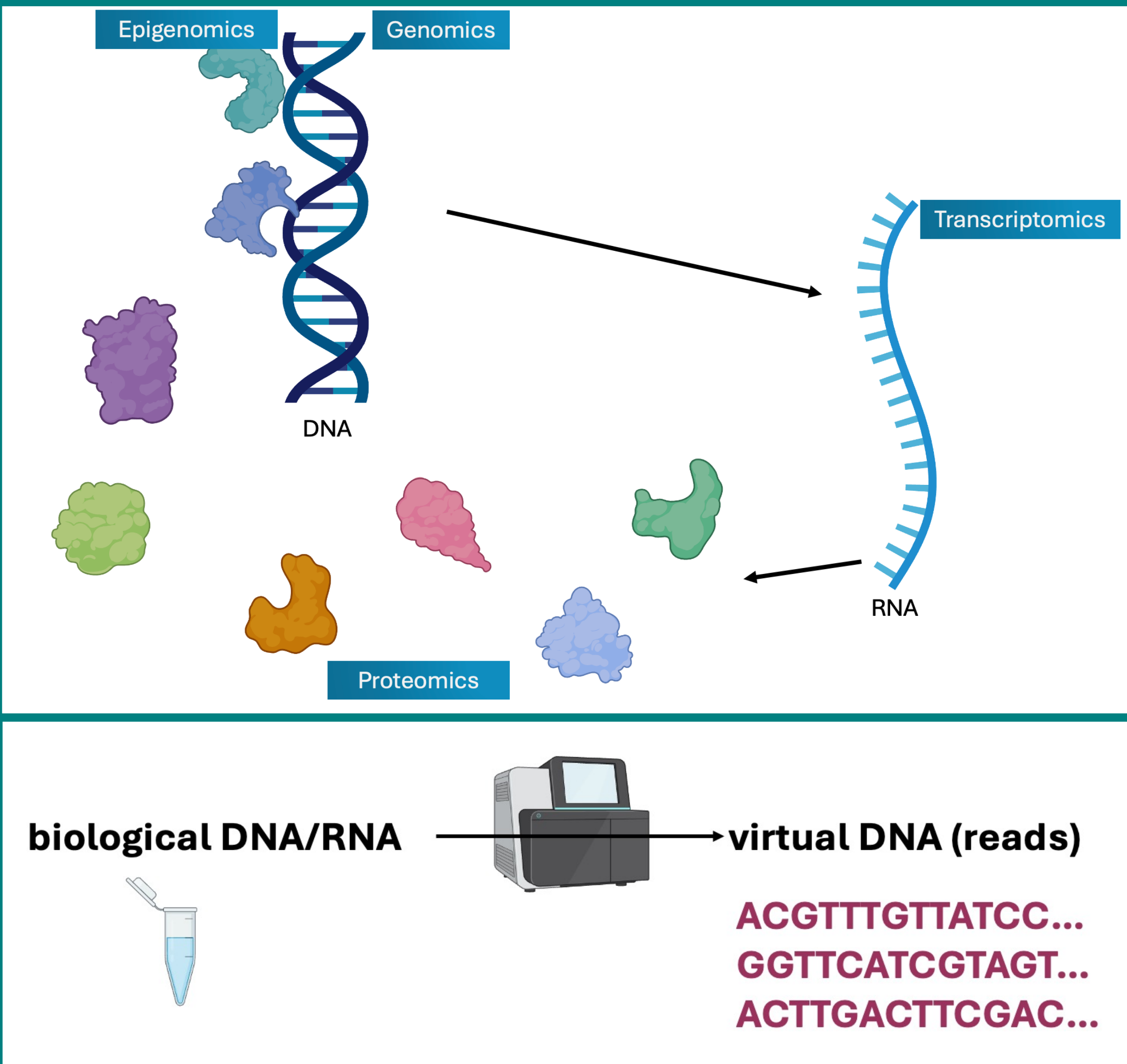


Bridging Bench and Code: RSEs Driving Omics

Ruxandra Neatu

Introduction



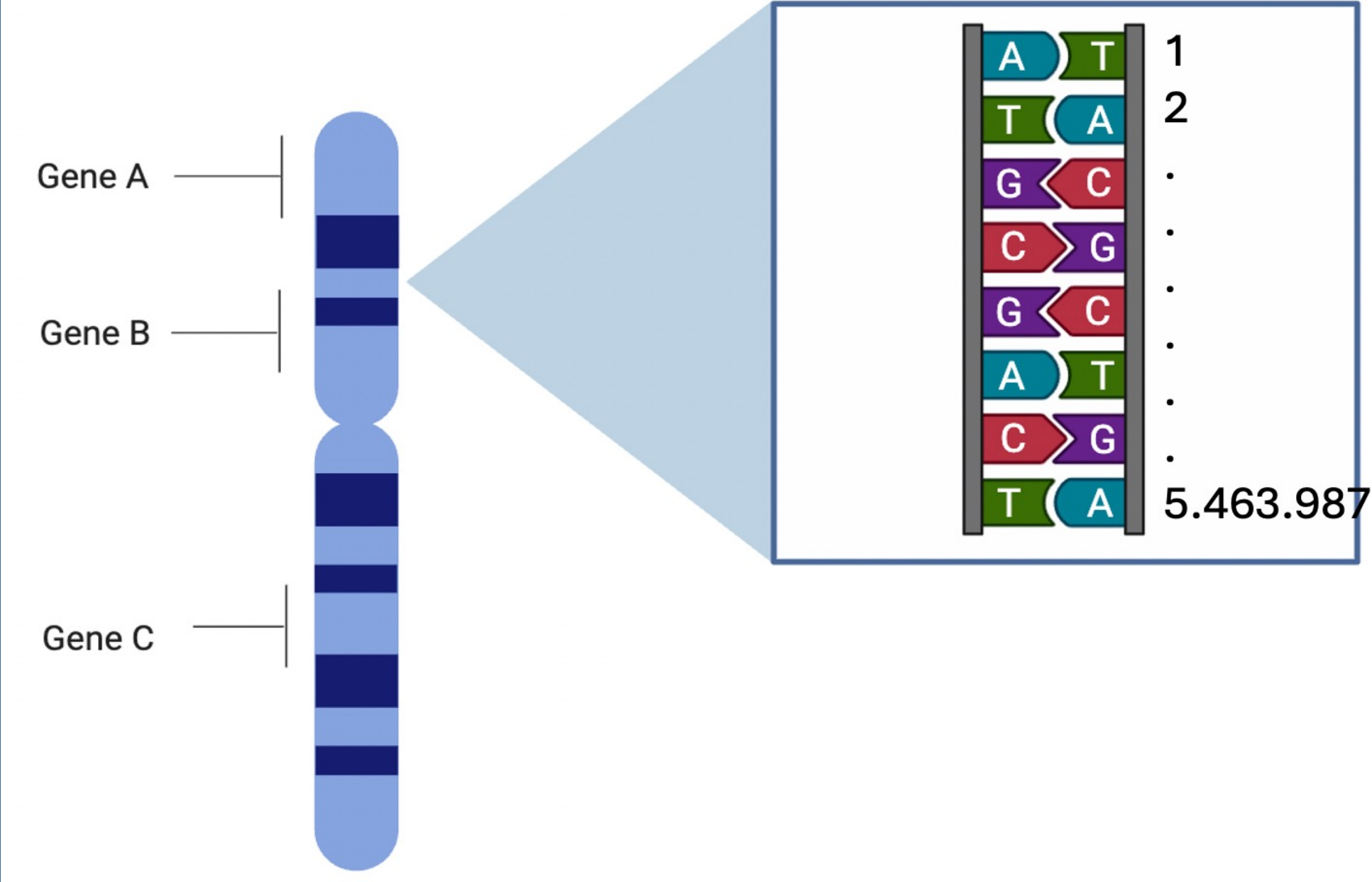
"Omics" are large-scale studies of biological molecules:

- Genomics - DNA, our genetic blueprint.
- Transcriptomics - RNA, copied from active DNA.
- Proteomics - proteins, built from RNA.
- Epigenomics - chemical modifications affecting gene activity.

Sequencing machines read short DNA fragments ("reads"), which bioinformaticians reassemble – like a puzzle – to identify genes and variations in the genome.

Context

Software Development



Once reads are assembled, scientists have a digital replica of the genome. Each chromosome is reconstructed end to end, with every base assigned a position (no.), allowing precise reference for analysis.

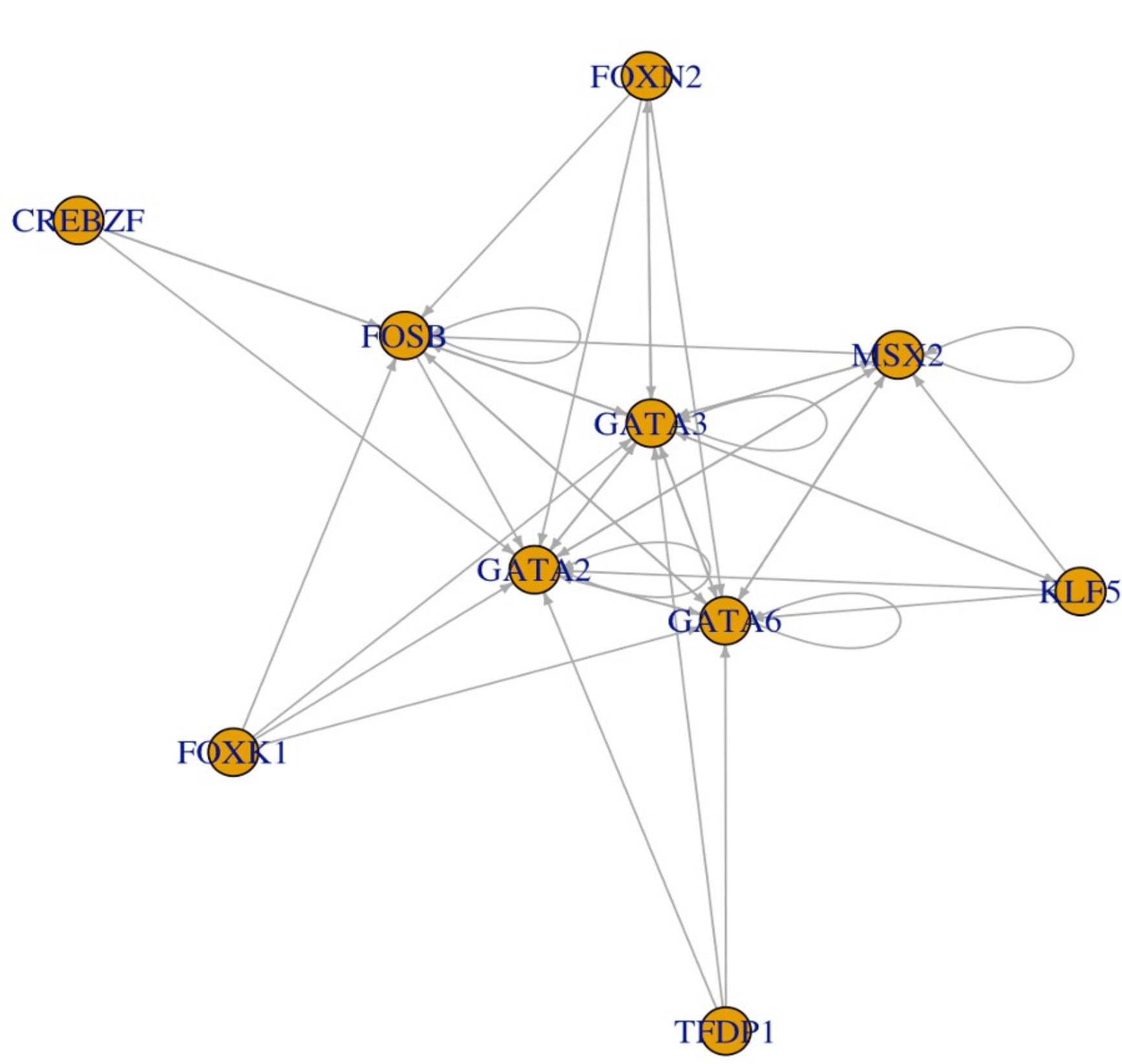
Downstream analysis varies widely, driven by evolving tools and software, as well as specific research questions. As a result, methods and best practices shift rapidly over time.

Data Manipulation

chromosome	Index number	unique id	reference base	mutation	quality score
chr10	44911522	chr10_44911522_C_A	C	A	45
chr10	44911523	chr10_44911523_A_C	A	C	42
chr10	44911525	chr10_44911525_G_A	G	A	35
chr10	44911527	chr10_44911527_G_T	G	T	40
chr10	44911532	chr10_44911532_G_A	G	A	45
chr10	44911533	chr10_44911533_G_T	G	T	39
chr10	44911537	chr10_44911537_G_A	G	A	45
chr10	44911548	chr10_44911548_C_T	C	T	50
chr10	44911550	chr10_44911550_C_T	C	T	43
chr10	44911555	chr10_44911555_C_T	C	T	39
chr10	44911556	chr10_44911556_G_A	G	A	49
chr10	44911560	chr10_44911560_C_T	C	T	45
chr10	44911573	chr10_44911573_A_G	A	G	47
chr10	44911574	chr10_44911574_C_T	C	T	45
chr10	44911579	chr10_44911579_G_A	G	A	45
chr10	44911585	chr10_44911585_A_G	A	G	37
chr10	44911589	chr10_44911589_C_T	C	T	41
chr10	44911593	chr10_44911593_G_A	G	A	47

Genetic variations are often studied using VCF files (see above), which record mutations across the ~3 billion bases of the genome. Files range from megabytes to gigabytes, with multi-sample sets even larger. Compression saves space but slows analysis, posing challenges for timely clinical diagnosis.

Domain-Specific Insight



Cells operate across several layers: Genomics, Transcriptomics, Proteomics, and Epigenomics.

For example, Gene Regulatory Networks (GRNs) (see above) integrate transcriptomics and epigenomics. GRNs can be modelled mathematically, enabling network analysis, statistics, and machine learning to reveal regulatory mechanisms and disease pathway.

RSEs in life sciences

- Evaluate and integrate new tools into workflows efficiently.
- Build adaptable, modular pipelines that evolve with emerging methods.
- Ensure reproducibility and robustness despite rapid change.

- Optimise storage and handling of large genomic datasets (e.g. VCFs).
- Develop faster parsing and analysis tools for timely results.
- Implement scalable workflows for both research and clinical use.

- Integrate multiple omics layers for deeper biological insight.
- Design intuitive tools for data exploration and visualisation.
- Apply mathematical models (e.g. GRNs) to reveal regulatory mechanisms.