



DEEP  
LEARNING  
INSTITUTE



DLI Accelerated Data Science Teaching Kit

# Lecture 19.1 - Introduction to Genomics



The Accelerated Data Science Teaching Kit is licensed by NVIDIA, Georgia Institute of Technology, and Prairie View A&M University under the [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

# Genomics

Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes.

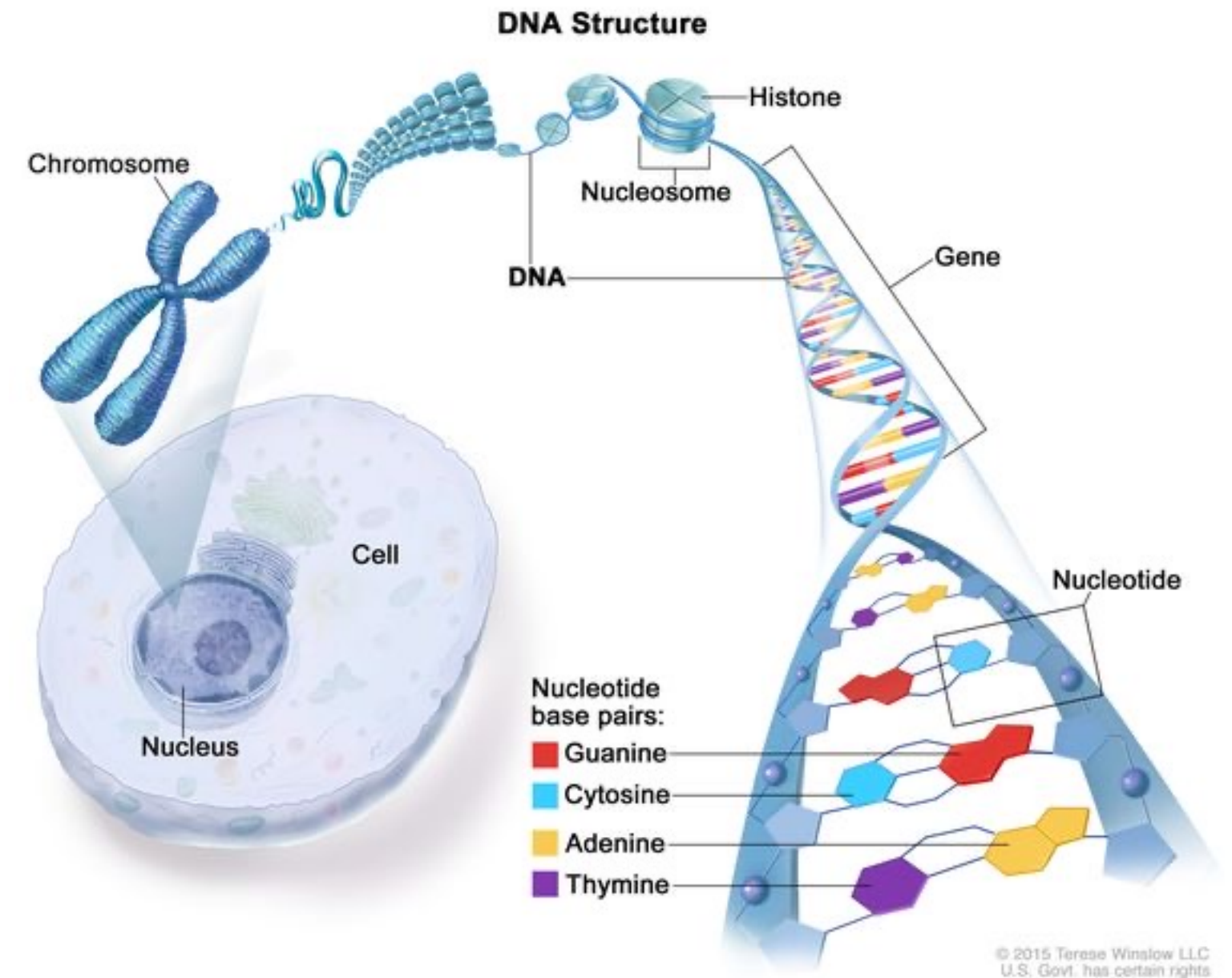
A genome is an organism's complete set of DNA, including all of its **genes**.

In contrast to genetics, which refers to the study of individual genes and their roles in inheritance, genomics aims at the collective characterization and quantification of all of an organism's genes, their interrelations and influence on the organism.

# What is gene?

Based on Central Dogma of Molecular Biology, proposed by *Francis Crick*, a **gene** is a genomic sequence of nucleotides which can be transcribed to an RNA transcript of protein-coding sequences which is translated using the genetic code into an amino acid sequence.

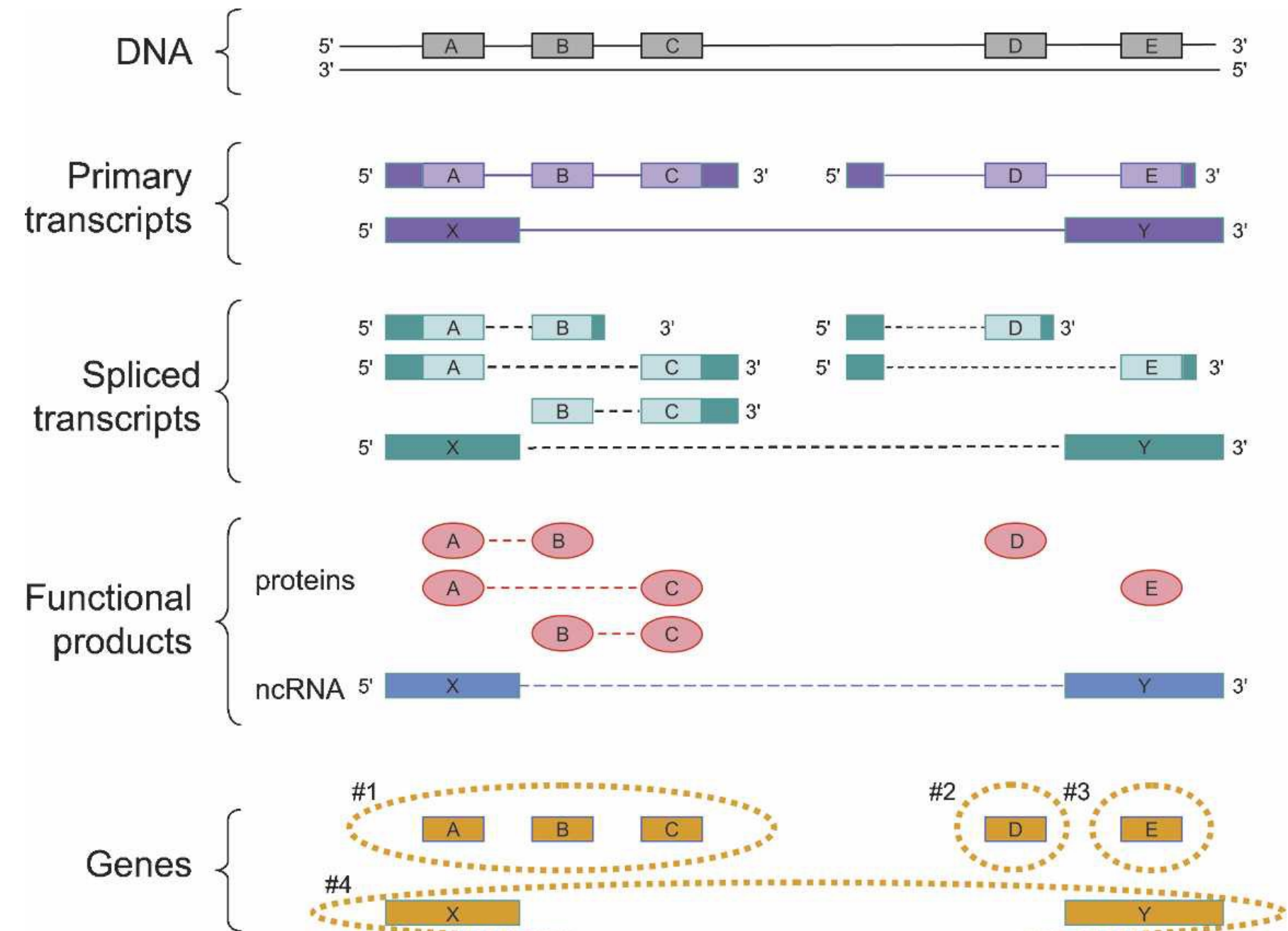
A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions.



# What is gene?

The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.

1. A gene is a genomic sequence (DNA or RNA) directly encoding functional product molecules, either **RNA** or protein.
2. In the case that there are several functional products sharing overlapping regions, one takes the union of all overlapping genomic sequences coding for them.
3. This union must be coherent—i.e., done separately for final protein and RNA products—but does not require that all products necessarily share a common subsequence.



# Gene Expression Regulation

Gene expression is the process by which the information encoded in a **gene** is used to direct the assembly of a protein molecule.

The process of turning on a gene to produce RNA and protein.

The cell reads the sequence of the gene in groups of three bases.

Each group of three bases (codon) corresponds to one of 20 different amino acids used to build the protein.

Gene expression regulation is how a cell controls which genes, out of the many genes in its genome, are “turned on”.

# DNA Sequencing

Determining the number and order of nucleotides that make up a given molecule of DNA

1. How many base pairs (bp) are there in a human genome?

**~3 billion (haploid)**

2. How much did it cost to sequence the first human genome?

**~\$2.7 billion**

3. How long did it take to sequence the first human genome?

**~13 years**

4. When was the first human genome sequence complete?

**2000 - 2003**

5. Whose genome was it?

**Several people's, but actually, mostly a person from Buffalo**



# DNA Sequencing

DNA sequencing is a laboratory technique used to determine the exact sequence of bases (A, C, G, and T) in a DNA molecule.

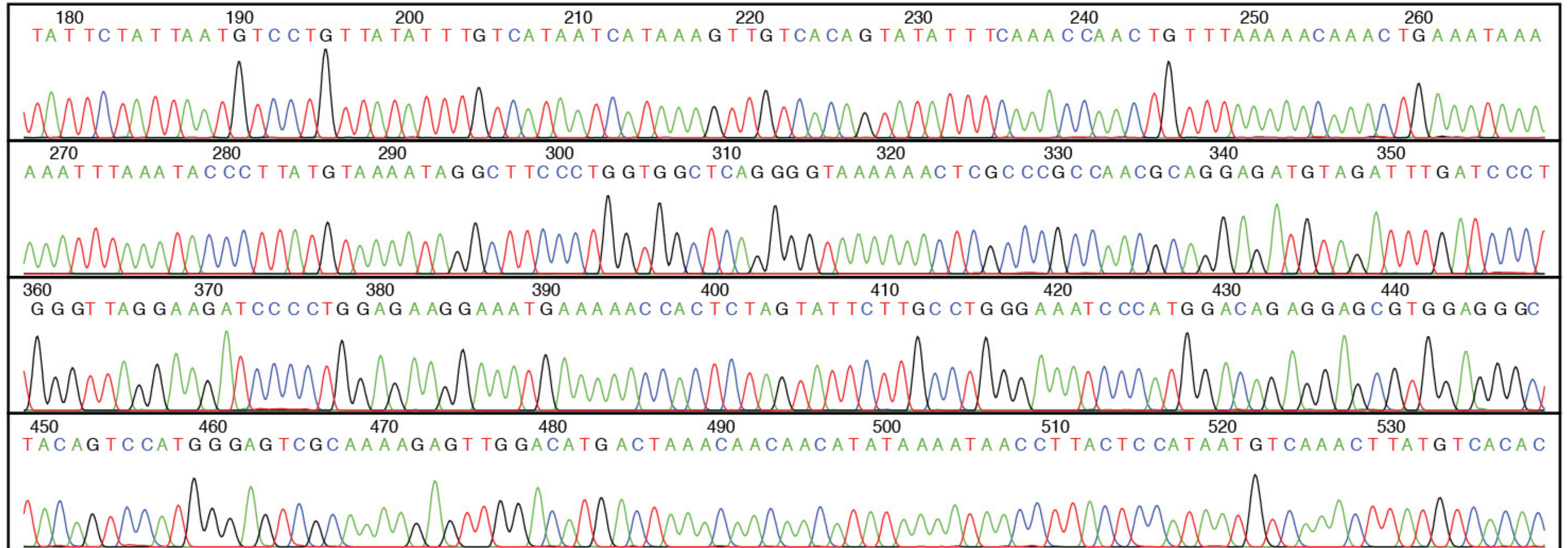
The DNA base sequence carries the information a cell needs to assemble protein and RNA molecules.

DNA sequence information is important to scientists investigating the functions of genes.

The technology of DNA sequencing was made faster and less expensive as a part of the Human Genome Project.



# DNA Sequence Data



DNA sequence data from an automated sequencing machine


# Data Format


FASTQ Format: The FASTQ format stores DNA sequence data as well as associated **Phred** quality scores of each base.

**Phred** scores define quality

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
::3::::::::::::7::::::::88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
::::::::::::7:::::-:::3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
::::::::::::9;7;.:7;393333
```

 DNA read

 Base quality score

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%



# FASTQ File

# FASTQ files

Line1: sequence identifier

Line2: raw sequence

Line3: meaningless

Line4: quality values for the sequence

[illegible]

# Sequence Alignment/Map Data Format (SAM)

SAM is a common format having sequence reads and their alignment to a reference genome.

BAM is the binary form of a SAM file.

Aligned BAM files are available at repositories (Sequence Read Archive at NCBI, ENA at **Ensembl**)

SAMTools is a software package commonly used to analyze SAM/BAM files.

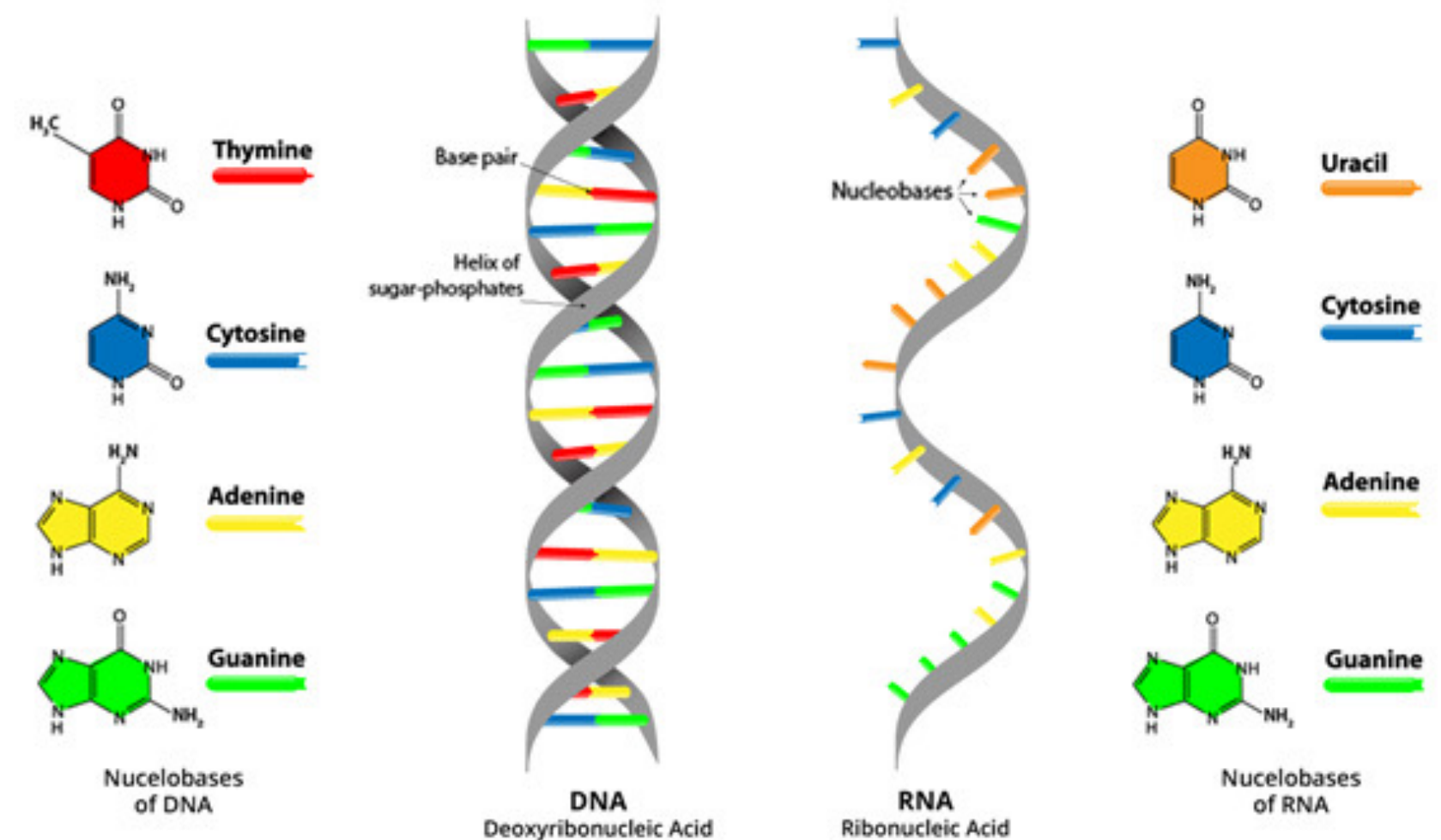
Column	Description
1	QNAME Query (pair) NAME
2	FLAG bitwise FLAG
3	RNAME Reference sequence NAME
4	POS 1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ MAPping Quality (Phred-scaled)
6	CIGAR extended CIGAR string
7	MRNM Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS 1-based Mate POSition
9	ISIZE Inferred insert SIZE
10	SEQ query SEquence on the same strand as the reference
11	QUAL query QUALity (ASCII-33 gives the Phred base quality)
12	OPT variable OPTional fields in the format TAG:VTYPE:VALU



# RNA-seq and Data Analysis

## RNA: Ribonucleic Acid

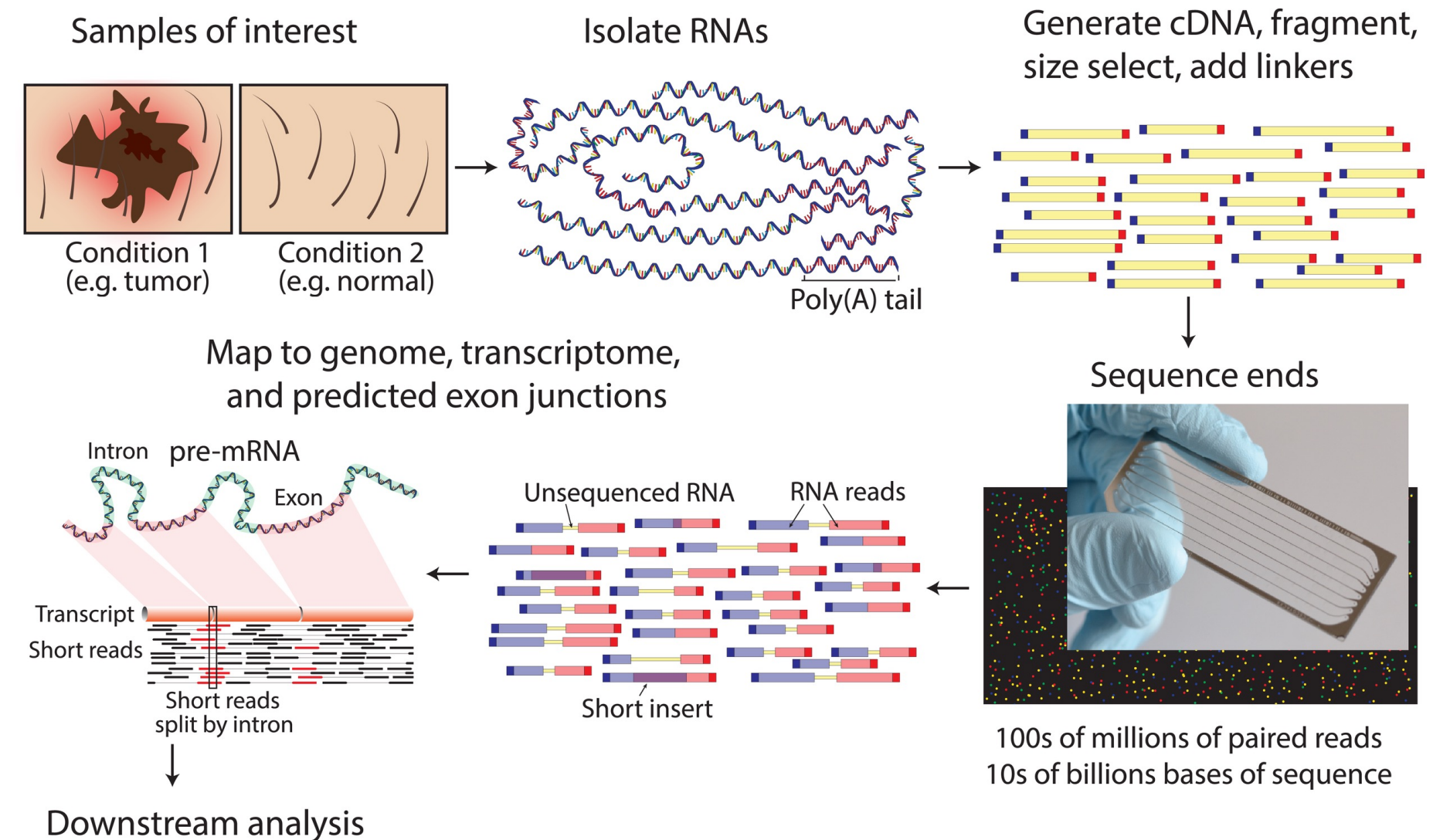
1. RNA converts the genetic information contained within DNA to a format used to build proteins, and then moves it to ribosomal protein factories.
2. RNA only has one strand, but like DNA, is made up of nucleotides.
3. RNA strands are shorter than DNA strands.
4. RNA sometimes forms a secondary double helix structure, but only intermittently.



# RNA-seq and Data Analysis

RNA-Seq (named as an abbreviation of "RNA sequencing") is a sequencing technique which uses next-generation sequencing (NGS) to detect the presence and quantity of RNA in a biological sample at a given moment, analyzing the continuously changing cellular transcriptome.

The quantification of the amount of RNA produced or "expressed" (i.e., RNA transcribed) at a given time in a biological sample can be done by several technologies, but RNA-seq is currently considered the most powerful, robust and adaptable technique for measuring gene expression and transcription activation at genome-wide level.



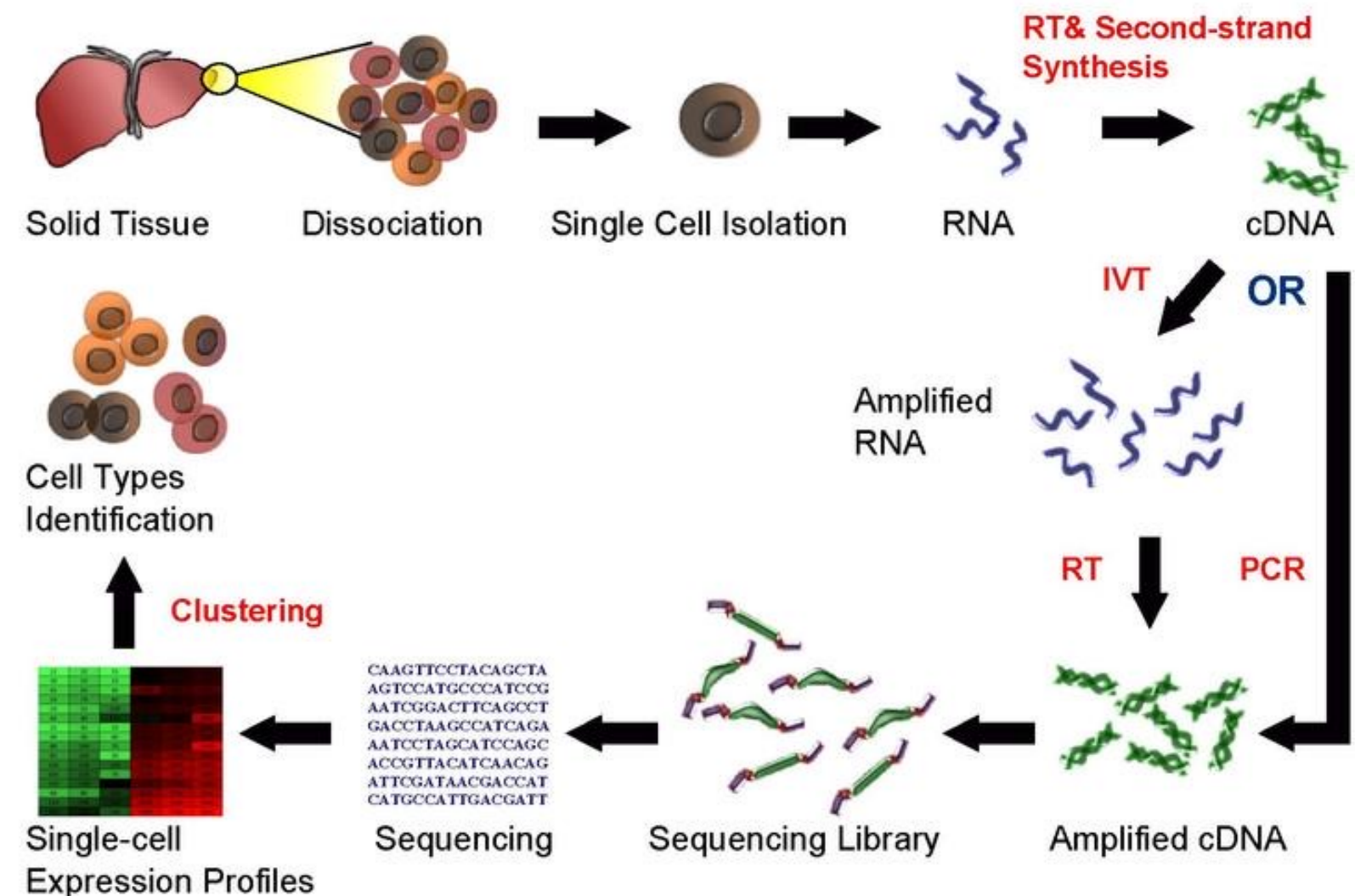
# RNA-seq and Data Analysis

Single-cell RNA sequencing (**scRNA-Seq**) provides the expression profiles of individual cells.

Although it is not possible to obtain complete information on every RNA expressed by each cell, due to the small amount of material available, patterns of gene expression can be identified through gene clustering analyses.

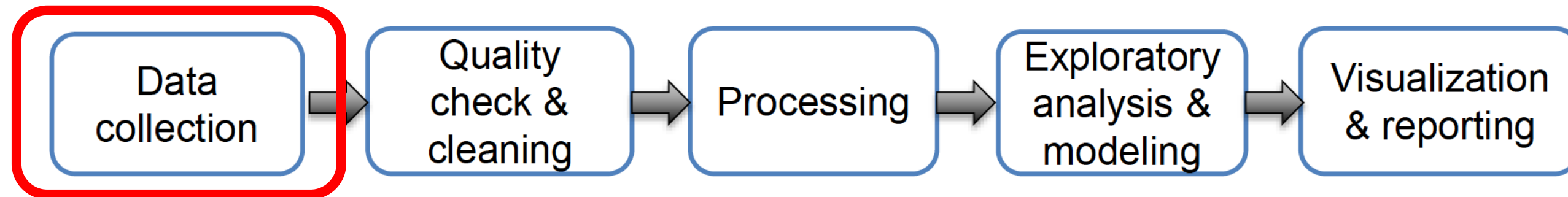
This can uncover the existence of rare cell types within a cell population that may never have been seen before.

## Single Cell RNA Sequencing Workflow





# Data Analysis Workflow

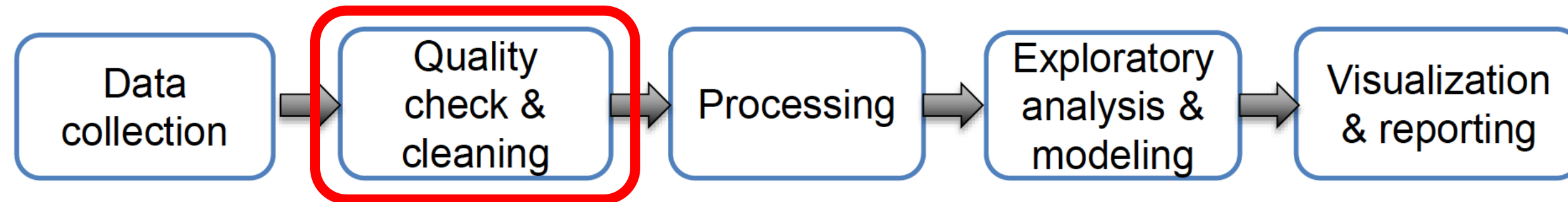


## Data collection

1. Which sequencing technology you are using ?
2. What kind of experiments are you doing ?
3. How many samples ?
4. How many replicates ?
5. Which public data you will include in your analysis ?

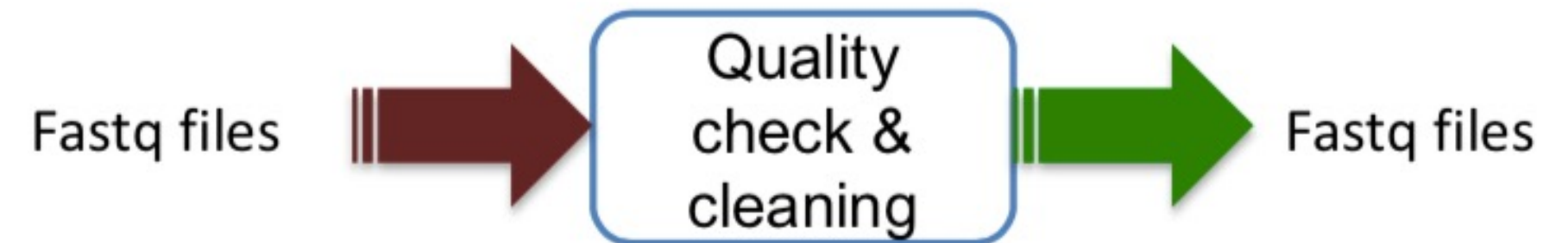


# Data Analysis Workflow



## Quality check and clean up

1. Quality check is mostly about checking read quality.
2. Involve removing low quality bases from reads
3. Involve removing adapter/barcode sequences from reads

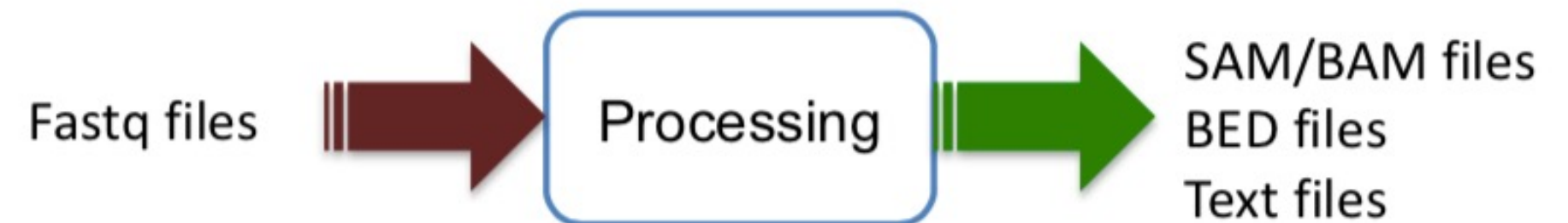


# Data Analysis Workflow

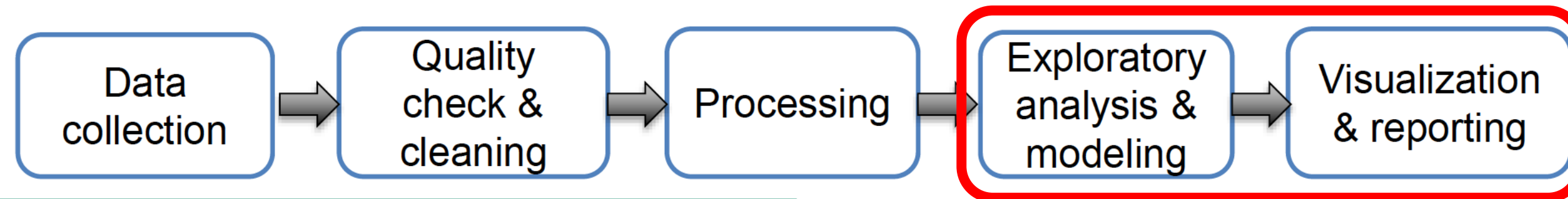


## Data Processing

1. Transforming raw data to a state where modeling or exploratory data analysis can start
2. Include making a tabular data structure from raw data
3. Include data transformations such as taking logs or normalization
4. Alignment + Quantification



# Data Analysis Workflow



## Exploratory analysis and modeling

In general,

- How samples or variables relate to each other
  - clustering & dimension reduction (PCA, etc.)
- Prediction of variable of interest:  $Y \sim X_1 + \dots + X_n$
- Statistical models including hypothesis testing

In genomics,

- Annotation with gene sets/pathways
- Looking at genomics data with special browsers, such as UCSC genome browser or IGV

## Final visualization and reporting

Final figures, tables and text that describes the outcome of analysis

Jupyter notebook or Rmarkdown go-to tool for compiling reports these days



DEEP  
LEARNING  
INSTITUTE



PRAIRIE VIEW  
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

# Thank You