# Computational tools and statistical approaches for transcriptomic and Whole-genome sequencing (WGS) data analysis in cancer

**Katsiaryna Davydzenka**

National PhD AI for Health and Life Sciences, XXXVIII cycle Bis

Biomedical Campus University of Rome

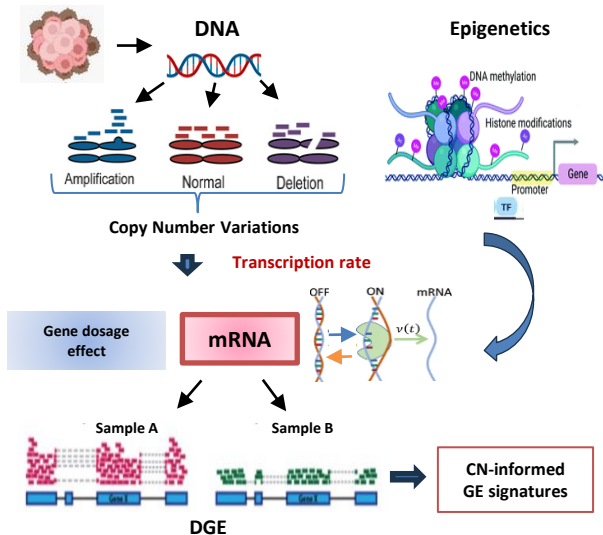International School of Advanced Studies (SISSA), Trieste

Machine Learning and Systems Biology group (SISSA) & Cancer Data Science Laboratory (UniTS)

EMBO Practical Course: Integrative analysis of multi-omics data
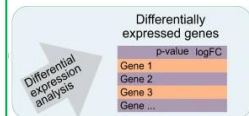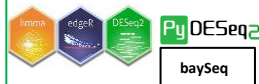27.02.2024

EMBL

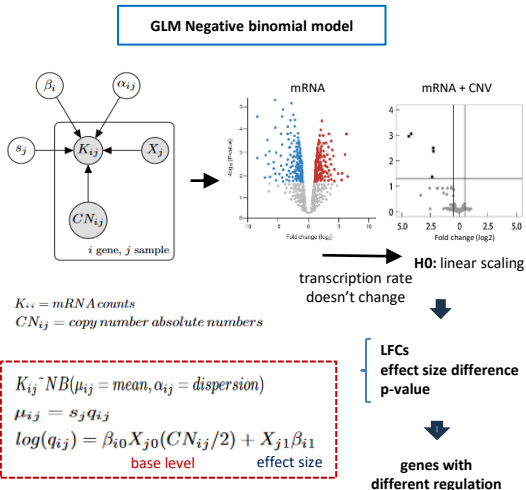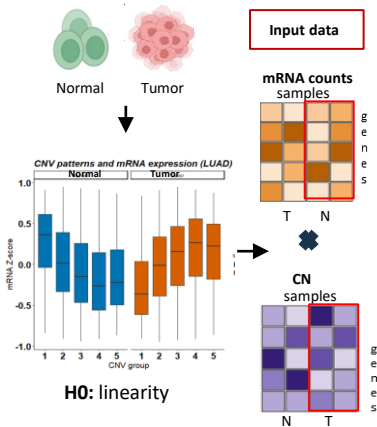# Copy-number-aware Differential Gene Expression (DGE) in cancer

## Background



**DNA**

**Epigenetics**

Amplification   Normal   Deletion

**Copy Number Variations**

**Transcription rate**

**Gene dosage effect**

**mRNA**

OFF   ON   mRNA

$v(t)$

Sample A   Sample B

**CN-informed GE signatures**

**DGE**

## State of the art statistical tools

**Empirical Bayes inference approach**

limma   edgeR   DESeq2   PyDESeq2

**baySeq**

Differentially expressed genes

Differential expression analysis

| | p-value | logFC |
|---|---|---|
| Gene 1 | | |
| Gene 2 | | |
| Gene 3 | | |
| Gene ... | | |

Katsiaryna Davydzenka, National Phd student in AI & Health and Life Sciences, UCBM, SISSA, Italy

**Goal:** combining copy number into DE analysis to separate genes affected by CNVs



**Input data**

Normal    Tumor

*CNV patterns and mRNA expression (LUAD)*

**H0:** linearity

**mRNA counts** samples

**CN** samples

**GLM Negative binomial model**

mRNA    mRNA + CNV

*i* gene, *j* sample

$K_{ij} = mRNA\ counts$

$CN_{ij} = copy\ number\ absolute\ numbers$

$$K_{ij} \sim NB(\mu_{ij} = mean, \alpha_{ij} = dispersion)$$
$$\mu_{ij} = s_j q_{ij}$$
$$log(q_{ij}) = \beta_{i0} X_{j0}(CN_{ij}/2) + X_{j1}\beta_{i1}$$

base level    effect size

transcription rate doesn't change

**H0:** linear scaling

LFCs
effect size difference
p-value

genes with different regulation

Katsiaryna Davydzenka, National Phd student, AI & Health and Life Sciences, UCBM, SISSA, Italy

# Preliminary results, future research directions

**Differential regulation of gene expression**
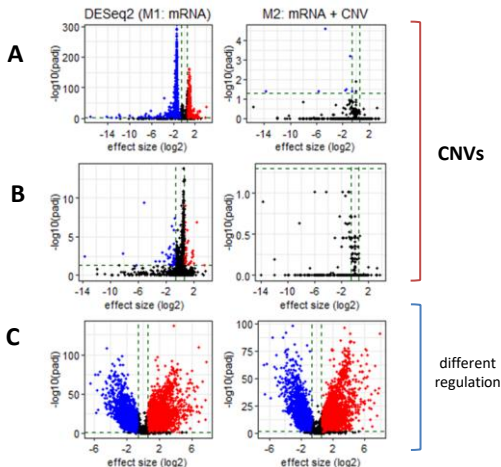


A-B: synthetic data test - homogeneous and heterogeneous CNVs across samples
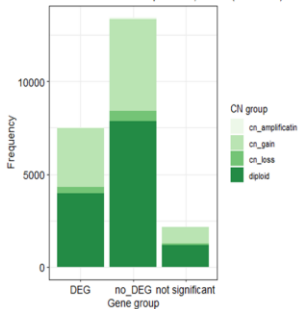C: real data test

**Challenges:**
- mRNA levels do not appear to always scale with the gene CN;
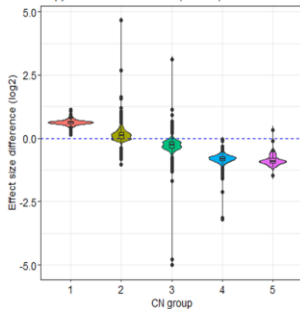- genomic and transcriptomic profile heterogeneity.

**What is next?**

- ❖ further develop the method (Bayesian inference techniques);
- ❖ integrate other omics data (e.g. epigenetics);
- ❖ possible integration into existing pipelines.

Katsiaryna Davydzenka, National Phd student in AI & Health and Life Sciences, UCBM, SISSA, Italy
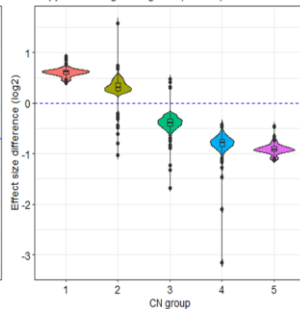
p-value
Effect size difference (log2)