RNA-Seq

UCSC genome browser

- https://genome.ucsc.edu/
- Authored by Jim Kent
- Came out of human genome project
- Major browser for vertebrate data
- Specific instances exist

UCSC browser

- Select Genome
- Select Version
- Hide All
- Tracks
- Display options

Problem

Your lab work on the SCN5A gene. The SCN5A gene belongs to a family of genes that provide instructions for making sodium channels. The sodium channels produced from the SCN5A gene are abundant in heart (cardiac) muscle and play a major role in signaling the start of each heartbeat, and maintaining a normal heart rhythm. Your lab has identifed 5 SNPs in SCN5A in Human which are related to heart failure Your advisor as asked you to choose 2 SNPs for mouse testing. Choose 2 of the 5. Hint: The bed file for these snps is available in the folder Lect_04/SCN5A_hg19.bed.

Also find:

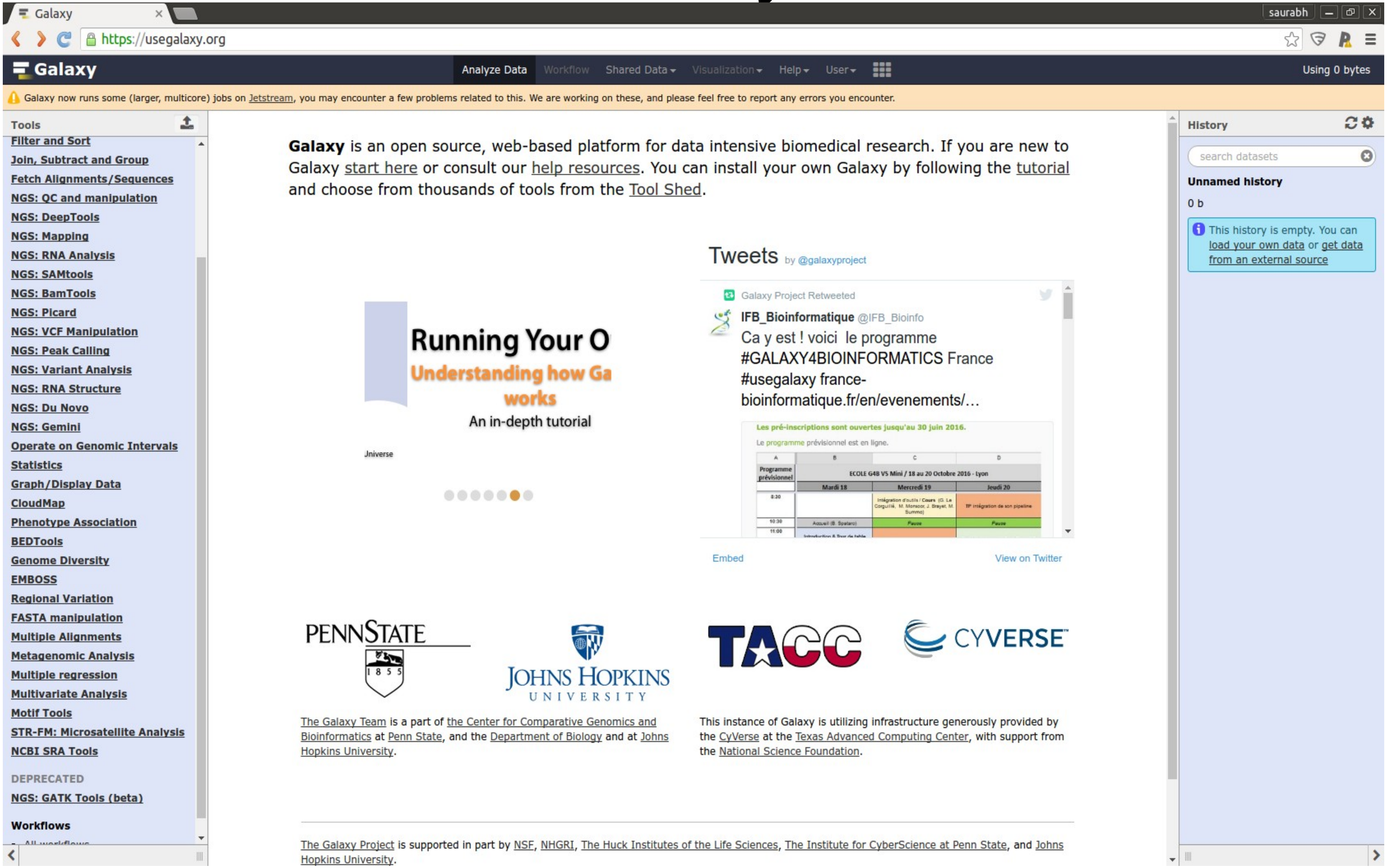
How many splice variants does SCN5A have ?

Do all of them code for a protein ?

Which Strand is SCN5A in ?

Disclaimer: This problem is completely imaginary! Any resemblance to real data is purely coincidental!

Galaxy



Galaxy now runs some (larger, multicore) jobs on Jetstream, you may encounter a few problems related to this. We are working on these, and please feel free to report any errors you encounter.

Tools

- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- NGS: QC and manipulation
- NGS: DeepTools
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAMtools
- NGS: BamTools
- NGS: Picard
- NGS: VCF Manipulation
- NGS: Peak Calling
- NGS: Variant Analysis
- NGS: RNA Structure
- NGS: Du Novo
- NGS: Gemini
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- CloudMap
- Phenotype Association
- BEDTools
- Genome Diversity
- EMBOSS
- Regional Variation
- FASTA manipulation
- Multiple Alignments
- Metagenomic Analysis
- Multiple regression
- Multivariate Analysis
- Motif Tools
- STR-FM: Microsatellite Analysis
- NCBI SRA Tools

DEPRECATED

- NGS: GATK Tools (beta)

Workflows

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Tweets by @galaxyproject

Galaxy Project Retweeted

IFB_Bioinformatique @IFB_Bioinfo

Ca y est ! voici le programme #GALAXY4BIOINFORMATICS France #usegalaxy france-bioinformatique.fr/en/evenements/…

Les pré-inscriptions sont ouvertes jusqu'au 30 juin 2016.

Le programme prévisionnel est en ligne.

A	B	C	D
Programme prévisionnel	ECOLE G4B V5 Mini / 18 au 20 Octobre 2016 - Lyon		
	Mardi 18	Mercredi 19	Jeudi 20
8:30		Intégration d'outils / Cours (G. Le Corguillé, M. Monsoor, J. Brayet, M. Summo)	TP intégration de son pipeline
10:30	Accueil (B. Spataro)	Pause	Pause
11:00	Introduction & Tour de table		

Embed View on Twitter

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, and the Department of Biology and at Johns Hopkins University.

This instance of Galaxy is utilizing infrastructure generously provided by the CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

History

search datasets

Unnamed history

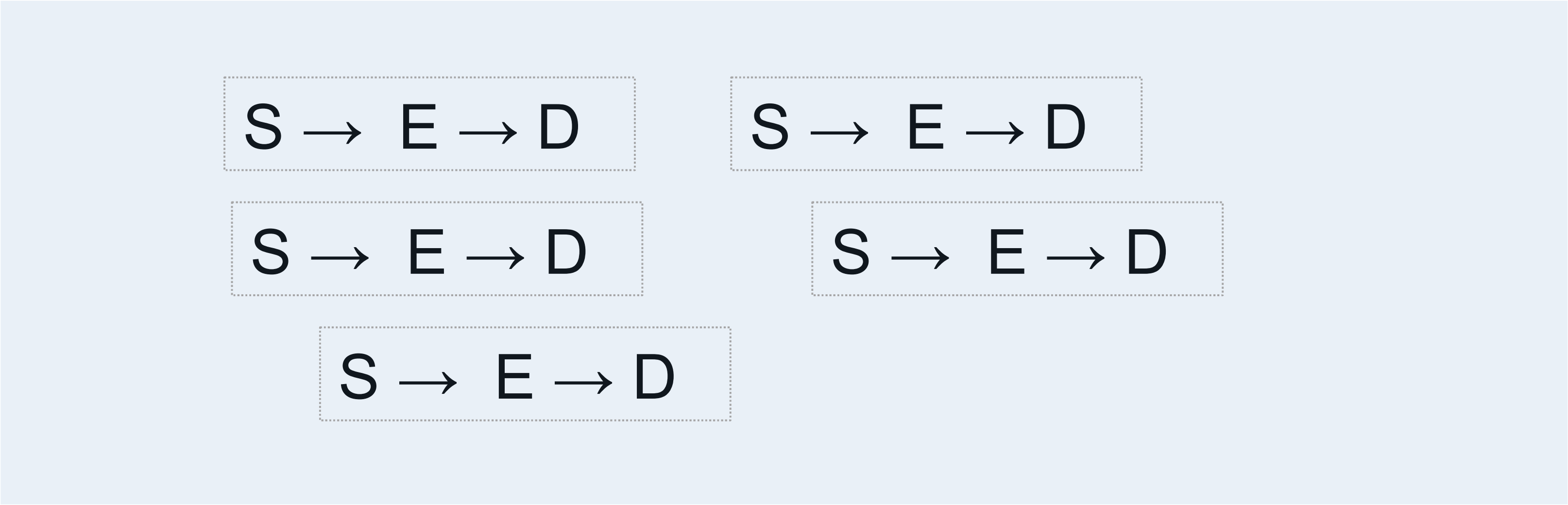
0 b

This history is empty. You can load your own data or get data from an external source

Exercise 1

- Create a Galaxy account
- Try **not** using TUM-ids instead use real name based email (firstname.lastname@tum.de) or any other email

Study Structure



S → E → D

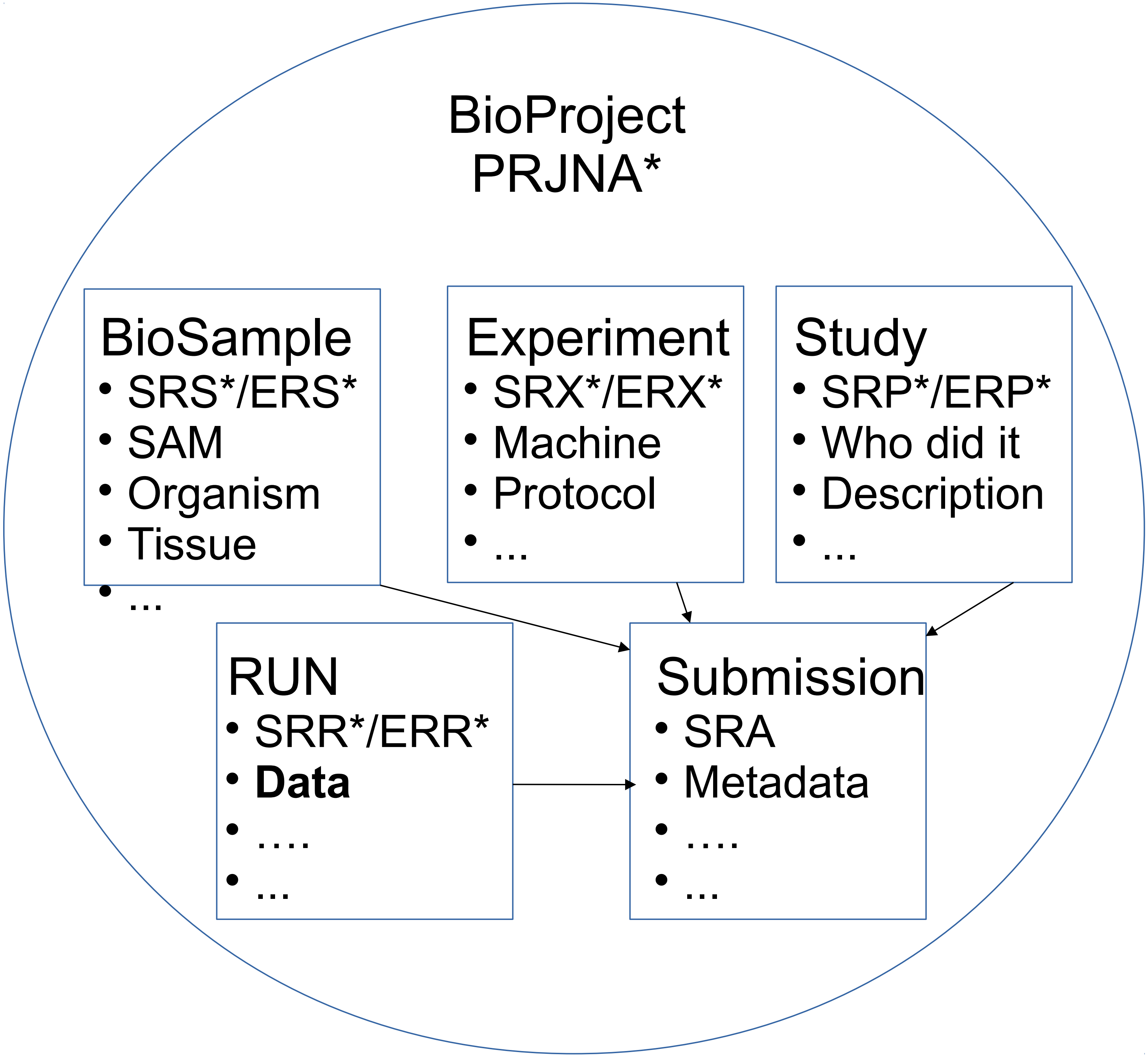
S → E → D

S → E → D

S → E → D

S → E → D

Sample → Experiment → Data



BioProject
PRJNA*

BioSample

- SRS*/ERS*
- SAM
- Organism
- Tissue
- ...

Experiment

- SRX*/ERX*
- Machine
- Protocol
- ...

Study

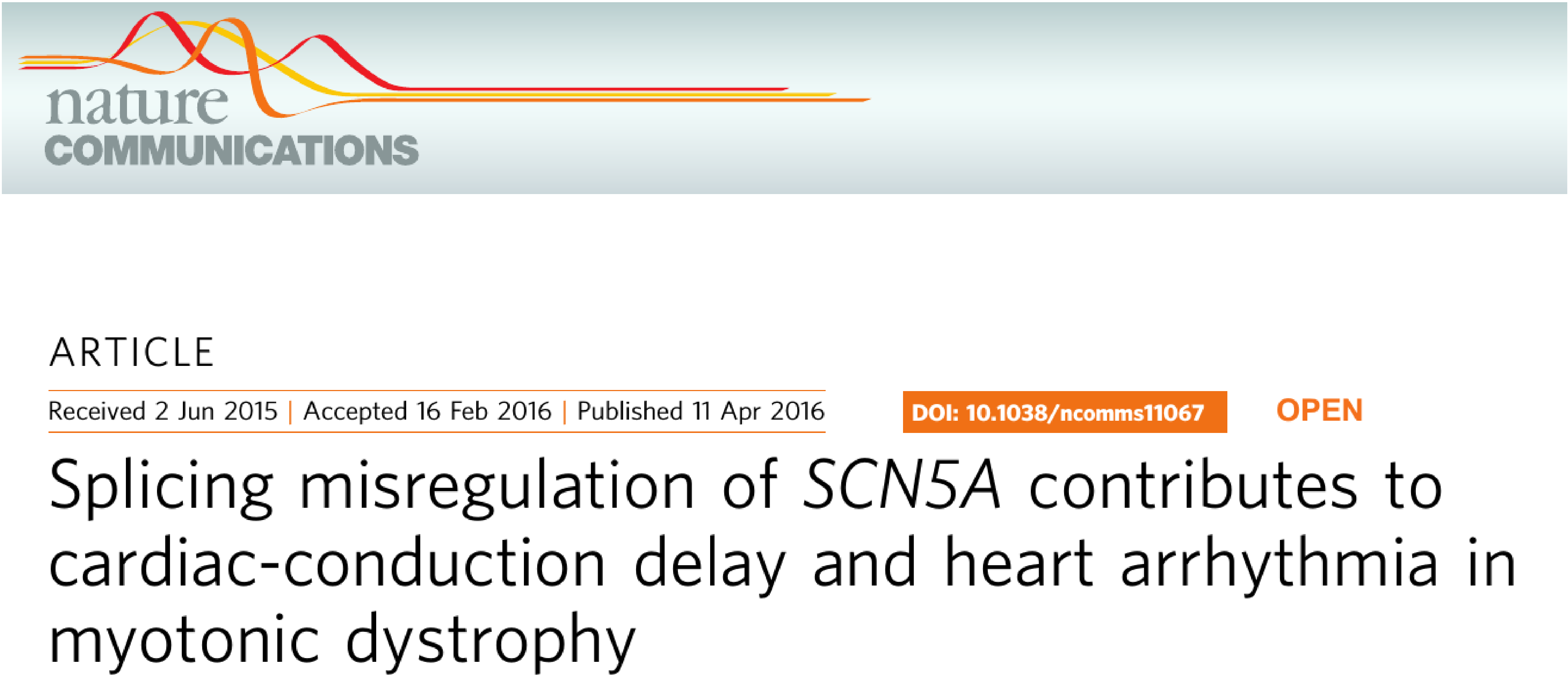
- SRP*/ERP*
- Who did it
- Description
- ...

RUN

- SRR*/ERR*
- **Data**
- ….
- ...

Submission

- SRA
- Metadata
- ….
- ...



nature COMMUNICATIONS

ARTICLE

Received 2 Jun 2015 | Accepted 16 Feb 2016 | Published 11 Apr 2016

DOI: 10.1038/ncomms11067

OPEN

Splicing misregulation of *SCN5A* contributes to cardiac-conduction delay and heart arrhythmia in myotonic dystrophy

ARTICLE

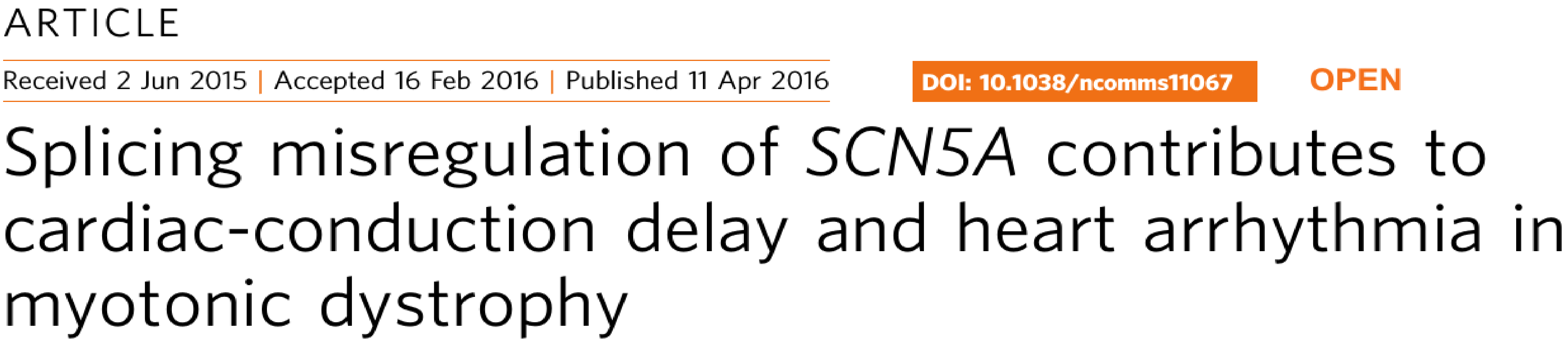
Received 2 Jun 2015 | Accepted 16 Feb 2016 | Published 11 Apr 2016

DOI: 10.1038/ncomms11067

OPEN

Splicing misregulation of *SCN5A* contributes to cardiac-conduction delay and heart arrhythmia in myotonic dystrophy

Myotonic dystrophy (DM) is caused by the expression of mutant RNAs containing expanded CUG repeats that sequester muscleblind-like (MBNL) proteins, leading to alternative splicing changes. Cardiac alterations, characterized by conduction delays and arrhythmia, are the second most common cause of death in DM. Using RNA sequencing, here we identify novel splicing alterations in DM heart samples, including a switch from adult exon 6B towards fetal exon 6A in the cardiac sodium channel, *SCN5A*. We find that MBNL1 regulates alternative splicing of *SCN5A* mRNA and that the splicing variant of *SCN5A* produced in DM presents a reduced excitability compared with the control adult isoform. Importantly, reproducing splicing alteration of *Scn5a* in mice is sufficient to promote heart arrhythmia and cardiac-conduction delay, two predominant features of myotonic dystrophy. In conclusion, misregulation of the alternative splicing of *SCN5A* may contribute to a subset of the cardiac dysfunctions observed in myotonic dystrophy.



ARTICLE

Received 2 Jun 2015 | Accepted 16 Feb 2016 | Published 11 Apr 2016

DOI: 10.1038/ncomms11067

OPEN

Splicing misregulation of *SCN5A* contributes to cardiac-conduction delay and heart arrhythmia in myotonic dystrophy

PRJNA280990

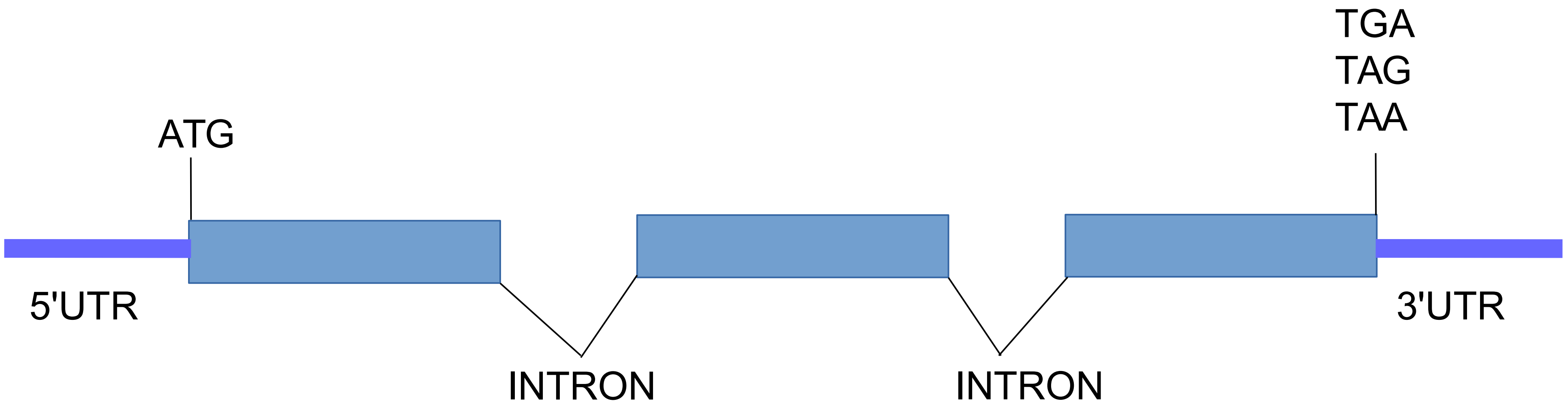
- Locate the study in SRA run selector
- How many case and control samples
- How many technical replicates ?

RNA-Seq

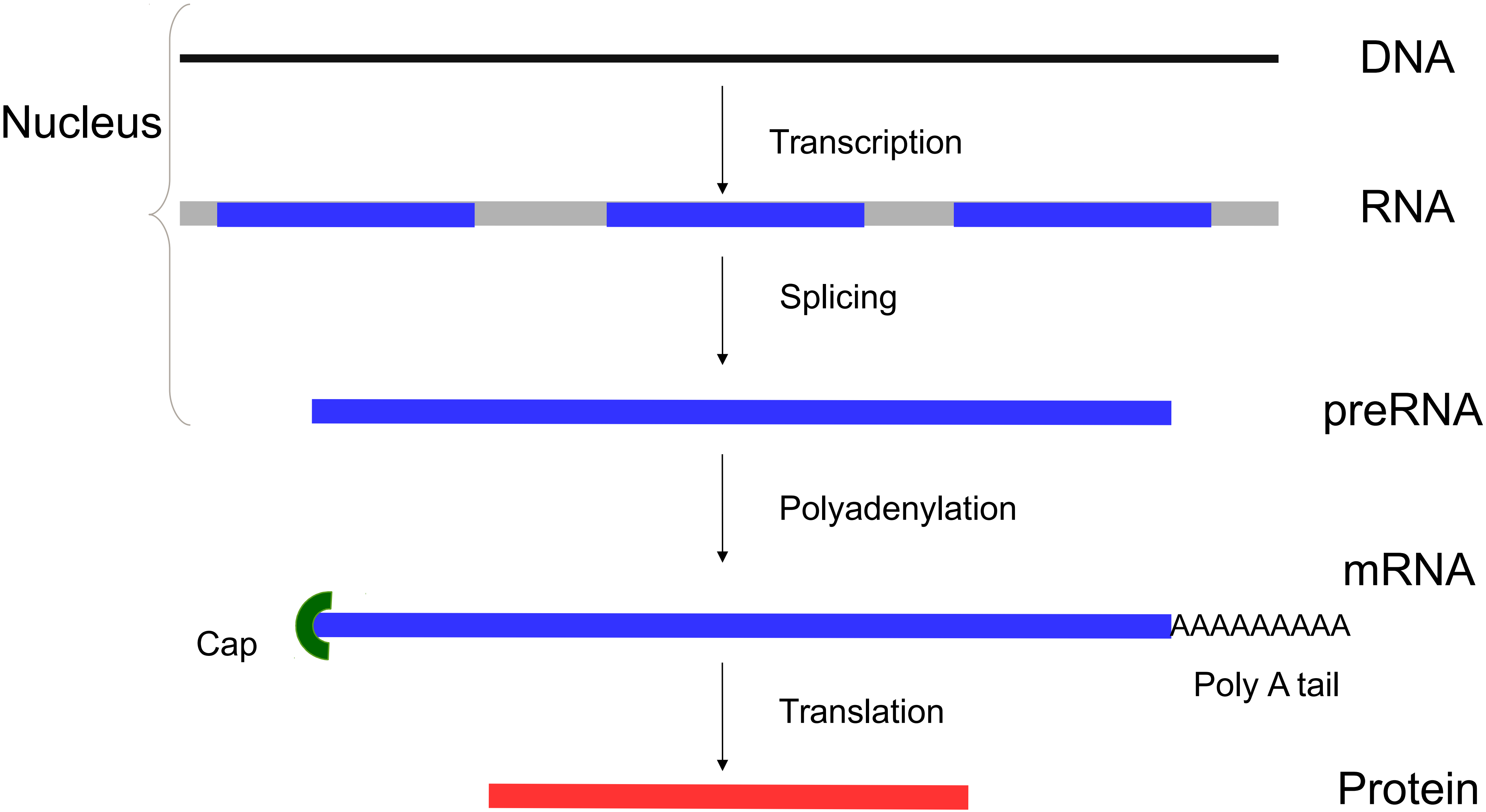
What can be done

- Unbiased Expression profiling
- Differential expression analysis
- Gene discovery
- Splice variant Profiling/Discovery
- Regulatory network
- Transcription start site discovery (CAGE)
- Single cell Genomics
- Strand specific RNA-Seq
- …..

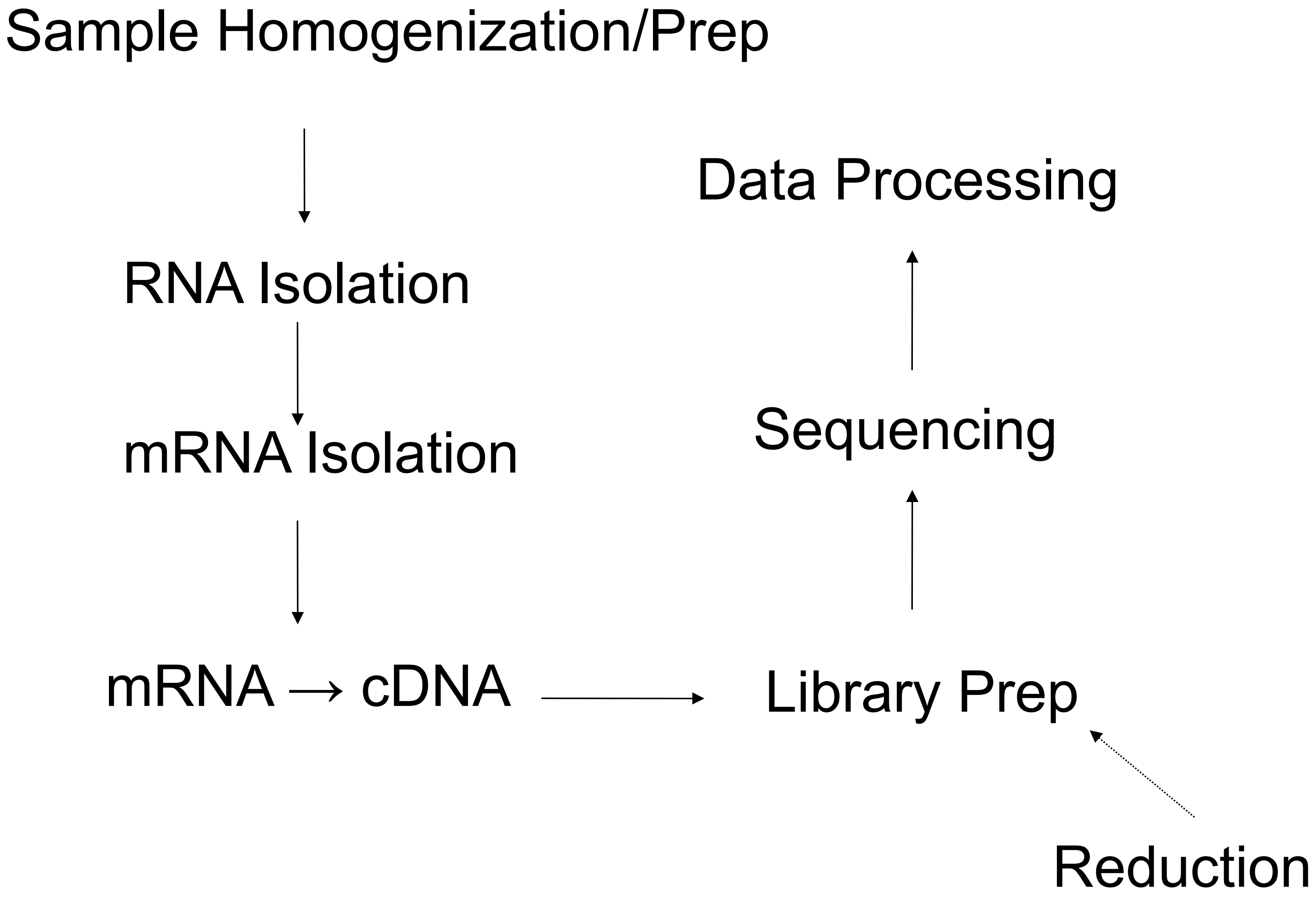
Canonical Gene



Central Dogma



RNA-Seq Protocol



Sample Homogenization/Prep

RNA Isolation

mRNA Isolation

mRNA → cDNA

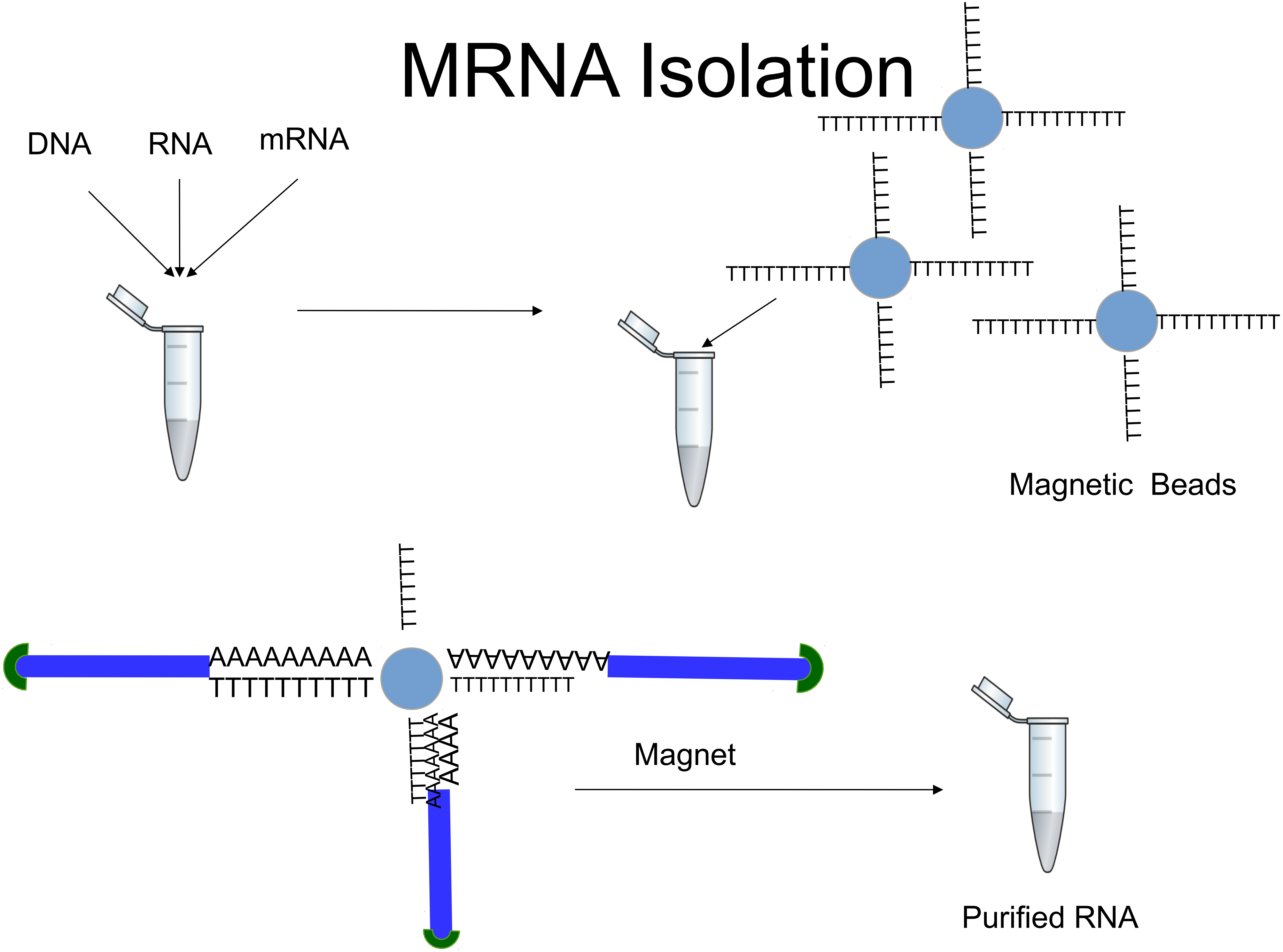
Library Prep

Reduction

Sequencing

Data Processing

MRNA Isolation



DNA

RNA

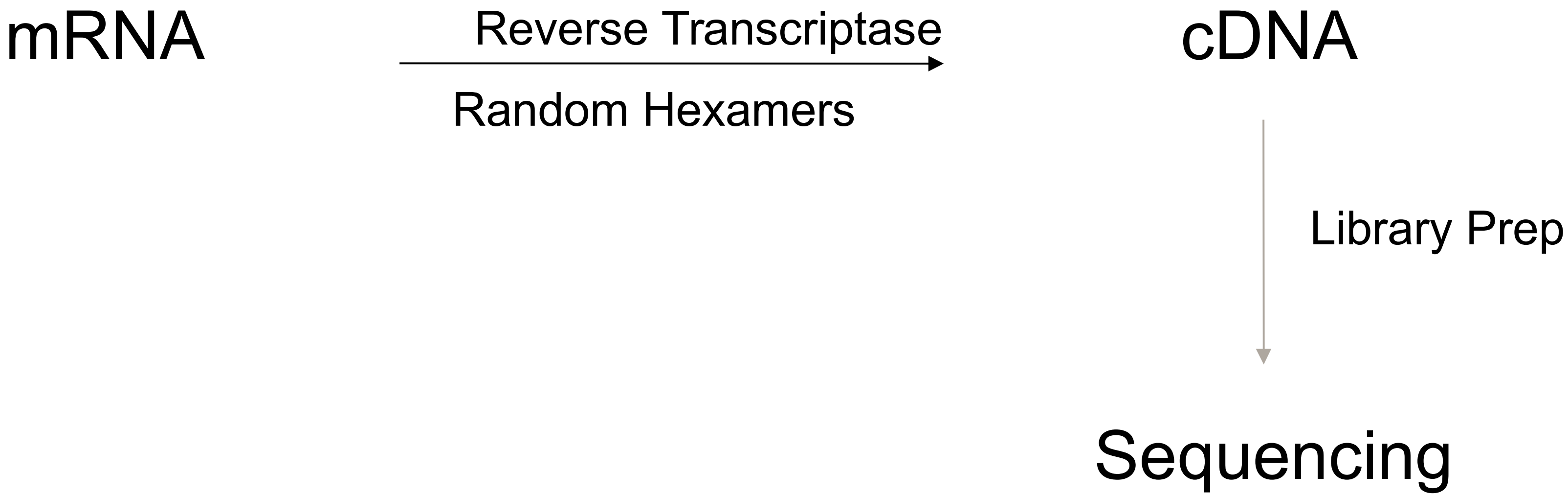
mRNA

Magnetic Beads

Magnet

Purified RNA

mRNA → cDNA



mRNA

Reverse Transcriptase

Random Hexamers

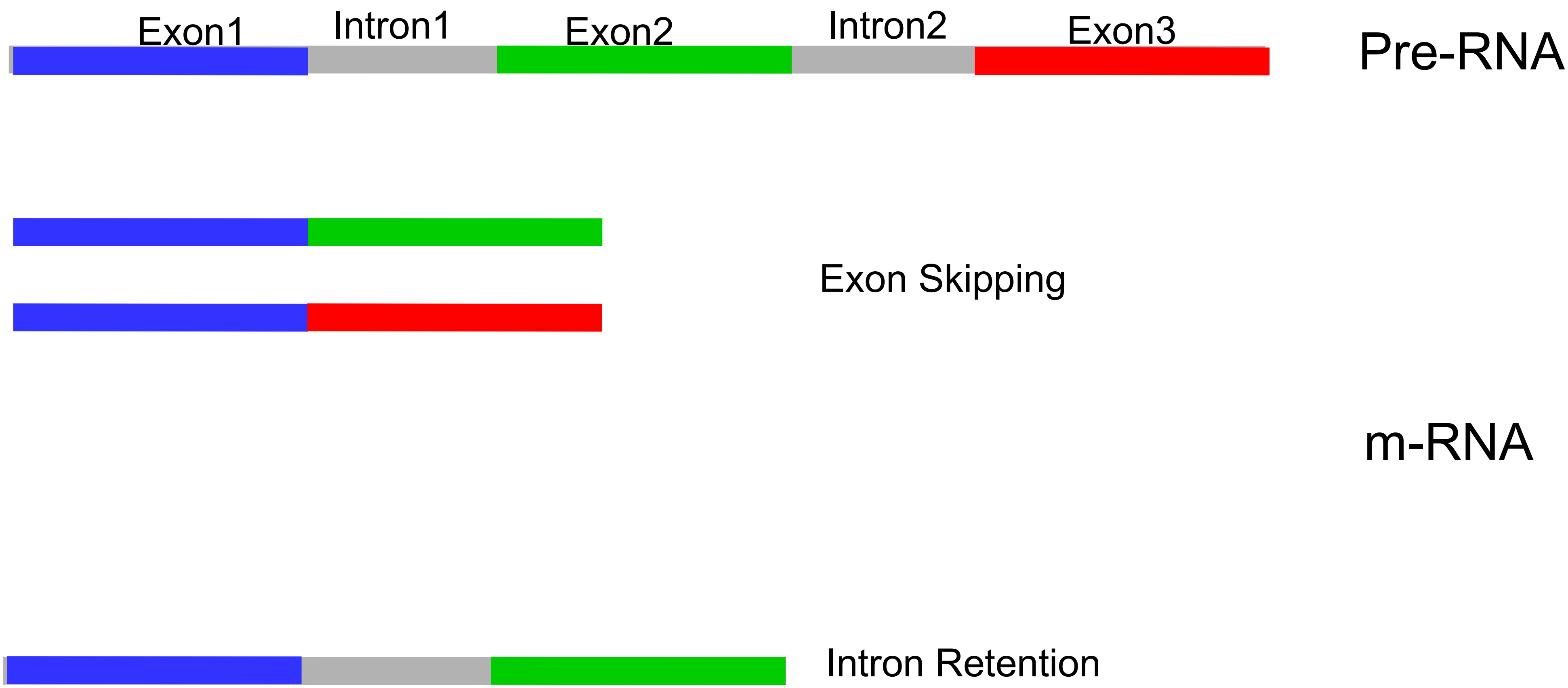
cDNA

Library Prep

Sequencing

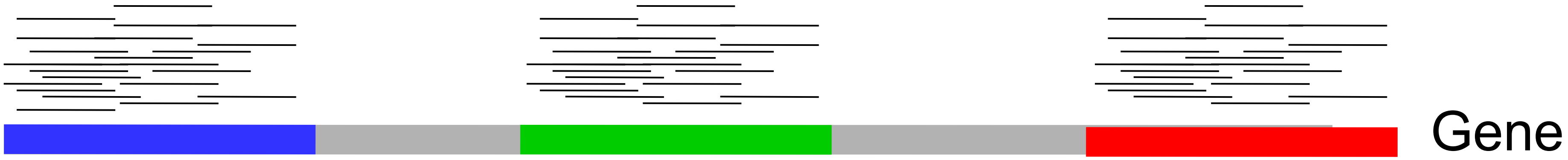
RNA-Seq is actually cDNA-Seq

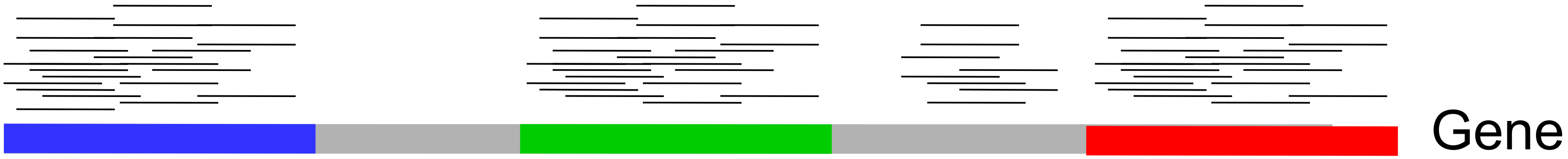
Splicing


Exon1
Intron1
Exon2
Intron2
Exon3
Pre-RNA
Exon Skipping
m-RNA
Intron Retention

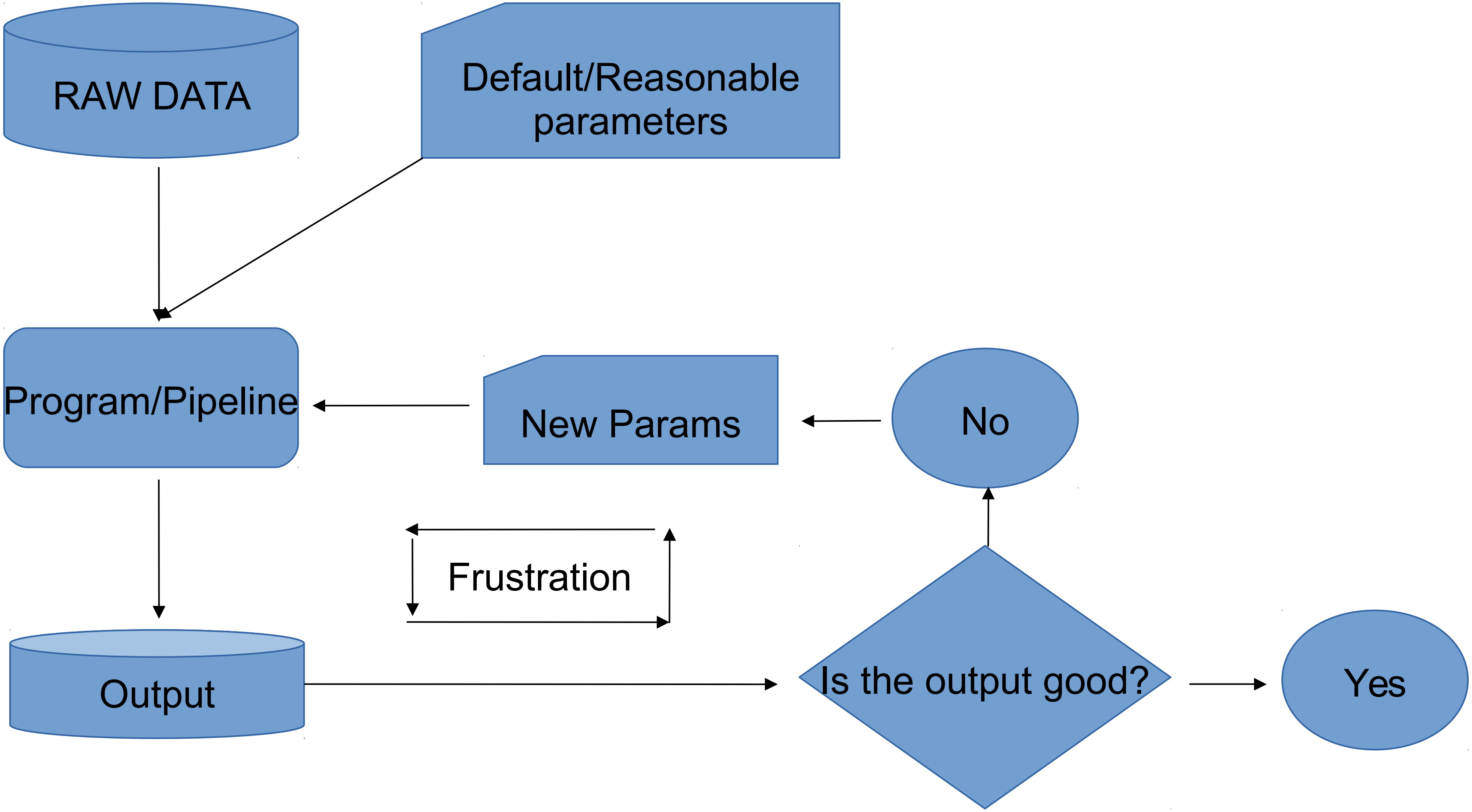
Problem

Which case has an intron retention event ?





Gold Standards



RAW DATA

Default/Reasonable parameters

Program/Pipeline

New Params

No

Frustration

Output

Is the output good?

Yes

Deciding Factors

- Study design
  - Differing read lengths for case and control
  - Differing insert sizes
  - Cases and Controls in different flowcells
  - Differing Library prep
- Technical
  - Differing FastQC results for cases and controls
  - Low mapping in one/more samples

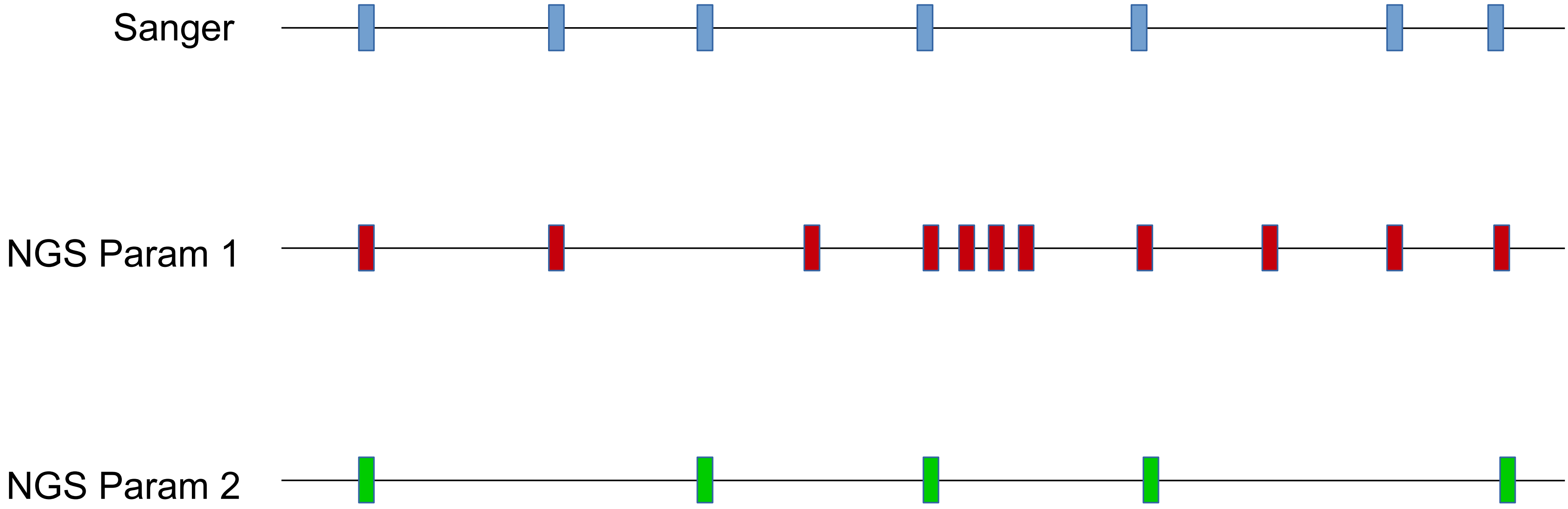
Deciding Factors

- Biological
  - Ts/Tv
  - Previous Knowledge
  - Hardy-Weinberg Eq
  - Too many non-sense mutations
  - Heterozygous SNPs
  - Unexpected Read Mappings
  - Housekeeping genes

Gold Standards

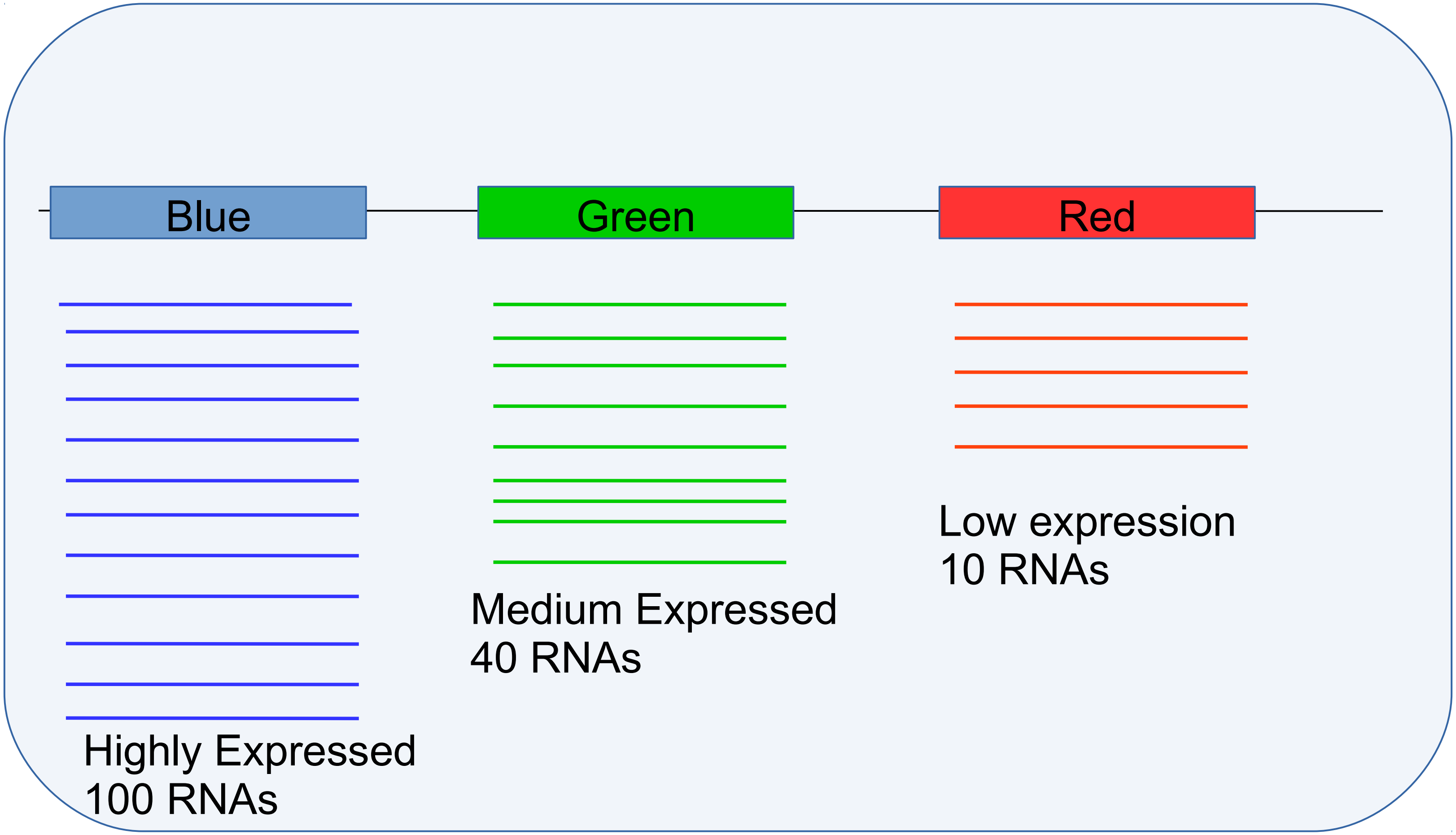
- Sanger Sequencing
- RT-PCR
- Calculate False Positives and False Negatives

Problem

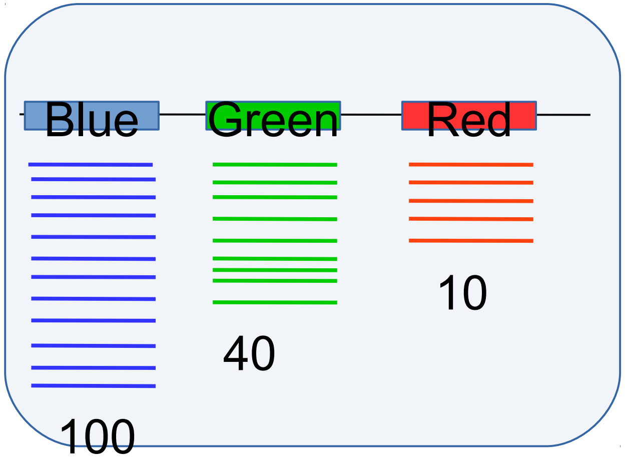


Identify the number of false positives and false negatives in both cases and try to guess the possible problems and their solutions

Sampling Issue with RNA-Seq



Problem

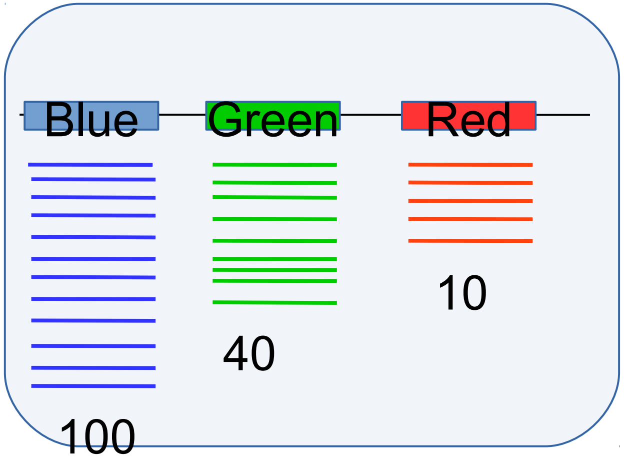


You sample 20 reads from the cell. Assuming that all 3 genes R G and B have same length. How many reads do you expect for each gene ?

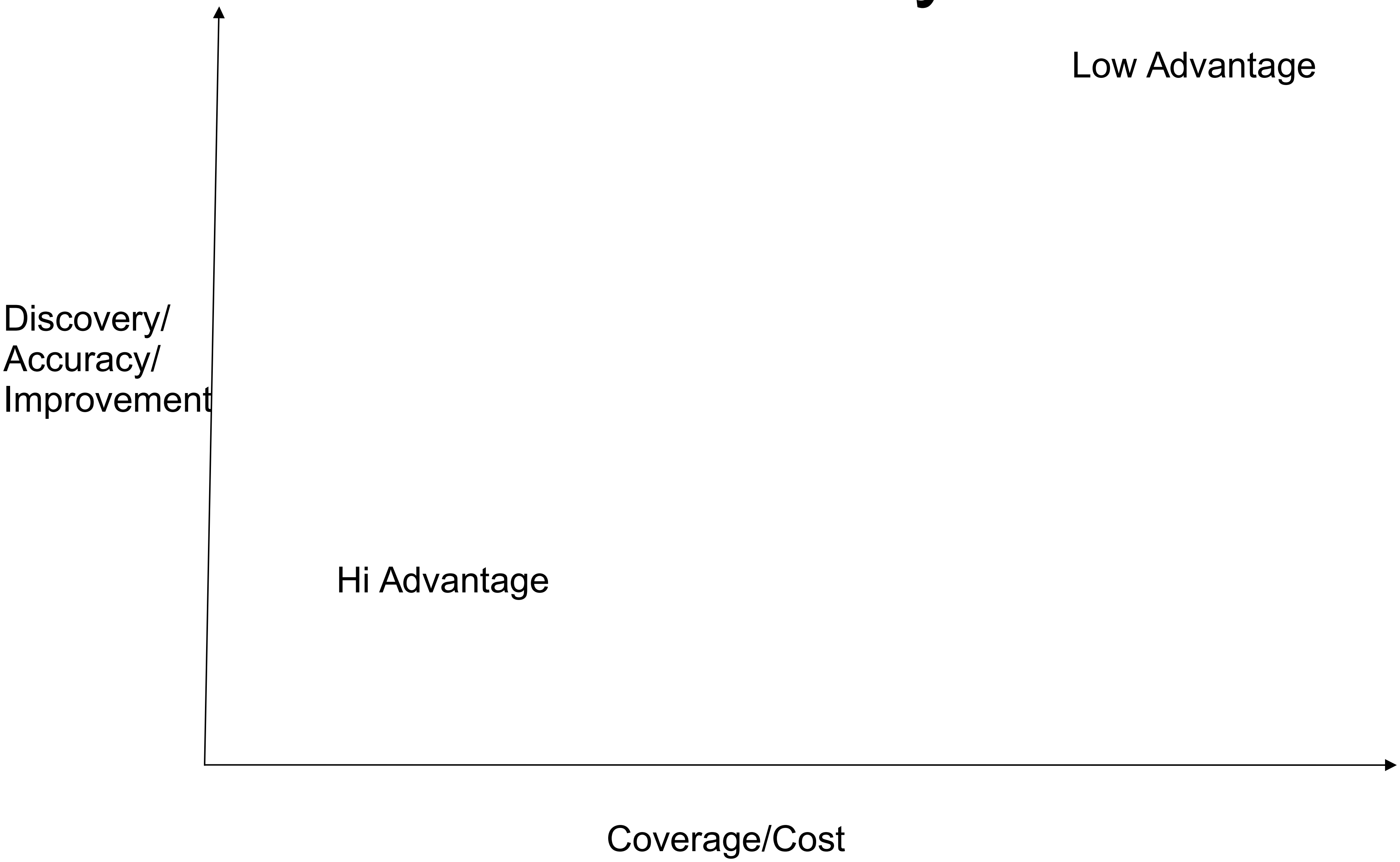
	20 Reads	50 Reads
A		
B		
C		

To Do

There is a file shuffle.sh in you Home directory. It randomly takes reads from each gene and prints the counts. The command is ./shuffle.sh n . Where N is the number of reads to sample. Run it for sample size 10, 50 ,100.



Coverage vs Discovery and accuracy



Low Advantage

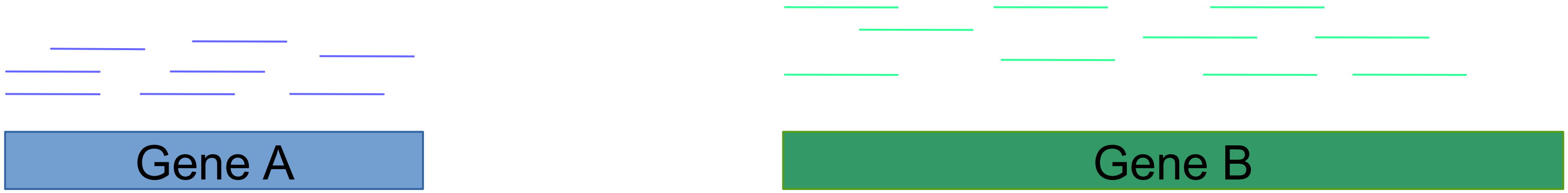
Discovery/
Accuracy/
Improvement

Hi Advantage

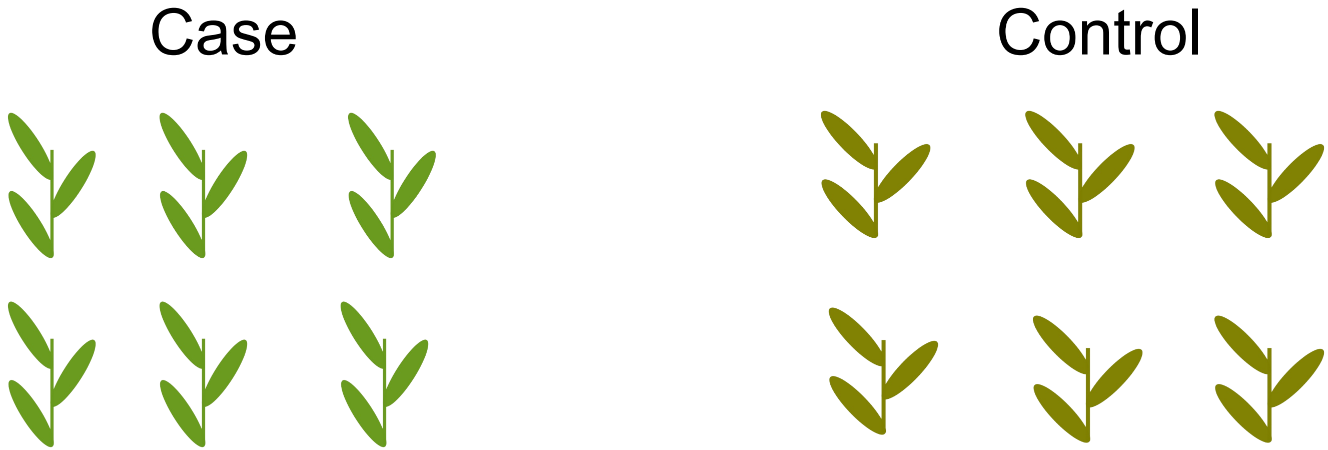
Coverage/Cost

Problem

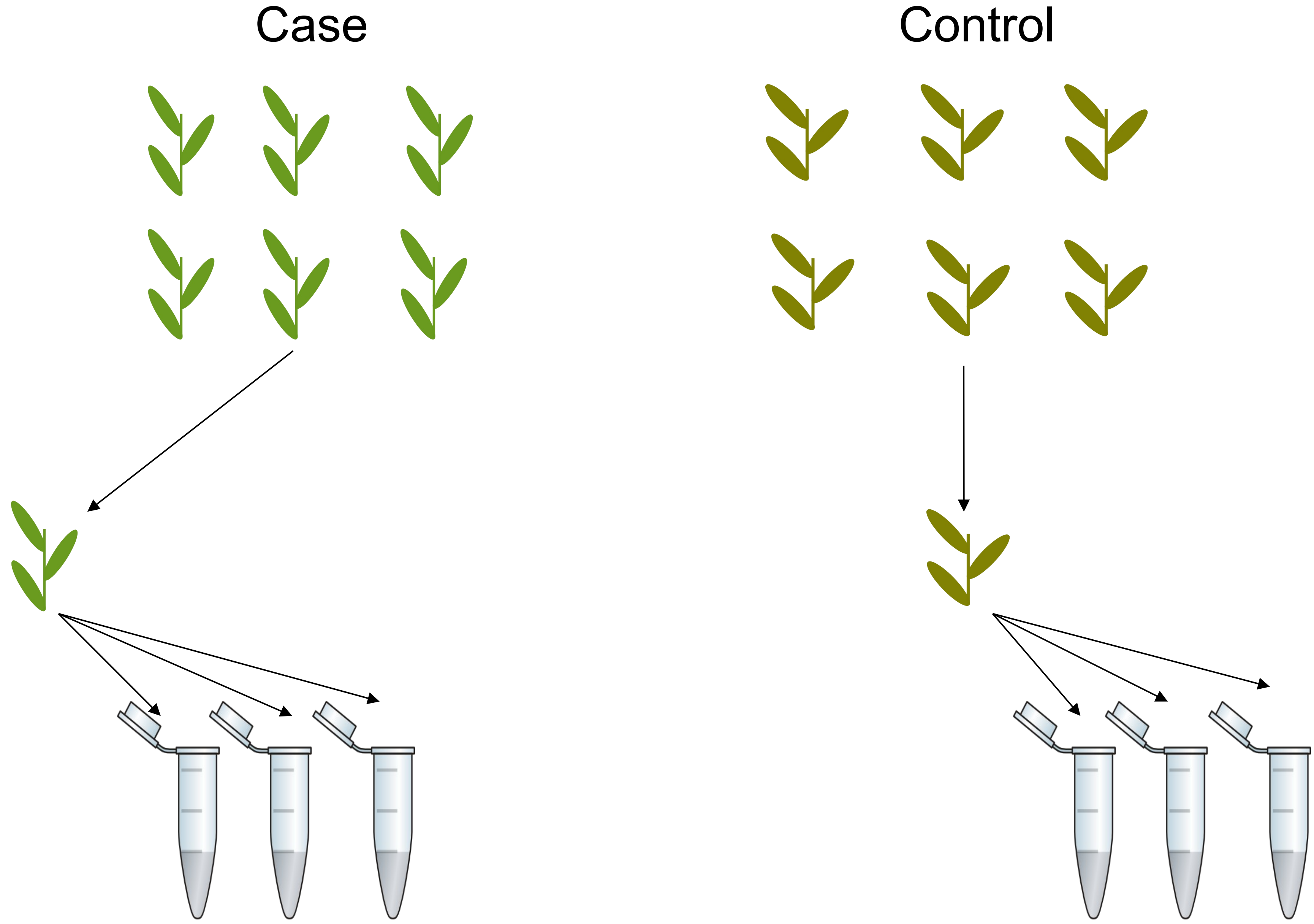
Two genes (A and B) were sequenced in the same library. Gene B is twice as long as gene A. Gene A has 100 reads mapped to it. How many reads would you expect for gene B to map ?



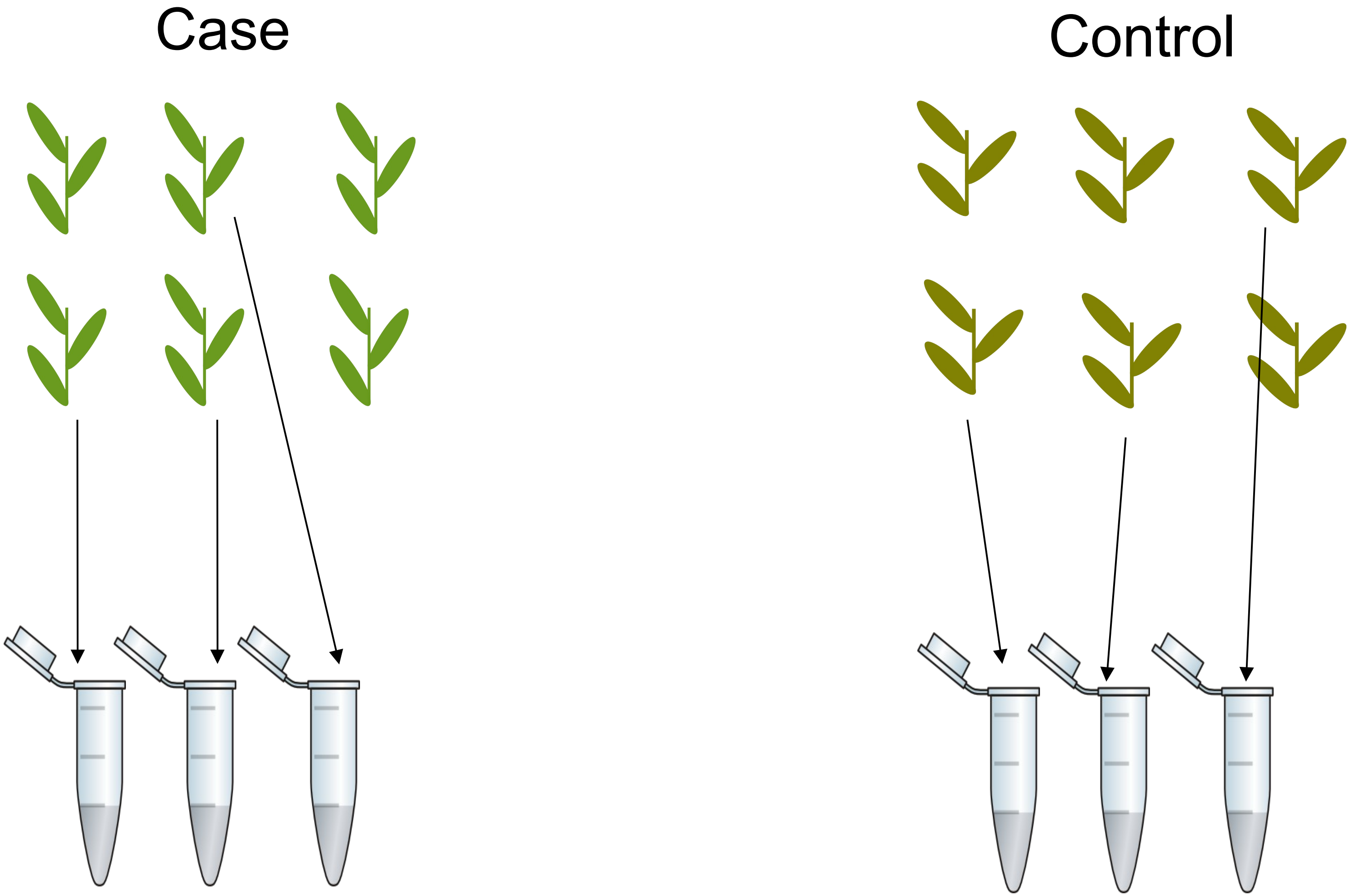
Replication



Technical Replication



Biological Replication



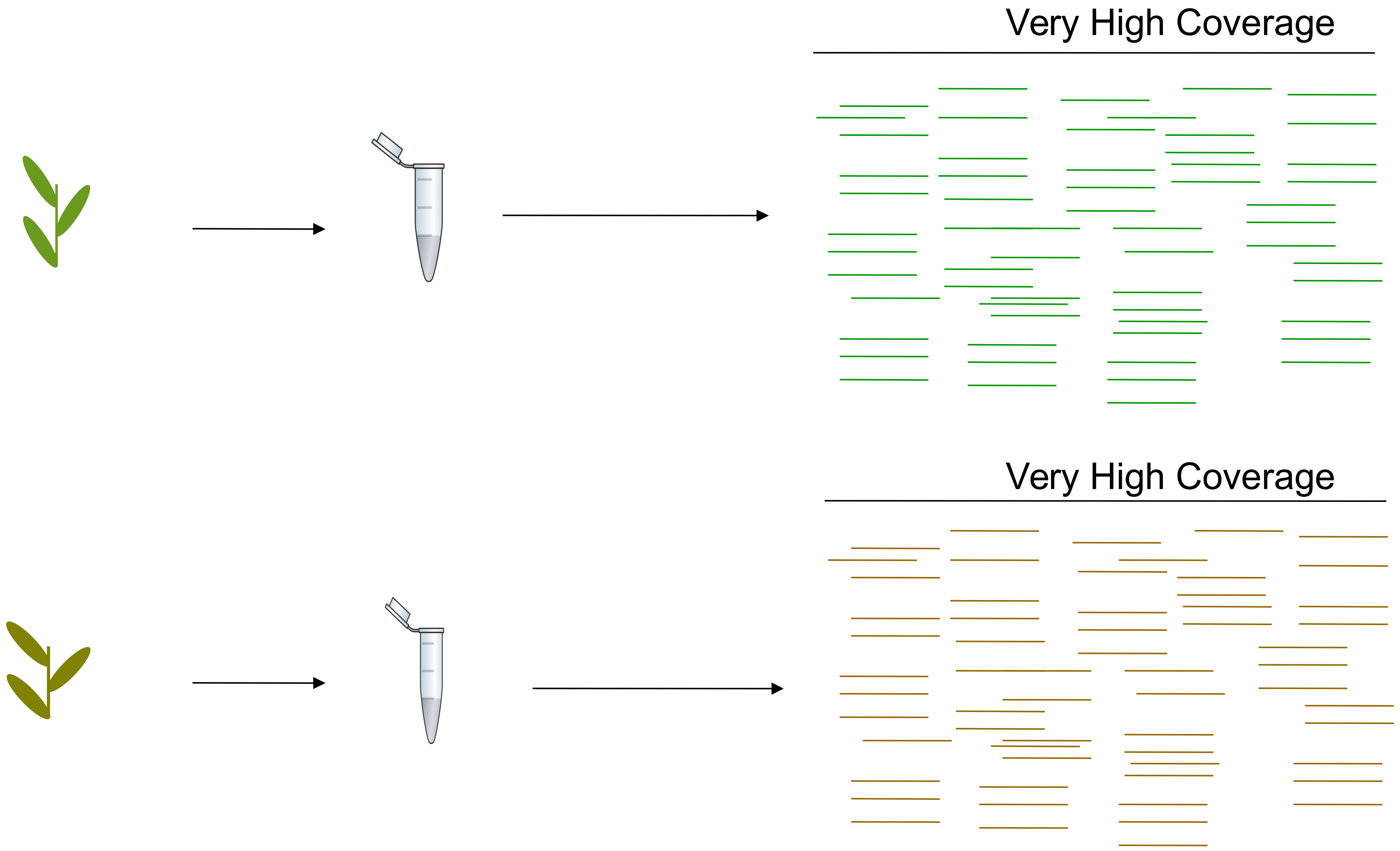
Case

Control

Replication

- Microarrays need technical replication
- Technical replicates for RNA-Seq are highly correlated
- Technical replication is generally not required
- Biological replication is considered good for differential expression
- Pooling by barcoding is highly recommended

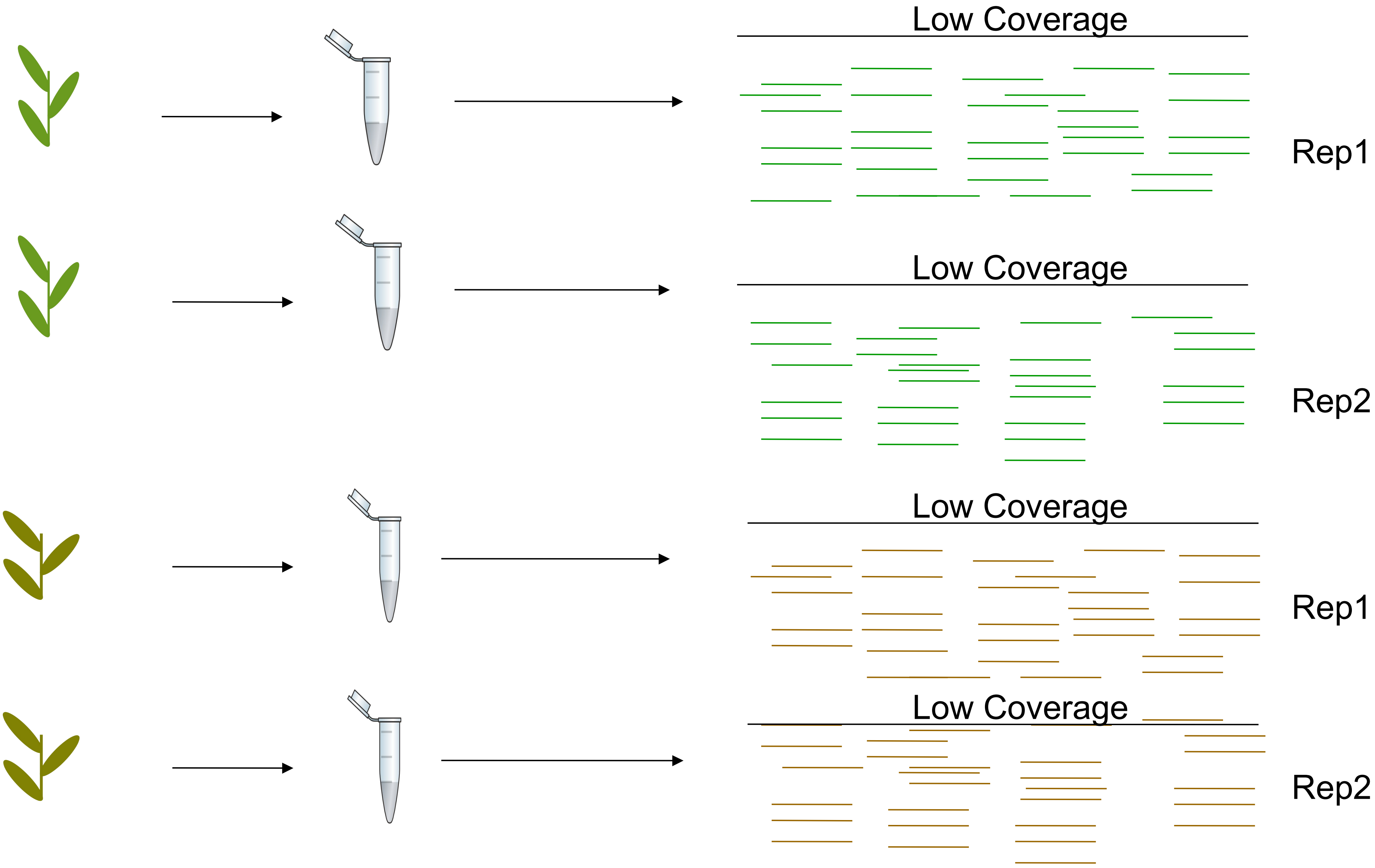
Coverage vs Replicates



Very High Coverage

Very High Coverage

Coverage vs Replicates



Which One to Choose ??

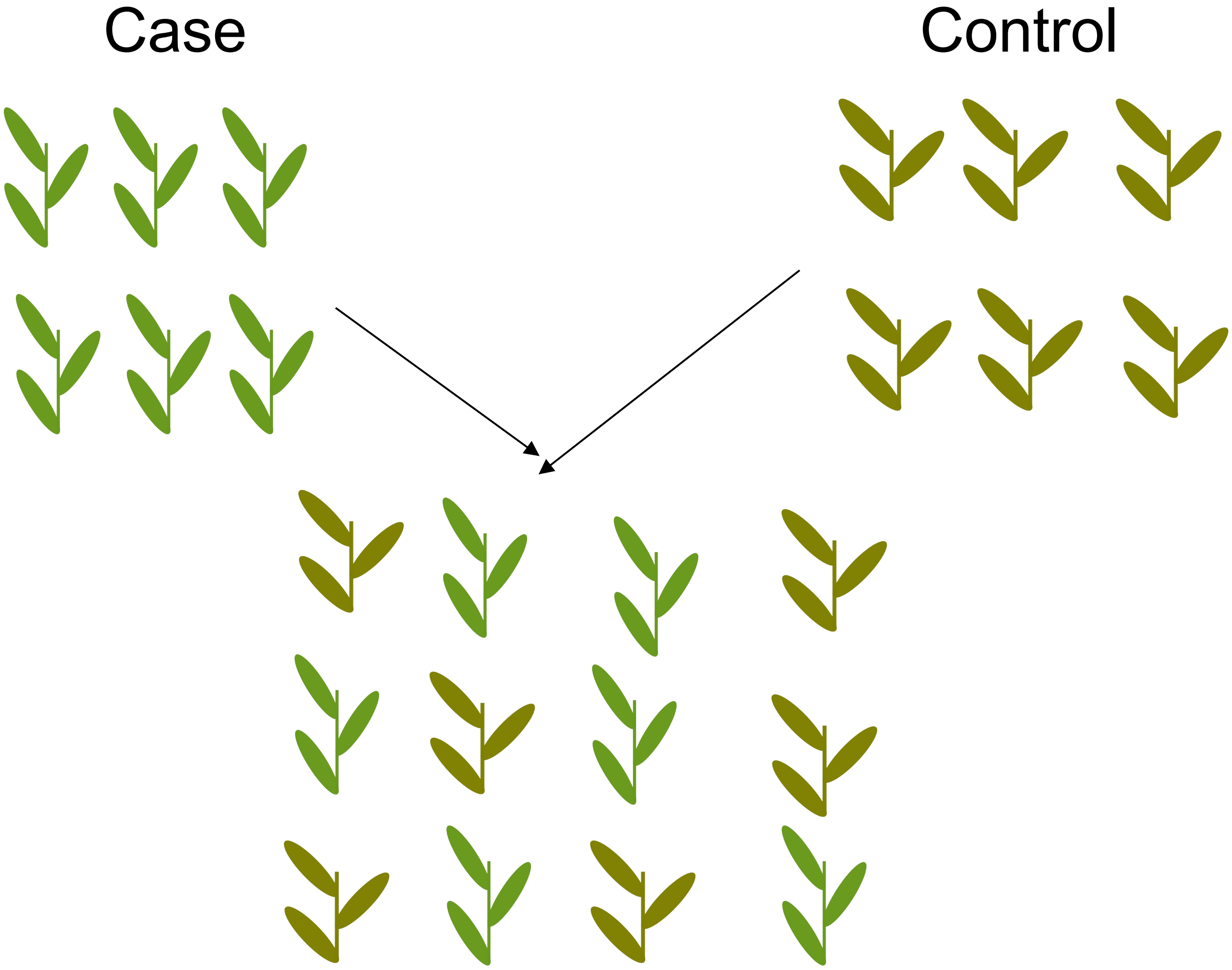
Think Over it!

Coverage vs Replicates

- How many results you want (Sensitivity) ?
- Is the project exploratory ?
- Monitory Considerations (Barcoding costs money)
- Type of sample
- Noise Level
- Quality of annotation
- More replicates is better
- Typically 3 replicates (more the better)

Randomizing

- Gene expression is very noisy
- Very plastic to conditions



Case

Control

Problem

You have 6 (C1 …. C6) case and 6 control samples (O1.... O6). You also have access to 3 lanes and 4 barcodes. How would you arrange the samples ?

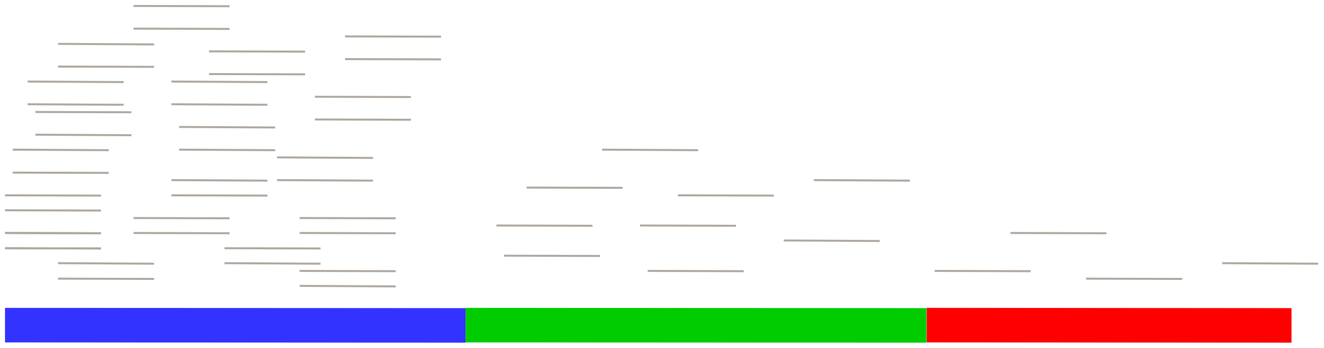
	Lane1	Lane2	Lane3
B1			
B2			
B3			
B4			

Checking for Splicing

- Discovery of splice variants
- Differential splicing between case and control

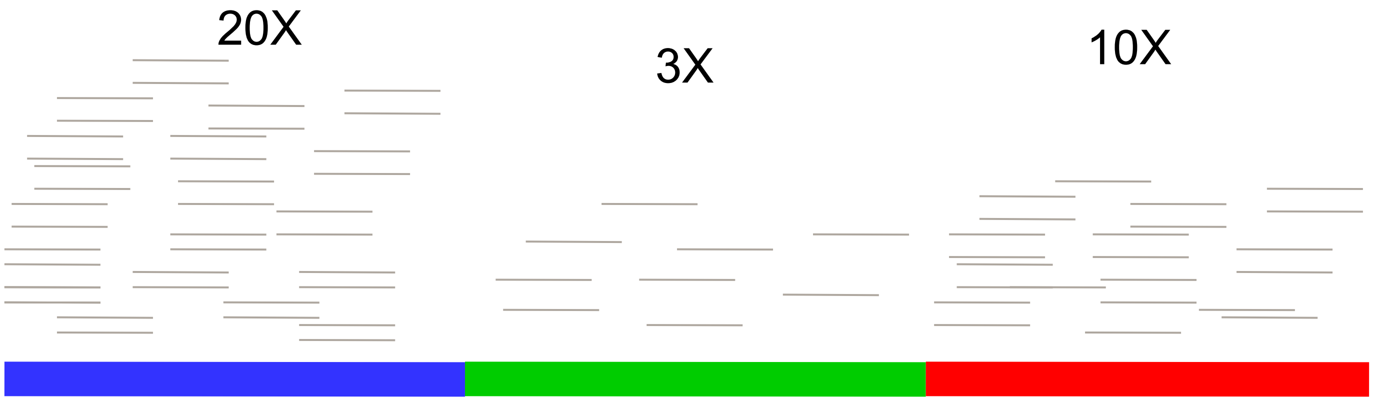
Coverage

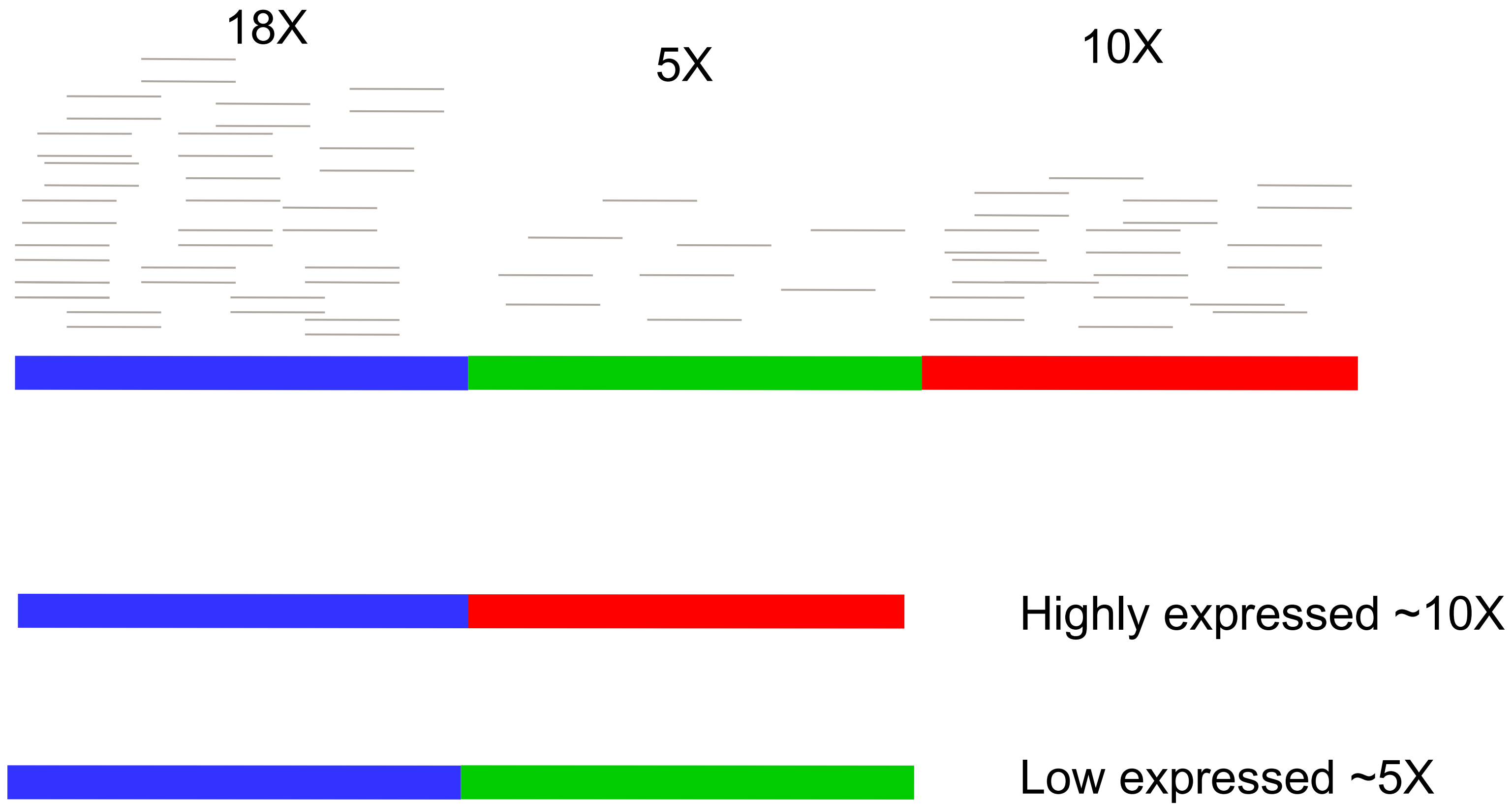
- Splice variants typically have varying expression
- Coverage can be used to predict a splicing event
- Tissue/Sample wise loss of coverage in an exon
- Differential splicing has been linked to variety of biological effects



Problem

Can You guess the splice variants and expression of each variant (Qualitatively) ?

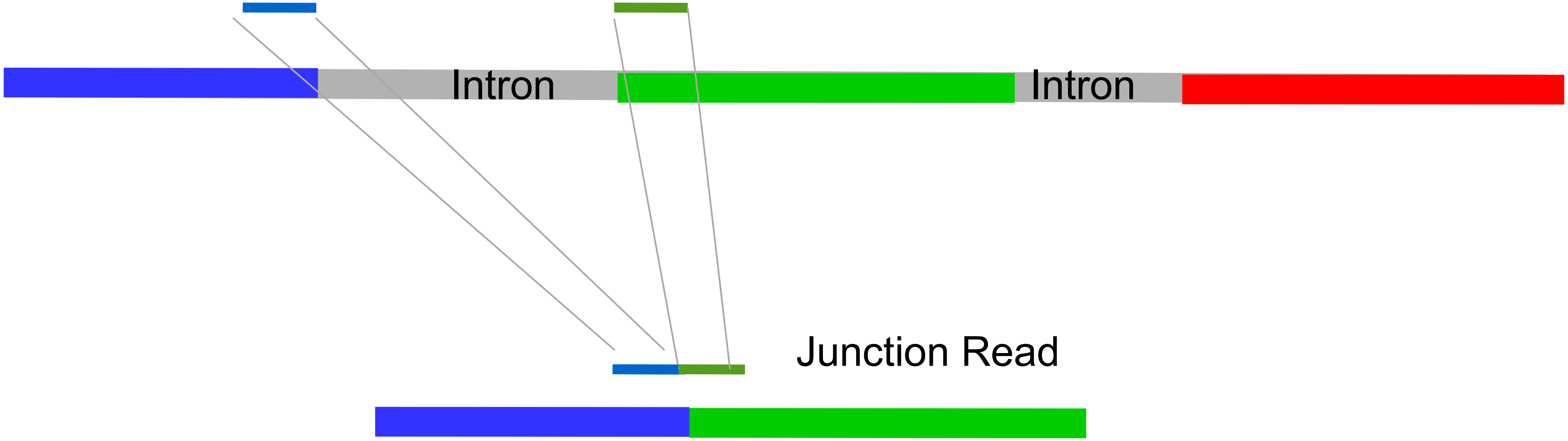




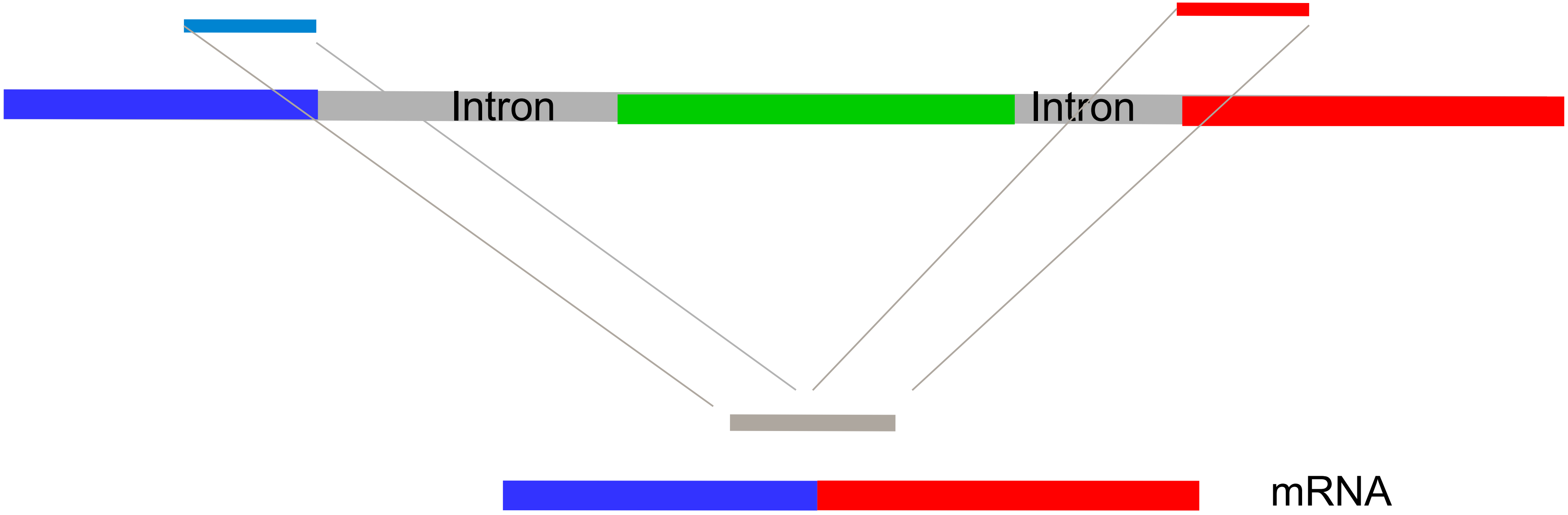
Highly expressed ~10X

Low expressed ~5X

Second Approach

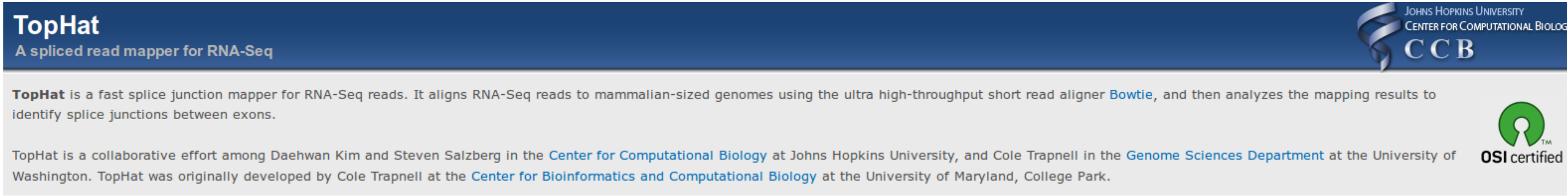


Junction Read



A well mapped Junction Read is a strong indicator of a splicing event

TOPHAT Aligner



TopHat

A spliced read mapper for RNA-Seq

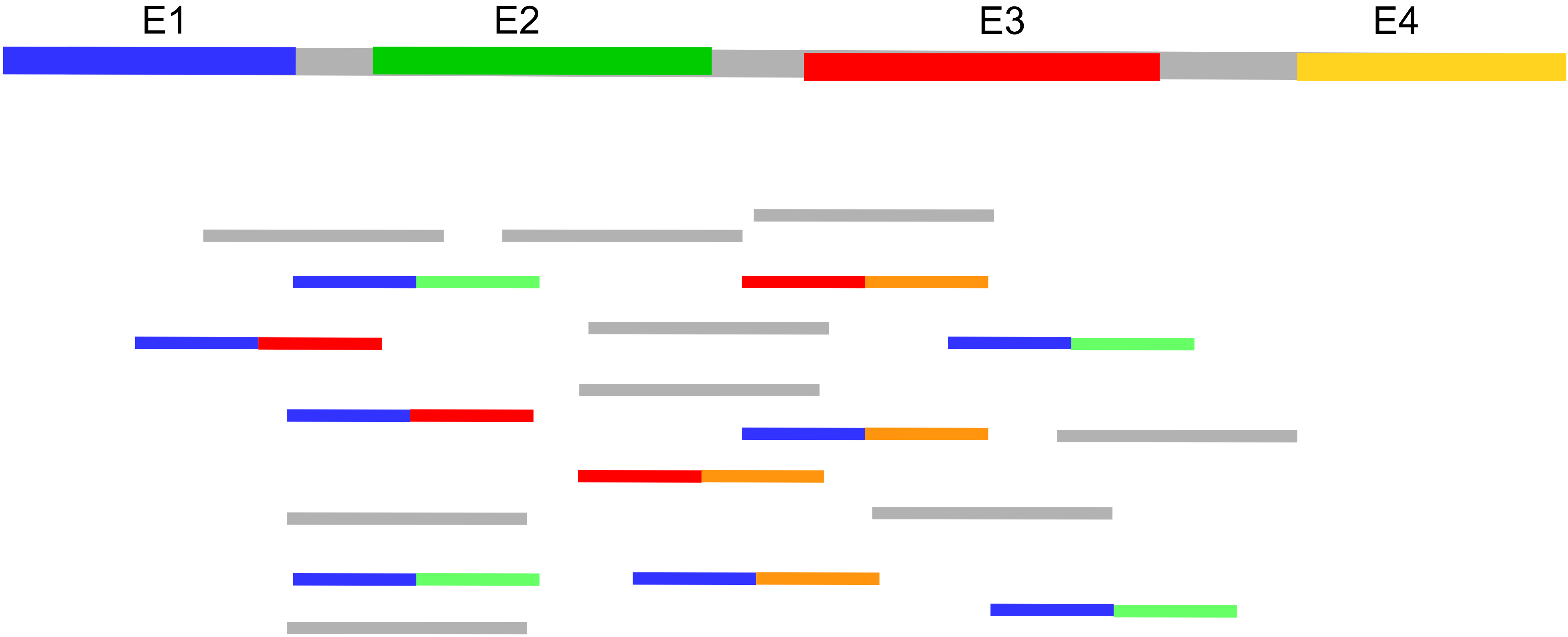
TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park.

- Splice Junction Mapper
- Will use bowtie to map reads normally
- The unmapped reads will then be realigned with large gaps

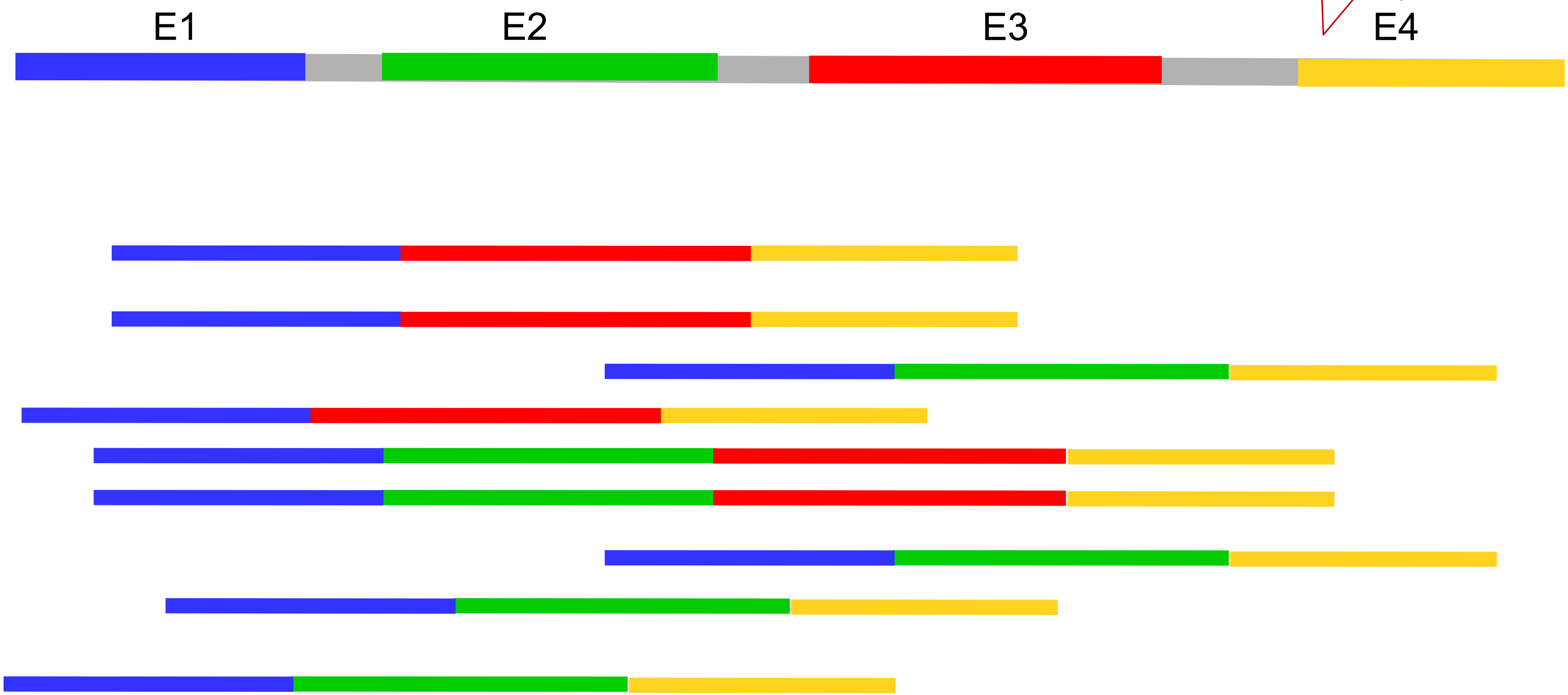
Problem

How many splicing events. Which splice pattern does not occur ?



Its already Obsolete :(

Long Read Sequencing



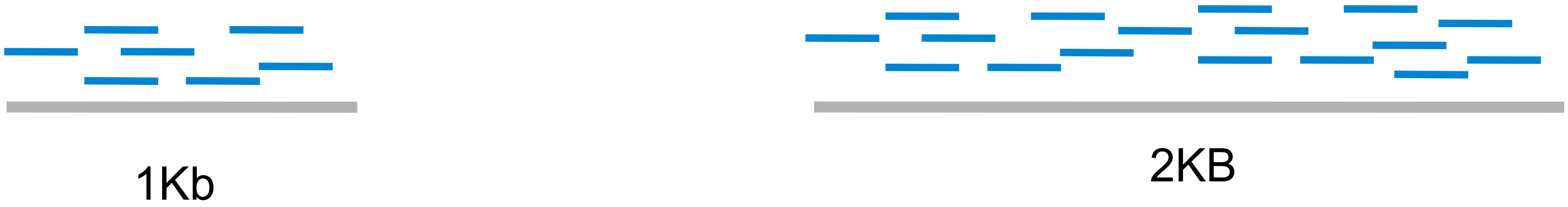
Expression Quantification

- One value which estimates the abundance of a gene/transcript
- Easy to compute
- Unbiased
  - Gene Length
  - Depth
- Comparable between different samples
- Easily understood

No of Reads

- Number of reads mapping to a given transcript is a good measure of abundance
- But larger the gene more the reads mapping to it
- More the sequencing depth more reads would be aligned

No of Reads

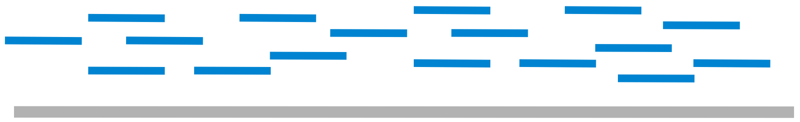


1Kb

2KB

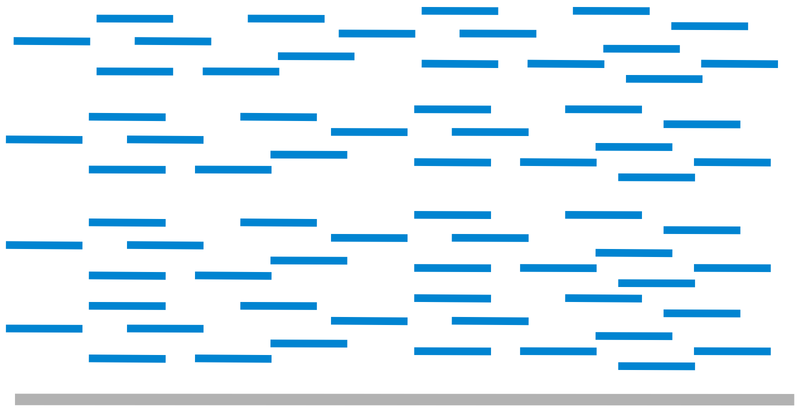
Total Sequencing Output

Sample1



1 Million Reads

Sample2



10 Million Reads

RPKM and FPKM

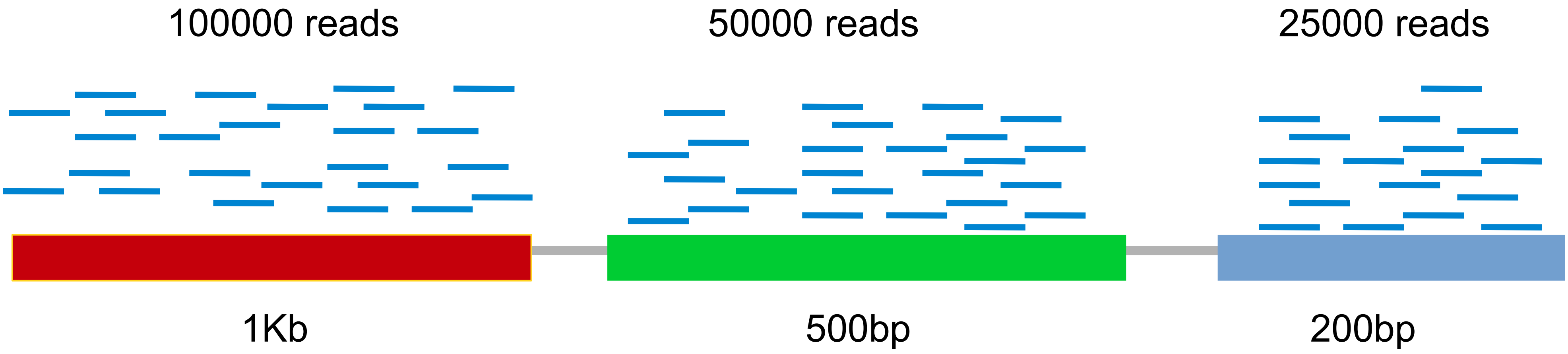
Reads Per Kilobase (Transcript) Per Million Mapped Reads.

Fragments Per Kilobase (Transcript) Per Million Mapped Reads.

$$\text{RPKM} = \frac{\text{Total Reads mapped to a transcript}}{\text{(Length of region (Kb)*(Total Mapped reads in Mb))}}$$

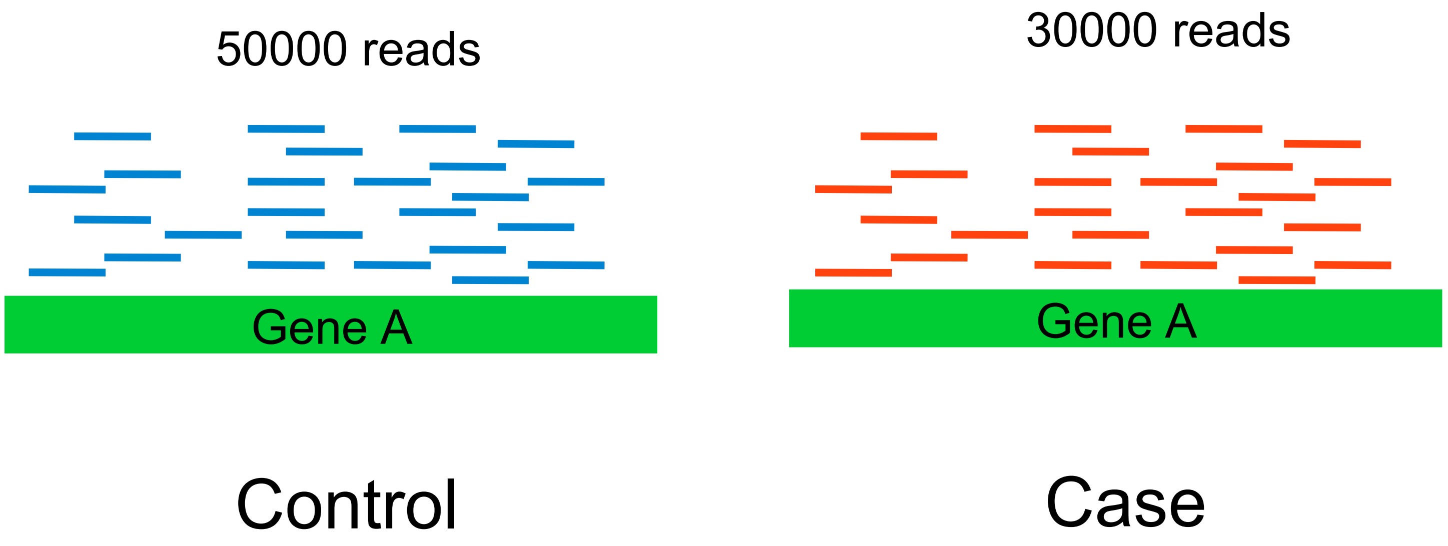
Problem

The genome consists of 3 genes. Calculate RPKM for the green gene.



Problem

You are analyzing a case control study. The case bam file has 2 million mapped reads whereas the control bam file has 1.5 million mapped reads. Is gene A differently expressed and what is the ratio of case/control expression ?



50000 reads

Gene A

Control

30000 reads

Gene A

Case