Module 5: Why do we need RNASeq?

Key concepts

- Quality control (continuation of Module 4)
- Gene-environment interactions
- Applications:
    - Alternative splicing events
    - Structural variant detection

Sequence Quality

Quality Score (Q) = -10 log10(e), where **e** is the estimated probability of the base call being wrong.

- **Higher Q scores** → Smaller probability of error
- **Lower Q scores** → Higher probability of error, leading to:
  - Unusable reads
  - False-positive variant calls
  - Inaccurate conclusions

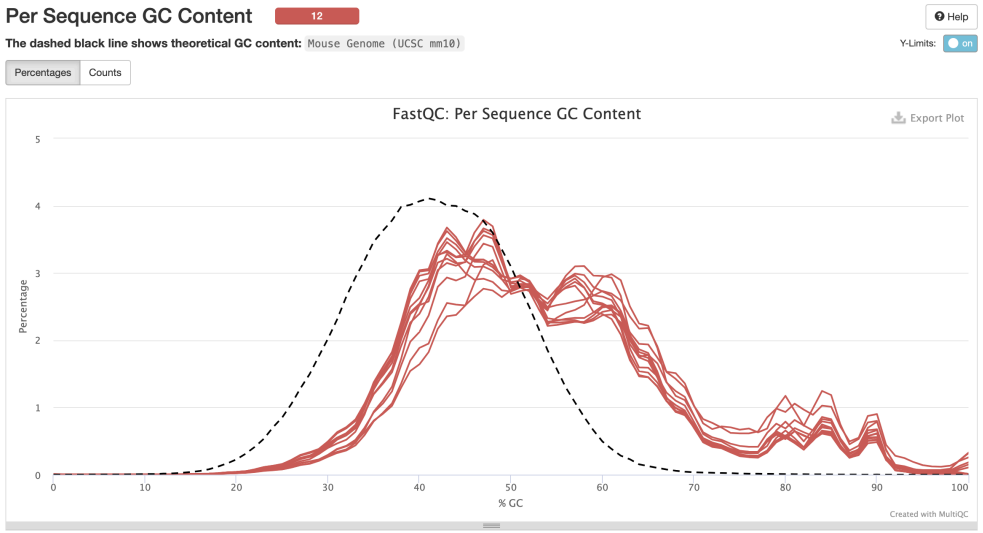
Quality Score Thresholds

- **Q20**: Error rate=1%; Call accuracy = **99%**
- **Q30**: Virtually all reads are error-free (**benchmark for NGS**)

GC Content

- The GC content of the reads should be similar to the expected % GC for the organism (see reference table here).
- The distribution should be normal unless over-represented sequences (sharp peaks on a normal distribution) or contamination with another organism (broad peak).

What happens when you align the reads to the wrong reference genome? [1]

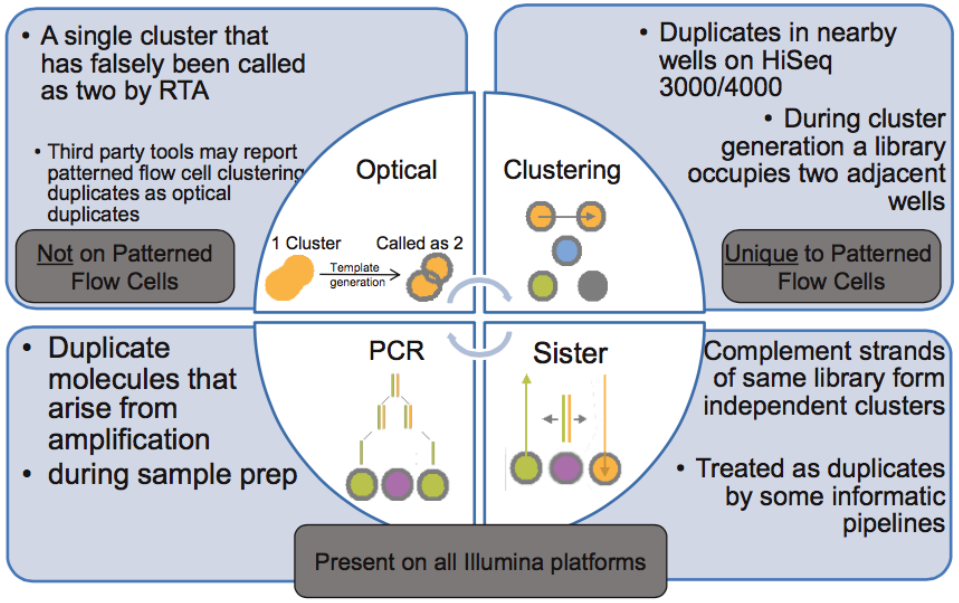


"[1]: (https://www.biostars.org/p/434543/)"

Overrepresented sequences (K-mers)

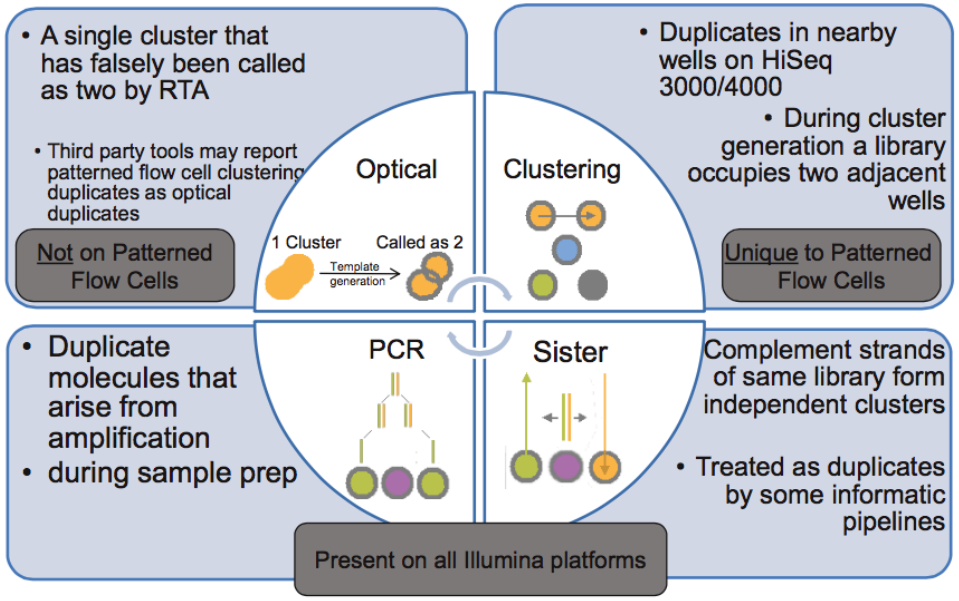
- The sequences (at least 20 bp) that occur in more than 0.1% of the total number of sequences
- Overrepresented sequences are usually adapters from the library prep and requires adapter trimming using tools like `fastp` or `trimgalore`

Read Duplication



- **Optical duplicate**: a duplicate read that occurs when a sequencer separates a single cluster into multiple clusters.
- Duplicates unique to patterned flow cells arise from when the library molecules from a cluster return to the surrounding solution and then act as a seed for second cluster in a nearby flow cell.[2]

Read Duplication



- A **PCR duplicate** is a duplicate read that arises from occur in library preparation during the PCR amplification stage.
- Sister duplicates arise when the same DNA strand is used to form multiple clusters on the flow cell.

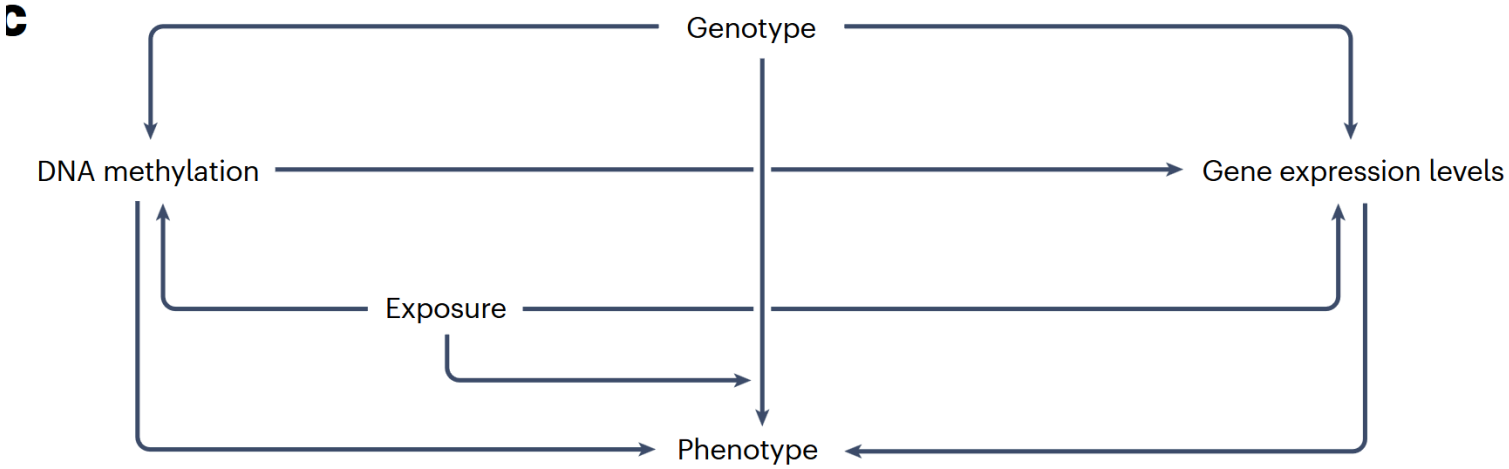
Read Duplication

- **Single-end** reads tend to have higher duplication rate than **paired-end** reads.
- **Higher** sequencing depth usually lead to **higher** duplication rate.
- **High** duplication rate does not necessarily indicate problems with the library preparation.
- **Highly** expressed genes have many copies of mRNA molecules, thus are expected to have **many** duplicate reads in the library, while **lowly** expressed genes should not.

Genotype × environment interactions (GxE)

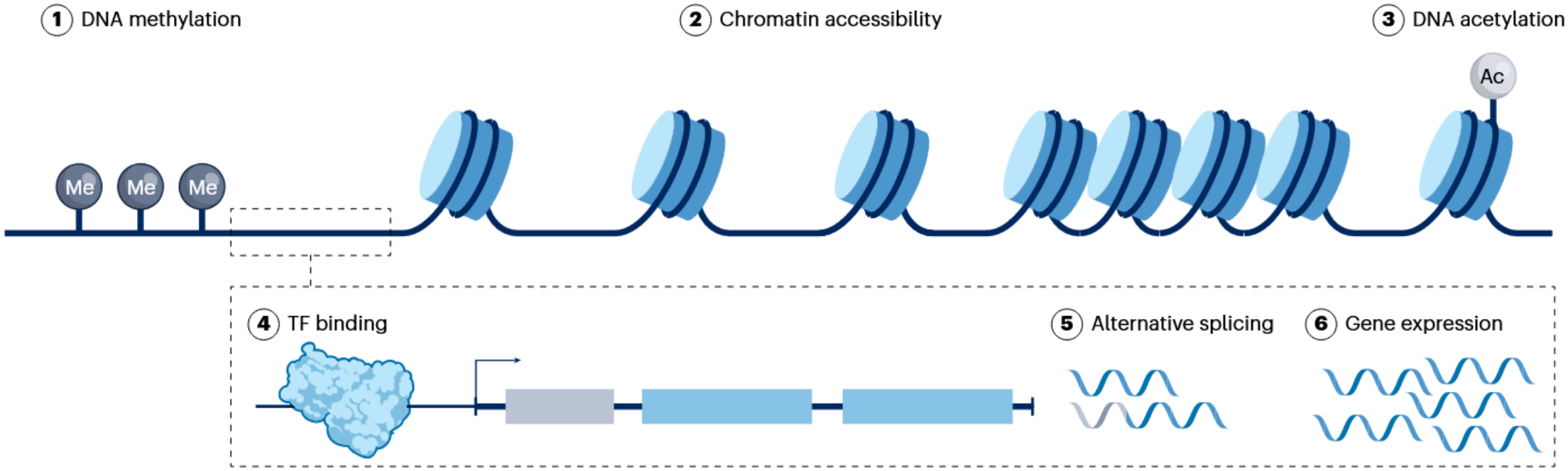
GxE occur when the genetic effect on a phenotype varies depending on the environment or, vice versa, the effect of the environment on a phenotype varies depending on the individual's genotype.

- Does a polygenic G × E interaction contribute to a human trait(and by how much) across one or more environmental contexts? [3]

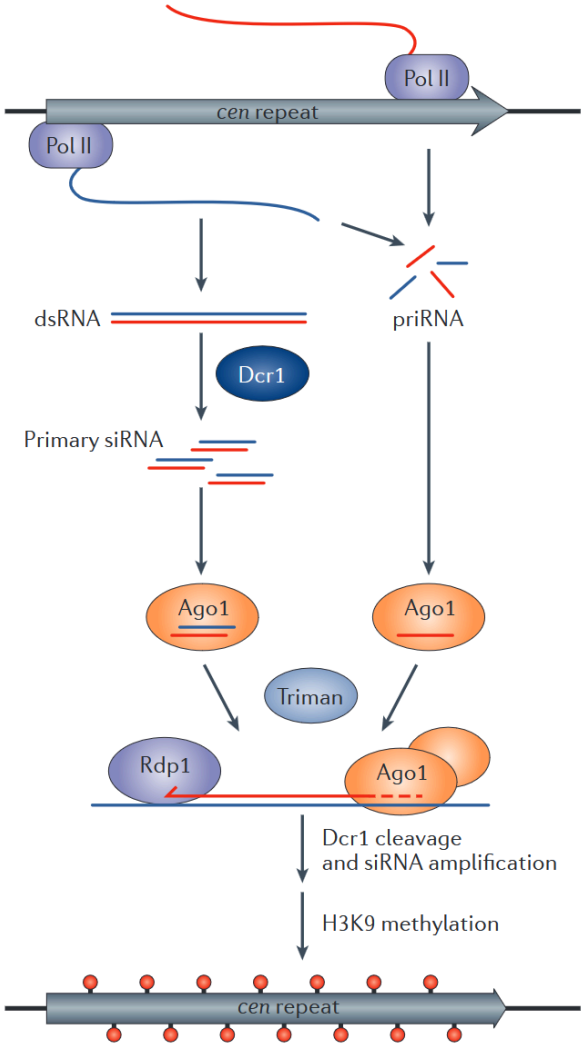


GxE interactions[4]

Different gene-regulatory mechanisms can be impacted by GxE.



Epigenetics in complex diseases[5]



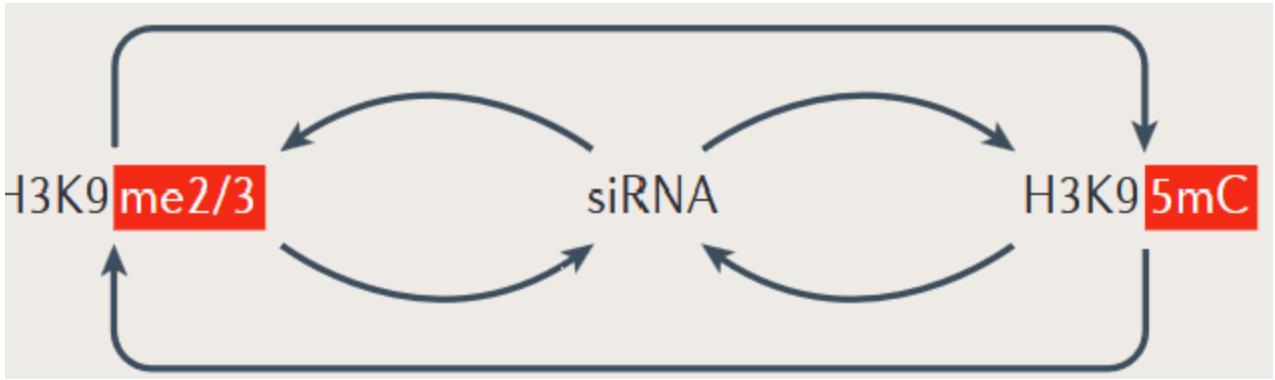
RNA interference (RNAi): various RNA silencing pathways that use small RNAs, together with a member of the conserved Argonaute (AGO) and PIWI family of proteins, to target genes for inactivation at the post-transcriptional or transcriptional levels

- *Drosophila melanogaster* and mammals have small RNAs--PIWI-interacting RNAs (piRNAs)-- which mediate RNA degradation in the cytoplasm and DNA or histone methylation in the nucleus

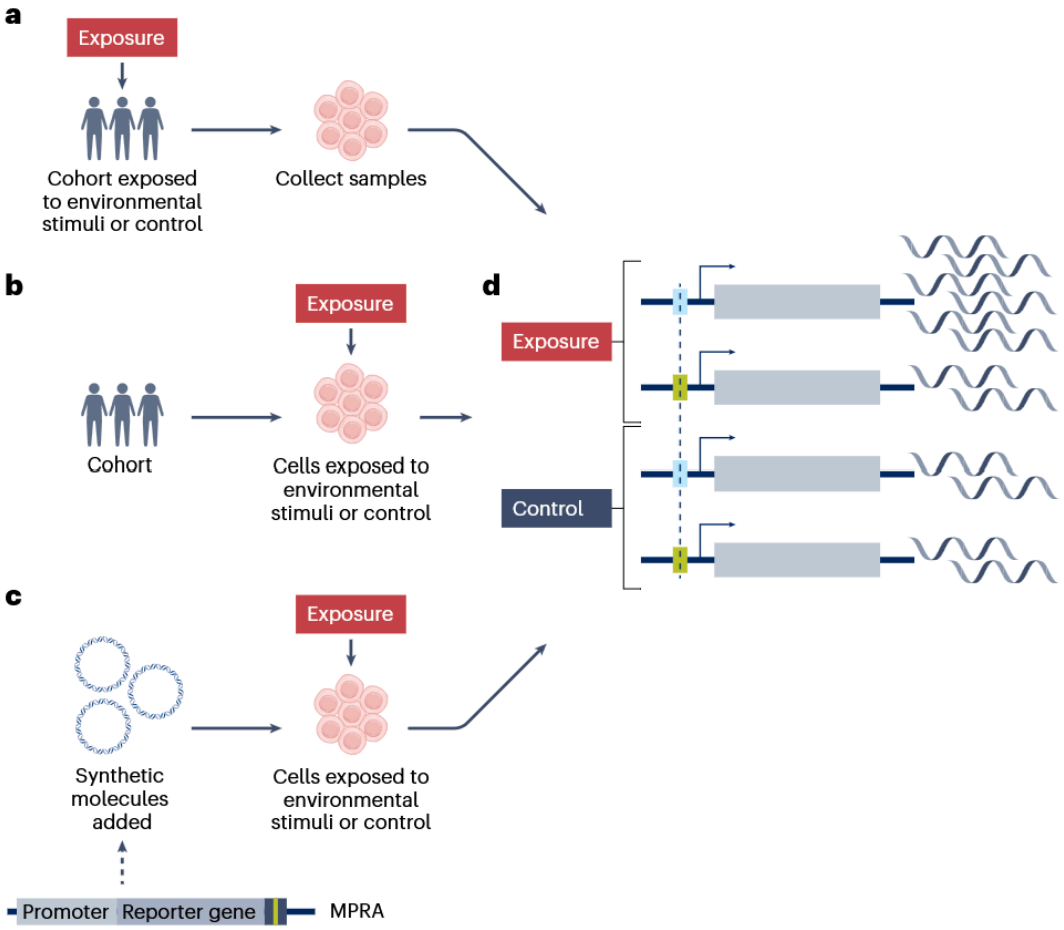
Epigenetics in complex diseases[5]

Self-reinforcing positive feedback loops are formed by the functional coupling of different types of signal generation events.

- You can understand the molecular basis of these coupling events by identifying the proteins that recognize either histone H3K9me2 and H3K9me3, or DNA cytosine methylation (DNA-5mC), and the proteins that recruit enzymes (e.g. siRNAs) which catalyse the other methylation events.



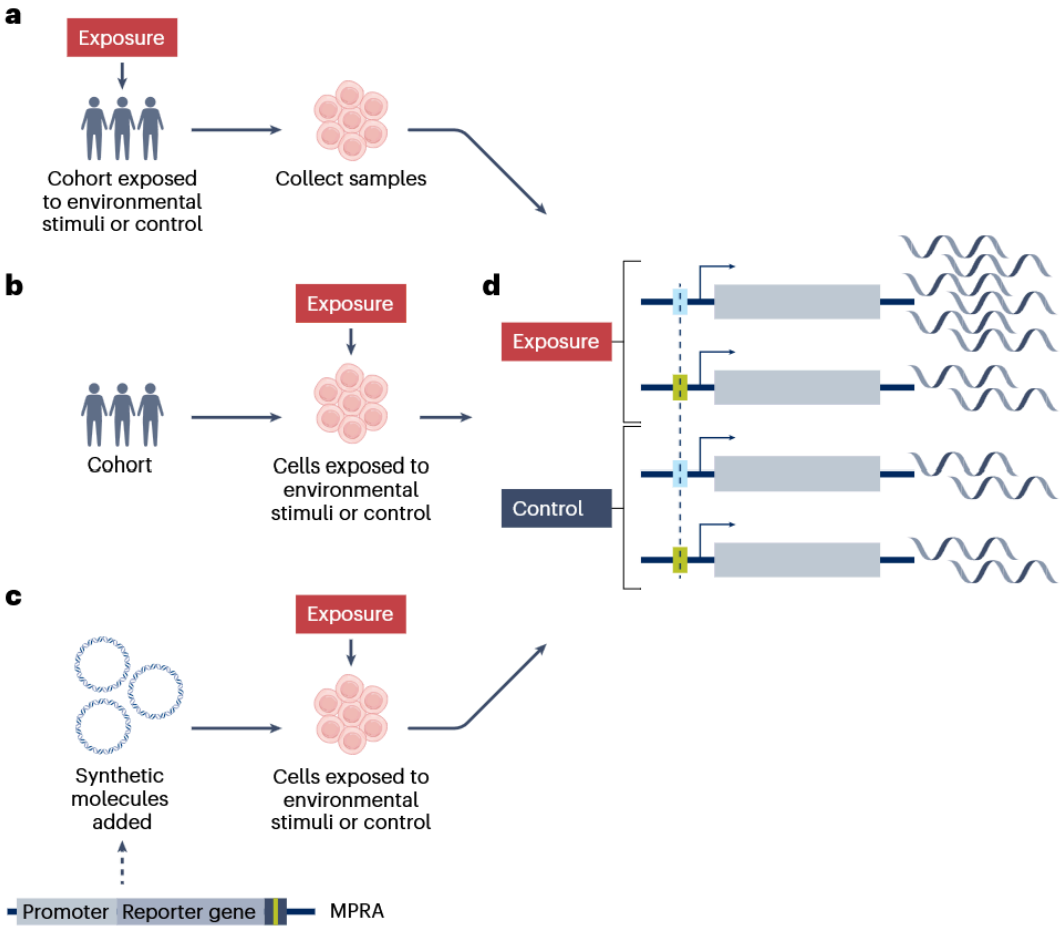
GxE interactions: Study Designs[4]



A. The environment is measured in a cohort, biological samples are collected, and both molecular readouts (e.g., gene expression) and genotypes are assessed.

B. Biological samples are collected from a cohort, then deliberately exposed to an environment; afterwards, molecular readouts and genotypes are measured.

GxE interactions: Study Designs[4]



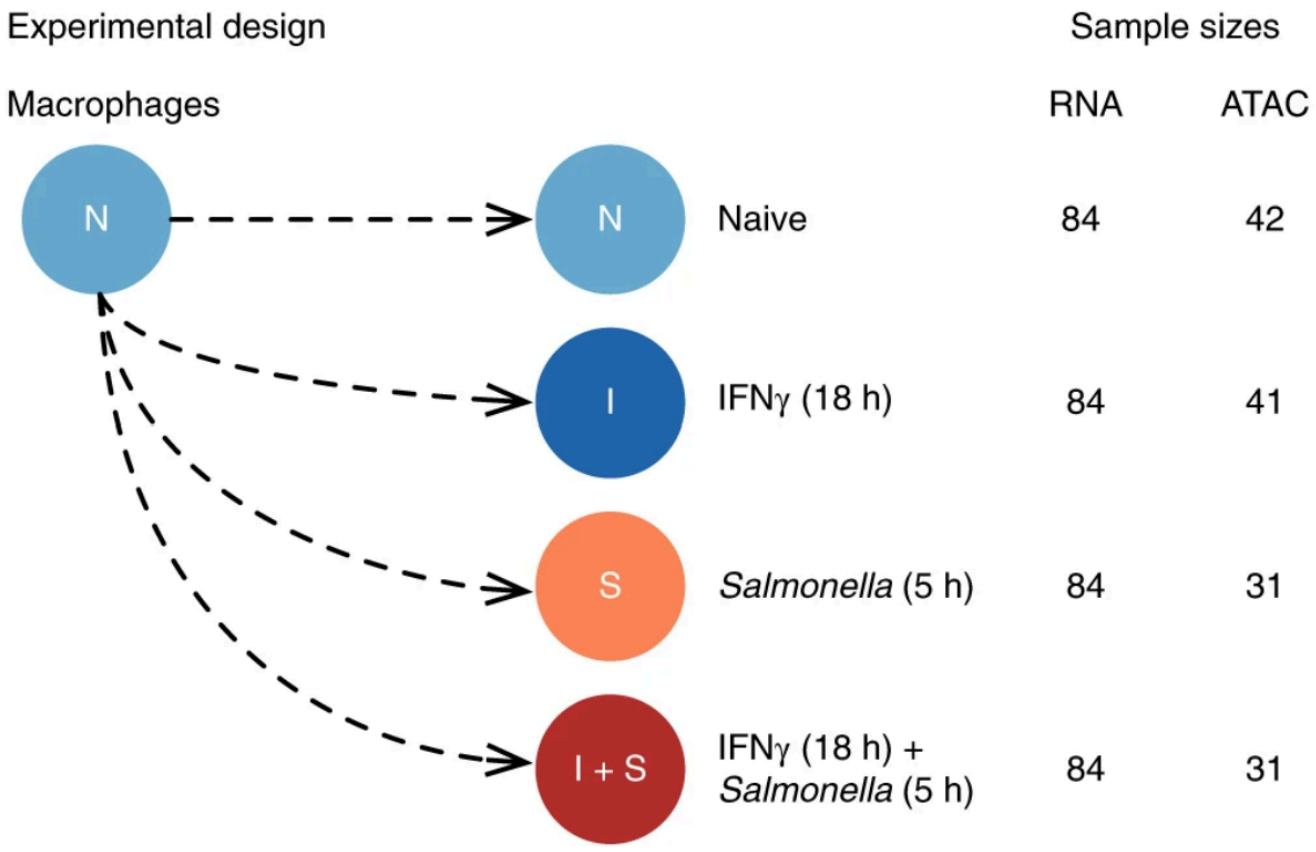
C. Engineered constructs (carrying a minimal promoter, a reporter gene, and a sequence of interest) are introduced into cells, and molecular readouts are compared across different alleles and environmental conditions.

Example: Molecular GxE Studies using Immune Stimuli[6]

Experimental objectives:

- Understanding molecular mechanisms such as binding of transcription factors that are activated in response to immune stimuli.
- Find complex interactions, such as GxE with sex or age, duration of the exposure or cell type
- Determine interindividual differences in pathogen response and autoimmune disorders

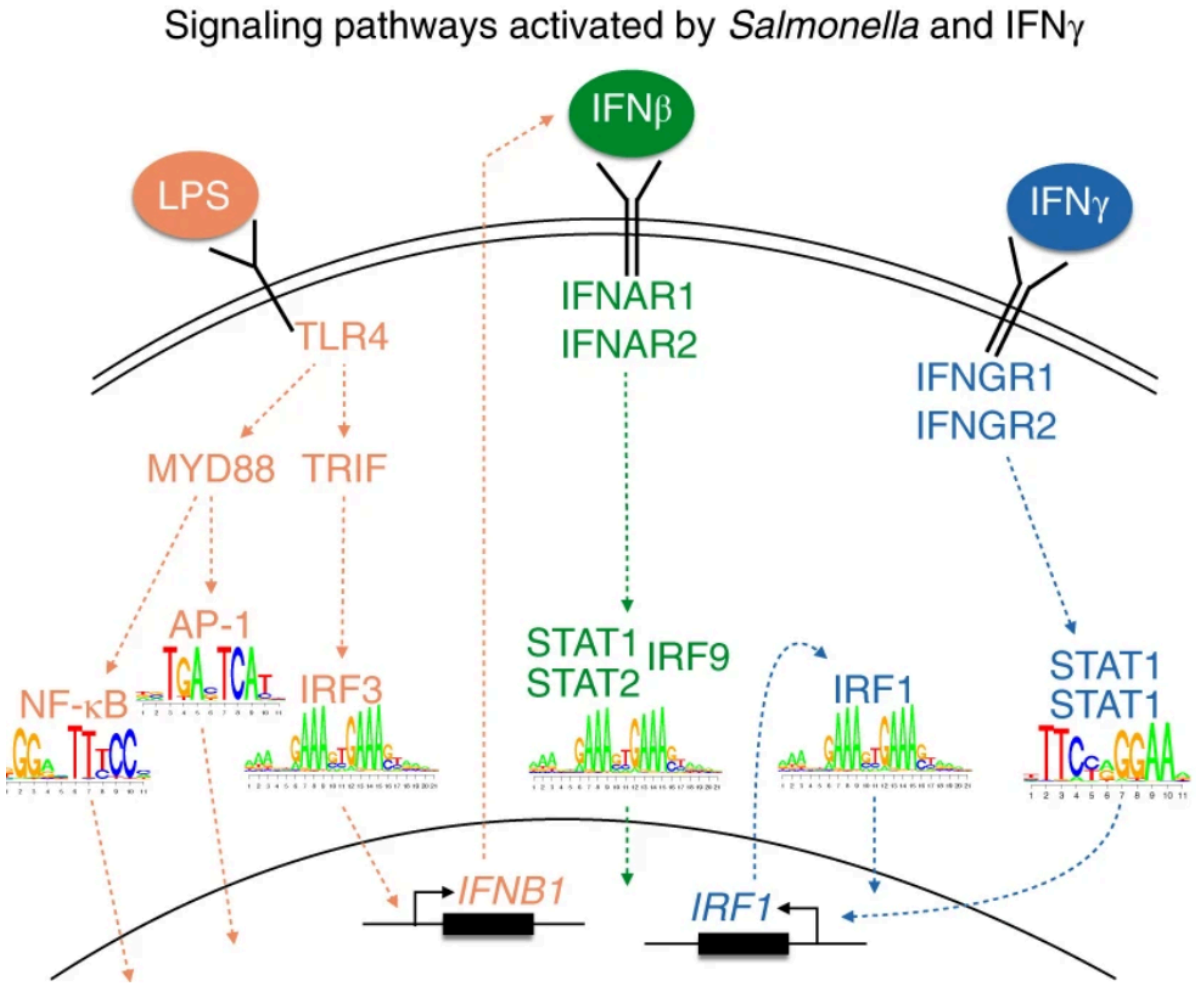
Example: Molecular GxE for environmental exposures[7]



What are the molecular mechanims for enhancer priming in immune response?

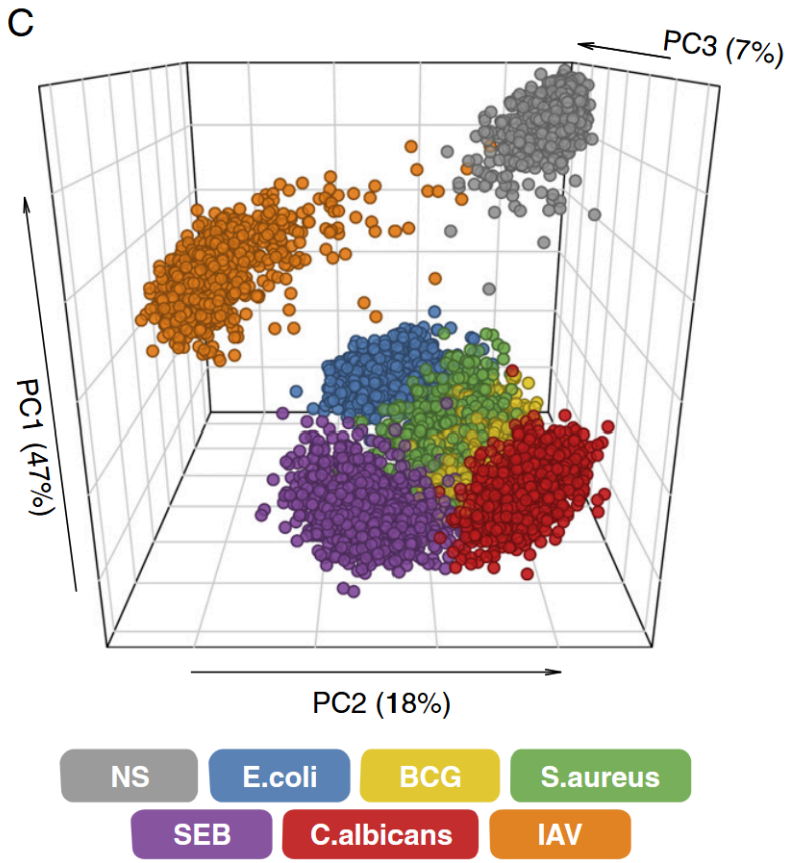
- Used ATAC-Seq and RNA-Seq to measure chromatin accessibility and gene expression in human macrophages exposed to IFNγ, Salmonella and IFNγ plus Salmonella

Example: Molecular GxE for environmental exposures[7]



Identified variants influencing the binding of cell-type-specific transcription factors, such as PU.1, which can then indirectly alter the binding of stimulus-specific transcription factors, such as NF-κB or STAT2.

Example: Molecular GxE Studies using Immune Stimuli[8]



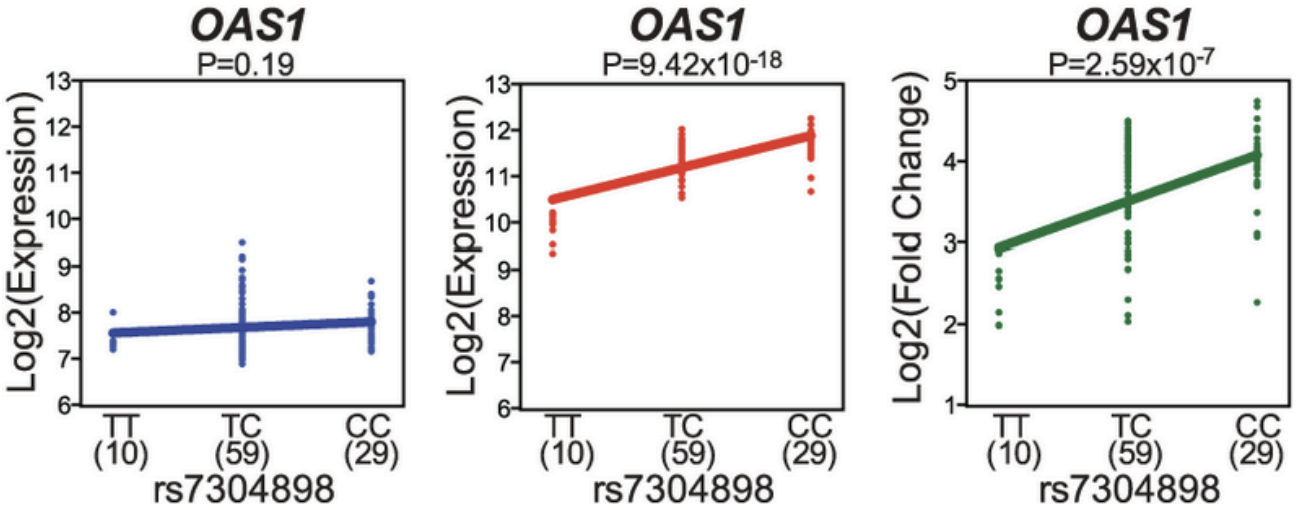
Do age and sex influence the gene expression of immune-related genes upon immune stimuli?

- Stimulated whole blood samples with three bacteria (*E. coli*, *S. aureus*, and *B. Calmette–Guérin*); a fungus (*C. albican*); a live virus, *influenza A virus* (IAV); and a superantigen, *staphylococcal enterotoxin B* (SEB)

Example: Molecular GxE Studies using Immune Stimuli[9]

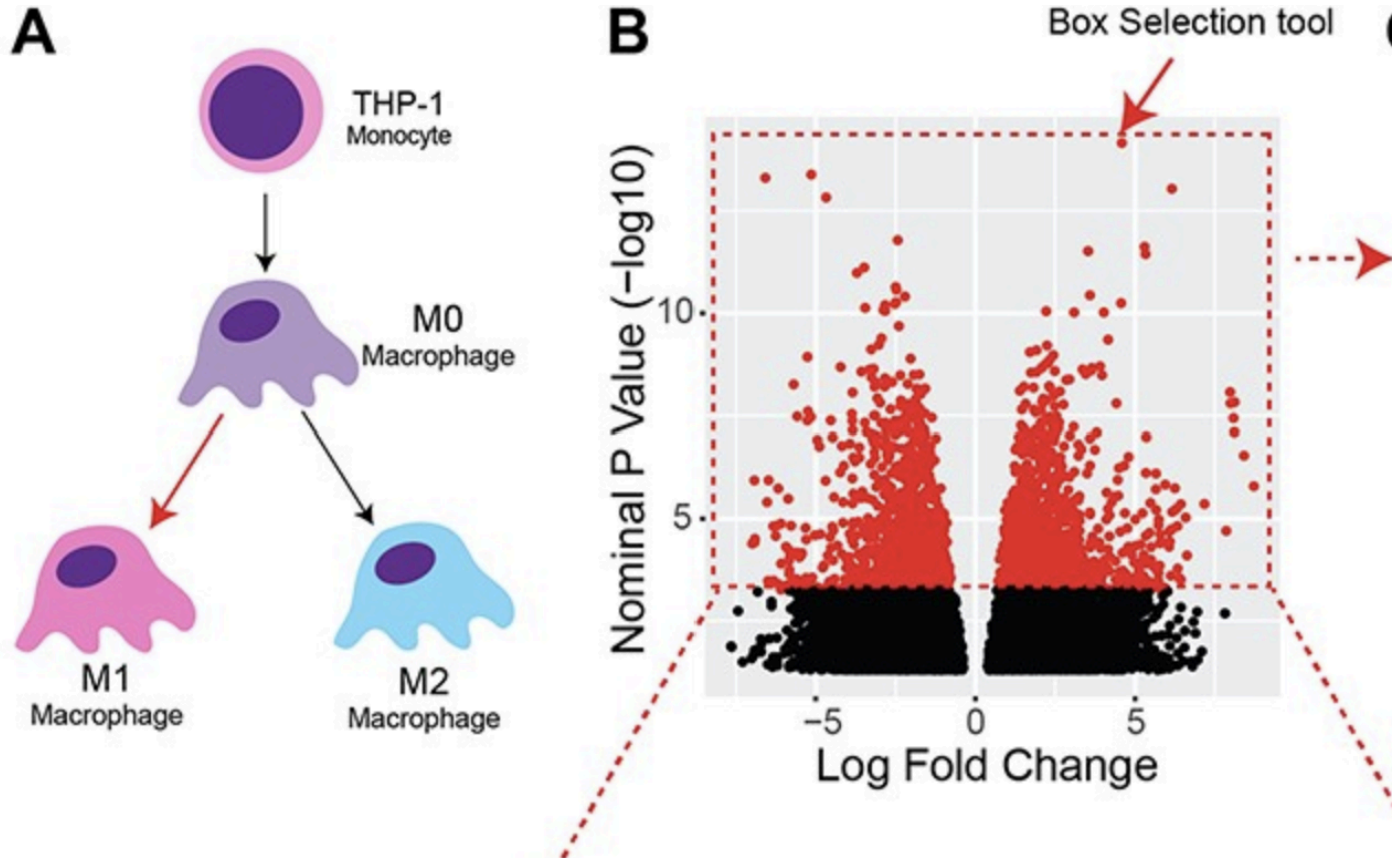
Is host genetic variation associated with variation in gene expression response to RV infections between individuals?

- OSA1 is a gene with known functions in viral response

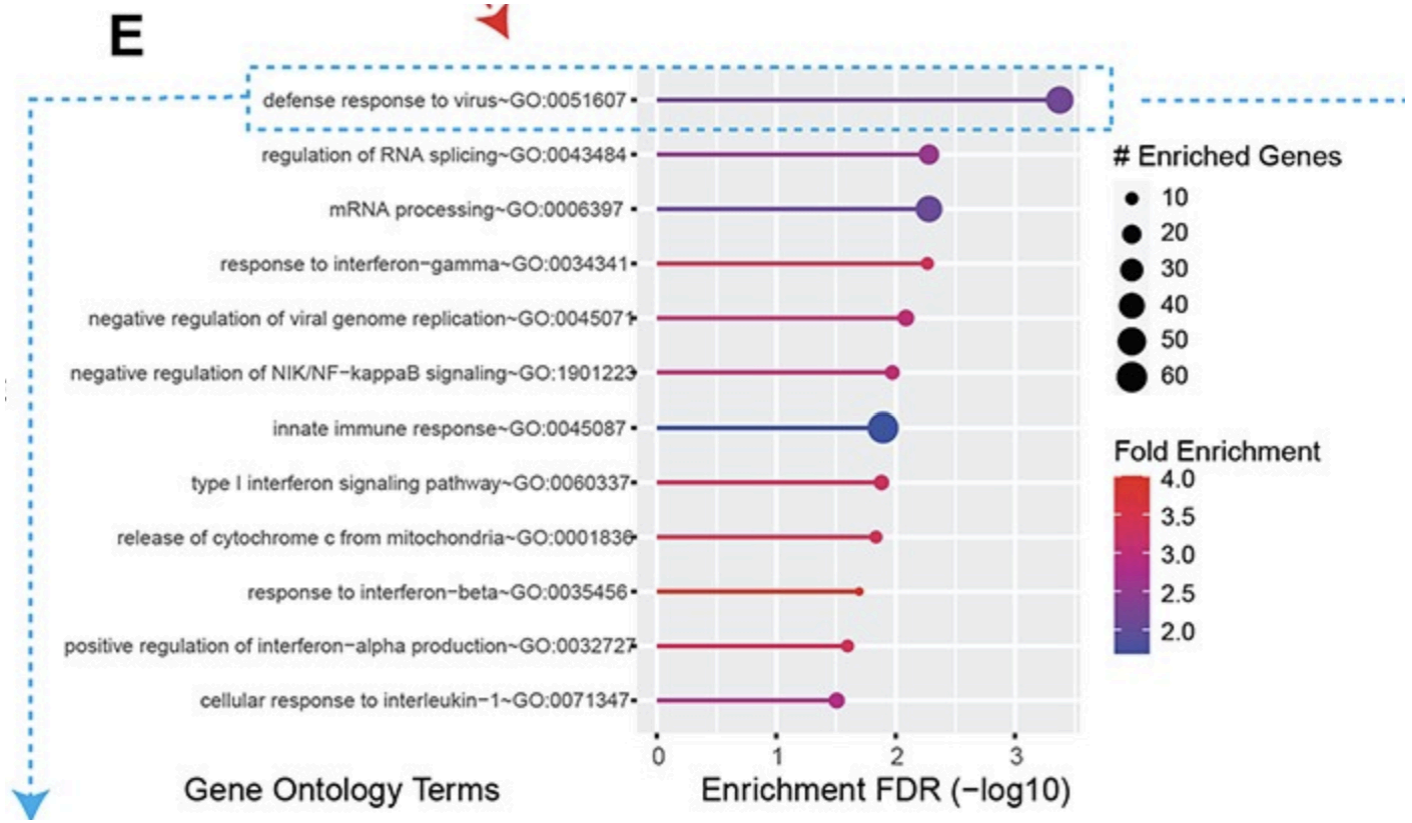


- uninfected cells (blue)
- RV-infected cells (red)
- reQTLs (green): differences in gene expression response

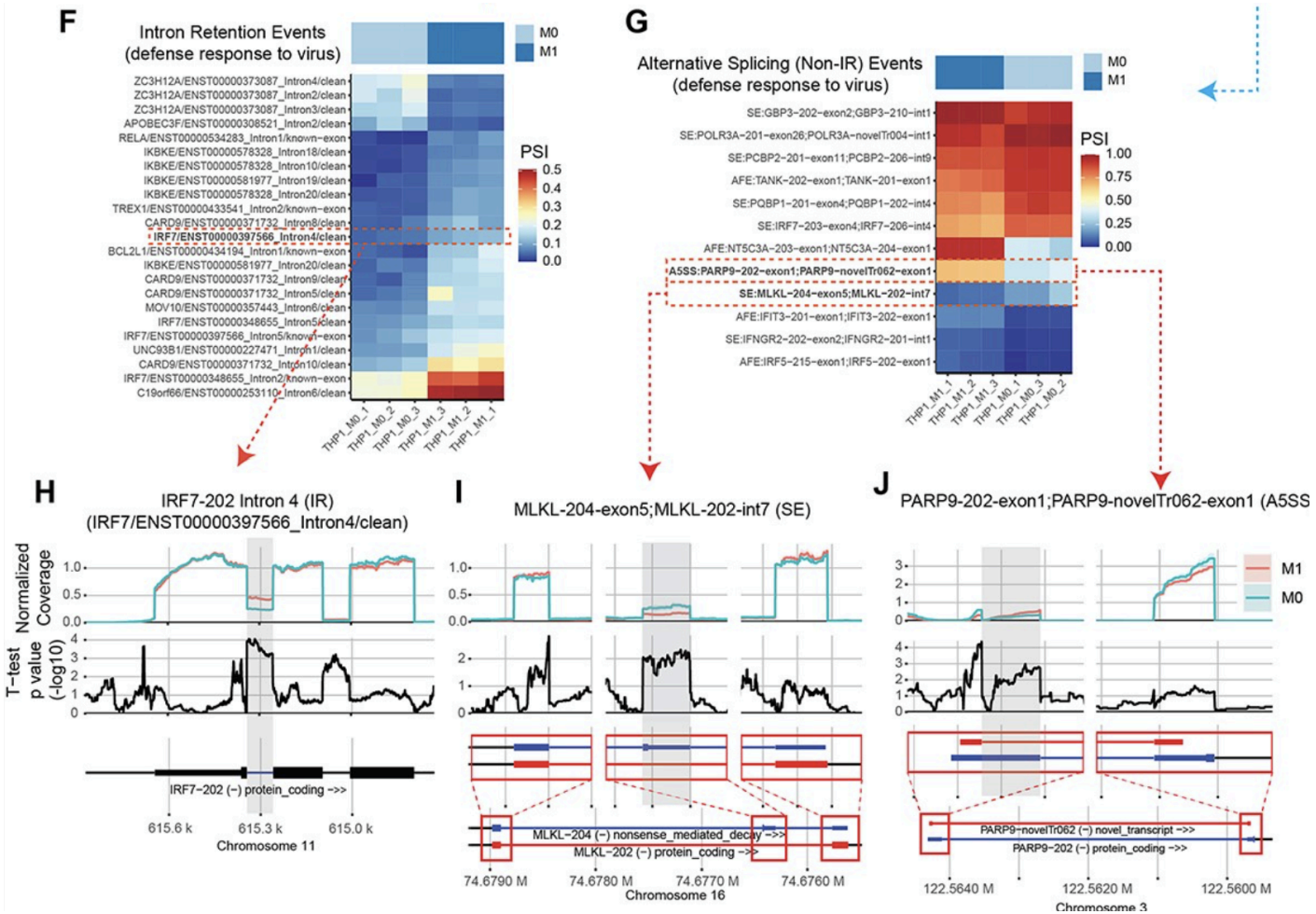
Applications: Alternative splicing events[10]



Applications: Alternative splicing events[10]

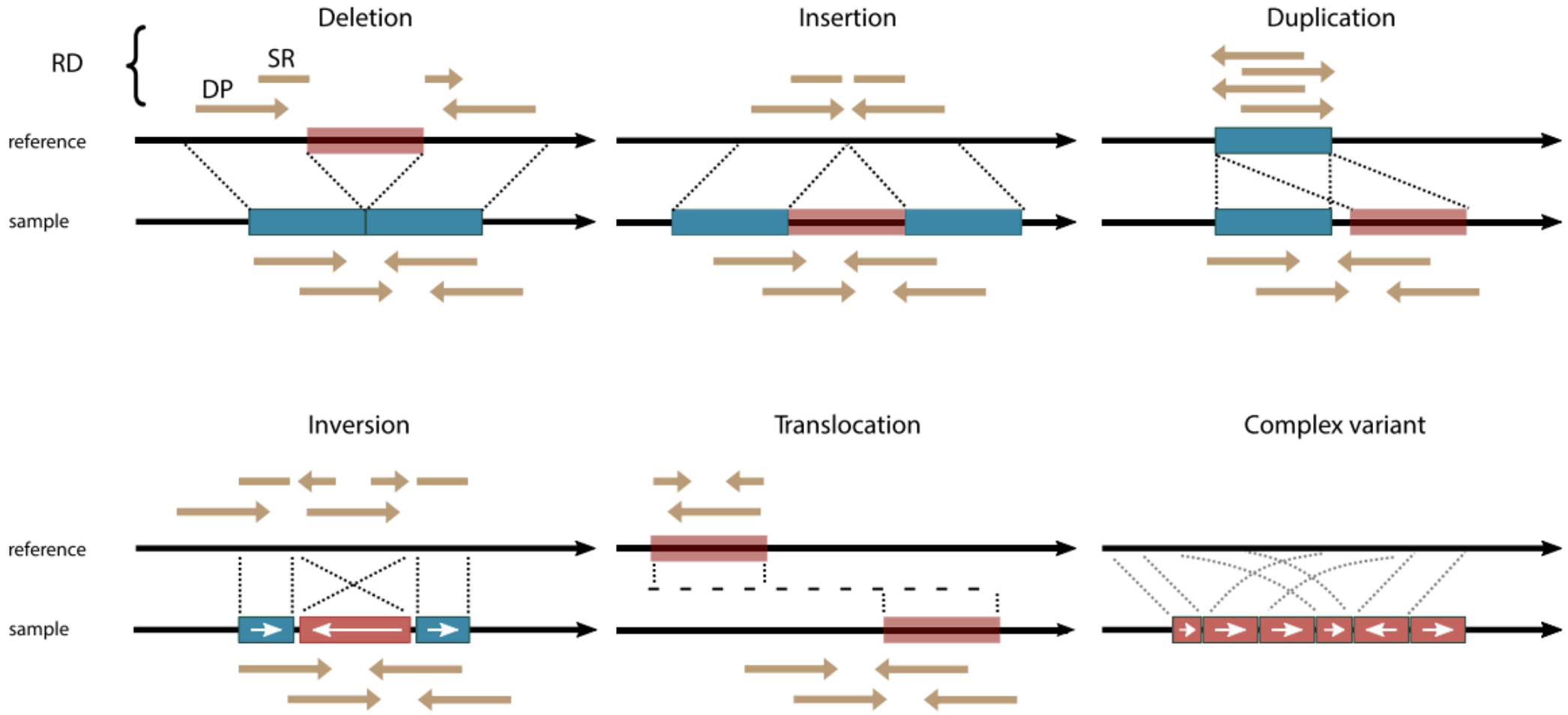


Applications: Alternative splicing events[10]



Applications: Structural variant detection

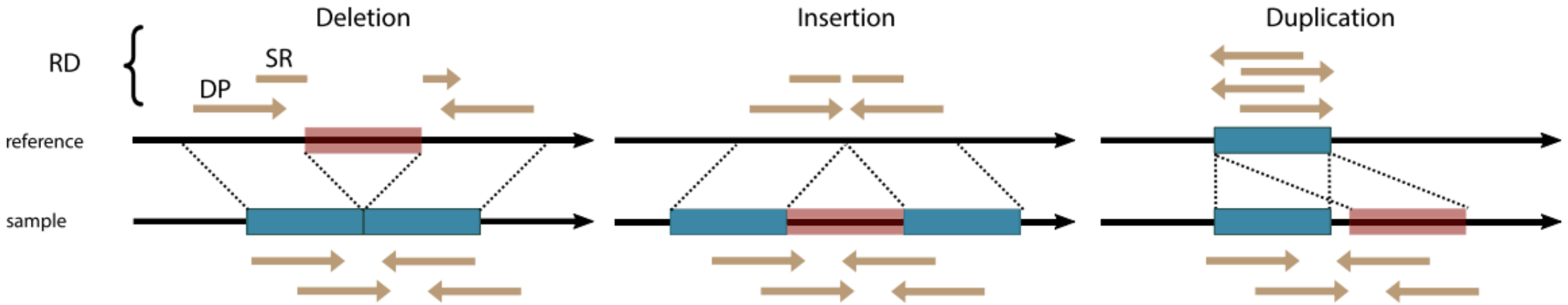
- SVs are genomic alterations **larger than 50 base pairs**



Applications: Structural variant detection

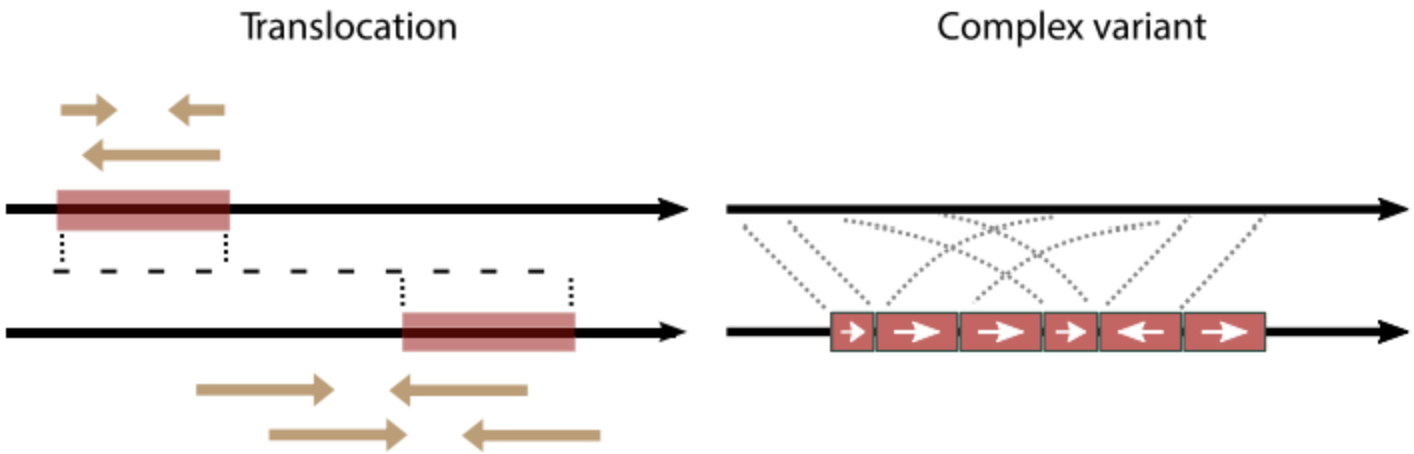
- SVs often indicate disruptions in genome stability
- Useful for identifying structural changes in genetic disorders and understanding complex genome rearrangements
- Cancer genomics and evolution

Applications: Structural variant detection (Patterns to look for)[11]



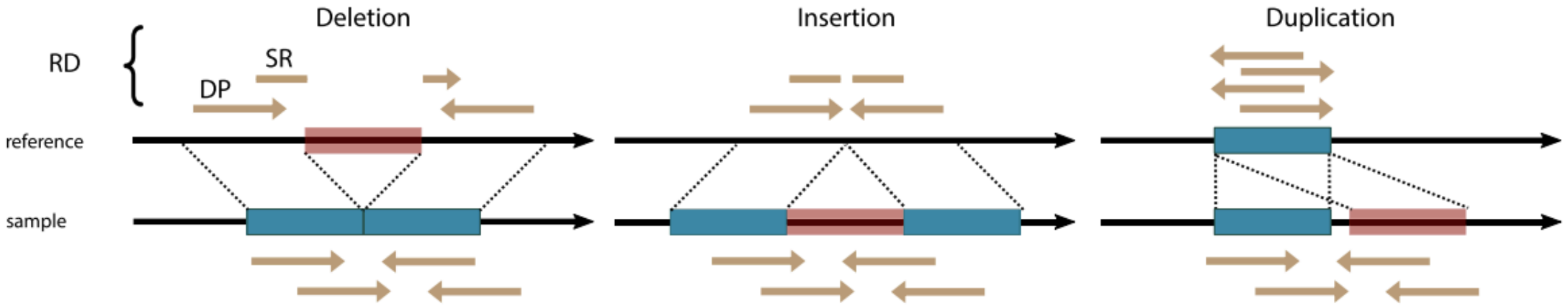
1. **Read Depth (RD):**
    - Changes in sequencing depth or coverage highlight larger duplications or deletions.
    - Primarily used to detect **Copy Number Variants (CNVs)**.

Applications: Structural variant detection (Patterns to look for)[11]



2. **Discordant Pairs (DP):**
    - Paired-end reads align to the reference at unexpected distances or orientations.
    - Best suited for detecting **large SVs** such as inter-chromosomal **translocations** and **inversions**.

Applications: Structural variant detection (Patterns to look for)[11]



3. **Split Reads (SR):**
    - Reads span breakpoint regions and align partially to reference sequences.
    - Detects **small SVs** with **base-pair resolution**.