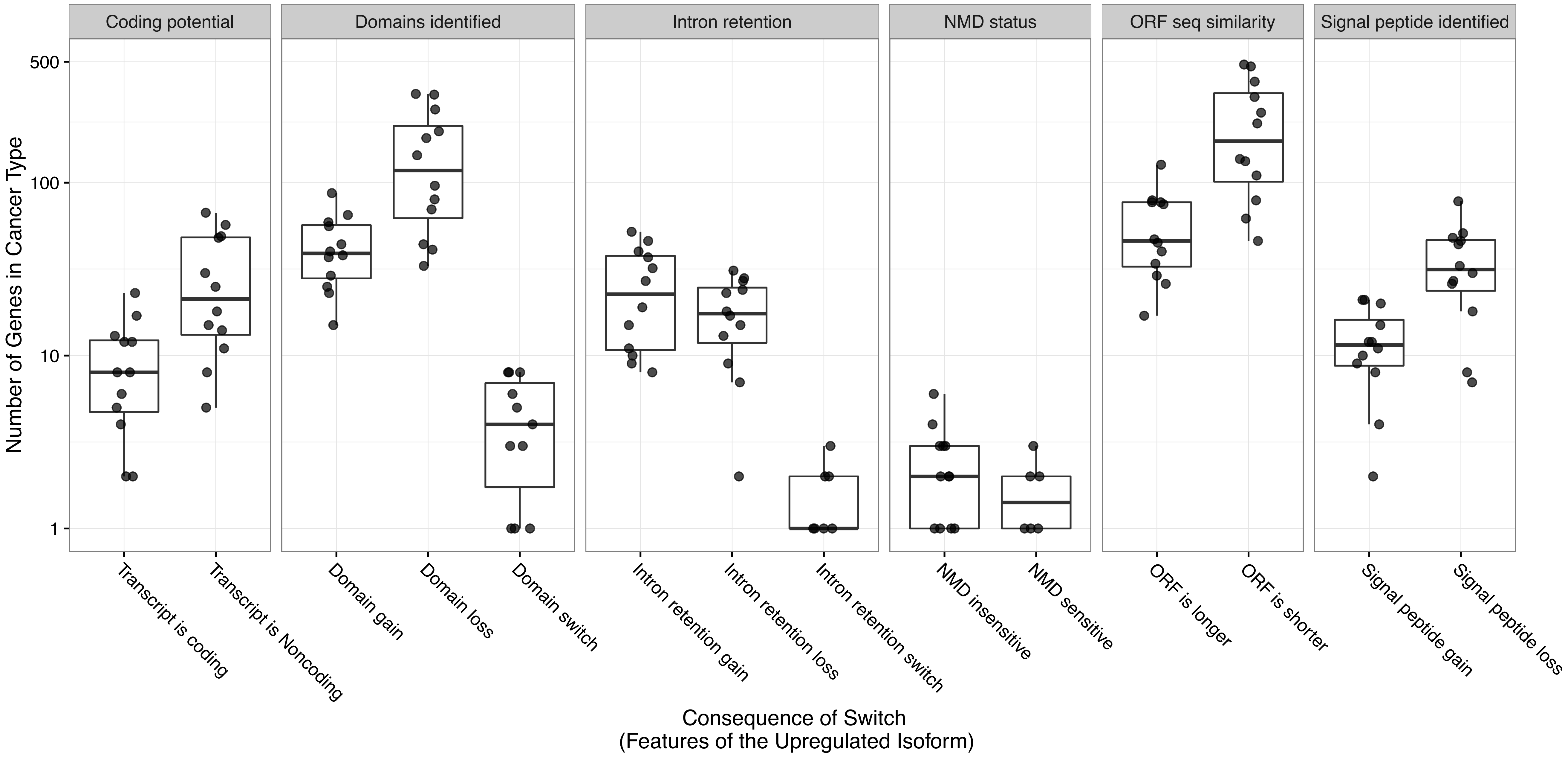
- RNA-seq data from ~6000 Cancer Patients and Healthy Controls
- Covering 12 Cancer Types

The Abundance of Isoform Switching in Cancers

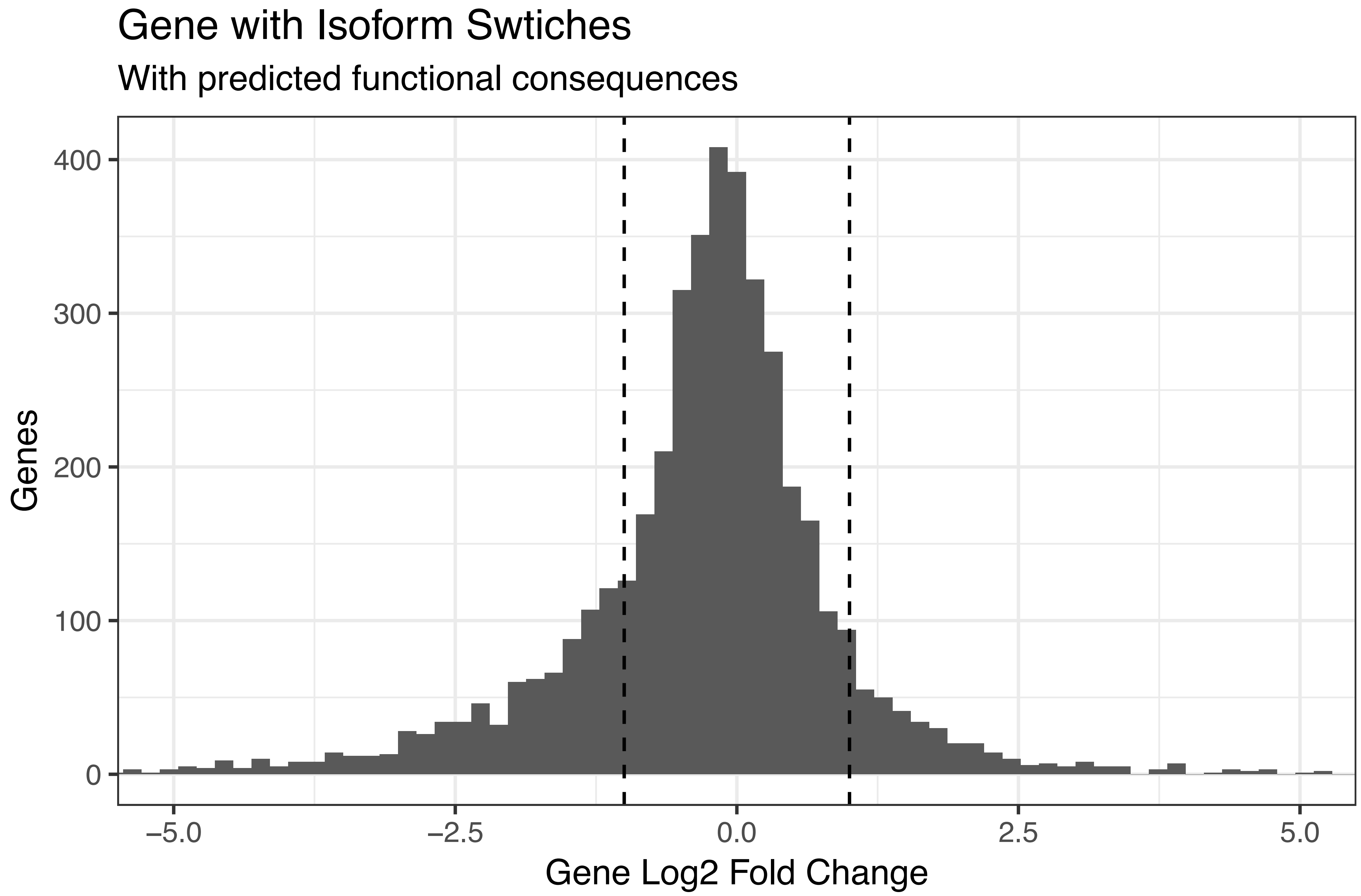


Across 12 cancer types 2334 different genes (18.81 % of tested) have significant changes in isoform usage with predicted functional consequences

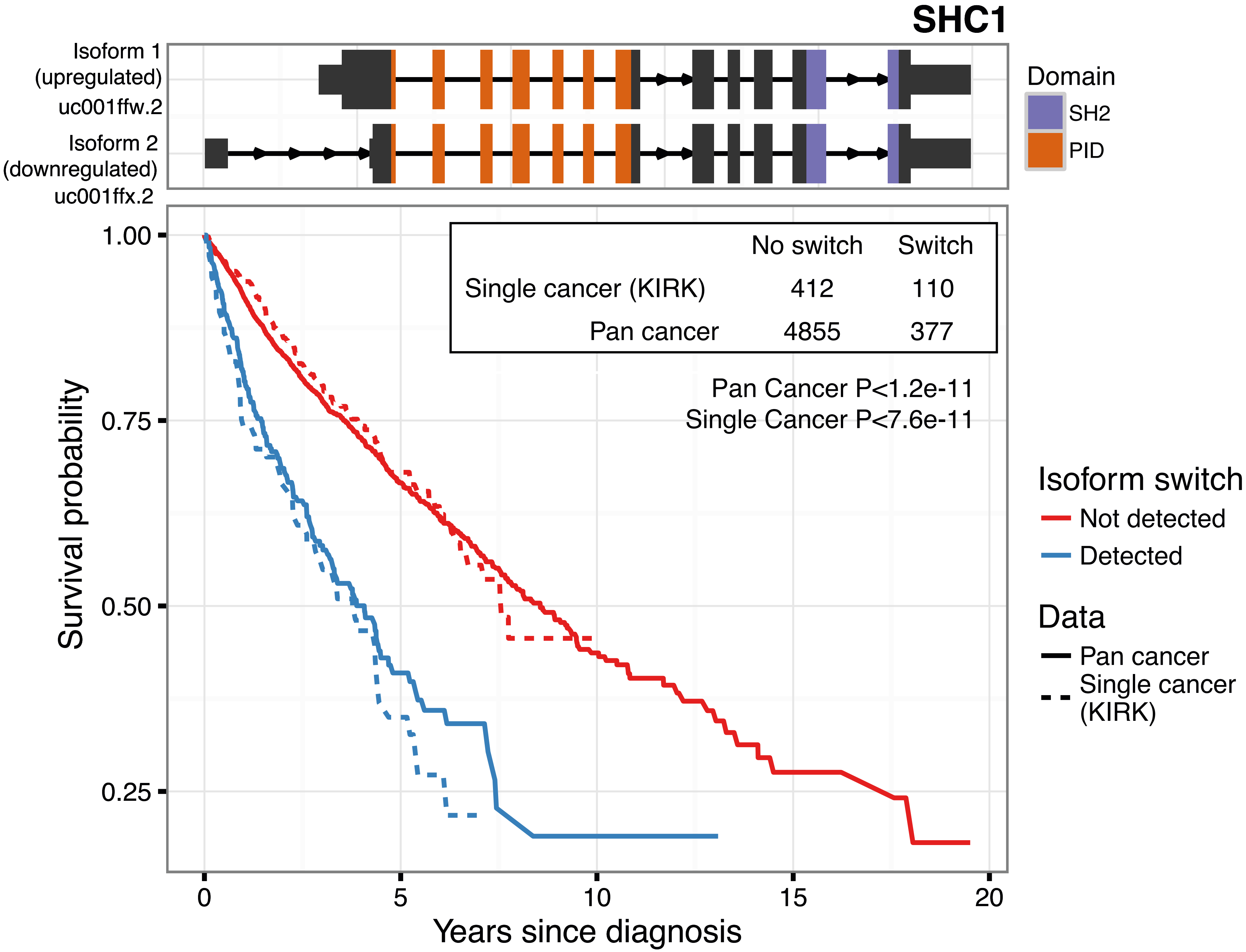
The Abundance of Isoform Switching in Cancers



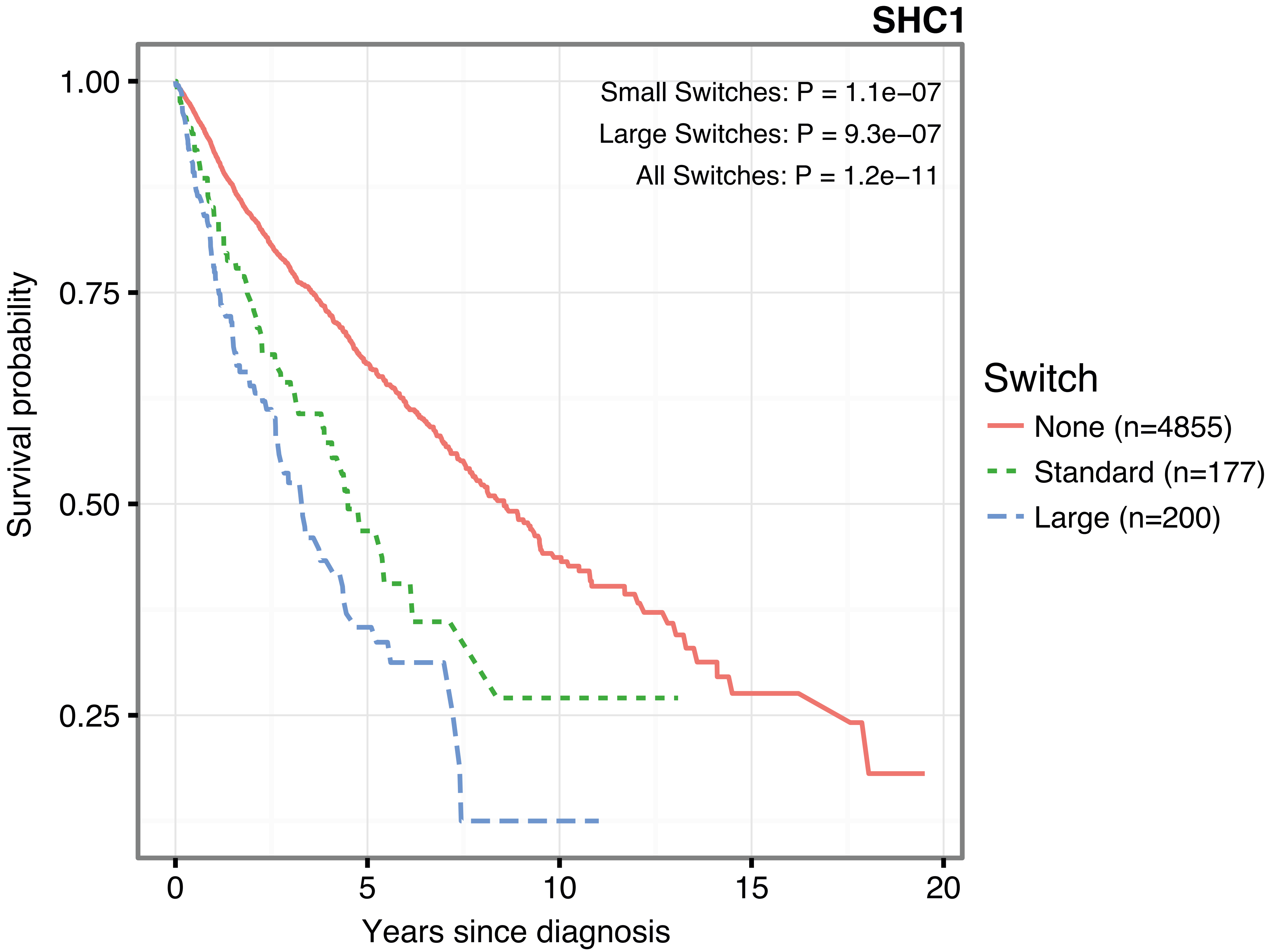
Isoform Switches vs Gene Expression



Isoform Switching vs Patient Survival



Isoform Switching vs Patient Survival



Summary

- You can do systematic high throughput analysis of isoform switches with functional consequences
- Isoform Switches (with functional consequences) are extremely common
- Isoform switches and changes in gene expression are NOT mutually exclusive
- Isoform Switches (with functional consequences) seems to be biologically relevant

Agenda

1. Introduction to RNA-seq
2. RNA-seq workflow
    1. Do-it-yourself exercise
3. Isoform Switch Analysis
    1. **Do-it-yourself exercise**
4. Perspective

Isoform Switch Analysis Exercise

Absalon / Files / RNA-seq / isoform_switch_excecise_wo_solutions.pdf

Agenda

1. Introduction to RNA-seq
2. RNA-seq workflow
    1. Do-it-yourself exercise
3. Isoform Switch Analysis
    1. Do-it-yourself exercise
4. **Perspective**

Nanopore/PacBio

- Is a new technology that allows for sequencing of full length RNA molecules
- Meaning no need to fragment the RNA during the library preparation
- Meaning no need for assembler tools (since we already would know the transcript structure) (although new tools will be needed)
- Prospect: Will revolutionise transcriptome profiling

*Seq

- In this course we have talked about DNA re-seqencing, CHiP-seq, CAGE-seq and RNA-seq

- But there are currently hundreds *-seq methods

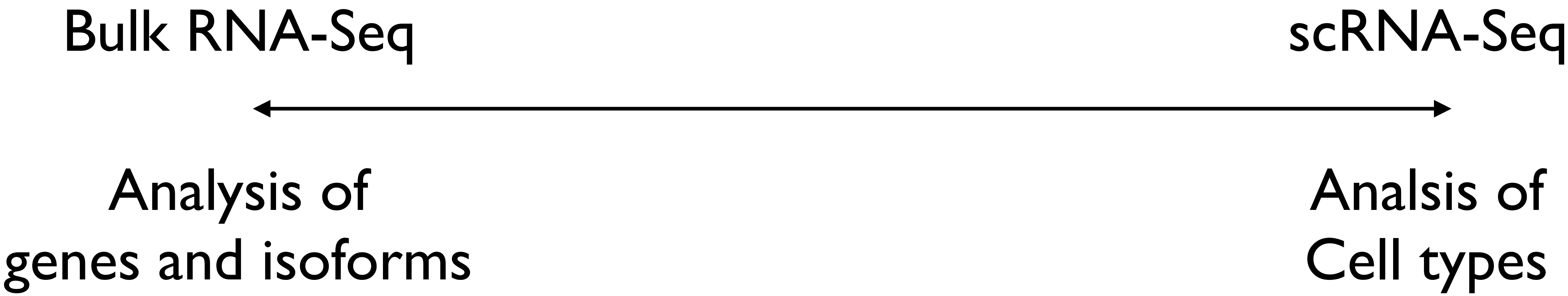
- The all profile different aspects of cell biology, ranging from “Identifying ribosome position”, over “RNA structure probing” to “long-range interaction of chromatin”

Continuos development and improvement of analysis tools

- A few years ago RNA-seq could only be used to find genes - now you have isoform resolution and analysis of alternative splicing
- The CAGE method was recently shown to also enable detection of active enhancers (http://www.nature.com/nature/journal/v507/n7493/full/nature12787.html)
- Systematic analysis of isoform switches

Single-Cell Sequencing

- Recent breakthroughs now allows us to do high throughput sequencing of single cells
- This really enables us to understand cell heterogeneity as well as the actual mechanisms behind diseases



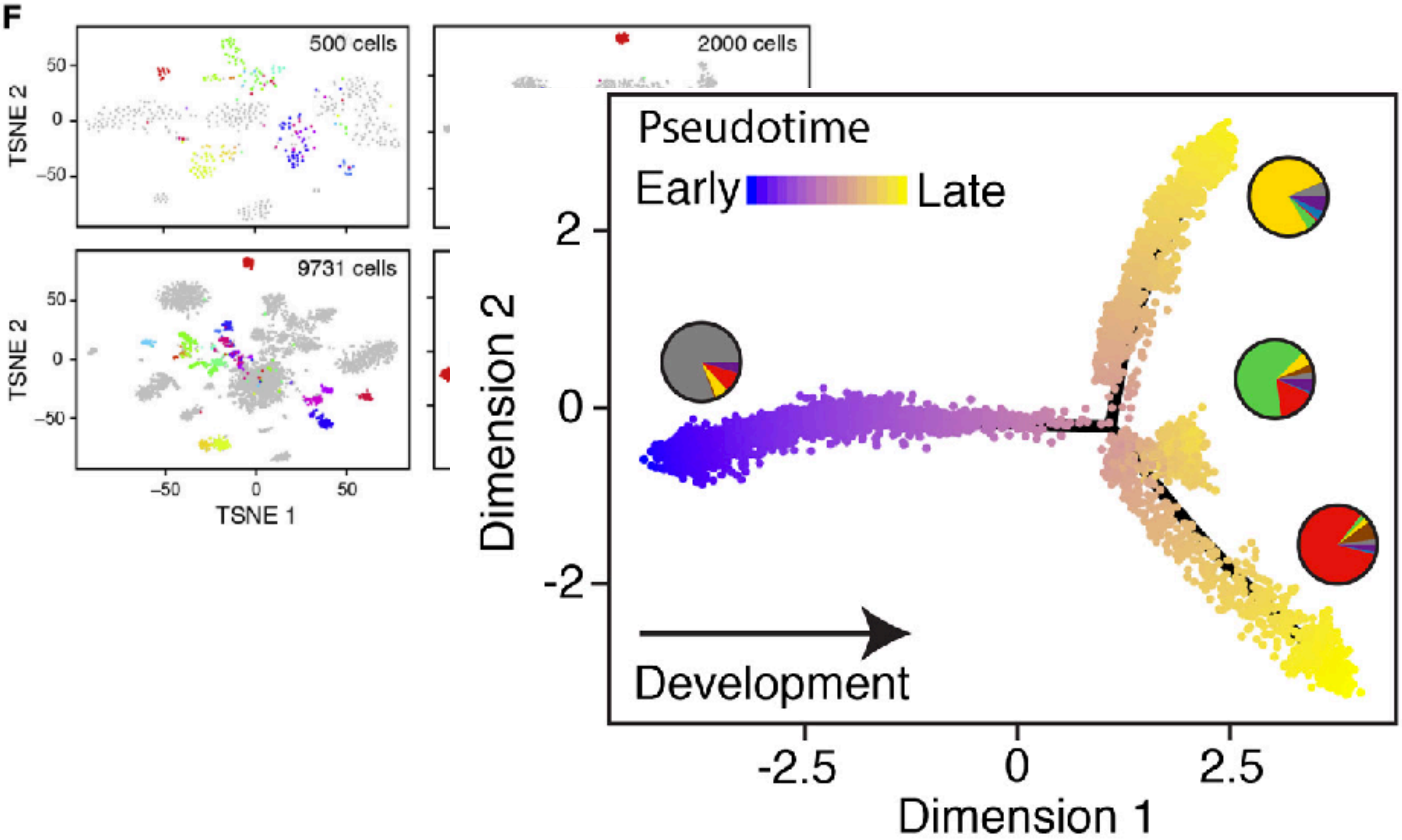
Bulk RNA-Seq

scRNA-Seq

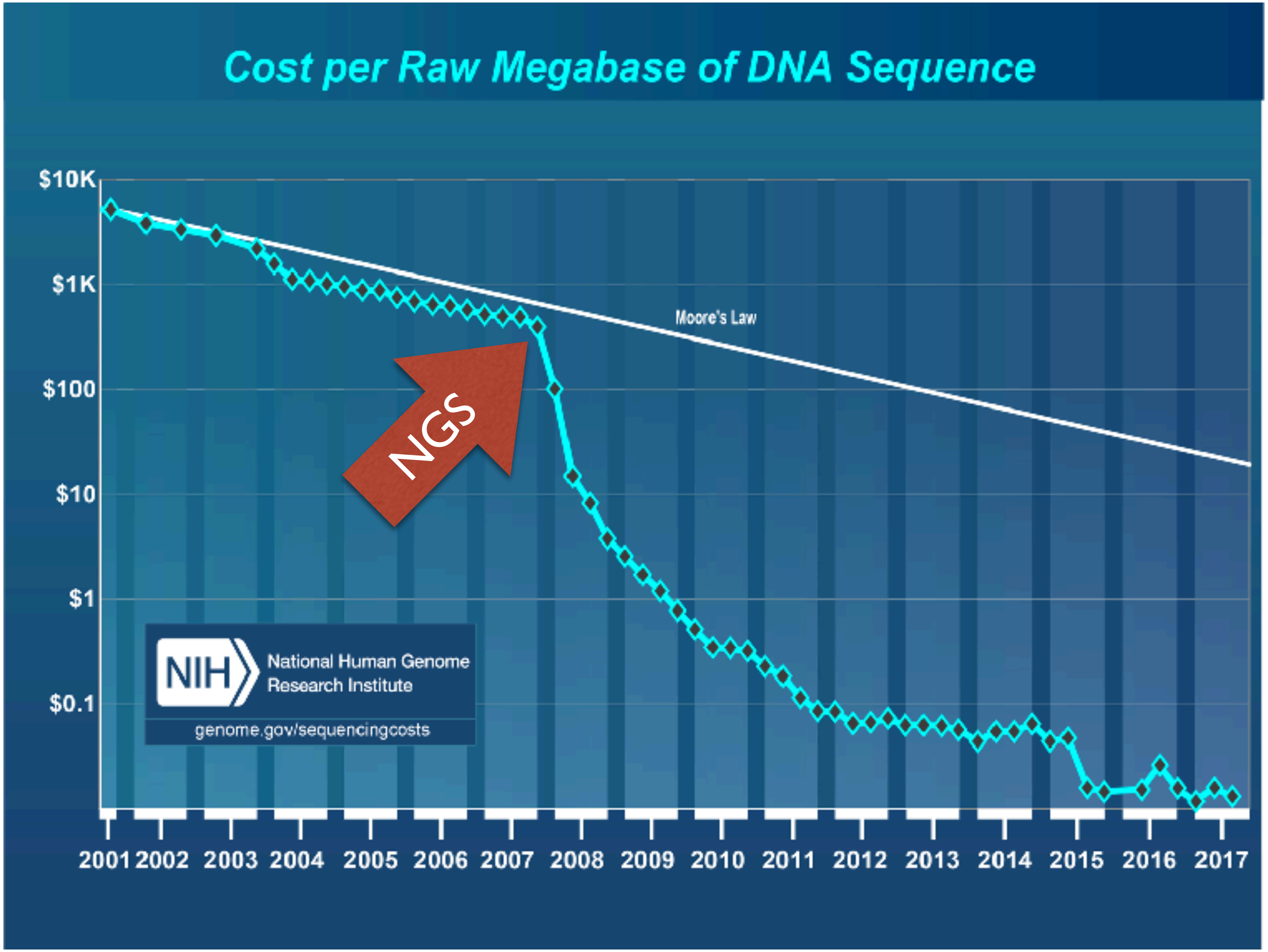
Analysis of
genes and isoforms

Analsis of
Cell types

Single-Cell Sequencing



Price of Sequencing



Summary

More and more sequenced based methods

+

Sequencing based methods become better and better

+

Analysis tools becomes better and better

+

Sequencing become cheaper and cheaper

=

High throughput methods is, and will continue to be even more so, a standard tool in all cell biology

The End