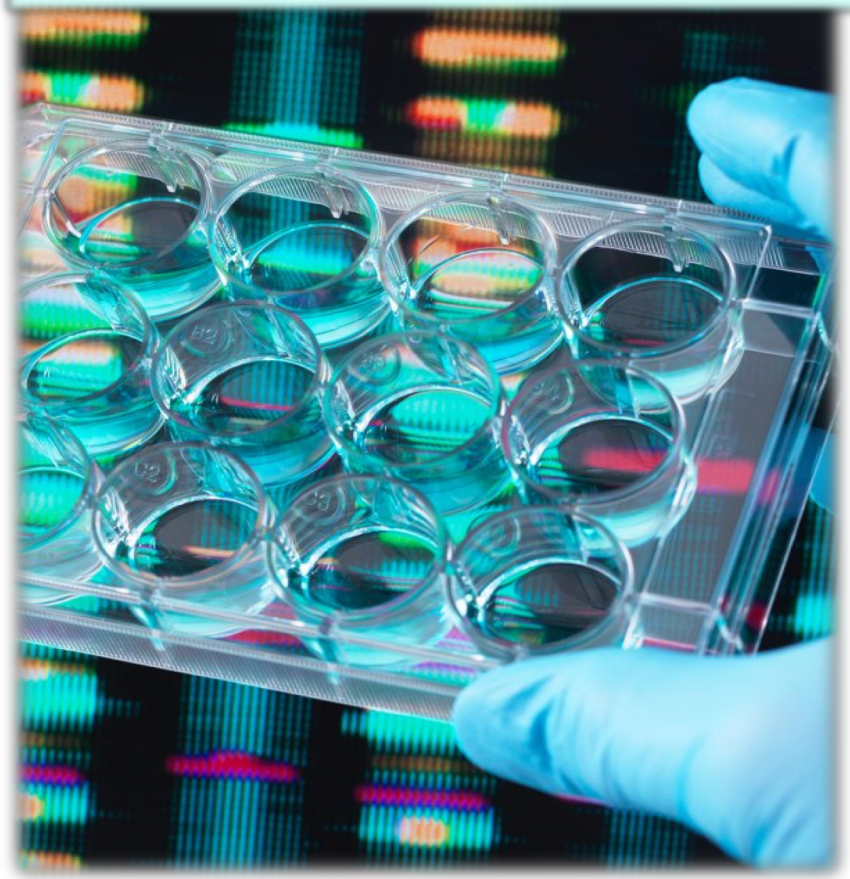
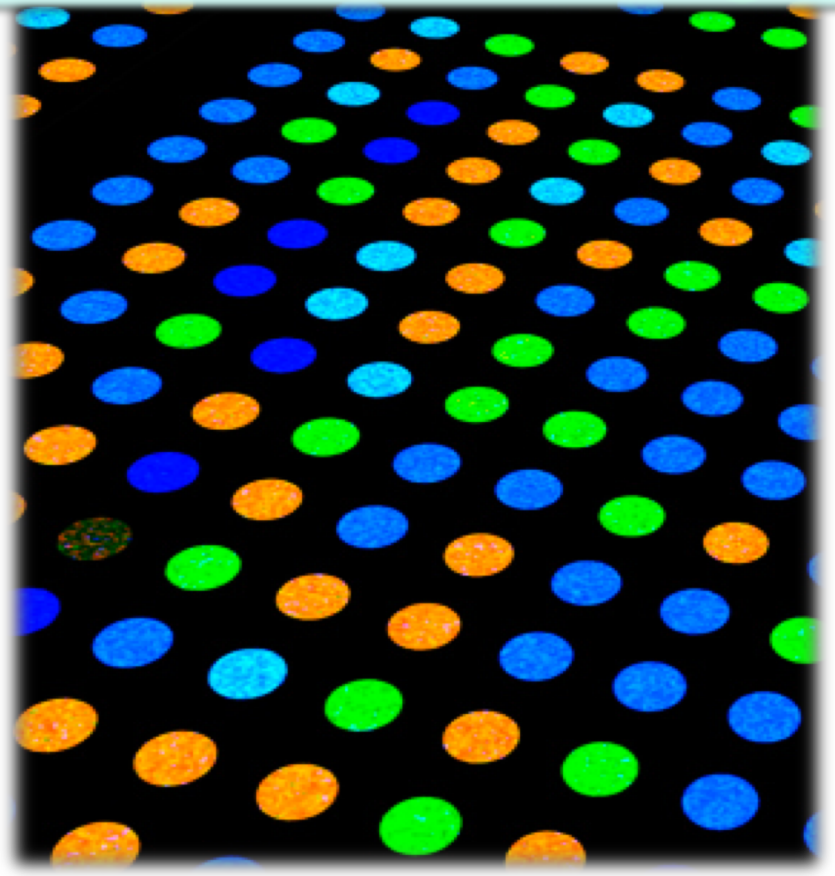
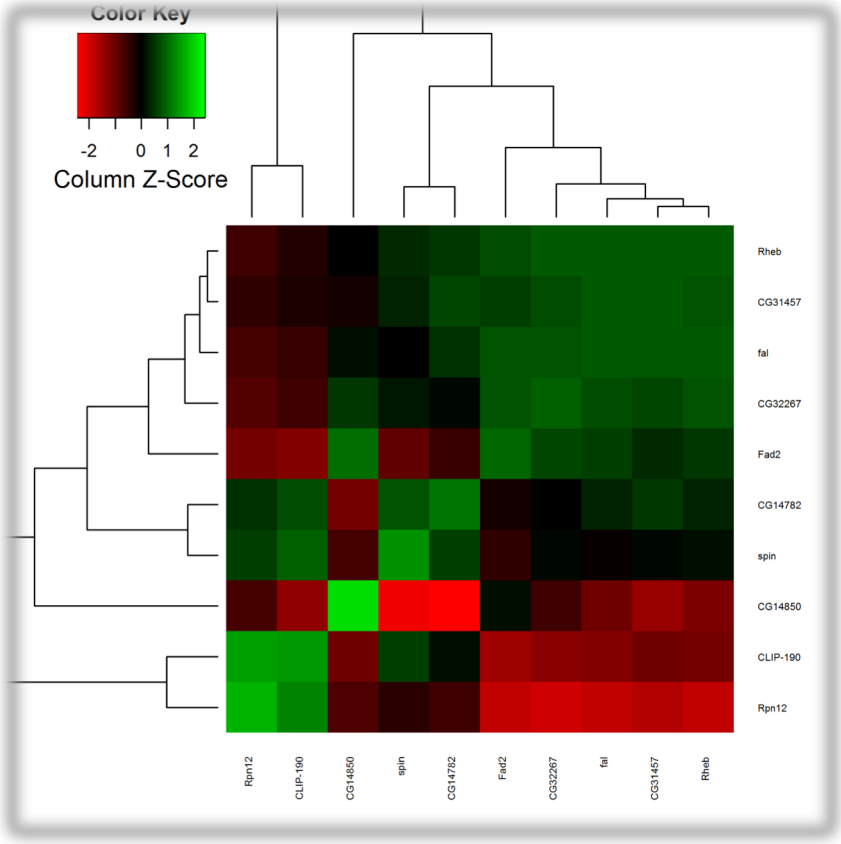
Children's Research Institute

Bioinformatics Unit (CBU)







https://bi-ctsicn.github.io/CBU/

bioinformatics@childrensnational.org

Overview of CBU

- A unique, interactive space, providing solutions to data analysis questions using state-of-the-art bioinformatics tools
- Established in November 2017.
- Generous support from:

  Dr. Vittorio Gallo, Chief Research Officer

  Center for Translational Science

  Clinical and Translational Institute at Children's National (CTSI-CN),

  Center for Genetic Medicine Research

  District of Columbia Intellectual and Developmental Disabilities Research Center (DC-IDDRC) at Children's National Medical Center.

People

- **Leadership**
  - Dr. Eric Vilain - Director of Center for Genetic Medicine Research
  - Dr. Lisa Guay-Woodford - Director of CTSI at Children's National
  - Dr. Hiroki Morizono - Director of CBU
  - Dr. Kazue Hashimoto-Torii - Center for Neuroscience Research liaison
  - Dr. Susan Knoblach - Director of Genomics Core
  - Dr. Michael Keller - Center for Cancer and Immunology liaison
  - Dr. Marius George Linguraru - Sheik Zayed Institute for Surgical Innovation liaison
- **Staff**
  - Payal Banerjee, MS
  - Surajit Bhattacharya, PhD
  - Hayk Barseghyan, PhD

Bioinformatics services provided by CBU

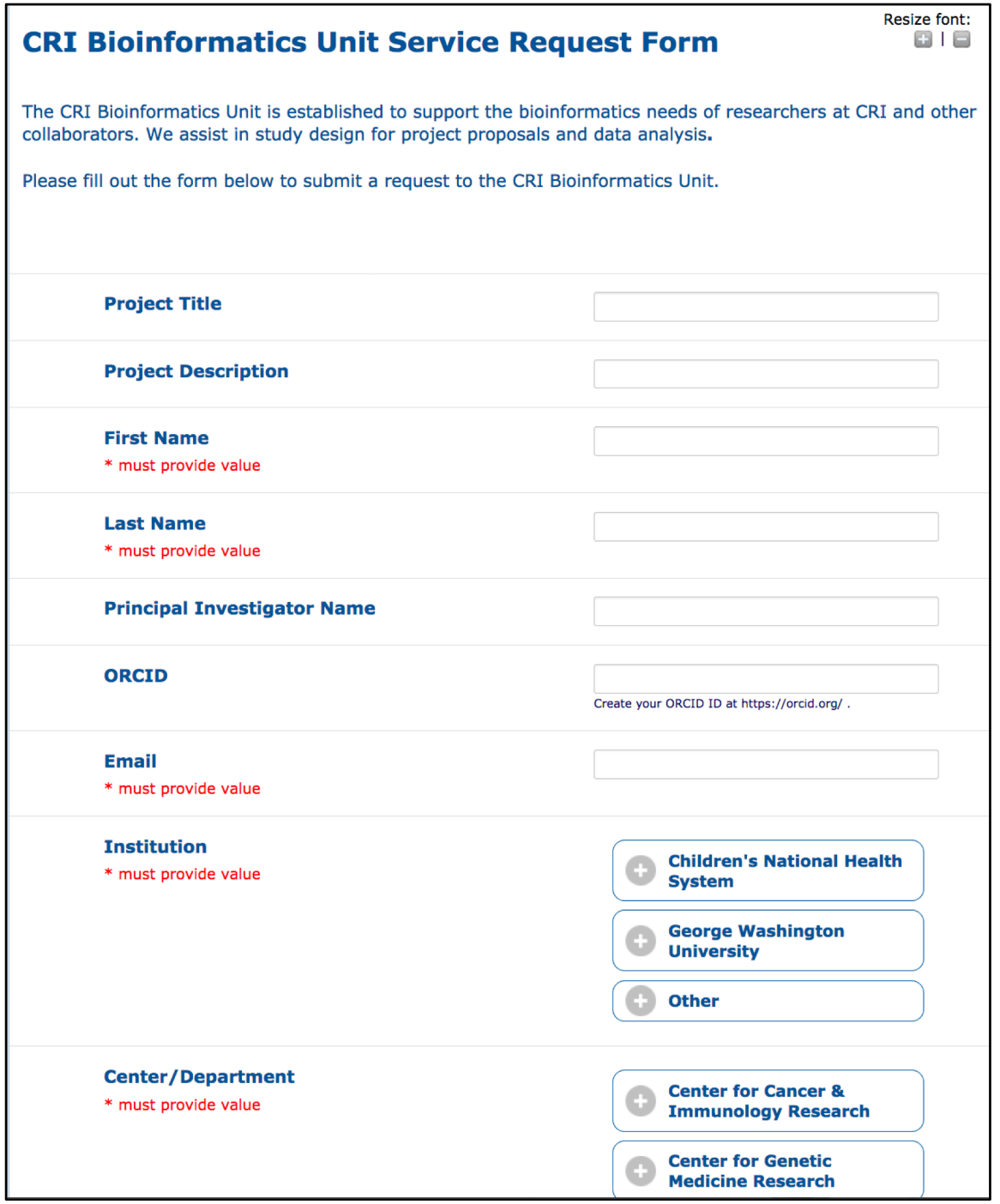
Experimental Design and Grant Proposal Support

- **RNA sequencing**
  - Differential Gene Expression
  - Single Nucleotide Variation Detection
  - Fusion Analysis
- **Whole Genome sequencing**
  - De Novo Assembly
  - Single Nucleotide Variation Detection
  - Copy Number Variation
  - Structural Variant Detection
- **Exome sequencing**
  - Single Nucleotide Variation Detection
  - Copy Number Variation
  - Structural Variant Detection
- **Microbiome Analysis (16s and metagenomics)**
- **Single Cell Sequencing**
- **T-cell receptor Sequencing**
- **Bisulphite Sequencing**
- **Chip Sequencing**
- **Microarray**
- **Other/Custom Analysis**

CBU Service Request Form

Email : bioinformatics@childrensnational.org

http://j.mp/2FSWz7s



Resize font:

CRI Bioinformatics Unit Service Request Form

The CRI Bioinformatics Unit is established to support the bioinformatics needs of researchers at CRI and other collaborators. We assist in study design for project proposals and data analysis.

Please fill out the form below to submit a request to the CRI Bioinformatics Unit.

Project Title

Project Description

First Name
* must provide value

Last Name
* must provide value

Principal Investigator Name

ORCID
Create your ORCID ID at https://orcid.org/ .

Email
* must provide value

Institution
* must provide value

- Children's National Health System
- George Washington University
- Other

Center/Department
* must provide value

- Center for Cancer & Immunology Research
- Center for Genetic Medicine Research

Take Home Message

- Who are we – CRI Bioinformatics Unit
- How to contact us – bioinformatics@childrensnational.org
- What we offer:
  - Data Analysis – Types available on website - https://bi-ctsicn.github.io/CBU/
  - Experimental Design Consultation (Wet Lab in collaboration with Genomics Core – Susan Knoblach Team)
  - Grant Technical Writing support

We generally recommend people either to share their data for analysis via Hard Drive (Mac compatible) or via Dropbox

Along with Data submission we request people to fill out a submission form - http://j.mp/2FSWz7s

Turnaround time – 2-3 weeks depending upon experiment and number of samples

Informatics Approach to Genomics

Introduction

General Steps of Sequencing Experiments

Experimental Design

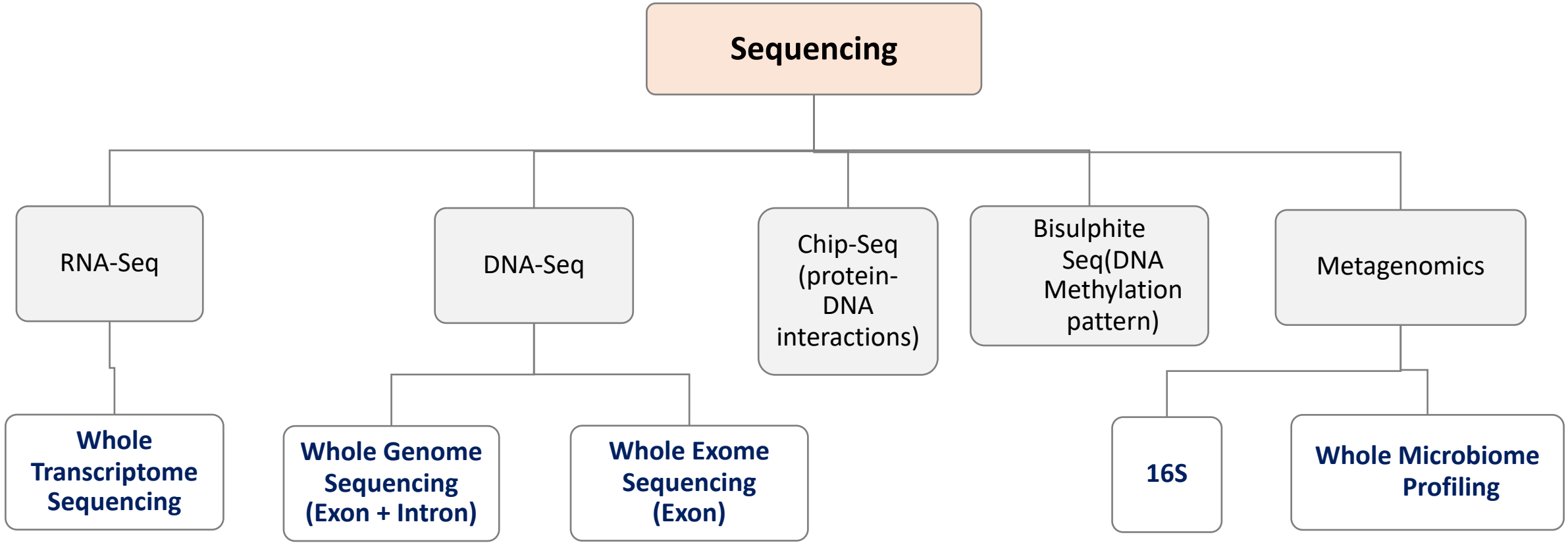
Sample extraction and preparation

Sequencing

Data Analysis

- ❑Preliminary Data Processing and Analysis
- ❑Further Downstream Data Analysis
  - ▪Annotation
  - ▪Visualization

Experimental Design



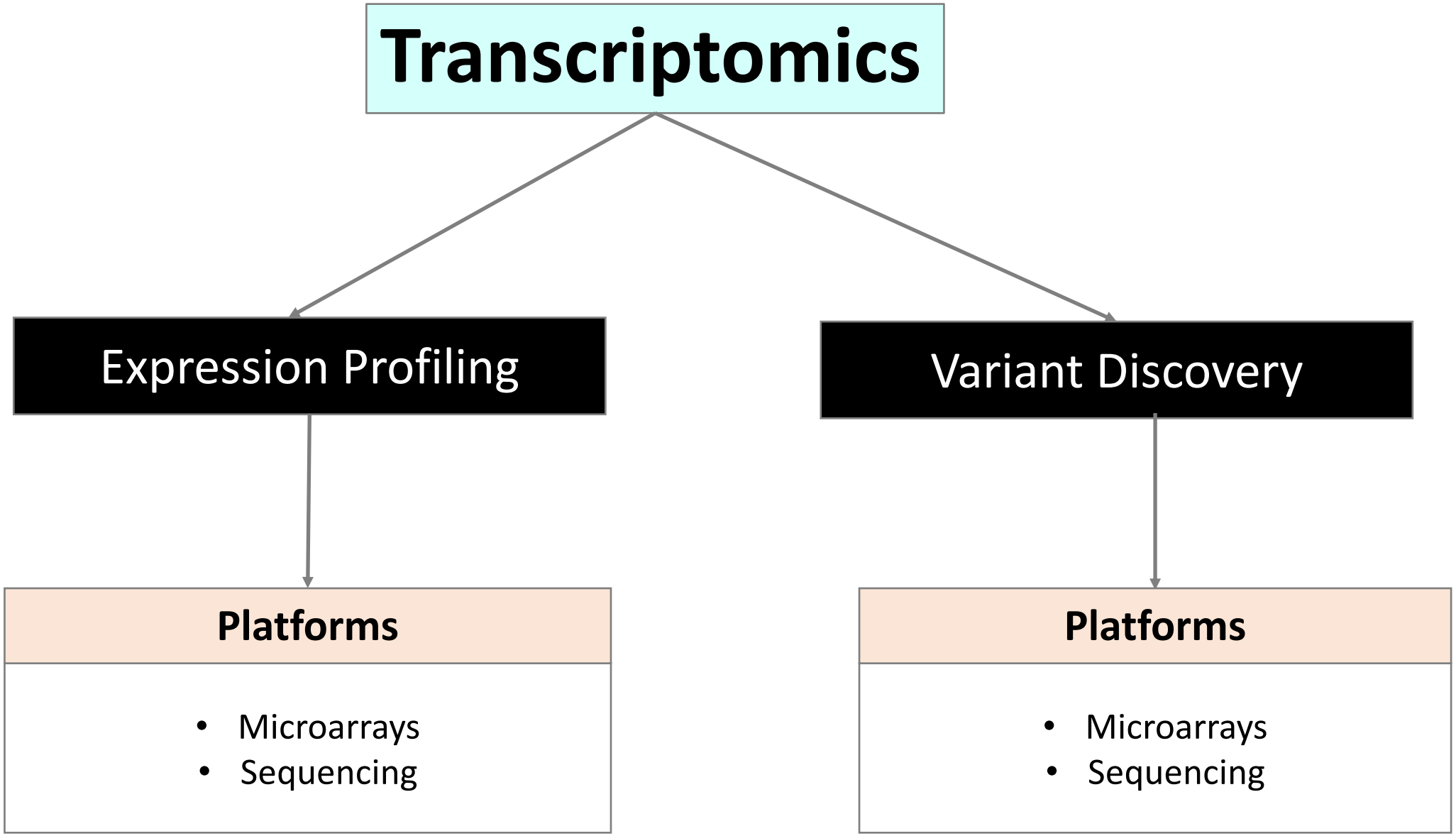
RNA Sequencing Applications

- Gene expression profiling between samples
- Identification of allele-specific expression, disease-associated single nucleotide polymorphisms (SNPs) and
- Novel transcript identification
- Gene fusions detection e.g. disease causal variants in cancer.

Single-cell RNA-seq has recently emerged as a way to study complex biological processes.

Single-cell RNA-seq Applications

- **Neurobiology** - classify neurons based on their transcriptional profiles
- **Cancer Research** - Delineating clonal diversity and understanding the role of rare cells during cancer progression.
- **Germline Transmission** - A novel approach to study the mechanisms that generate germline variation during the transmission of genetic material to offspring.
- **Embryogenesis** - transcriptional regulation and epigenomic reprogramming that occurs during the earliest stages of embryogenesis.
- **Organogenesis** - transcriptional profiling and identify groups of cells that share common expression programs, representing distinct cell types.
  - Eg: SCS was used to analyze gene expression patterns of single cells during kidney development in mice at E11.5, E12.5, and P4 (Brunskill et al., 2014)
- **Immunology** - To investigate heterogeneous transcriptional responses in immune cells after antigen activation.


Transcriptomics
Expression Profiling
Variant Discovery
Platforms
• Microarrays
• Sequencing
Platforms
• Microarrays
• Sequencing

Whole Genome and Exome Seq Applications

WGS (Whole Genome Sequencing)

- Coverage in coding as well as non-coding regions
- WGS for novel mutations (rare diseases), copy number variations, structural mutations
- assembly of a new genome
- More coverage

Exome Sequencing

- Exome covers 2% of whole genome
- Allows for genetic panels, and focus on only the protein coding regions
- Less cost and time

Popular WGS/Exome Seq approaches

- GWAS Studies
- Diagnostic screens

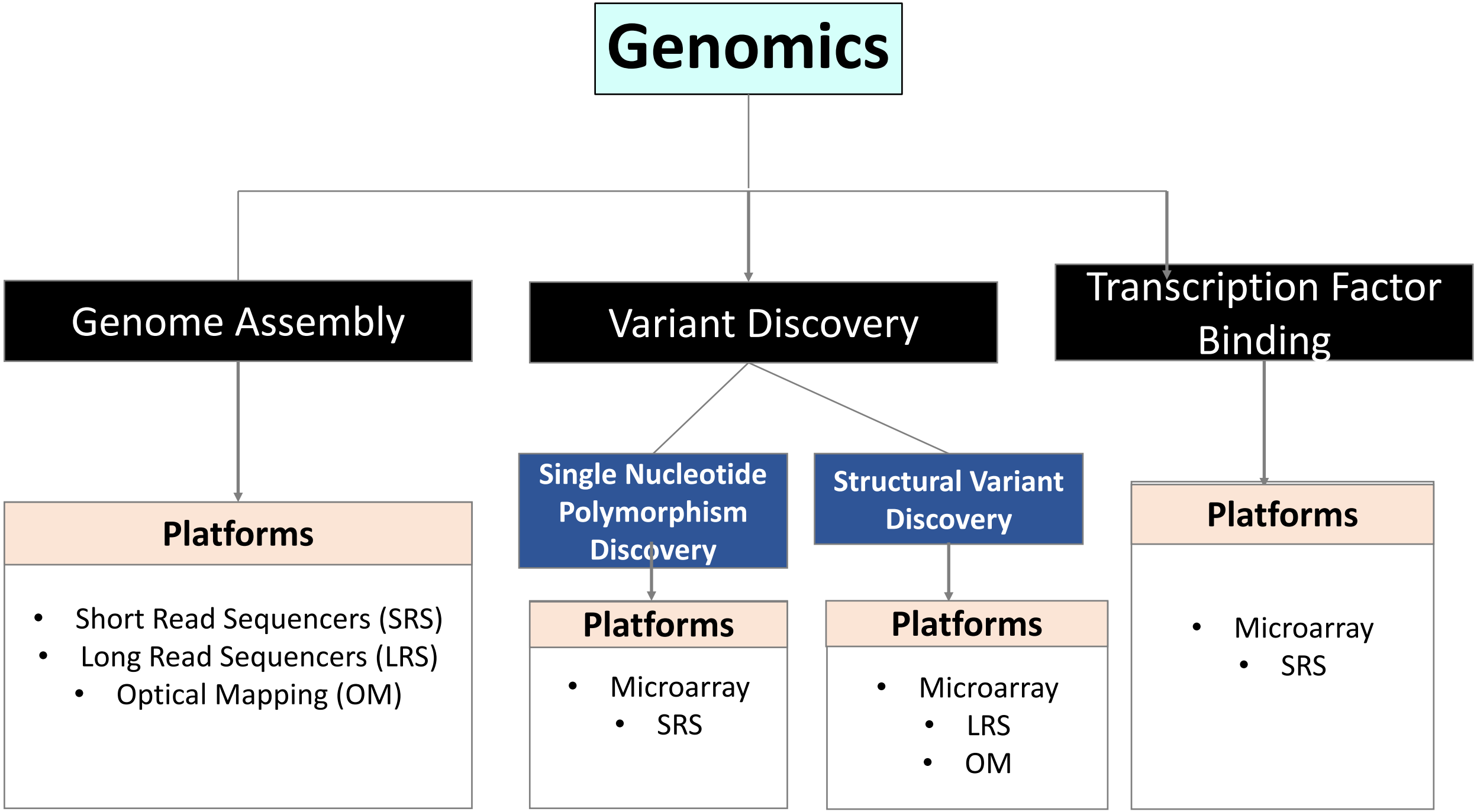
Chip Seq Applications

- Detection of protein-DNA interaction sequences

Eg:

- Determination of sequences associated with transcription factor
- Information about overall structure of chromatin

These interactions can strongly influence the regulation of gene expression and are therefore critical for complete understanding of cell operations.


Genomics
Genome Assembly
Variant Discovery
Transcription Factor Binding
Platforms
Short Read Sequencers (SRS)
Long Read Sequencers (LRS)
Optical Mapping (OM)
Single Nucleotide Polymorphism Discovery
Platforms
Microarray
SRS
Structural Variant Discovery
Platforms
Microarray
LRS
OM
Platforms
Microarray
SRS

Methylation Applications

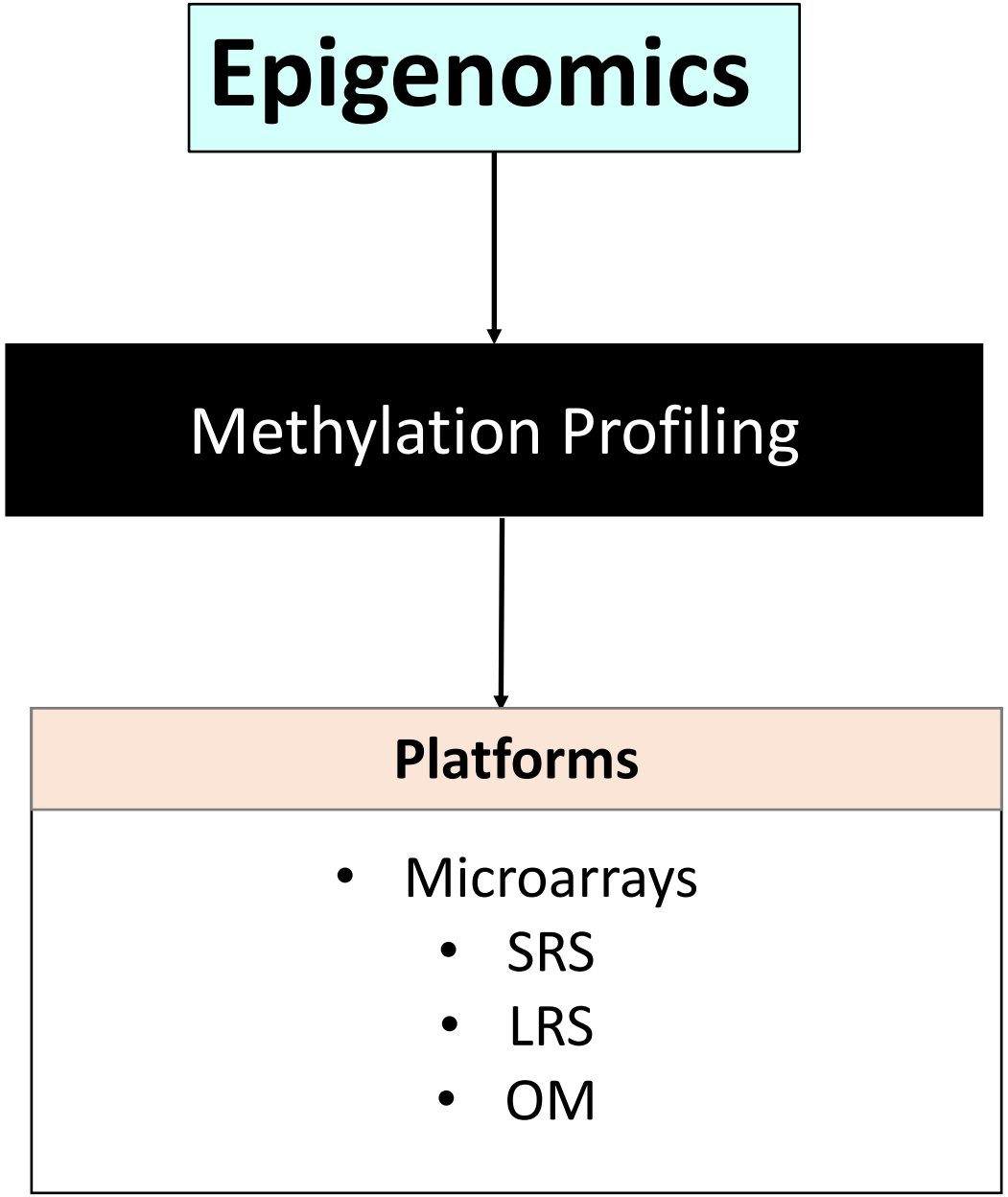
Types of Methylation Sequencing Experiments :

- Bisulfite sequencing
- Oxidative Bisulphite Sequencing

Applications:

- Study Methylation patterns of DNA
- DNA methylation implicated in repression of transcriptional activity

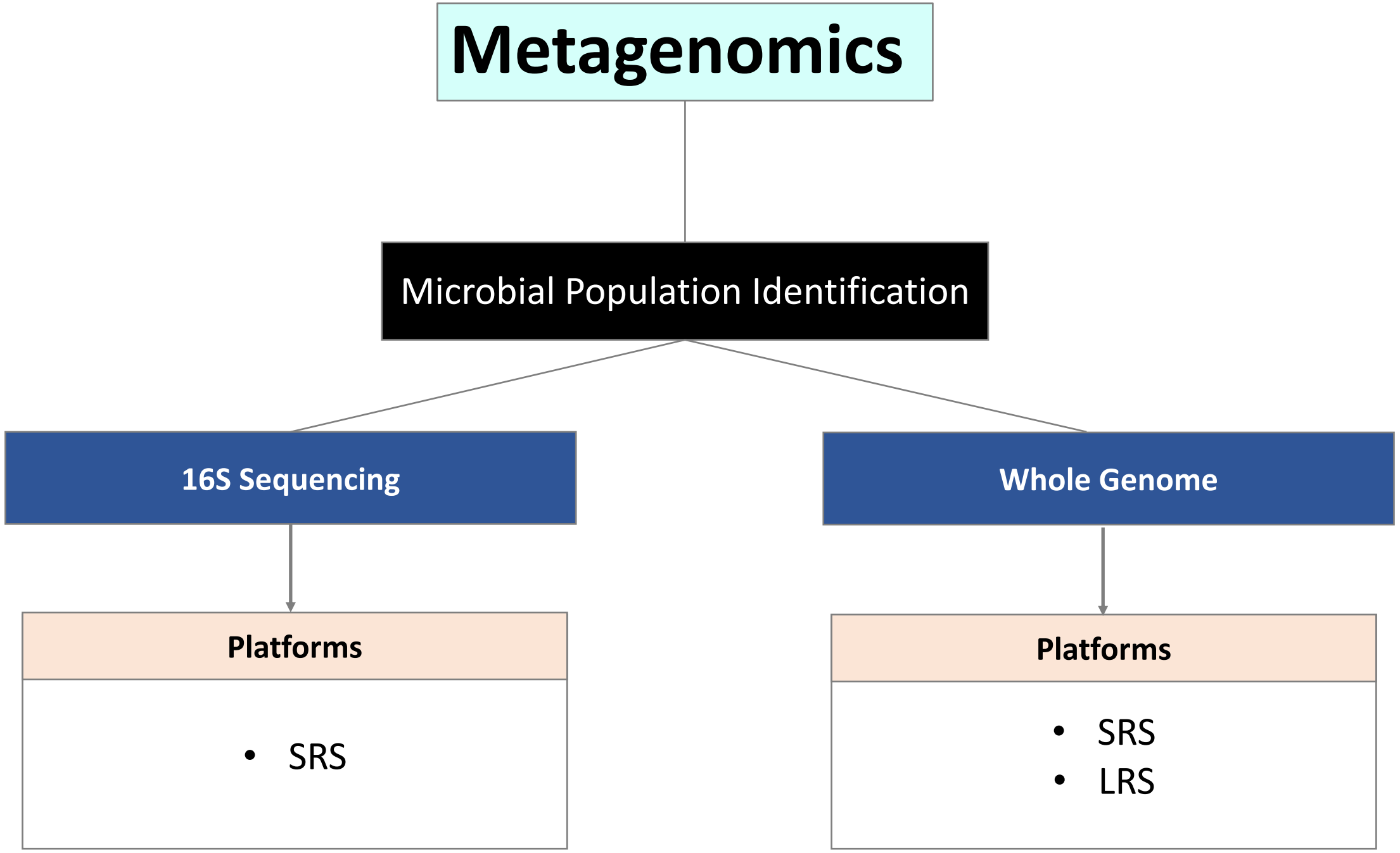
Sequencing over microarrays has enabled the transition from studying the methylation patterns of just a few genes or small regions to that of truly genome-wide studies, and this is leading to a far richer view of the methylome.


Epigenomics
Methylation Profiling
Platforms
Microarrays
SRS
LRS
OM

Metagenomics Applications

- **16S ribosomal RNA sequencing** –Identifies highly conserved 16S rRNA component of the 30S small subunit of a prokaryotic ribosome. Identifies mainly bacterial species in a sample.

- **Whole Microbiome profiling** - Multiple microbial genomes are analyzed at the same time.


Metagenomics
Microbial Population Identification
16S Sequencing
Whole Genome
Platforms
SRS
Platforms
SRS
LRS

General Steps of Sequencing Experiments

Experimental Design

Sample extraction and preparation

Sequencing

Data Analysis

- ❑Preliminary Data Processing and Analysis
- ❑Further Downstream Data Analysis
  - ▪Annotation
  - ▪Visualization

Topics to consider for Sample Preparation

❖**Sample preparation kits**

- ➢ DNA
- ➢ RNA

Depends on:

- Experiment (Bulk, Single Cell, Low input RNA etc)
- Instrument
- Organism – Human, Mouse etc
- Type of sample – Tissue, Cell Line, Blood etc

Read Length

- Read length refers to the number of base pairs that are read at a time.
  - For a read length of 50 base pairs, **single end reads** would read 50 base pairs from each fragment,
  - while **paired end reads** would consist of 2 x 50bp reads, covering up to 100 base pairs on the same fragment.

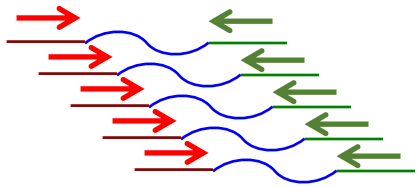
While longer read lengths give you more accurate information on the relative positions of your bases in a genome, they are more expensive than shorter ones.

Single end and Paired end

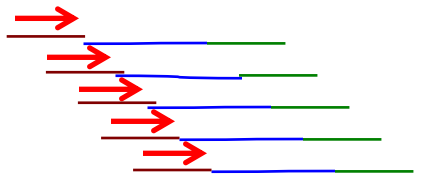
The sequencer instrument read from one end to the other end, and then start another round of reading from the opposite end.

The sequencer reads from one end of a fragment to the other end.

PE = paired end (mate pairs)



SE = single end



PE sequencing provides additional positioning information in the genome

Single end vs Paired end Applications

Single End	Paired End

• Profiling or counting experiments:
  ▪ RNA seq
  ▪ Chip seq
	• Structural re-arrangements such as deletions, insertions and inversions
	• SNP /SNV identification
	• Epigenetic modifications (methylation)

PolyA or Depletion? (RNA)

- Ribosomal RNA (rRNA) constitutes >70% of the purified total cell RNA.
- RiboDepletion removes specifically ribosomal RNA, leaving all other RNA transcripts, however it is not 100% efficient.
- PolyA selection is very efficient, but it will only select polyadenylated RNA, therefore many long, non coding RNAs will be lost.

Poly A	Depletion
Eukaryotes mostly	Prokaryotes
mRNA	mRNA along with non-coding RNA like lncRNA etc

Replicates

Ideally, if the experiment were repeated with new, independently obtained samples, the effect would likely be observed again.

Variation in data – are they actual biological changes or are just caused by chance?

Replication for Reproducibility!

Without replicates, what do you miss?

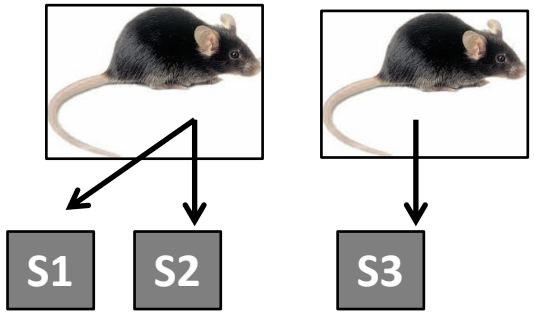
- Identification of random Variation
- Accuracy of measurement

Replicates

- **Biological replicates** measure a quantity from **difference sources** under the same conditions
  - e.g., Tumors from 5 different people with lung cancer may show similar gene expression patterns. These replicates are useful to show what is similar in your replicates and how they are different from a different set of conditions (ie. treated, normal).

- **Technical replicates** measure quantity from **1 source**. This measures the reproducibility of the results. The differences are based only on technical issues in the measurement.
  - e.g., sequence the same sample twice but get different results

Replicates



Technical replicates are: S1 /S2

Biological replicates are: S3 – S1/S2

To make inferences about the population you need ***biological replicates***

General Steps of Sequencing Experiments

Experimental Design

Sample extraction and preparation

Sequencing

Data Analysis

- ❑Preliminary Data Processing and Analysis
- ❑Further Downstream Data Analysis
  - ▪Annotation
  - ▪Visualization

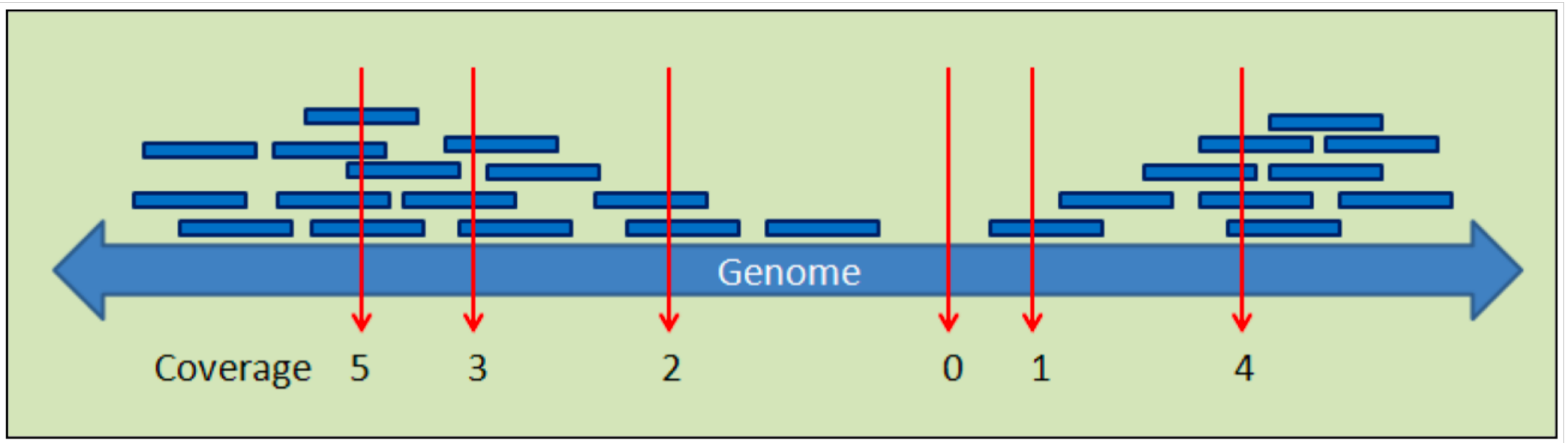
Topics to consider for Sequencing

- **Instruments (CRI)**
  - Illumina (Short Read)
    - Miseq
    - Nextseq
  - Nanopore (Long Read)
    - MinIon
    - GridIon
  - Bionano (Optical Mapping)

Factors : What kind of experiment, What type of samples, Number of samples, Coverage etc

Coverage

- Coverage refers to the average number of times a single base is read during a sequencing run.
  - Eg: If the coverage is 100 X, this means that on average each base was sequenced 100 times.
- The more frequently a base is sequenced, the more reliable a base is called, resulting in better quality of your data.



Coverage

Lander / Waterman equation

$$\text{Coverage (X)} = \frac{\text{L (read length)} * \text{N (number of reads)}}{\text{G (genome length)}}$$

Coverage Calculators

- http://core-genomics.blogspot.com/2016/05/how-many-reads-to-sequence-genome.html
- http://apps.bioconnector.virginia.edu/covcalc/

Coverage (DNA)

Estimate of Coverage Requirements by Application Type

Application Type	Coverage
DNA-Seq (Re-Sequencing)	30 - 80X
DNA-Seq (De novo assembly)	100X
SNP Analysis / Rearrangement Detection	10 - 30X
Exome	100 - 200X
ChIP-Seq	10 - 40X

Coverage (RNA)

A more useful metric for RNA-Seq is determining the total number of mapped reads.

- It is important to distinguish between total reads and mapped reads, as not all reads will map onto a reference genome

So, the number of usable reads will be less than the number of actual reads.

- The number of reads that will map depend on the
  - ❖library type
  - ❖quality of sample
  - ❖how complete the reference genome is

Coverage (RNA)

Coverage needed for a RNA is not always uniform:

- Different transcripts are expressed at different levels, meaning more reads will be captured from highly expressed genes while fewer reads will be captured by genes expressed at low levels
- Alternate expression

Coverage (RNA)

Recommended RNA-Seq Parameters

Optimal sequencing depth for RNA-Seq will vary based on the scientific objective of study but here are some general recommendations based on sample type and application:

Sample Type	Reads Needed for Differential Expression (millions)	Reads Needed for Rare Transcript or De Novo Assembly (millions)	Read Length
Small Genomes (i.e. Bacteria / Fungi)	5	30 - 65	50 SR or PE for positional info
Intermediate Genomes (i.e. Drosophila / C. Elegans)	10	70 - 130	50 – 100 SR or PE for positional info
Large Genomes (i.e. Human / Mouse)	15 - 25	100 - 200	>100 SR or PE for positional info

General Steps of Sequencing Experiments

Experimental Design

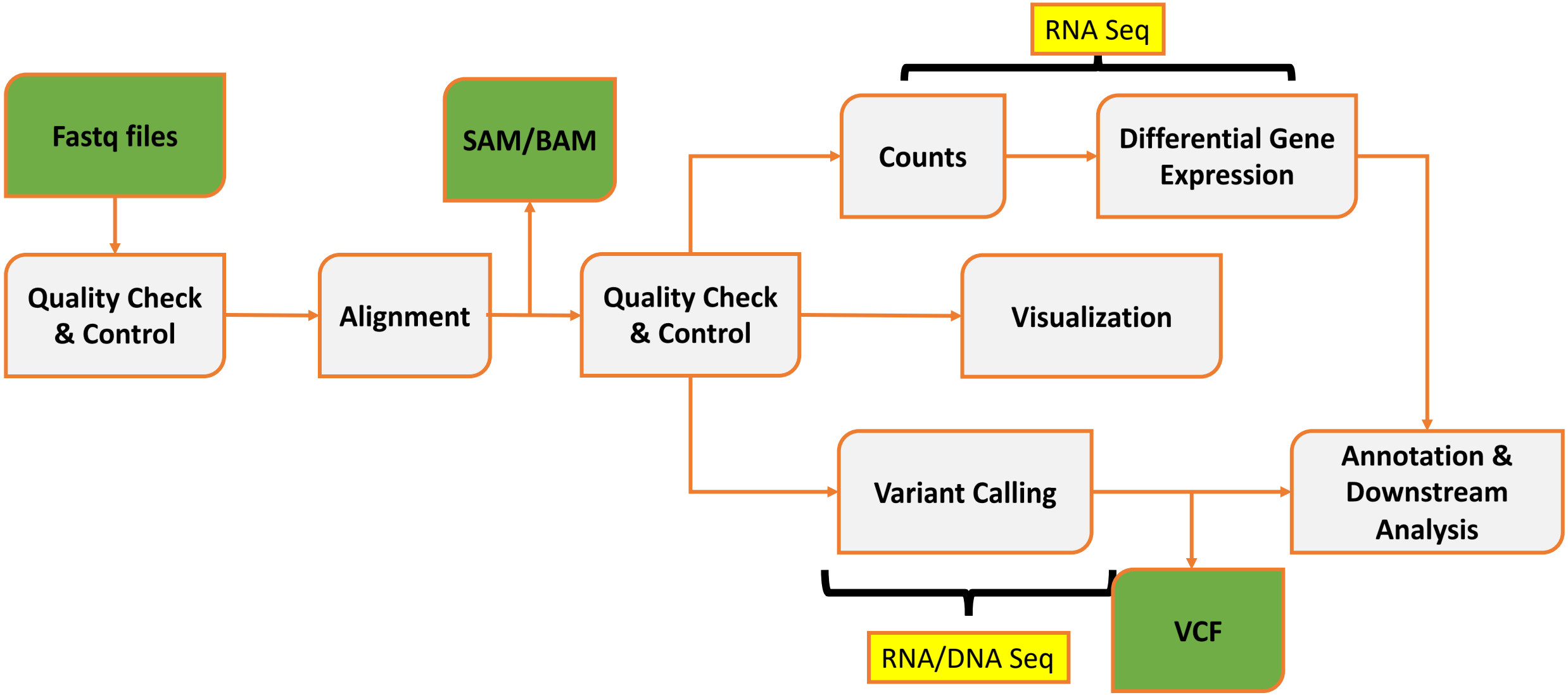
Sample extraction and preparation

Sequencing

Data Analysis

- ❑**Preliminary Data Processing and Analysis**
- ❑Further Downstream Data Analysis
  - ▪Annotation
  - ▪Visualization

Preliminary Sequencing Data Analysis Pipeline



Fastq Format

FASTQ format is a

- Text-based format
- Stores :
  - Biological sequence
  - **Corresponding quality scores**

Size of fastq files depend on:

- Type of experiment – WGS > Exome > RNA
- Type of Genome – Human > Mouse > Bacteria
- Coverage – more the coverage greater the size of fastq file

FASTA Format

```
>unique_sequence_ID My sequence is pretty cool
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAAGAATAC
```

FASTQ Format

sequence identifier and an *optional* description

raw sequence letters

```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAAGAATAC
+
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4DDD@@D
```

quality values for the sequence

Quality scores

Phred quality scores are used for:

- ❑ Assessment of sequence quality
- ❑ Represents the probably of a base-calling error
- ❑ Recognition and removal of low-quality sequence

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
  Quality score: 0........10........20........30........40
```

Quality scores

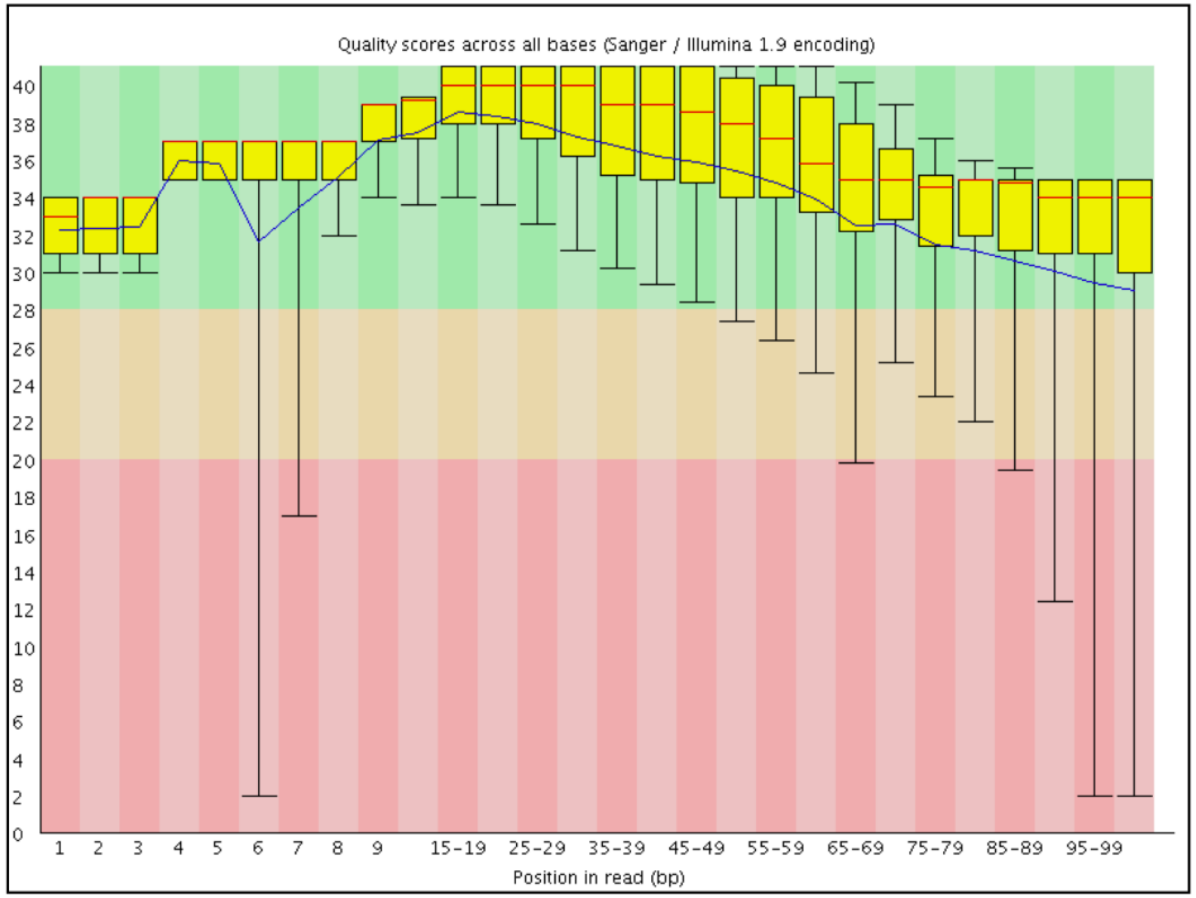
Phred Quality score	Probability base call = wrong	Accuracy of base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%

Paired End Fastq Format

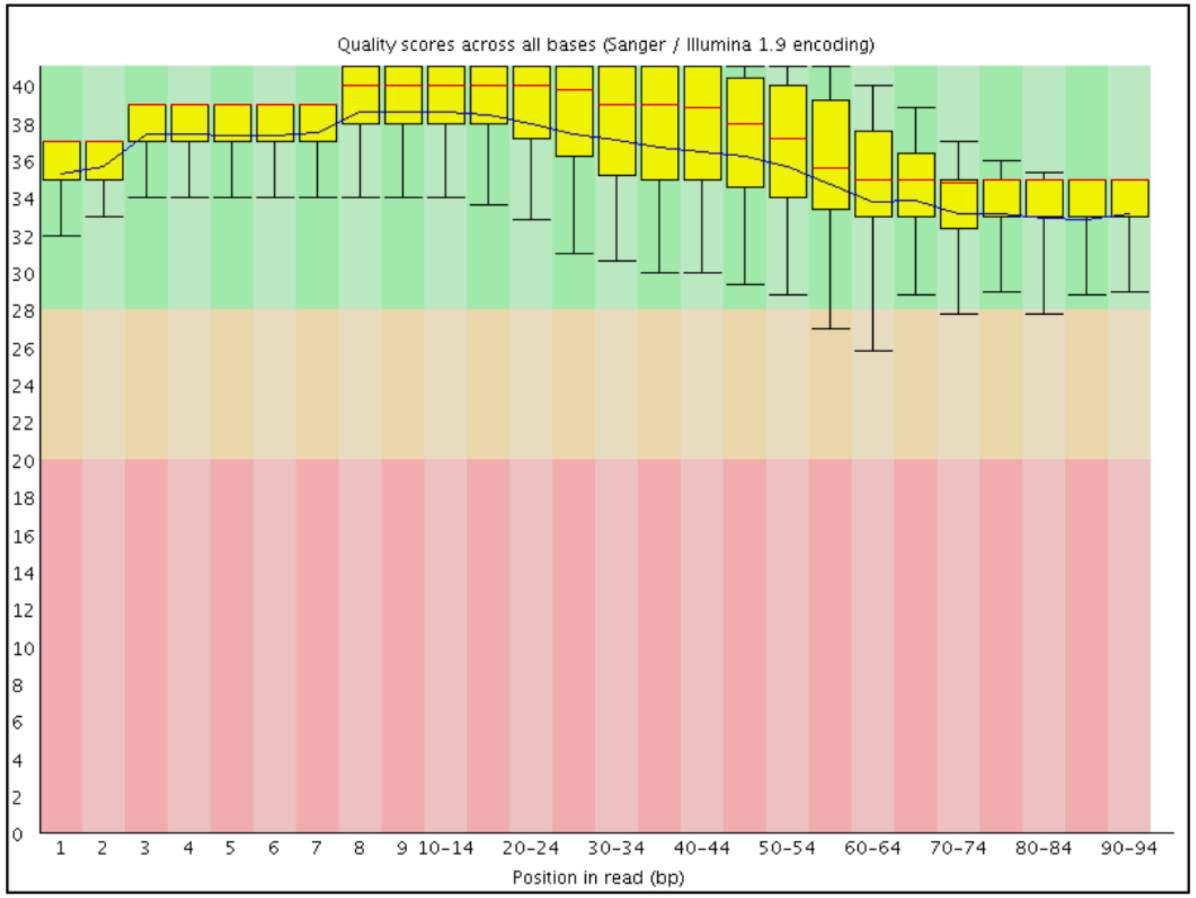
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<sample number>

Trimming

Before quality trimming

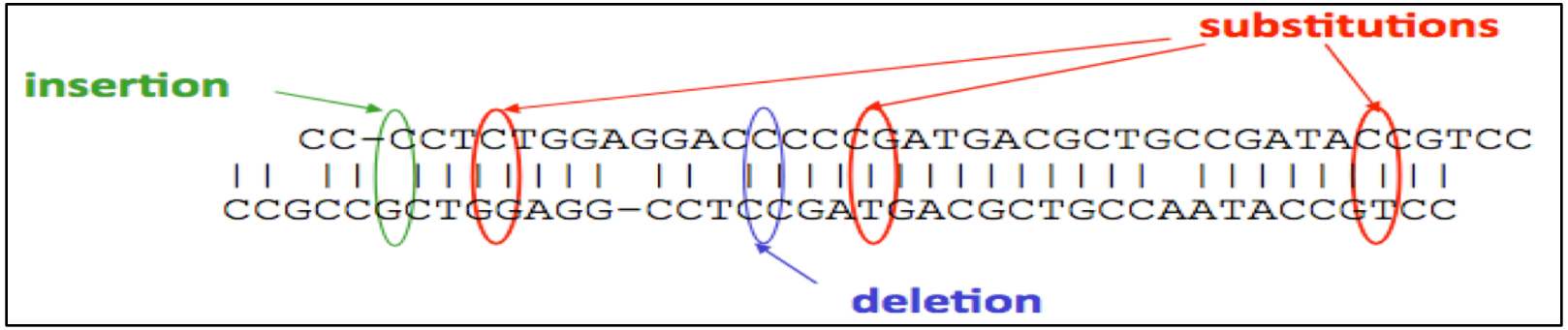


After quality trimming



Alignment

- Sequence a fragment of the RNA or DNA, then map (align) it to the genome
- Alignment = a mapping between the letters of the two sequences, with some spacers (indels)



Sequence Alignment Map (SAM) Format

- Sequence Alignment Map (SAM) format
  - Contains FASTQ reads
  - Quality information
  - Alignment information
  - Other information about samples (meta data) etc
- May be unsorted, or sorted by sequence name or genome coordinates
- Makes the alignment information easily accessible to downstream applications to extract specific features, e.g. genomic locations
- **Normally converted into BAM - binary form of SAM**

SAM format

SAM format specification - http://samtools.sourceforge.net/SAM1.pdf

```
@SQ SN:chr2 LN:181748087
@SQ SN:chr3 LN:159599783
@SQ SN:chr3_random LN:41899
@SQ SN:chr4 LN:155630120
@SQ SN:chr4_random LN:160594
@SQ SN:chr5 LN:152537259
@SQ SN:chr5_random LN:357350
@SQ SN:chr6 LN:149517037
@SQ SN:chr7 LN:152524553
@SQ SN:chr7_random LN:362490
@SQ SN:chr8 LN:131738871
@SQ SN:chr8_random LN:849593
@SQ SN:chr9 LN:124076172
@SQ SN:chr9_random LN:449403
@SQ SN:chrM LN:16299
@SQ SN:chrUn_random LN:5900358
@SQ SN:chrX LN:166650296
@SQ SN:chrX_random LN:1785075
@SQ SN:chrY LN:15902555
@SQ SN:chrY_random LN:58682461
@PG ID:bwa PN:bwa VN:0.6.1-r104
```

Header section

The header section starts with '@' and it contains information such as the name and length of the reference sequence.

```
2:2208:33.30:98.40:N    4   *       0          0   *    *   0   0   AATCAGAGNNNNNNNNNNCNNNNNNNNNNNNNNNNNNNNNNNNNNNNN   2(222..2!!!!!!!!!!(!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
2:2208:38.40:99.10:Y    0   chr3    96363649   37  47M  *   0   0   AGAGGCAATCACATTATCTCAGGGACTTGCGCATCGTGGCGTGGCAA     4422232222222223,333332222111111000000000-.---..
2:2208:63.30:98.20:Y    16  chr19   44471151   37  47M  *   0   0   AATCCAAGGAGGAAACTAAAGCAATGGATGGAGCCAACACCATGCGG     00*0000000*00111*1111114112222332+2222+22002244
2:2208:60.10:98.60:Y    0   chr4    133237527  37  47M  *   0   0   AAAGTGGGTCTGGATATGCTCAGTGGTATCCCCACCACCAATCCGCC     44222222023223233333332222221111111100000000000
2:2208:66.80:98.80:N    16  chr2    35163284   37  47M  *   0   0   CGCGCGTTTTTTTGTTTTTGTTTTTTACGGTAGCCAAAACAAGCCCT     ',,,---'...(000)000001111112+2332)22222022++222
2:2208:68.90:99.50:Y    0   chrM    6227       23  47M  *   0   0   CACACGAGCTTACTTTACATCAGCCACTATAATTATCGCAATTCCTA      44220000003222222222221111111111000100000000000/
2:2208:71.30:99.90:Y    0   chr2    121976626  37  47M  *   0   0   GTATACTCACGCCACCCACCGGAGAATGGGAAGCCGAACATACTGAA      44222222222222333322210100000000000/0----.......
2:2208:81.10:97.80:Y    16  chrM    11423      37  47M  *   0   0   AACCAACCATATAATTAACCTCCAACCCTCACACACACGAGAACTAA      .-0000000000000000111412222333332223222222222244
2:2208:93.30:97.90:Y    4   *       0          0   *    *   0   0   GGAGTCAGAGAGAGAGAGAGAGAGAGAATGAGAATGACAGGGGAAGA      44202222222222233222111111)*0**0*00000000/-"-.
```

Alignment section

SAM format - CIGAR

For example:

```
RefPos:      1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:   C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

Aligning these two:

```
RefPos:      1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:   C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                     A  C  T  A  G  A  A     T  G  G  C  T
```

With the alignment above, you get:

```
POS: 5
CIGAR: 3M1I3M1D5M
```

3 Match , 1 Insertion, 3 Match, 1 Deletion, 5 Match

General Steps of Sequencing Experiments

Experimental Design

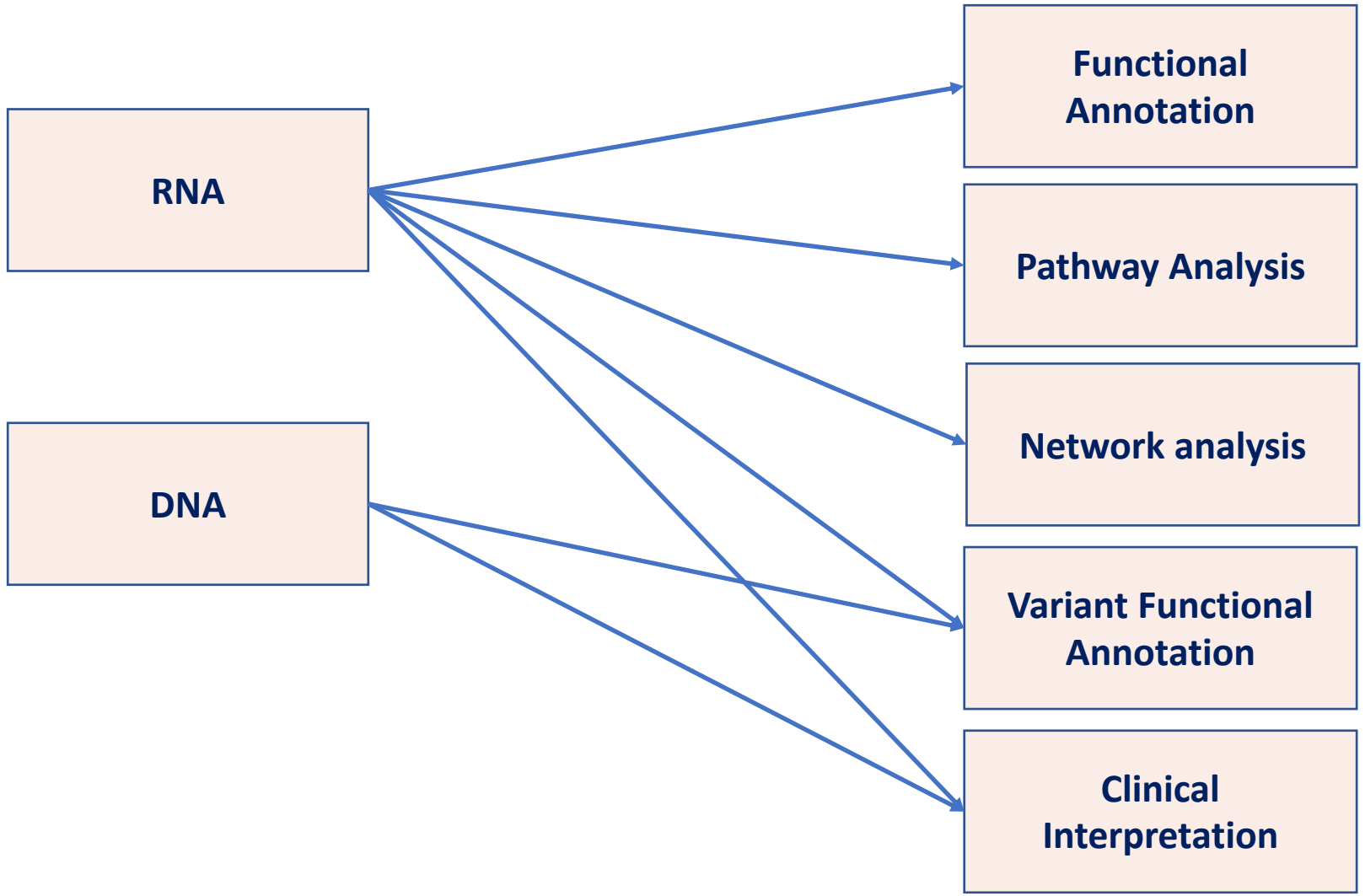
Sample extraction and preparation

Sequencing

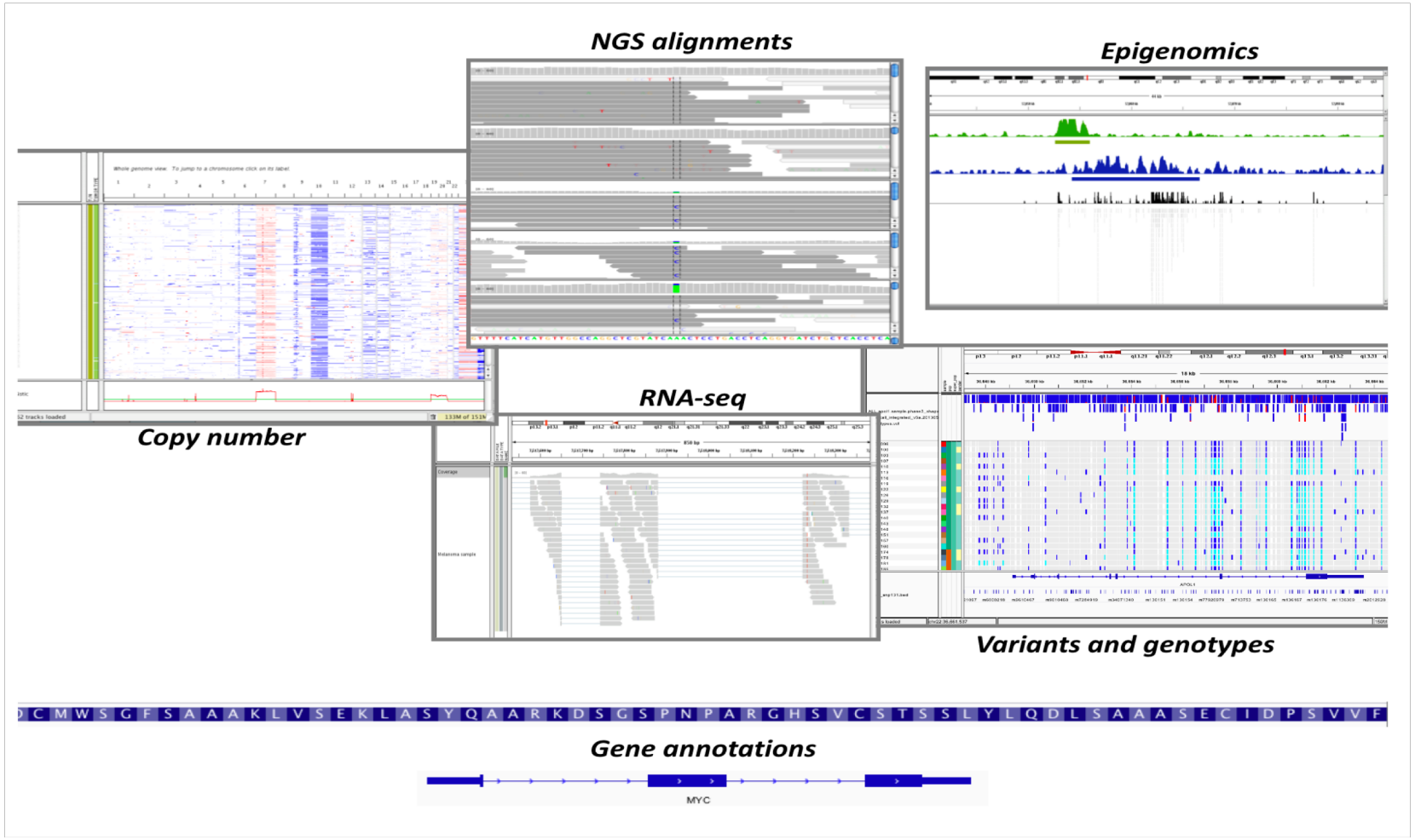
Data Analysis

- ❑Preliminary Data Processing and Analysis
- ❑**Further Downstream Data Analysis**
  - ▪**Annotation**
  - ▪**Visualization**

Further Downstream Data Analysis



Visualization



Take Home Message

- What **kind of sequencing** you want for your experimental question
- What kind of **sample preparation kit** you require for your sequencing experiment?
- What kind of library preparation you want to do? – **SE or PE**
- How much **Coverage** is required for your experiment?
- **How many samples** do you have ?
  - How many **replicates** you plan to do?
- What sequencing **instrument** will suit your experimental design?
- Data Analysis – Do you need just **preliminary data analysis** or **both preliminary and downstream data analysis**?

Coming up

Transcriptomics (March)

THANKS!