Chapter 10

Working with Omics Data: An Interdisciplinary Challenge at the Crossroads of Biology and Computer Science

Thibault Poinsignon, Pierre Poulain, Mélina Gallopin, and Gaëlle Lelandais

Abstract

Nowadays, generating omics data is a common activity for laboratories in biology. Experimental protocols to prepare biological samples are well described, and technical platforms to generate omics data from these samples are available in most research institutes. Furthermore, manufacturers constantly propose technical improvements, simultaneously decreasing the cost of experiments and increasing the amount of omics data obtained in a single experiment. In this context, biologists are facing the challenge of dealing with large omics datasets, also called "big data" or "data deluge." Working with omics data raises issues usually handled by computer scientists, and thus cooperation between biologists and computer scientists has become essential to efficiently study cellular mechanisms in their entirety, as omics data promise. In this chapter, we define omics data, explain how they are produced, and, finally, present some of their applications in fundamental and medical research.

Key words Genomics, Transcriptomics, Proteomics, Metabolomics, Big data, Computer science, Bioinformatics

1 Introduction

There are different types of omics data, each revealing an aspect of cell complexity. To illustrate this complexity, we propose in Fig. 1 an analogy between the functions of a cell and that of a factory. The different omics data types are replaced there, in their specific context. Cells are the building blocks of living organisms. They can be pictured as microscopic, automated factories, made up of thousands of biological molecules (or molecular components) that work together to perform specific functions. Basically, there are four main types of molecular components: DNA, RNA, proteins, and metabolites. The whole population of one type of cellular component is named with the suffix -ome, i.e., gen*ome* (DNA), transcript*ome* (RNA), prote*ome* (proteins), and metabol*ome* (metabolites) (*see* Fig. 1). The scientific fields, which aim at studying those respective populations, are named with the suffix -omics,

Olivier Colliot (ed.), *Machine Learning for Brain Disorders*, Neuromethods, vol. 197, https://doi.org/10.1007/978-1-0716-3195-9_10,

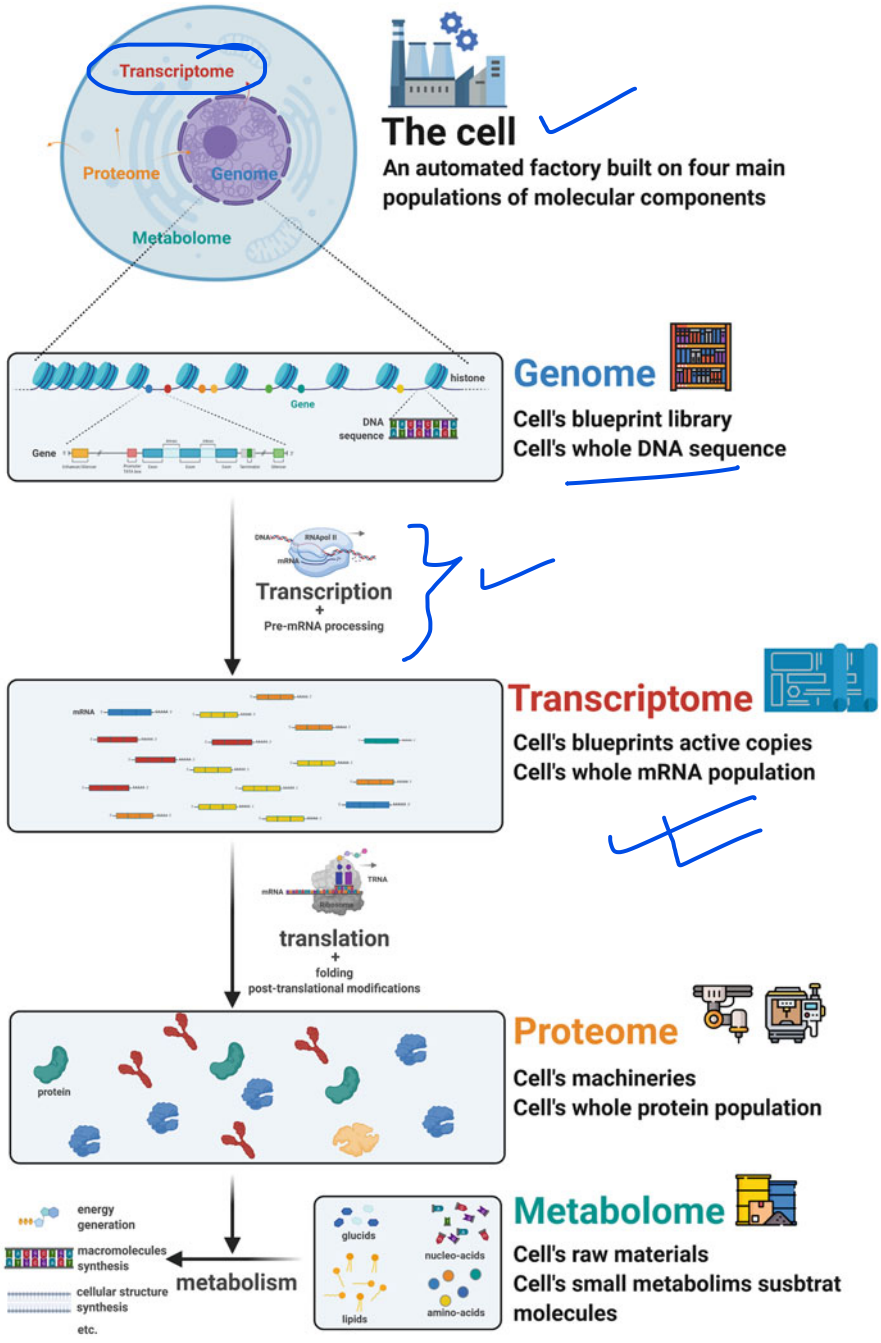

Fig. 1 The four main -omes and an analogy of their functions. The genome designates all cell's DNA molecules. The transcriptome, the proteome, and the metabolome refer, respectively, to the cell's whole set of RNA, proteins, or metabolites at a given time

i.e., gen*omics*, transcript*omics*, prote*omics*, and metabol*omics*. The common point between the different types of omics data is that they all arise from high-throughput experimental strategies that allow the simultaneous observation of all individual components that constitute either the genome, the transcriptome, the proteome, or the metabolome [1].

The genome is made of DNA molecules, which are the carrier of genetic information. It can be imagined as the blueprint library of the cell (*see* Fig. 1). From a chemical point of view, DNA molecules are polymers (or sequences) of simpler chemical units called nucleotides. There are four main types of nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). DNA molecules are organized into chromosomes, which are compacted in the cell nucleus. The genome is directly connected to the transcriptome and the proteome (*see* next sections). The information to synthesize RNA molecules (transcriptome) and proteins (proteome) is encoded in specific regions of the DNA sequence called genes (*see* Fig. 1). Genes are made of successive nucleotides (clustered into codons), which correspond to amino acids, i.e., the molecules that constitute the proteins. The correspondence between nucleotides, codons, and amino acids is known as the genetic code. To summarize, a genomics dataset thus contains the sequences of DNA molecules present in a cell (or a population of cells) and can be seen as a copy of the cell's blueprint library (its genome) written as a long sequence of A, T, C, and G.

The transcriptome is made of RNA molecules. Multiple types exist, and they can be roughly classified into messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNA (ncRNA). Transcriptomics datasets mainly focus on mRNAs, which are the intermediate messengers between the genome and the proteome (*see* previous paragraph). The transcriptome is thus intimately connected to the genome and the proteome (*see* Fig. 1). Notably, the RNA polymerase is required to generate mRNA, reading the genome during transcription. In eukaryotes, mRNAs exit the nucleus to be used as templates by ribosomes (a macromolecular complex made of rRNA and proteins), to synthesize proteins by assembling amino acids (following the genetic code) during translation. Compared to the genome, the transcriptome is much more dynamic. The cell population of mRNA molecule varies according to cell requirement in proteins, and a transcriptomics dataset lists all sequences of mRNA present at a given time. They can be seen as snapshots of which parts of the genome are currently transcribed and in which proportion. Following up on the genome analogy presented in Fig. 1, mRNAs can be seen as active copies of the cell's blueprints that are more or less actively used.

The proteome is made of proteins, i.e., macromolecules made with one or several polymers of amino acids. Proteins are extraordinarily diverse in their three-dimensional (3D) conformations and associated functions. To illustrate this diversity, some proteins constitute the backbone of the cell structure, others detect or transmit external or internal chemical signals, and a large portion of them (enzymes) catalyze chemical reactions of the metabolism (the whole set of chemical reactions sustaining the cell). Proteins are also responsible for the regulation and expression (transcription and translation) of the genetic information (*see* previous paragraph). Protein functions are closely linked to their 3D spatial conformation, and all processes of the cells are based on protein activities (*see* Fig. 1). The proteome is as dynamic as the transcriptome because the set of proteins present at a given time in a cell varies accordingly to the current state and function of this cell. Proteomics datasets give a snapshot of which proteins are present at a given moment in the life of the cell. Genomics, transcriptomics, and proteomics resume the classical central dogma of biology, as first stated by Francis Crick in 1957. Even if it has been further detailed since, with, for instance, a better understanding of epigenomics, it still effectively summarizes the principal flow of information between the main molecular components of the cell: DNA is transcribed into RNA which is translated into proteins.

To end this description of omics data types, we believe it is important to mention the metabolome (*see* Fig. 1). The metabolome is made of metabolites, small molecules that are protein substrates in chemical reactions. Nucleotides and amino acids, cited before, are metabolites, as well as other molecules like lipids (forming bilayer membranes that compartmentalize the cell) or ATP (a molecule used as intracellular energy transfer). To extend, again, the analogy, metabolites can be seen as the raw materials used by the automated microscopic factory (*see* Fig. 1). Metabolomics datasets peek into the population of metabolites in a cell at a given time. Again, it is important to specify that if each cited "omics" field gives an assessment of its associated "ome" population, it is quite a "blurred" one. Everything is intertwined in a cell. Moreover, most omics studies give only an average observation on a population of cells. Multi-omics and single-cell techniques are trying to overcome these limitations.

In this chapter, we detail the different types of files used for omics data and present examples of databases where they are stored. We introduce different methods for generating omics data and finally provide some applications of omics data in fundamental research, cancer research, and pandemic response.

2 What Are Omics Data?

2.1 Results from High-Throughput Studies Written in Multiple Binary and Text Files

To describe the files used to store omics information, it is necessary to consider genomics and transcriptomics on one side and proteomics and metabolomics on the other side. Indeed, these files are generated by different experimental techniques, which are, respectively, sequencing (for genomics and transcriptomics) and mass spectrometry (for proteomics and metabolomics) (*see* Fig. 2). For each group, two types of files must be distinguished: the ones that are directly obtained after the applications of experimental protocols, i.e., the raw omics data files, and the ones that are generated by downstream informatic analyses, i.e., the processed omics data files (*see* Fig. 2). Experimental protocols and the informatic treatments applied to raw data files will be detailed in the next section.

Genomics and transcriptomics raw data files are essentially nucleotide sequence files. In that respect, the FASTA and the FASTQ text formats are commonly used. FASTA was created by Lipman and Pearson in 1985 as an input for their software [2] and became a de facto standard, without any clear statement acknowledging it [3]. This probably explains the absence of a common file extension (e.g., .fasta, .fna, .faa) even if FASTA is a unified file type. FASTA files contain one or several sequences. A sequence begins with a description line starting with the character ">". NCBI databases (*see* next sections) have unified rules to write this line.[1]

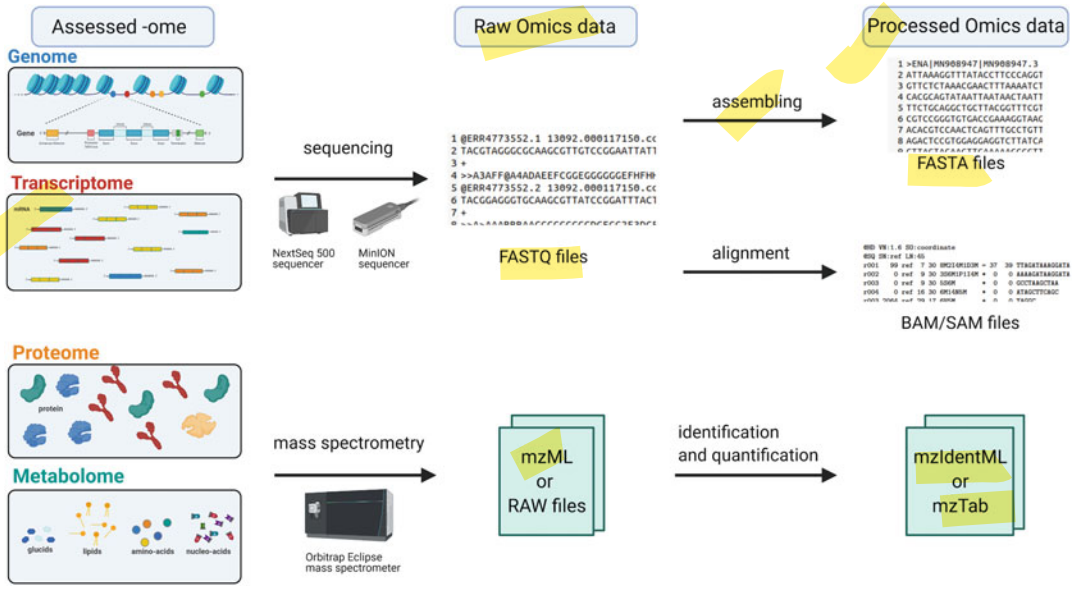

Fig. 2 Omics data are assessments of -ome populations. Raw omics data are generated through sequencing (for DNA and cDNA) or mass spectrometry (for proteins and metabolites)

[1] https://www.ncbi.nlm.nih.gov/genbank/fastaformat/

Subsequent lines contain the sequence itself split into multiple blocks of 60 to 80 characters (one per line). With nucleic acid sequences, the sequence lines are a series of A/T/C/G/U characters, representing the nucleic acids: adenine, thymine, cytosine, guanine, and uracil (the latter replacing thymine in RNA). FASTQ is the file format for the raw data generated by the sequencer in genomics and transcriptomics (*see* Fig. 2). The first two lines are similar as with FASTA file: identification line starts with "@" instead of ">" and the second line contains the nucleic sequence, but a quality score is associated with each position of the sequence (i.e., each letter in the sequence line). This score is called "Phred score," and it codes the probability of error in the identification of this nucleotide [3]. It goes from 0 to 62 and is coded in ASCII symbols. This allows to code any score using a single symbol, keeping the same length as the sequence line. FASTA and FASTQ files can be opened with any text editor software. FASTQ files are mainly lists of short sequences called "reads" (between 50 and 200 nucleic acids), which need to be processed (aligned or assembled) to be further analyzed. Alignment data files are one type of processed data. Indeed, reads in FASTQ files can be aligned to a reference genome sequence to allow further analyses (*see* below for pipeline description and example of applications). The text file format used in this case is the SAM[2] (sequence alignment and mapping) format [4, 5]. It can be further compacted into its binary equivalent, which are BAM or CRAM formats [6].

The file formats for proteomics and metabolomics data are not as homogeneous as for genomics and transcriptomics. At least 17 types of formats exist for mass spectrometry files (*see* below) [7]. Each machine manufacturer created its own, adapted to proprietary software to read and analyze it, thus multiplying formats. In an effort to facilitate data exchange and to avoid data loss (in case of no more readable old file formats), HUPO [8] and PSI[3] created the open-source mzML[4] format (XML text file with specific tag syntax) in 2011 [9]. In the main databases that host mass spectrometry result files, most of the files are in the RAW format, developed by Thermo Fisher Scientific. These binary files contain retention time, intensity, and mass-to-charge ratios (*see* later sections). Software like Peaks, Mascot, MaxQuant, or Progenesis [10, 11] use these files to identify proteins present in the sample and to quantify them. Results from these analyses are shared through two other text file formats: mzIdentML[5] and mzTab.[6]

[2] Sequence Alignment/Map Format Specification

[3] HUPO Proteomics Standards Initiative

[4] mzML 1.1.0 Specification | HUPO Proteomics Standards Initiative|mzML 1.1.0 Specification | HUPO Proteomics Standards Initiative

[5] mzIdentML | HUPO Proteomics Standards Initiative|mzIdentML | HUPO Proteomics Standards Initiative

[6] mzTab Specifications | HUPO Proteomics Standards Initiative|mzTab Specifications | HUPO Proteomics Standards Initiative

Note that many other file formats exist. One of the most critical for omics data analyses concerns the annotations of features on a DNA, RNA, or protein sequence. They are shared through the General Feature Format (GFF[7]) that is a text file with nine tabulated separated fields: sequence, source of the annotation, feature, start of the feature on the sequence, end of the feature, score, strand, phase, and attributes.

2.2 Results from High-Throughput Studies Shared Through Multiple Public Databases

The set of public biological databases hosting omics data is large and constantly evolving. Omics terminology started being regularly used in the 2000s. Between 1991 and 2016 (25 years), more than 1500 "molecular biology" databases were presented in publications, with a proliferation rate of more than 100 new databases each year [12]. These numbers are only the visible part of existing databases. How many have been created without being published? Around 500 of those databases are roughly co-occurrent with the apparition of the World Wide Web, the very Internet application allowing the creation of online databases. The availability of molecular biology databases decreased by only 3.8% per year from 2001 to 2016 [12]. This shows a sustained motivation from the community to create and maintain public platforms to share data. But it also highlights that this motivation comes more from a shared need for easy access to data rather than a supervised effort to coordinate approaches and unify sources. Such efforts indeed exist, for example, the ELIXIR project started in 2013 as an effort to unify all European centers and core bioinformatics resources into a single, coordinated infrastructure [13]. This notably produces the ELIXIR Core Data Resources (created in 2017), a set of selected European databases, meeting defined requirements, and the website "bio.tools," i.e., a comprehensive registry of available software programs and bioinformatics tools. The US National Center for Biotechnology Information (NCBI[8]) databases are also main references.

Given the "raw" nature of omics dataset, they are stored in archive data repositories: raw data from scientific articles, shared on databases easily accessible for reproducibility. Except for the Sequence Read Archive (SRA), the databases cited here are mixed ones: they host raw archive data and knowledge extracted from them. For genomics dataset, NCBI database Genome [14] and EMBL-EBI (member of ELIXIR) database Ensembl [15] are references. They organize genome sequences together with annotations and include sequence comparison and visual exploration tools. Transcriptomics data can be deposited into several databases, like Gene Expression Omnibus (GEO) [16] initially dedicated to microarray datasets, which is structured into samples forming

[7] GFF/GTF File Format
[8] NCBI

datasets. Tools are available to query and download gene expression profiles. The Sequence Read Archive (SRA) [17] accepts raw sequencing data. PRIDE [18] is a reference database for mass spectrometry-based proteomics data. Raw files containing spectra are available with associated identification and quantification information. For metabolomics data, MetaboLights [19] is an archive data repository and a knowledge database. It lists metabolite structures, functions, and locations alongside reference raw spectra. Those databases are generalist references, and many more specialized databases exist: 89 new databases are reported in the 2021 NAR database issue, and a dozen of them are omics specific [20]. For example, AtMAD is a repository for large-scale measurements of associations between omics in *Arabidopsis thaliana*, and Aging Atlas gathers aging-related multi-omics data [21, 22]. Finally, noteworthy is the existence of general-purpose open repositories like Zenodo,[9] which allow researchers to deposit articles, research datasets, source codes, and any other research-related digital information. Researchers thus receive credit by making their work more easily findable and reusable and hence support the application of the FAIR (findable, accessible, interoperable, reusable) data principles.[10]

Consistent efforts are made to cross-reference biological components (genes, proteins, metabolites) through the diversity of databases. Each database represents terabytes and petabytes of biological information (43,000 terabytes of sequence data just for SRA[11]), and the scale of the network they form through cross-reference is hard to conceptualize. This is the "big data" in biology and even more are generated every day.

3 How to Generate Omics Data?

Genomics started in 1977 with the application of the gel-based sequencing method developed by Sanger, to sequence for the first time the whole genome of a virus: the phage phiX. Only 13 years later, in 1990, the Human Genome Project began, aiming at sequencing three billion bases of the human genome, using capillary sequencing [23]. More than 10 years and almost three billion dollars later, this titanic task was accomplished [24]. When we think of omics analyses, microarray technology remains emblematic [25]. In the 2000s, the microarray represented the keystone of a discipline then called "post-genomics" [26]. Behind this terminology, the idea was that once the genomes are entirely sequenced,

[9] https://zenodo.org/

[10] https://www.go-fair.org/fair-principles/

[11] NCBI Insights: The wait is over... NIH's Public Sequence Read Archive is now open access on the cloud

new studies could be performed to understand their functioning. Microarrays thus emerged as a promising tool to monitor gene expression. They allow the quantification of the abundances of transcripts, which are associated with several thousands of different genes, simultaneously. Briefly, microarrays are slides, made of glass, on which probes have been attached. These probes are small DNA molecules, which have the particularity of being specific to one (and only one) gene. The experiment then consists of extracting mRNA molecules from a population of cells and transcribing them into complementary DNA (cDNA), labeled with a fluorescent molecule. These cDNAs are then hybridized on the glass slide and end up attached to the probes which are specific to them. They create a local fluorescent signal there. The higher the amount of mRNA, the more fluorescent signal is measured at each probe location position. Microarrays have been used to successfully study many biological processes, some fundamental such as the cell cycle [27] and others directly related to health issues such as human cancer [28]. It thus paved the road to new applications for sequencing technologies (*see* below).

3.1 High-Throughput Sequencing Technologies

From 2007, new methods called next-generation sequencing (NGS) [29] helped to considerably reduce cost, technical difficulties, and duration of the process.

Illumina is the currently predominant NGS method (*see* Fig. 3). After extraction, the DNA molecules are sequenced by synthesis (SBS) on a flow cell. Thanks to sequence adaptors, each DNA molecule is amplified by bridge amplification as a cluster of copies on the flow cell. The reading of the flow cell is based on optical detection: each time a DNApol adds a new nucleotide, a flash of light is detected. NGS advantage, compared to older Sanger techniques, is to allow massive parallel sequencing of large numbers of short sequences (between 50 and 250 nucleotides) called "reads." The limit of this technique is the size of the fragments, but Illumina technology has very high fidelity (very low error rate).

MinION of Oxford Nanopore is another well-established NGS technology [30]. It is based on electronic detection through a nanopore (*see* Fig. 3). When there is an electric potential around a membrane (measurable as a voltage between the two sides), the passage of a macromolecule through a nanopore (a modified biological protein canal) triggers small changes in this electric potential. The changes are distinctive in function of the current nucleotide in the nanopore. So, the succession of electronic potential variation can be associated as the nucleotide sequence. This is the fundamental concept behind MinION technology, and the main advantage is the length of the sequenced molecules. Without the technical necessity of flow cells, the sequence passing through the nanopore can be very long (order of magnitude of a thousand instead of a hundred base pairs) [31]. But given that the physical

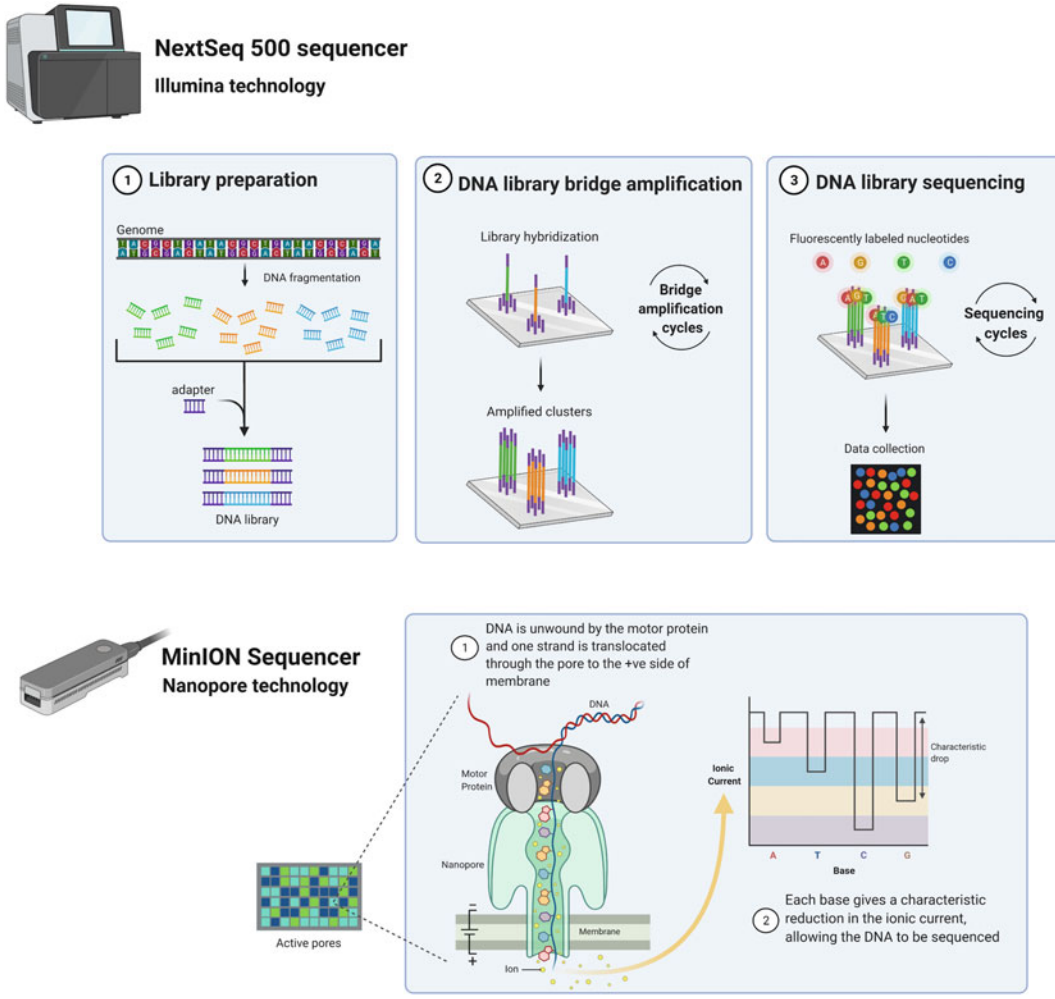

Fig. 3 Illumina and MinION sequencing technologies. Illumina is a sequencing by synthesis technology that allows massive parallel sequencing of small DNA molecules. MinION is a nanopore-based technology that allows the sequencing of longer DNA molecules

signal detected is small variations of an electric potential, the sequencing is less reliable (higher error rate). Depending on the fidelity of the sequencing or the size of the sequence needed, SBS and nanopore-based techniques are complementary.

The sequencing machine output is a group of FASTQ files (*see* previous section). For genomics data, fragments must be assembled to obtain a single sequence of the genome. For transcriptomics data, fragments can be aligned on a reference genome to observe which genes are transcribed at a given time (transcriptome de novo assembly is also possible but still very challenging). Therefore, to extract information from the FASTQ files produced by the sequencer, two main processing steps are needed. The numerous small sequences (reads) stored in the file must be aligned to a

reference genome (mapping), and then the count of reads aligned to a gene sequence gives an estimation of its level of transcription (quantification). Dozens of bioinformatics tools have been developed over the years for mapping (STAR [31], TopHat [32], HISAT2, Salmon [33]) and quantification (featureCounts [34], Cufflinks [35]). Benchmarking studies highlight similar performance for most of them [36–38]. Interestingly, TopHat2 exhibits an alignment recall on simulated malaria data that varies from under 3% using defaults to over 70% using optimized parameters [39]. This underlines the impact of parameter optimization on result quality. Quantification tools generate a text file summarizing the level of transcription of each gene in each condition into a matrix of counts.

3.2 Mass Spectrometry Technologies

Since the first use of a mass spectrometer for protein sequencing in 1966 by Biemann,[12] the improvement of mass spectrometer is closely linked to proteomics and metabolomics development [40]. Metabolites and proteins cannot be read as templates like DNA or RNA, and so they neither can be amplified nor sequenced by synthesis. To access their sequence, the main tool is the mass spectrometer. In the classical bottom-up approach, proteins are digested into small peptides, which pass through a chromatography column. They are then sequentially sprayed as ions into the spectrometer. Migration through the spectrometer allows separation of the peptides according to their mass-to-charge ratio. For each fraction exiting the column, an abundance is calculated. In a data-dependent acquisition (DDA), a few peptides with an intensity superior to a given threshold are isolated one at the time. They are fragmented, and additional spectra (mass-to-charge ratio and intensity) are generated for each fragmented ion. In a data-independent acquisition (DIA), a spectrum is generated for all fractions coming out of the chromatography column. Obtained spectra are a combination of spectra corresponding to each peptide present in each original fraction. Comparison with a peptide spectrum library generated in silico is therefore required to allow the deconvolution of those complex spectra. All this information (abundances in fractions, mass-to-charge ratios, intensities) is stored into .raw files, which can only be read by dedicated software (*see* Subheading 2.1).

3.3 Single-Cell Strategies

Most omics experiments are bulked, and they are an average measure done on a population of cells, which is more or less homogeneous. Single-cell omics allow a more precise measurement, highlighting the plasticity of the cell system. Single-cell techniques started with manual separation of a single cell under a microscope in 2009 [41] and quickly evolved toward techniques allowing the

[12] HUPO—Proteomics Timeline

parallel sequencing of thousands of cells [42]. Plate-based techniques use flow cytometry to separate isolated cells into the different wells of a plate, allowing processing of hundreds of cells. The introduction of nanometric droplets to separate isolated cells allowed the parallel processing of thousands of cells thanks to individual barcoding [43, 44]. Cells isolated from tissues are mixed with microparticles in a buffer that forms droplets in oil. Most droplets are empty, but some contain both a microparticle and a cell. After cell lysis, oligonucleotide primers on the microparticles allow the capture of the cell mRNA (by oligo-dT and polyA tail complementarity). Primers on the same microparticle are barcoded, thus creating a cell tag on each sequence. Amplification and sequencing can be bulked without losing the cell of origin for each transcript. Several bioinformatics tools are specialized for single-cell transcriptomics data [45]. For example, Cell Ranger and Loupe Browser are, respectively, four pipelines (mapping, quantification, and downstream analysis) and a visualization tool developed by 10× Genomics [44]. Single-cell transcriptomics data are challenging for bioinformatics analysis because of their high level of technical noise and the multifactorial variability between cells [45]. Transcriptomics is the more advanced single-cell omics, but single-cell genomics is also used in SNP and copy number variation screening (*see* Subheading 4.2).

Proteomics and metabolomics data are still challenging to obtain at a single cell level: one cell yields only 250–300 pg [46] of proteins when MS in-depth measurement still necessitates population scale yield. But thanks to innovations in sample preparation and experimental design, single-cell proteomics assessments scaled up from a few hundred to more than a thousand identified proteins in just 4 years [47].

4 Which Applications for Omics Data?

4.1 In Fundamental Research

Describing biological systems implies to identify, quantify, and functionally connect their individual molecular components. Given the diversity of cellular components and their multiple interlocking functions, the large scale of omics data empowers the characterization of biological systems. As stated before, each type of "omics" is an assessment of a specific subpopulation of molecular components. Mining omics data thus allows bulk identification of the nature (sequence and structure), location, function, and abundance of molecular components in those subpopulations.

Genomics data are making the genome sequences of thousands of species accessible. The first direct application of these resources is the annotation of genomic features onto those genomic sequences: protein-coding genes, tRNA and rRNA genes, pseudogenes, transposons, single-nucleotide polymorphisms, repeated regions, telomeres, centromeres... Genomic features are numerous, and DNA sequences alone can be enough to recognize

patterns specific to some of them. For example, specific tools exist to detect protein-coding genes, like Augustus[13] [48]. The annotation can be based only on sequence patterns or also on comparison with another sequence. Comparative genomics, i.e., the comparison of genome sequences, allows the transfer of knowledge for homolog genes (evolutionarily related genes) between species. Bioinformatics tools exist to infer evolutionary relationships between genes based on their sequence similarity [49]. Understanding the evolution of the genome helps to understand the dynamics behind phenotypic convergence, population evolutions, speciation events, and natural selection processes. For example, the study of 17 marine mammals' genomes offered insight into the macroevolutionary transition of marine mammal lineages from land to water [50].

Transcriptomics data give insight on the levels of gene transcription. The resulting count matrix (*see* previous section) is mainly used to carry out differential expression analysis (DEA) of genes between conditions. Conditions differ by the variation of a single factor: a mutation, a different medium, or a stimulus. Basic DEA is a multi-step workflow [51] that allows the detection of statistically significant variations in expression across conditions. The final goal is to deduce insight on the gene's functions from the observed variations. Transcriptomics data are also used to increase the quality of genome annotation. The presence of hypothetical genes can be verified by their transcription, the exact structure of known genes can be refined (size of UTRs and exons; *see* Fig. 1), and previously undetected genes can be observed [52].

Proteomics data allows the identification and quantification of proteome. Proteome does not totally correlate with transcriptome. RNA can be spliced (assembly of the mRNA from exons, not always the same and in the same order), and proteins undergo several post-translational modifications (minor changes in the chemical structure of the protein) and re-localization [53]. Cellular pathways and phenotypes thus cannot be fully understood only through transcriptomics assessments. Proteomics completes the information given by genomics and transcriptomics. It describes the third -ome of the central dogma of biology (*see* Fig. 1).

Multi-omics analysis, taking advantage of several omics insights in the same experimental approach, comes with several challenges. Generating several types of omics data comes with a significant investment in time, skilled manpower, and money [1]. Even if generated in the same experimental approach, omics data are heterogeneous by nature, thus complexifying their integration. If challenging, multi-omics datasets are also a step toward the systemic description of biological systems [54].

[13] Augustus/ABOUT.md at master

4.2 In Medical Research

An early application of genomics in medical research is the genome-wide association studies (GWAS). By comparing genome sequences from a large population of individuals (both healthy and sick), GWAS highlight SNPs (single-nucleotide polymorphisms) that are significantly more frequent in individuals with the disease. Correlation does not mean causality, but GWAS can give a first clue of the metabolic pathways or cellular components involved in the disease [55]. This strategy has proven to be efficient in the case of "common complex diseases." Unlike Mendelian diseases (which are rarer), the heritability (genetic origin) of these diseases depends on hundreds of SNPs with small effect sizes, which GWAS studies help identify [56]. Alzheimer's disease and cancers are examples of "common complex diseases" whose genetic underpinnings have been explored through GWAS [55, 57].

Most cancers emerge from the successive alteration of cell functioning (by accumulation of mutations), leading to abnormal growth causing tumors and metastasis. Multi-omics studies can highlight the underlying molecular mechanisms of cancer development, better explain resistance to treatment, and help classify cancer types. Screening cohorts of patients helps assess alleles associated with the development of certain types of cancer. The different subtypes for breast cancer are a well-documented example [58].

Single-cell genomics is the only way of characterizing rare cellular types such as cancer stem cells [59]. Single-cell omics data are also used to follow the rapid evolution of cancer cell population inside tumors. Understanding and describing cancer cell population dynamics is crucial: the characteristic accelerated rate of mutation can be the cause of treatment resistance. Omics data specific to cancer cell lines are shared on specific databases driven and maintained by global consortium such as the Cancer Genome Atlas Program[14] (over 2.5 petabytes of genomics, epigenomics, transcriptomics, and proteomics data) or the International Cancer Genome Consortium [60].

Omics data proved to be a priceless resource in pandemic response. The virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing the COVID-19 disease quickly spread around the world, causing more than six million deaths (as of March 2022) and a global health crisis. Its RNA sequence was obtained in January 2020 and allowed the development of detection kits and later RNA-based vaccines. Since the beginning of the pandemic, the genomic evolution of the virus is followed almost in real time, as new variants (with mutations affecting mostly the spike protein of the virus envelope) are sequenced. Variant profiling allows the World Health Organization to closely monitor variants of concern. The precise characterization of the virus structure

[14] https://www.cancer.gov/tcga

opens the research of therapeutic targets. Multi-omics studies helped specify the COVID-19 biomarkers, pathophysiology, and risk factors [61].

Getting omics data in brain tissue studies is promising but challenging because of brain specificity. Indeed, except in a few specific diseases where in vivo resections are performed (brain tumors, surgically treated epilepsy, etc.), human brain samples are collected postmortem, when the less stable molecule populations are already significantly altered. For example, studies of the brain transcriptome are deeply impacted. On the other hand, some omics studies target peripheral fluids (e.g., plasma, cerebrospinal fluid, etc.) with the aim to find biomarkers, but the relationships between observations in peripheral fluids and pathophysiological mechanisms in the brain are far from clear. Moreover, the brain is organized as a network of intricate substructures, constituted of several cell types (glial cells and different neuron types) with distinct function and thus different omics landscape [62]. Nonetheless, multi-omics exploratory studies are describing complex diseases in a systematic paradigm, highlighting diversity of cellular dysregulations linked to complex pathologies like Alzheimer's disease [57].

5 Conclusion

Genomics, transcriptomics, proteomics, and metabolomics are arguably the most developed and used omics, but they are not the only ones. Other omics describe other sides of the functioning of the cell, which require intricate relationships between omics levels. For example, epigenomics describes the transitory chemical modifications of DNA, and lipidomics looks at the lipidic subpopulation of metabolites (*see* Fig. 1). Omics diversity mirrors the complexity of cell systems. With the constant improvement of measurement techniques, possibilities to assess ever larger subsystems of the cells are increasing. Omics dataset generation is paired with the development of software, essential tools to generate, read, and analyze them. By design, computer science is therefore omnipresent in modern "big data" biology. The need for more gold standard analysis pipelines and file formats grows with the scale and complexity of produced datasets.

Acknowledgments

This work was funded by the Agence Nationale pour la Recherche (MINOMICS, grant number ANR-19-CE45-0017).
The authors are grateful to Sarah Cohen-Boulakia for reviewing this chapter.
Figures were made on BioRender, using icons from Flaticon.

References

1. Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18:83
2. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441
3. Cock PJA, Fields CJ, Goto N et al (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771
4. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
5. Danecek P, Bonfield JK, Liddle J et al (2021) Twelve years of SAMtools and BCFtools. GigaScience 10:giab008
6. Hsi-Yang Fritz M, Leinonen R, Cochrane G et al (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res 21:734–740
7. Deutsch EW (2012) File formats commonly used in mass spectrometry proteomics. Mol Cell Proteomics 11:1612–1621
8. Deutsch EW, Lane L, Overall CM et al (2019) Human proteome project mass spectrometry data interpretation guidelines 3.0. J Proteome Res 18:4108–4116
9. Martens L, Chambers M, Sturm M et al (2011) mzML—a community standard for mass spectrometry data. Mol Cell Proteomics 10(R110): 000133
10. Ma B, Zhang K, Hendrie C et al (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17:2337–2342
11. Välikangas T, Suomi T, Elo LL (2017) A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. Brief Bioinform
12. Imker HJ (2018) 25 years of molecular biology databases: a study of proliferation, impact, and maintenance. Front Res Metr Anal 3:18
13. Harrow J, Drysdale R, Smith A et al (2021) ELIXIR: providing a sustainable infrastructure for life science data at European scale. Bioinformatics 37:2506–2511
14. Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2010) GenBank. Nucleic Acids Res 38: D46–D51
15. Howe KL, Achuthan P, Allen J et al (2021) Ensembl 2021. Nucleic Acids Res 49:D884–D891
16. Barrett T, Wilhite SE, Ledoux P et al (2012) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41: D991–D995
17. Leinonen R, Sugawara H, Shumway M et al (2011) The sequence read archive. Nucleic Acids Res 39:D19–D21
18. Perez-Riverol Y, Csordas A, Bai J et al (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 47: D442–D450
19. Haug K, Cochrane K, Nainala VC et al (2019) MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res gkz1019
20. Rigden DJ, Fernández XM (2021) The 2021 *Nucleic Acids Research* database issue and the online molecular biology database collection. Nucleic Acids Res 49:D1–D9
21. Lan Y, Sun R, Ouyang J et al (2021) AtMAD: *Arabidopsis thaliana* multi-omics association database. Nucleic Acids Res 49:D1445–D1451
22. Aging Atlas Consortium, Liu G-H, Bao Y et al (2021) Aging Atlas: a multi-omics database for aging biology. Nucleic Acids Res 49:D825–D830
23. Karger BL, Guttman A (2009) DNA sequencing by CE. Electrophoresis 30:S196–S202
24. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945
25. Schena M, Shalon D, Davis RW et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470
26. Gershon D (1997) Bioinformatics in a post-genomics age. Nature 389:417–418
27. Spellman PT, Sherlock G, Zhang MQ et al (1998) Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9:25
28. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 14(4):457–460
29. Metzker ML (2010) Sequencing technologies — the next generation. Nat Rev Genet 11:31–46

30. Deamer D, Akeson M, Branton D (2016) Three decades of nanopore sequencing. Nat Biotechnol 34:518–524
31. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21
32. Kim D, Pertea G, Trapnell C et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36
33. Patro R, Duggal G, Love MI et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14: 417–419
34. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930
35. Trapnell C, Roberts A, Goff L et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578
36. Teng M, Love MI, Davis CA et al (2016) A benchmark for RNA-seq quantification pipelines. Genome Biol 17:74
37. The RGASP Consortium, Engström PG, Steijger T et al (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10:1185–1191
38. Schaarschmidt S, Fischer A, Zuther E et al (2020) Evaluation of seven different RNA-Seq alignment tools based on experimental data from the model plant Arabidopsis thaliana. Int J Mol Sci 21:1720
39. Baruzzo G, Hayer KE, Kim EJ et al (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods 14: 135–139
40. Biemann K, Tsunakawa S, Sonnenbichler J et al (1966) Structure of an odd nucleoside from serine-specific transfer ribonucleic acid. Angew Chem Int Ed Engl 5:590–591
41. Tang F, Barbacioru C, Wang Y et al (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382
42. Svensson V, Vento-Tormo R, Teichmann SA (2018) Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 13: 599–604
43. Macosko EZ, Basu A, Satija R et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214
44. Zheng GXY, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049
45. Stein CM, Weiskirchen R, Damm F et al (2021) Single-cell omics: overview, analysis, and application in biomedical science. J Cell Biochem 122:1571–1578
46. Jehan Z (2019) Single-cell omics: an overview. In: Single-cell omics. Elsevier, pp 3–19
47. Kelly RT (2020) Single-cell proteomics: progress and prospects. Mol Cell Proteomics 19: 1739–1748
48. Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465–W467
49. Quest for Orthologs consortium, Altenhoff AM, Boeckmann B et al (2016) Standardized benchmarking in the quest for orthologs. Nat Methods 13:425–430
50. Yuan Y, Zhang Y, Zhang P et al (2021) Comparative genomics provides insights into the aquatic adaptations of mammals. Proc Natl Acad Sci 118:e2106080118
51. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love M, Patro R, Robinson MD (2019) RNA sequencing data: Hitchhiker's guide to expression analysis. Annu Rev Biomed Data Sci 2:139–173
52. Chen G, Shi T, Shi L (2017) Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci 60:116–125
53. Arivaradarajan P, Misra G (eds) (2018) Omics approaches, technologies and applications: integrative approaches for understanding OMICS data. Springer Singapore, Singapore
54. Veenstra TD (2021) Omics in systems biology: current progress and future outlook. Proteomics 21:2000235
55. Tam V, Patel N, Turcotte M et al (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20:467–484
56. Uitterlinden A (2016) An introduction to genome-wide association studies: GWAS for dummies. Semin Reprod Med 34:196–204
57. Hampel H, Nisticò R, Seyfried NT et al (2021) Omics sciences for systems biology in Alzheimer's disease: state-of-the-art of the evidence. Ageing Res Rev 69:101346
58. Kohler BA, Sherman RL, Howlader N et al (2015) Annual report to the nation on the status of cancer, 1975–2011, featuring

incidence of breast cancer subtypes by race/ethnicity, poverty, and state. JNCI J Natl Cancer Inst 107

59. Liu J, Adhav R, Xu X (2017) Current progresses of single cell DNA sequencing in breast cancer research. Int J Biol Sci 13:949–960
60. Zhang J, Bajari R, Andric D et al (2019) The international cancer genome consortium data portal. Nat Biotechnol 37:367–369
61. Overmyer KA, Shishkova E, Miller IJ et al (2021) Large-scale multi-omic analysis of COVID-19 severity. Cell Syst 12:23–40.e7
62. Naumova OY, Lee M, Rychkov SY et al (2013) Gene expression in the human brain: the current state of the study of specificity and spatio-temporal dynamics. Child Dev 84:76–88