Introduction to Cancer genomics

Author:
Aurélie GABRIEL

This document comes from the first chapter of my PhD thesis, supervised by Dr. Matthieu Foll and Dr. James McKay and entitled "Integrating somatic and germline multi-omics data to improve our understanding of lung cancer: a computational biology perspective"

in the

Genetic Cancer Susceptibility group, IARC-WHO

November 18, 2020

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The biology of cancer

Cancer was the second cause of death worldwide, with almost 10 million deaths, in 2018 [1] and could in a near future become the leading cause [2]. The disease can affect different parts of the body, although some tissues are more frequently altered than others. Lung cancer is one of the most common cancers and the deadliest according to the 2018 GLOBOCAN database (a project of the International Agency for Research on Cancer (IARC) providing worldwide cancer statistics) [1]. Cancer is a complex disease that is highly controlled by the genome [3, 4]. It originates from normal cells whose genetic information has been altered. Those alterations can result from endogenous processes as well as from exogenous processes like environmental exposures and lifestyle [5, 6]. As a result of these alterations, tumor cells have acquired specific capabilities that allow them to grow in an uncontrolled way as opposed to normal cells. These capabilities are referred to as the hallmarks of cancer and are listed in Figure 1.1 [7].

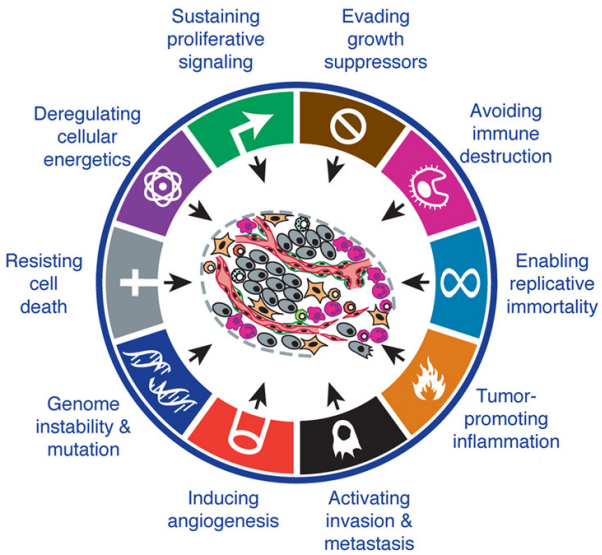

FIGURE 1.1: **The hallmarks of cancer**. From Hanahan *et al.* [7]

The first part of the introduction describes how genomic changes can influence cancer development and how the technological advances in the genomics area have enabled to shed lights on the mechanisms involved.

1.1.1 The central dogma of molecular biology

At the beginning of the 19th century, Avery and colleagues isolated and identified the Deoxyribonucleic acid (DNA) as the molecule constituting our chromosomes, defined previously as carriers of our hereditary material by Thomas Morgan [8, 9]. In 1953, Watson and Crick proposed a new structure for the DNA molecule, the double helix structure [10] (See Figure 1.2A). Five years later, Francis Crick formulates how the information contained in the sequence of nucleic acids is processed to produce the proteins needed by our cells in what is called the central dogma of molecular biology (Figure 1.2B-C).

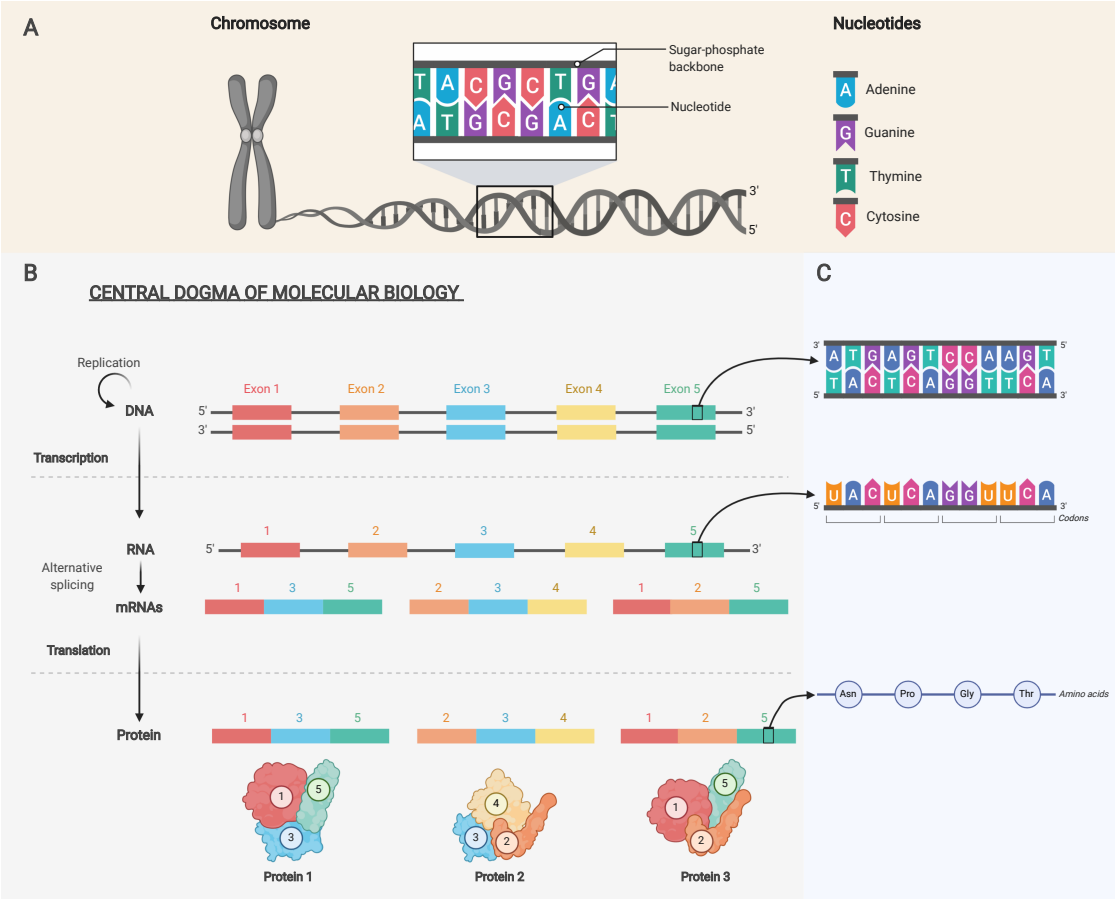

FIGURE 1.2: **The DNA molecule and the central dogma of molecular biology**. A) The structure of DNA: the double helix molecule is composed of two complementary strands of nucleotides. B) Representation of the steps described by the central dogma of molecular biology. C) Illustration of the molecules resulting from the central dogma transfers at a higher resolution. Created with BioRender.com

Three main transfers are described by the central dogma: replication, transcription and translation (See Figure 1.2B). During replication, the DNA molecule is duplicated to provide the needed information to progeny cells. Through the two other steps, the information contained in DNA is used to generate proteins. Firstly, the process of transcription consists in reading the DNA sequence to synthesize a single-stranded molecule of the same length, the Ribonucleic acid (RNA). During translation, the transcribed molecule is then read using a reading frame of three nucleotides that form what is called a codon encoding for one amino acid, the unit of a protein (See Figure 1.2C). Note that the genetic code is redundant; multiple codons can encode an amino acid. The conversion of the information encoded in our genes to functional gene products like proteins is referred to as gene expression.

Since the statement of the central dogma, other mechanisms have been identified as determinant for the expression of a protein. Firstly, the RNA molecule resulting

from the transcription process, containing regions coding for the final amino acids sequence (exons) and non-coding regions (introns), is actually a precursor messenger RNA (pre-mRNAs). The step transforming precursor RNA to mature messenger RNAs (mRNAs) is called alternative splicing and consists in truncating intronic regions and joining different exons together (See Figure 1.2B). Hence, one pre-mRNA can lead to multiple mRNAs that are then transported outside of the nucleus to be translated into different proteins. While around 20,000 genes are described, much more proteins are generated as a result of alternative splicing.

Although all of our cells share the same genetic information and follow the same dogma, it is known that cells in distinct tissues differentiate and do not express the same proteins, at the same time. Such differences can be explained by the fact that several regulatory processes control gene expression levels. Firstly, genes transcription is dependent on transcription factors that represent around 7% of the genes [11]. They specifically bind to control regions of genes, provide or prevent access to the DNA and can control multiple genes [11]. The fact that genes, for example the transcription factors, can influence multiple genes and thus multiple possibly unrelated phenotypes is referred to as pleiotropy. After transcription, mRNAs can also be regulated through other RNA molecules, like the micro RNAs (miRNAs), that can degrade mRNAs. Besides, differences in gene expression can be controlled via non-genetic mechanisms like epigenetic processes, including histone modifications and DNA methylation. Histones are proteins around which the DNA is wrapped and hence control DNA accessibility (Figure 1.3). For example, histone phosphorylation leads to the condensation of the chromatin and inhibits gene expression [11]. DNA methylation consists in the addition of a methyl group to cytosine nucleotides located in cytosine–phosphate–guanine (CpG) dinucleotides sites (cytosine followed by a guanine nucleotide). Such positions are not homogeneously distributed across the genome and are more frequently observed in what is called CpG islands, themselves mainly observed in regulatory regions of genes, the promoters. It has been observed that the methylation of CpG sites in promoters can repress gene expression while methylation of positions in the gene body positively correlates with gene expression [12].

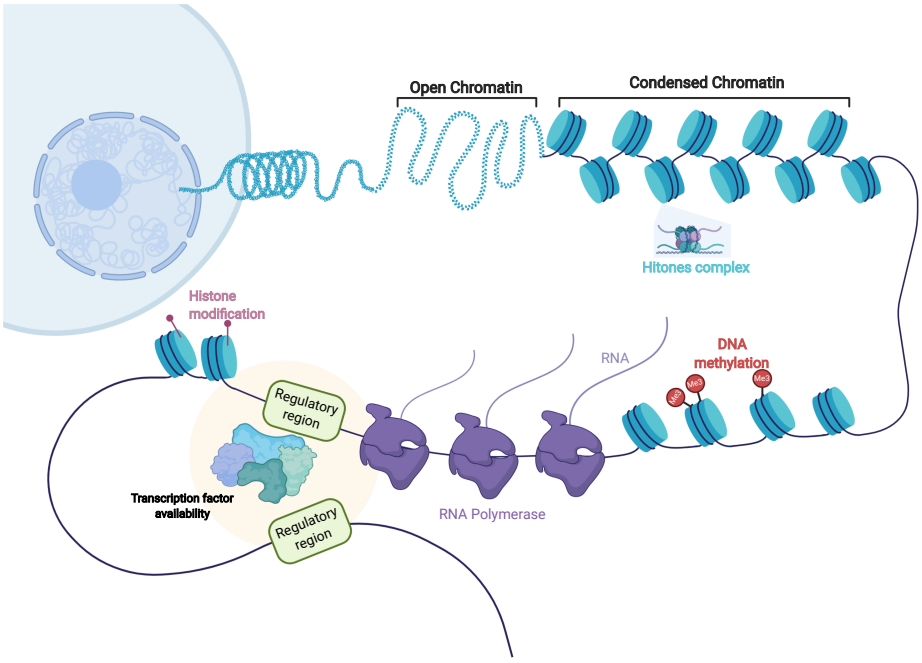

FIGURE 1.3: **Regulation of transcription**. The figure represents different configurations of DNA packaging. The DNA molecule is wrapped around histones proteins that themselves are gathered in complexes called nucleosomes. This packaging forms the chromatin structure. This structure can be more or less compact (open versus condensed chromatin), which is influencing gene expression. When the chromatin is open, transcription factors can access the DNA molecule, and RNA polymerases can initiate the transcription. Note that the structure of the chromatin can be influenced by histones modifications and DNA methylation events. Created with BioRender.com

Finally, post-translational events like enzymatic modifications of proteins or protein cleavage can occur and increase the number of proteins that can be generated in human cells, hence adding an additional layer of complexity.

As such, the numerous steps of transferring the DNA sequence information to proteins reflect the complexity behind protein expression. Any of these steps can be disrupted and result in altered molecules and proteins, leading to cancer development.

1.1.2 Cancer: a genomic disease

Our DNA continuously undergoes diverse alterations and their accumulation over time can cause cancer. Researchers started to investigate the role of genomes in cancer at the end of the 19th century. In 1890, David von Hansemann, by observing cancer cell division under a microscope, identified for the first time abnormal chromosomes. This observation, among others, led Theodor Boveri 20 years later

to suggest that cancer was a consequence of alterations in our inherited DNA [3]. His hypothesis was supported in the mid 20th century by the identification of a recurrent alteration resulting in a peculiar chromosome 22 (the Philadelphia chromosome), in chronic myelogenous leukemia (CML). While those alterations have been observed at the chromosomal level, genomes can be impacted by a multitude of alterations detectable at a lower resolution, the modification of one nucleotide in the DNA sequence being the lowest resolution.

At any position of the genome, the nucleotides might vary from an individual to another as well as between cells of an individual; those variations are called Single Nucleotide Variations (SNVs). Also, larger events like nucleotides insertions or deletions (indels) of up to 1,000 bases and structural variations (chromosomal rearrangements or large indels) can alter the DNA sequence. All of these genomic changes are called mutations.

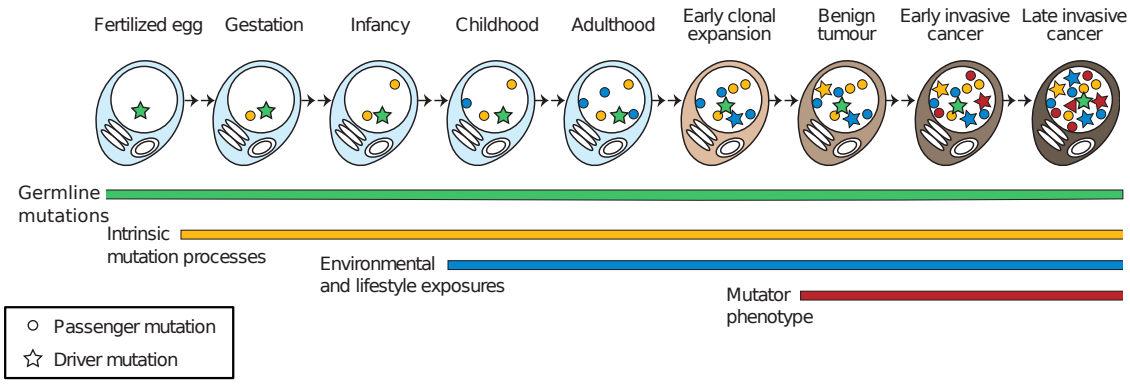

FIGURE 1.4: **The timing of somatic mutations acquisition**. Mutations can be inherited at birth (germline mutations, in green) or acquired during life course (somatic mutations, in yellow, blue and red). They can have little to no impact (passenger mutations represented by circles) or confer an advantage to the cell (driver mutations represented by stars). Adapted from Stratton *et al.* [3]

Mutations can occur at different moments in life (See Figure 1.4). Some mutations are inherited at birth since they are present in the germ line cells (sperm and egg) transmitted by parents to the offspring. They are called germline mutations and are found in all the cells of an individual, normal cells as well as tumor cells. Such mutations are observed at different frequencies in different populations and are called Single Nucleotide Polymorphism (SNP)s. Another category of mutations can also be found in all cells of the body even if they were not transmitted by our parents, if they occur early in life during the development, at gestation. They are

called *de novo* mutations. Finally, the rest of the mutations found in humans are acquired later in life as a result of errors in the DNA maintenance or exogenous damages (See next section). Those mutations occur in cells outside the germ line and are called the somatic mutations.

Also, whether they are germline or somatic, mutations can have different impacts. Most mutations have, due to the redundancy of the genetic code, little to no impact on the genes encoded around them, they are the passenger mutations [13]. Others though alter the gene product and confer a selective advantage to the cell, *e.g.* a faster proliferation or a better survival in comparison to neighbour cells [3]. Those mutations are called driver mutations as they are thought to contribute to "driving the carcinogenic process" and are preserved by positive selection. In 2018, the Cancer Gene Census described more than 700 driver genes (genes carrying driver mutations). Among them, 90% were associated with somatic mutations and 20% contained germline mutations [14, 15]. Generally, two types of driver genes exist, oncogenes and Tumor Suppressor Genes (TSG). Oncogenes are genes whose functions are to promote cell growth, proliferation or inhibit apoptosis and usually result from a gain of function. A mutation in an oncogene can thus lead to a deregulation of one of these processes, hence resulting in uncontrolled proliferation and cancer. The first mutation identified as causing cancer was discovered in 1982 by Reddy *et al.* and was activating an oncogene named *HRAS* [16]. Besides mutations, other processes like over-expression of genes via amplification or chromosomal translocations can activate this category of genes. In contrast to oncogenes, TSGs are refraining cellular growth and proliferation and are often referred to as the "gatekeepers" genes. Mutations in TSGs tend to result in a loss of function; the latter genes are inactivated, and their negative regulation of cell proliferation is cancelled, which leads to abnormal growth. In 1971, Knudson proposed the two hits hypothesis which stipulates that both alleles (versions of a gene inherited by our mother and father, identical alleles leading to the homozygous state while two different alleles to the heterozygous state) of a TSG must be inactivated or lost for the gene to lose its normal functions [17]. This hypothesis seemed to explain familial cancer cases [18]. Indeed when the first hit is an inherited germline mutation, the cancer susceptibility of a person increases since only one alteration is needed to alter the TSG functions. The second alteration can result from different events: a mutation in the second allele, the loss or translocation of chromosome pieces or the loss of an entire chromosome. The two latter events causing what is called loss of heterozygosity (LOH) [5].

In the case of the two hits hypothesis, two mutations in the same gene are required for cancer initiation. However, it has been described that cancer is rather a multi-step process, meaning that multiple mutations and more than one gene are usually involved. A certain number of alterations in key pathways are necessary, and it can take several years for cancer to develop [11]. However, the multi-step process can be accelerated. Firstly, as mentioned previously, the inheritance of germline mutations speeds up the cancer development as one driver mutation might be present from birth, increasing the probability that the remaining necessary events, which generally follow a stochastic process, will also occur. [11]. Also, even if multiple DNA repair mechanisms fix most of the alterations that a genome endures, the DNA repair pathways themselves can be disrupted, leading to an acceleration of the accumulation of alterations. Such an event increases the mutation rate of an individual and generates what is called a "mutator phenotype" [3, 19]. Finally, driver genes can also be altered by epigenetic changes that are more frequent, which increases the chance of disrupting key biological pathways for cancer development.

1.1.3 Cancer: an environmental disease

Mutations can arise from endogenous processes, for example, errors happening during DNA replication. In that regard, the appearance of mutations across the genome seems random, and the advent of a driver mutation leading to cancer development seem associated with bad luck. This idea has been developed by Tomasetti *et al.* [20] in a controversial paper, published in 2015, suggesting that the majority of cancer mutations were due to "bad luck". In 2017, the same authors confirmed that mutations due to random errors represent a large proportion of mutations in multiple cancers while specifying that if luck and randomness do play a role in cancer development, other factors like exogenous processes also impact our DNA and contribute to cancer development. [21]

Cancer incidence varies depending on the countries considered. Lung cancer incidence, for example, is much higher in Asia, Europe and North America than in Africa [22]. Those differences can be explained by the fact that cancer has a heritable component that differs in different parts of the world and by the fact that environmental exposures are different across countries. It has been shown, though, in studies exploring cancer rates in migrants populations, that the differences observed among populations could not be explained only by the genetic component [23]. In the second half of the 20th century, epidemiological studies have indicated that several environmental exposures were associated with cancer incidence, showing that many cancers could be prevented. One of the most striking findings was

that of Doll *et al.* showing that smokers had a twenty-fold higher risk of developing lung cancer than non-smokers [24]. At the same period, chemical agents have been identified as being able to induce cancer, *i.e.* being carcinogenic [25]. Some of these agents were also defined as mutagenic agents, *i.e.* agents inducing DNA damages.

Some carcinogens can impact cancer evolution without causing DNA alterations; they are non-mutagenic agents and are considered as tumor promoters. One example of tumor promoter is alcohol which is a cytotoxic substance. Its consumption leads indeed to the death of epithelial cells in the mouth and throat, which triggers the division of the stem cells to regenerate the epithelium. If tobacco consumption precedes this event, tobacco-induced mutations might be present in the dividing cells, and clonal expansion of these mutations may lead to cancer [11]. In that case, smoking acts as a tumor initiator and alcohol as a promoter by stimulating cell proliferation. Such interaction between alcohol and smoking is observed in head and neck cancers. Note, however, that alcohol can also have a mutagenic effect due to metabolites generated during ethanol oxidation like acetaldehyde [26]. Other examples of tumor promoters are steroid hormones acting as mitogenic agents or chronic inflammation (*e.g.* due to viruses).

We have seen that mutations in our genome can result from endogenous processes like replication errors or DNA repair defects and from exposition to carcinogens. Observing these mutations across the whole genome have revealed patterns. Indeed, each of these processes can generate what is called mutational signatures, *i.e.* specific combinations of mutations [27]. The first studies of mutational signatures focused on single base nucleotide substitutions (six possible substitutions: C>A, C>T, C>G, T>A, T>C, T>G) and their tri-nucleotide contexts (the 5′ and 3′ nucleotides flanking the substitution) leading to 96 possible classes of mutations. The classification of all mutations found in cancer genomes in those 96 groups and the use of mathematical methods (See section 1.4) to decompose the mutational processes enable the identification of a limited but diverse set of signatures. In the case of lung cancers, comparing the DNA of smokers with that of non-smokers revealed an increase of mutations in smokers mainly due to an elevation of C to A (C>A) mutations, probably caused by the tendency of tobacco carcinogens to induce this particular change [28]. In melanoma samples, an increase of C>T substitutions has been identified as a result of Ultraviolet (UV) light exposition [29]. In 2015, COSMIC provided a curated set of 30 mutational signatures based on previously published studies on different cancer types [30]. Recently the methods to disentangle mutational signatures in human genomes have been extended. In 2020, Alexandrov *et al.* have considered higher context to classify single base substitutions by considering

two flanking bases around the positions of the mutations and analyzed as well other types of mutations like double base substitutions and indels. This work led to an expansion of the repertoire of mutational signatures with more than 60 signatures in total [31].

Although some signatures are resulting from endogenous processes, like defects in DNA repair or unknown processes, multiple signatures have been associated with preventable exposures. Considering the important impact of environmental exposures, Wild *et al.* suggested in 2005 the concept of the *exposome* which corresponds to all the exposures encountered by an individual during his lifetime (*e.g.* life-style, exposures to chemicals). He expressed the need to improve the measurement of such exposures at the same scale of the genomic events measurements [32]. Indeed on the genome side, remarkable technological advances were made in the past decades allowing researchers to explore the human genome at high resolution. The evolution of these technologies is described in the next section.

1.2 The era of genomics

1.2.1 From arrays to next generation sequencing

The identification of the genomics variations leading to cancer has been enabled by multiple technical and technological advances that occurred after the discovery of the DNA structure. Since that discovery, researchers have attempted to decipher the hidden information contained in the double helix molecule. One fundamental advancement in genomics has been the development of the first generation sequencing by Frederic Sanger in the 1970s. After automatization, this technique led indeed to the sequencing of the first human genome in the context of the Human Genome Project (HGP) that started in the 1980s, took 13 years and cost around 3 billion dollars to lead, in 2003, to the sequencing of the 3 billion nucleotides that our DNA constitutes. At that time, the largest genome sequenced was the 20,000 times smaller genome of the Epstein-Barr virus [33]. While many researchers thought it was impossible, the project completed and delivered the first version of the human genome reference which, after being revised and improved, is now used on a day-to-day basis in genomics. However, the first generation sequencing technology was too long and costly to be applied in larger research projects aiming in that period to catalogue the genetic variations involved in human diseases.

The array technology

At the same period, the microarrays technologies were far less expensive. This technique consists in disposing, on an array, DNA sequences, called probes, designed to bind (by hybridization) to target sequences in a sample. The target sequences are labelled to measure the hybridization and quantify the target molecules.

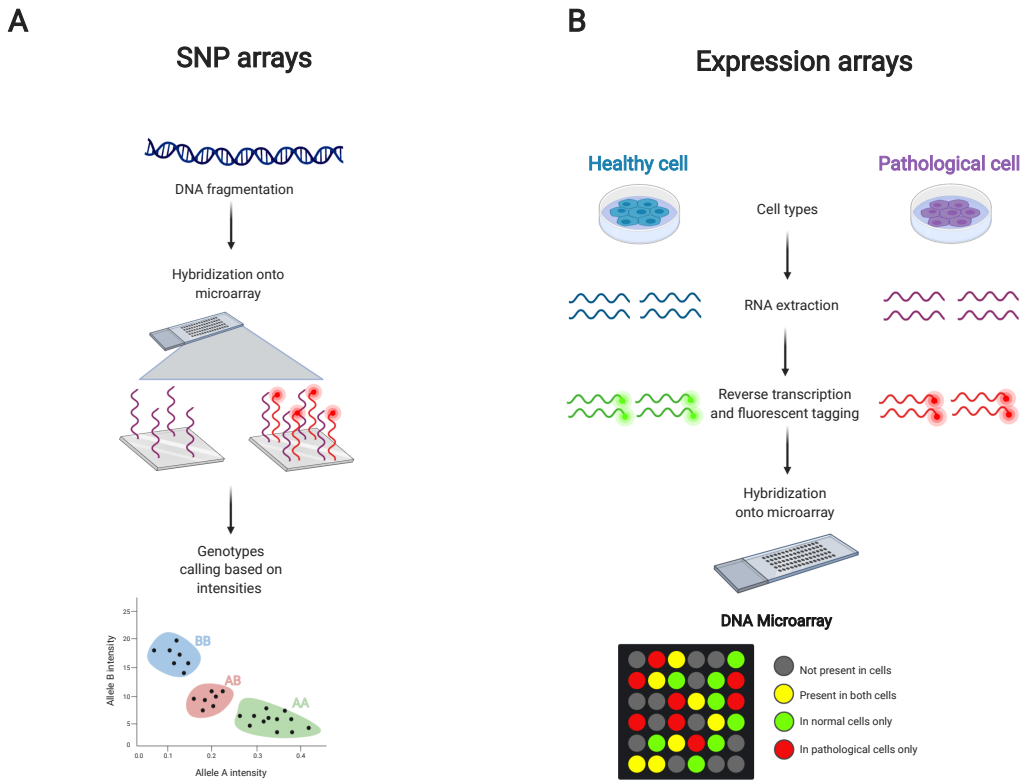

FIGURE 1.5: **Microarrays**. A) SNP arrays: fragmented DNA sequences bind to designed probes on the microarray, which generates an intensity signal that varies depending on the allele carried by the DNA sequences. B) Expression arrays: tagged complementary DNA, reverse-transcribed from mRNAs molecules, bind to gene-specific probes, which generates a fluorescence signal used to compare expression levels in different cell conditions. Created with BioRender.com

In order to study genomic variations across the genome, specific microarrays were developed, the genotyping or SNPs arrays. Those arrays contain unique probe sequences, targeting specific positions of the genome, which hybridize to single-stranded DNA that has been fragmented. This generates intensities signals varying depending on the allele carried by the DNA sequence binding to each probe. This intensity, indicating the presence or absence of each allele, is then converted into

genotypes [34] (See Figure 1.5A). The SNP arrays developed for commercial purposes have evolved, interrogating from 10,000 to millions of sites simultaneously in a given individual [35]. Key produces of these technologies were developed by Affymetrix and Illumina inc. Those arrays have been used so far for different purposes. They allowed the identification of copy number changes or, for arrays with high marker density regions, the detection of LOH events by identifying regions without heterozygous positions [36, 37]. They have also been used to identify germline variants that associate with a certain disease through Genome-Wide Association Studies (GWAS) [38]. As illustrated in Figure 1.6, GWAS interrogate millions of positions across the genome by testing their association with a specific trait, like smoking traits, individually and reveal positions significantly associated with that trait.

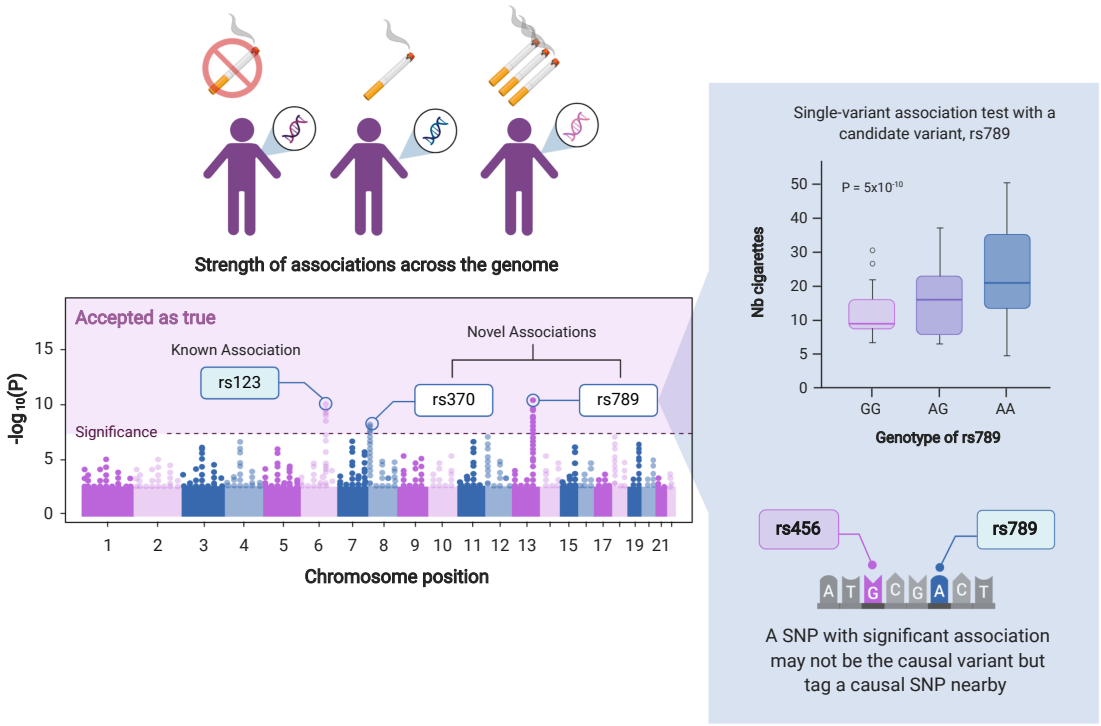

FIGURE 1.6: **Genome-wide association studies**. The figure illustrates a GWAS identifying SNPs associated with the number of cigarettes smoked per day. For each position, the association between the variant genotypes and the number of cigarettes per day is tested (rs789 example). The associations p-values are represented in a Manhattan plot (left panel). SNPs reaching the genome-wide significance threshold of 5.10^{-8} are considered as true associations. Those SNPs do however not always correspond to the causal variant but often tag a nearby SNP in linkage disequilibrium. Created with BioRender.com

Although SNPs arrays are limited to the positions assayed, much more positions can be studied based on the arrays. Indeed, SNPs are transmitted to the offspring linked to other close SNPs in blocks called haplotypes. This relationship between SNPs is called linkage disequilibrium (LD). Knowing the SNPs composition of a haplotype enables to predict the genotype of SNPs that were not assayed by the array by using the information of the assayed positions in the haplotype. Hence, genotyping hundred thousands of SNPs allows actually to impute the genotype of millions of other variants thanks to LD. The definition of the haplotypes required though to study such genomic structure in different samples to build a map as reference. Those were the goals of the Haplotype Map project (HapMap) started in 2002 [39, 40].

Micro-arrays platforms have also been used to study the other molecular layers like the transcriptome and the methylome. For the analysis of the expression profile, micro-arrays have enabled to measure and compare the expression levels of specific genes in cells under different conditions, *e.g.* diseased versus healthy cells or treated versus non-treated cells. Figure 1.5B describes the main steps of an expression array experiment. The extracted mRNAs molecules from both types of cells, after being reverse-transcribed to complementary DNA (cDNA) and labelled with fluorescent dye, hybridize to the genes specific probes fixed on the array. The array is then scanned using fluorescent imaging [41]. The fluorescence amount detected at each probe is proportional to the amount of mRNAs in cells. While these measures do not provide absolute quantification of gene expression levels, they enable to compare the expression levels in the different conditions. Arrays have also been used to study the epigenome by allowing the detection and the analysis of methylation events. The most commonly used methylation arrays are the Illumina arrays [42]. As for the SNPs arrays, probes are designed to target specific loci of the human genome, in this case, CpG positions. The number of positions interrogated by such arrays can vary from 25,000 to 850,000 positions depending on the array (*e.g.* Illumina 25K, 450K and 850K arrays). Probes are designed and fixed to the array to bind to both methylated and unmethylated loci (Figure 1.7). This binding is enabled by a chemical process called bisulfite conversion, which converts unmethylated cytosines to uracil and leave methylated cytosine unchanged. At the hybridization step, a single-based extension is performed with labelled nucleotides, allowing to distinguish for each locus a methylated vs non-methylated signal (Figure 1.7). The ratio between the two signals at a locus provides a value, called β value, which indicates the level of methylation. This value ranges between 0 and 1, 0 corresponding to a non-methylated and 1 a methylated position.

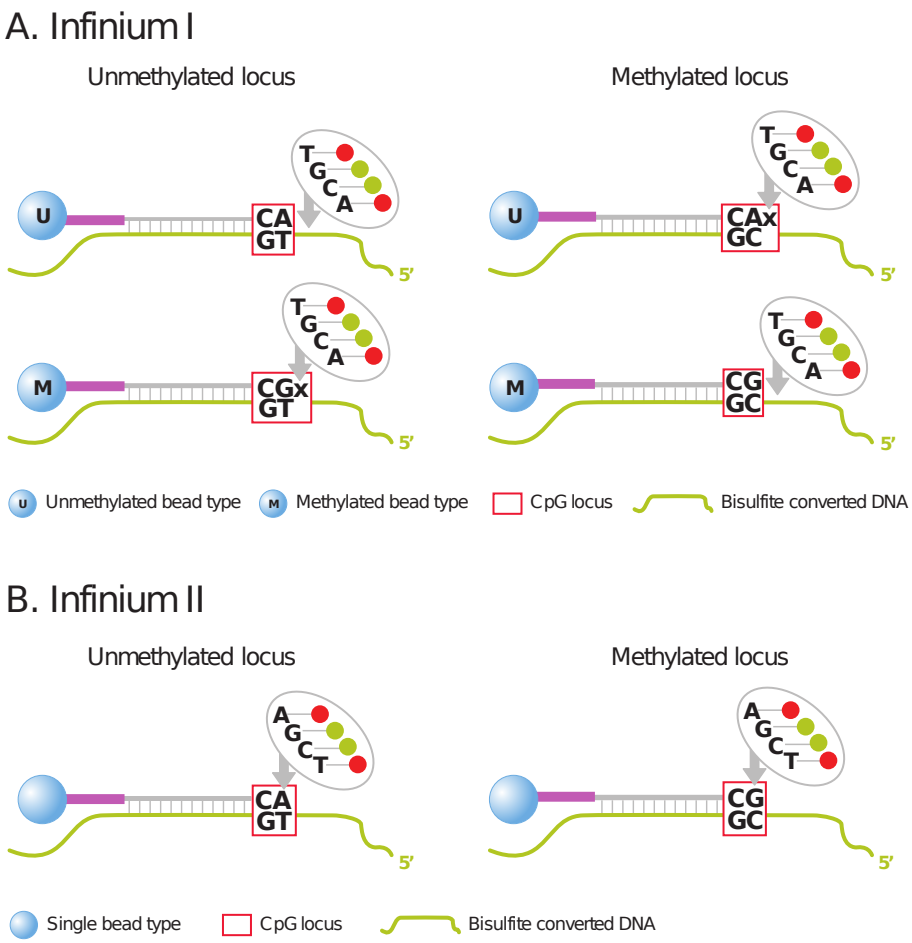

FIGURE 1.7: **The Illumina Infinium methylation assay** (From [42]). This figure represents the probes used for methylation profiling by Illumina. A) Infinium type I probes. Two site-specific probes are found on the array: probes allowing methylated sites with the preserved cytosine to bind (methylated bead M) and probes designed for the unmethylated site with the thymine nucleotide resulting from bisulfite conversion and whole-genome amplification (methylated bead U). B) Infinium type II probes. Only one probe per locus is required to bind to both methylated and unmethylated sites. In that case, single-base extension with labelled nucleotides is used.

Next-generation sequencing

While the SNP arrays enabled to access the genotype information of millions of positions, there was still a need to re-sequence human genomes more efficiently and access the complete DNA sequence to better identify genetic variations. Around 2005, the second generation of sequencing methods called Next Generation Sequencing (NGS) has been developed.

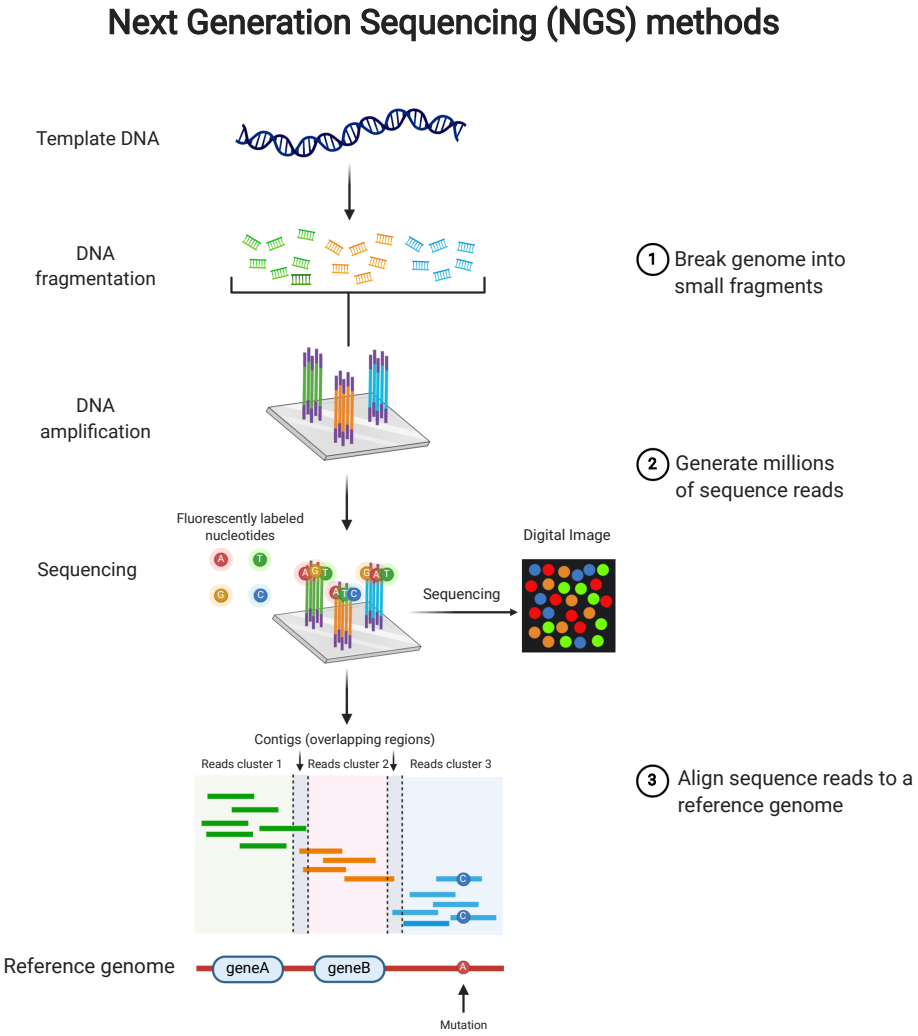

FIGURE 1.8: **Next Generation Sequencing methods**. The figure describes the NGS steps consisting in: i) fragmenting the nucleic acid molecule, ii) amplifying the fragments (using Polymerase Chain Reaction (PCR)), iii) sequencing the resulting copies using single-base extension that adds one after the other labelled nucleotides whose signals are detected using digital imaging. The sequencing reads are then aligned to a reference genome to assemble the reads in a single sequence or to detect mutations across the genome. In the case of RNA sequencing, the reads align to exonic regions of the genes and they are counted to quantify gene expression levels. Created with BioRender.com

The main change in these new methods in comparison to the first one was the parallelization of the sequencing, which allowed to produce millions of sequences, called reads, at the same time and hence to decrease drastically the time of sequencing as well as its cost [43] (Figure 1.8). NGS methods enabled the rapid resequencing of different parts and lengths of the genome. The entire genome sequence (except some highly problematic regions) can be accessed with Whole Genome

Sequencing (WGS). The restricted sequencing of coding regions (exonic regions) can be performed with Whole Exome Sequencing (WES). Finally, it is possible to sequence specific regions of the genome, usually genes, using targeted sequencing. Based on these techniques, bioinformatics methods have been developed to detect germline as well as somatic variants. They consist in mapping (or aligning) the sequenced reads to a reference genome, and positions that vary from the reference are identified as variations (Figure 1.8). A mismatch between a sequenced genome and the reference genome is expected around every 1,000 bases. To distinguish somatic from germline mutations, both tumor and normal cells DNA from the same individual have to be sequenced. The tumor DNA is compared to the normal DNA and variations found in the tumor cells only are classified as somatic mutations. Somatic mutations are expected every 1,000,000 bases approximately depending on the cancer type [27].

While the DNA sequencing techniques have been used to detect DNA mutations, they do not explore the expression or methylation layers. In 2008, the sequencing of the RNA molecule (RNA Sequencing (RNA-Seq)) had been performed to study expression profiles. In this technique, the mRNAs molecules are fragmented and converted to complementary DNA before sequencing, and the resulting reads are aligned to the reference genome [44]. After the alignment step, the reads can be assigned to genes and the abundance of reads mapped on a gene, quantified using the number of mapped reads, reflects the expression level of the gene (Figure 1.8). A high read count value indicating that a gene is active and transcribed in that sample. The comparison of the read counts distributions in samples from different conditions, *e.g.* samples with and without disease or diseased samples under different treatment, can be used to identify genes involved in or causing a specific condition. RNA-Seq can also be used to identify different transcripts of a gene as well as gene rearrangements like translocations.

Note that other recent techniques, while not described in the thesis, also exist to access different omics layers. A new sequencing technique has been developed for the analysis of the methylome, the bisulfite sequencing, which in contrast with the methylation arrays, can interrogate millions of CpGs positions across the whole genome as well as positions in targeted regions. Also, the study of chromatin accessibility and DNA-binding proteins is possible thanks to Assay of Transposase Accessible Chromatin sequencing (ATAC-seq) and Chromatin immunoprecipitation experiments followed by sequencing (Chromatin immunoprecipitation Sequencing (ChiP-Seq)) respectively [45, 46]. Finally, while the sequencing methods presented so far process DNA coming from a bulk of cells, single-cell sequencing methods

have been developed to perform molecular characterization at the cell level. These methods allow the identification of distinct populations of cells in a tumor and hence the study of tumor heterogeneity and tumor microenvironment [47, 48].

The decreasing costs of genotyping and sequencing methods have enabled the establishment of genomics studies involving large cohorts [43]. Sequencing a human genome today costs less than 1,000 dollars using NGS methods while it would still cost millions if the Sanger method was chosen. Multiple research groups have coordinated their efforts to create large consortia for that purpose and in many cases have shared the resulting data to the scientific community. The next section provides an overview of some of these initiatives.

1.2.2 Large public databases

The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a public database providing access to 10,000 patients whose tumors have undergone multi-omics characterization. The project was launched in 2005 by the National Institutes of Health (NIH) and aimed at characterizing the genomic alterations underlying several cancer types. For that purpose, multiple omics data were generated [49]. The tumor and normal samples from most of the TCGA participants have been sequenced using WES. Based on these data, multiple variant callers have been used to catalogue the germline and somatic mutations present in each sample. Genotyping has been performed to analyze copy number variations. The transcriptome of most samples has also been sequenced, using RNA and miRNAs sequencing. The methylation profiles of the tumors were explored with the use of 25K or 450K methylation arrays. Finally, protein expression profiling has been performed based on Reverse-Phase Protein Array (RPPA). In addition to the molecular data, clinical and environmental exposures data were collected when possible. The TCGA projects also delivered the histopathological images associated to each tumor. Based on these diverse omics and clinical datasets, "marker papers" describing the molecular landscape of each tumor type have been published. While the tissues explored at the beginning of the initiative were limited to lung, brain and ovaries, the TCGA data encompasses today molecular data from 33 different cancer types. Those cancer-specific studies led to the identification of genomics alterations causing each cancer type, hence the discovery of new driver genes and potential cancer biomarkers, *i.e.* molecules found in the body as an indicator of a disease or specific condition. Also, cancer subtypes were characterized

on the molecular level and subtype-specific alterations were identified, which resulted in new clinical managements of tumors [50]. In parallel to the cancer-specific studies, the TCGA research network launched, in 2012, the Pan-Cancer Atlas initiative aiming at exploring the commonalities between cancer types, distinguishing tissue-specific determinants of cancer as well as increasing the statistical power for the identification of genomic alterations [50]. This initiative was completed in 2018 and the data have been released and associated to 27 papers, published in Cell, focusing on three main topics: i) cell-of-origin patterns and cancers subgrouping, ii) oncogenic processes, and iii) signaling pathways involved in cancer [51].

The International Cancer Genome Consortium (ICGC) initiatives

The TCGA studies focused their efforts on the characterization of the cancer exomes. However, exomes represent only 1% of the human genome and much more can be discovered by exploring the remaining 99% of the genome. In 2007, the International Cancer Genome Consortium (ICGC) project was launched to study more than 20,000 whole genomes from 50 cancer types having an impact in multiple regions of the world (the 25k initiative). The international consortium aimed at generating a catalogue of the somatic mutations in those cancer types, sharing the resulting datasets and complementing them with transcriptomic and epigenomic datasets [52, 53]. Based on the samples included in the TCGA and the ICGC projects, the Pan-Cancer Analysis of Whole Genomes (PCAWG) project, an ICGC initiative also know as the Pan-Cancer project, has arisen [54]. The project relied on more than 2600 samples from 38 different tumor types and aimed at meta-analyzing whole-genome data across cancers along the same lines as the PanCancer Atlas project. The first results from these data have been released in 2020 in a series of publications in Nature [53]. While the TCGA initiative enabled the study of the coding regions of the samples, the PCAWG project, thanks to the use of whole genome sequences, was designed to explore broader mutational patterns in the coding and non-coding regions, from small to large events like structural variations. For example, chromoplexy and chromothripsis events, which are complex chromosomal rearrangements resulting from catastrophic genomic events, have been observed in more cancers than expected, 17.8% and 22.3% of the tumors, respectively [53]. Also, one major result from the PCAWG project has been the expansion of the mutational signatures mentioned in section 1.1 [31], as well as the discovery of 16 structural variants signatures [55].

UKbiobank

The previously described projects mainly targeted the somatic landscape of genomes. Other large projects have enabled the research community to explore the germline component of human disease. The largest public dataset, focusing on germline genetics, has been generated by the UKbiobank project, which started in 2010 in the UK. This project gathered data from a population-based cohort of around 500,000 participants between 40 and 69 [56] and had as main objective to improve our understand of the interaction between genetics and multiple human diseases. For that purpose, all participants were genotyped. Besides, multiple other biological samples, like urine, blood and saliva as well as physical measures, *e.g.* brain Magnetic Resonance Imaging (MRI), heart and eye measurements, were collected. It is a prospective cohort; participants are followed up and are linked to electronic health records [57]. The genotyping data of the full cohort were released in 2017. Based on this dataset and the large panel of phenotypes, a multitude of GWAS studies related to human diseases have been performed and their resulting summary statistics were made available. In 2019, around 100 GWAS studies resulting from the UKbiobank data were available on the GWAS catalogue, which provides curated GWAS summary statistics results [58]. The follow-up of the patients has established that, in 2018, 79,000 of the participants were diagnosed with cancer [57], which means that cancer-related traits can also be studied using this dataset. After the release of the genotyped and imputed data, WES and WGS sequencing of the samples have been initiated. Part of the exome data, around 50,000 exomes, have already been released and about 200,000 exomes should be expected by the end of 2020. These data foreshadow future key findings in genomics, a better understanding of molecular and phenotypic interactions and probably an improvement of the translation of those findings in the clinic.

Data sharing

With the increasing number of genomics studies, public repositories, like the Database of Genotypes And Phenotypes (dbGAP), the European-Genome Phenome Archive (EGA) or Gene Expression Omnibus (GEO), have been established to store petabytes of genomics data that can be accessed by the research community. In addition, large projects, like the TCGA and ICGC, have worked on solutions to improve data storage and accessibility. One of the goals of those projects was to promote open-access data and the development of tools to foster the reuse of the data by the research community [50, 52]. In 2010, the TCGA provided the data in open access for the first

time [59] and updated and extended the content of the open access data over the years. In 2016, the Genomic Data Common (GDC) was launched by the National Cancer Institute (NCI) to store all the TCGA data [60]. For each omics, the data are categorized by levels: low-level data (raw and unnormalized data) that generally enable individuals re-identification are under controlled access, while higher-level data (processed data, clinical data) that do not permit re-identifiability are available without any requirement. In addition to providing the data storage, the GDC also aimed at harmonizing and sharing the bioinformatics pipelines used to process the data [60, 61]. The processed data resulting from the PanCancer Atlas papers are also available via the NIH GDC website [62] and allow researchers to explore broader genomic features like immune variables [63] or biological pathway measures [64]. Also, cloud computing solutions have been developed to facilitate the analyses of large public genomic datasets while avoiding the download and duplication of the data. The TCGA and ICGC data are available and can be analyzed on the cloud, for example via the Cancer Genomics Cloud (CGC) [65] or the ISB Cancer Genomics Cloud (ISB-CGC) [66]. Also, the ICGC consortium, to process the PCAWG data, has developed a computational tool, Butler, which simplifies genomic analyses that have to be run on clouds environments (academic or commercial) [67].

In the past decades, the development of genomics technologies and the implementation of large consortia have enabled to characterize human cancers on the molecular level. The understanding of cancer causes and the biological mechanisms underlying tumor development has been improved. Also, due to the identification of correlations between molecular events and patient's prognosis and response to treatments, molecular studies have impacted the way that tumors are classified and managed in the clinic.

1.3 The example of lung cancer

1.3.1 Lung cancer subtypes and etiology

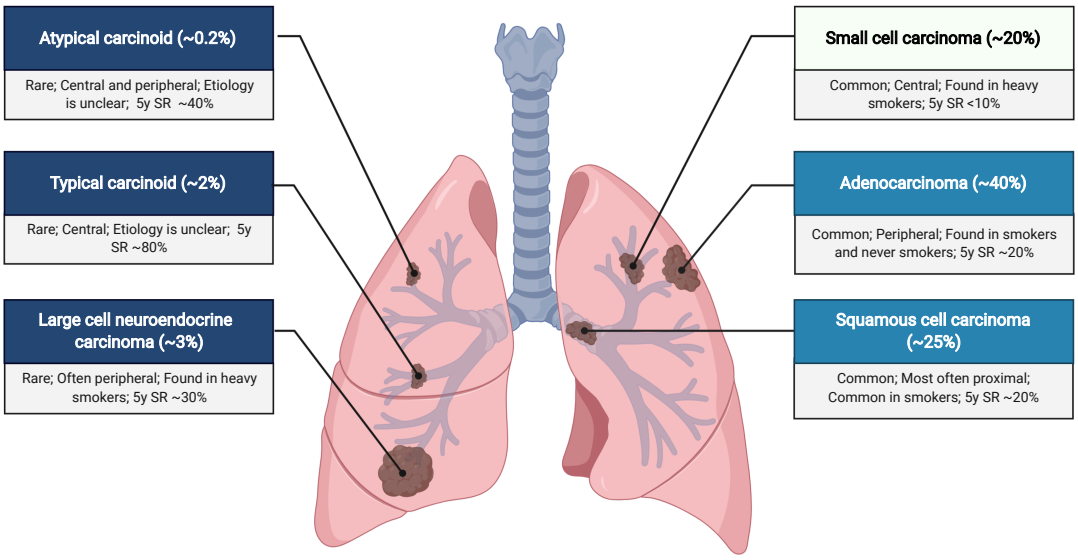

FIGURE 1.9: **Lung cancer subtypes.** Each lung cancer type occurs at different frequencies as well as at distinct locations in the lung (from proximal to distal locations). Each box on the figure is associated to one cancer type and provides their characteristics (frequency, localisation, etiology and overall 5-year survival rate (5y SR)) [68, 69, 70, 71]. Figure created with BioRender.com

As mentioned at the beginning of the manuscript, lung cancer is one of the most common and deadliest cancer worldwide. Several subtypes of lung cancers have been identified (Figure 1.9). The most common lung cancers are usually divided into two groups: the Small Cell Lung Cancer (SCLC) and the Non Small Cell Lung Cancer (NSCLC) samples, representing respectively around 20 and 75% of the lung cancers [72]. The second group is further separated into two main subgroups: the Lung Adenocarcinomas (LUAD) and the Lung Squamous Cell Carcinomas (LUSC). Also, rarer forms of lung cancer exist. Multiple lung cancer subtypes, including such rarer cancers, were grouped in one category named the lung neuroendocrine tumors by the World Health Organization (WHO) 2015 classification [73]. This group comprises the pulmonary carcinoids, including the typical and atypical carcinoids,

Large Cell Neuroendocrine Carcinoma (LCNEC) as well as the previously mentioned SCLC tumors. Each lung cancer type can be distinguished by different etiologies, histopathological characteristics, molecular profiles and clinical outcomes (See Figure 1.9).

The strongest risk factor for lung cancer is smoking. Indeed, SCLCs and LCNECs are frequently found in heavy smokers. Smoking is also a major risk factor for LUAD and LUSC cancers [74]. However, lung cancer can also develop in non-smokers. In particular, the LUAD category corresponds to the lung cancer type most commonly found in never smokers. Although the etiology of the pulmonary carcinoids is not clear, the majority of these tumors are found in nonsmokers [69]. In addition, around only 15% of smokers develop lung cancer suggesting than other factors mediate lung cancer risk.

1.3.2 Lung cancer susceptibility

While exposures other than smoking like air pollution, radon, heavy metals or asbestos have been identified as lung cancer risk factors [75], genetics is also contributing to the disease risk. In line with this hypothesis, it has been shown that having a family history of lung cancer confers a 2.5 fold lung cancer risk increase [76]. Further evidence of lung cancer germline susceptibility has been revealed by GWAS studies, with the identification of common variations associated with lung cancer. Genes involved in nicotine addiction (*CHRNA* genes), telomere activities (*TERT*) as well as genes related to the DNA repair and cell-cycle pathways (*e.g. Check2*, *RAD52* or *CDKN2A*) have been identified [77]. Also, some lung cancer associated variants were identified as related to the propensity to smoke [78, 79] and genetic correlations between lung cancer and smoking traits, like smoking initiation, smoking cessation or smoking intensity have been described [79]. Such observations provided evidence that susceptibility variants could influence lung cancer risk through environmental exposures. Hence, GWAS studies have enabled to gain insights on lung cancer etiology as well as on the biological pathways involved in the disease. However, the variants identified so far do not account for most of the heritability of lung cancer, estimated at 18% and remaining today largely unexplained [79].

1.3.3 Lung cancer molecular profiling

In the past decades, molecular profiles of human tumors, including lung tumors, have also been explored thanks to the development of NGS studies. Such studies have, for example, established that lung cancers are among the cancer types with the

highest mutational burden (total number of mutations for a given part of DNA) [80]. As mentioned in Section 1.1, in smoking-related cancers, those mutations revealed a signature associated with tobacco consumption. Among the Catalogue Of Somatic Mutations In Cancer (COSMIC) signatures identified by Alexandrov *et al.* [27, 31], the smoking signature corresponds to the Signature 4 (COSMIC version 2) and SBS 4 (COSMIC version 3). Those signatures are the results of DNA damages caused mainly by benzo[α]pyrene, which is a mutagenic compound found in tobacco smoke and whose effects on DNA has been shown in experimental mutagenesis studies [28]. Even though smoking does heavily impact the lung tissue, it has been shown that quitting smoking can restore the damaged tissue [81].

In addition, molecular analyses of lung tumors have identified cancer driver genes in the different cancer types. Among those genes, the *Epidermal Growth Factor Receptor (EGFR)* gene, which is part of the protein kinase family currently known to be mutated in around 15% of the LUAD samples [82], has been related to therapeutic response in 2004 [72]. Indeed LUAD samples, carrying activating mutations in the *EGFR* gene, are responsive to tyrosine kinase inhibitor therapy and have an improved survival in comparison to other cancer patients treated with chemotherapy. Such molecular studies largely influenced the way that lung tumors are classified since it is only since those discoveries that NSCLC are further sub-classified. Guidelines were published in 2013 to include molecular testing, mainly based on *EGFR* and *ALK* alterations testing, in the clinical practice for the NSCLC patients. In 2018, those guidelines were updated and new alterations, like rearrangements in the tyrosine kinase *ROS1*, are now recommended for molecular testing [83]. In 2012 and 2014, the TCGA marker papers on the two lung cancer cohorts (LUAD and LUSC) were published. The authors expanded the molecular profiling of these tumors and hence the list of drivers genes, improving the understanding of the biological mechanisms involved and providing new opportunities for patients management [84, 82]. Those studies also explored the transcriptomic, methylation and proteomic data from the lung tumors. Based on their expression profiles, the LUAD tumors, were divided into subtypes that could help to refine those tumors classification [82].

The identification of driver genes in lung cancer has also led to the proposal of molecular targets for early detection. The molecular profiling of SCLCs is an example of such an application. SCLCs are characterized by universal inactivation of both *RB1* and *TP53* genes [85, 86, 87]. In 2016, Fernandez-Cuesta *et al.* analyzed circulating tumor DNA (ctDNA), which are fragments of tumor DNA released in the bloodstream that can be used as molecular biomarkers, in SCLCs. They showed that *TP53* mutations were detectable in the ctDNA of the SCLC cases [88]. ctDNA

applications are viable for multiple cancer types. In 2018, Cohen *et al.* described a blood test called CancerSEEK, detecting proteins and mutations in cell-free DNA for the early detection of eight different cancer types, including lung cancer [89]. Such tests face though sensitivity issues due to the low abundance of mutated DNA in body fluids, hence adapted bioinformatics tools are needed. Needlestack, a highly sensitive multi-sample variant caller has been for example developed in this context [90].

Even though rare forms of lung cancers are less explored than the common lung cancers, recent molecular studies have started to characterize the lung neuroendocrine tumors as well [91, 92, 93, 94]. Those studies have revealed that, on top of their histopathological differences, the lung neuroendocrine neoplasms were also distinct molecular entities [87]. Low mutational burden has been observed in the atypical and typical pulmonary carcinoids in contrast to the highly mutated LCNECs and SCLCs [69]. Also, the transcriptomic profiling of those tumors has been investigated. These analyses identified molecular subgroups in different cancer types, revealing the molecular heterogeneity in those tumors [92, 95].

The discoveries described in this section were enabled thanks to the large amount of data generated during the era of genomics (See Section 1.2). However, the analyses of these data have raised multiple challenges that required the use and development of specific computational methods. The next section intends to describe those aspects.

1.4 Interpreting high dimensional data

The evolution of genotyping and sequencing technologies led to the generation of high dimensional datasets. In Section 1.2, we have seen for example that arrays can interrogate thousands to millions of positions across the genome and that sequencing techniques can provide the entire genome sequence or the expression levels of thousands of genes. While the amount of information unveiled by these methods is colossal, it can also bring about several challenges and adapted computational methods are required to analyze and interpret the data. The issues resulting from high dimensionality are associated to what is called the curse of dimensionality, firstly introduced by Bellman in 1961 and stipulating that the number of samples needed to interpret high dimensional data analyses appropriately increases exponentially with the number of dimensions [96]. In omics datasets, even though large cohorts have been implemented (see section 1.2), the number of variables (also known as

features), p, to analyze can be largely superior to the number of samples, n, included in the study. This introduces the $n << p$ problem, which leads to multiple issues. Firstly, usual statistical models like regression models need to be adapted since they require $p < n$. There is also a substantial amount of noise in the generated data that can mask the true signal in the data, *i.e.* not all the measured features are of interest [97, 98]. In addition, when the number of dimensions increases, the data points can occupy a more voluminous space and a larger proportion of this space will be empty, we say that the data are sparse (See Figure 1.10) [96]. High data sparsity influences basic properties to which we are used to in two or three dimensions like distances. In high dimensions, distances between points increase and all points seem at the same distance from each other [98, 96]. Also, the higher the dimensions, the lower the correlations between the features will be. For those reasons, it is thus statistically more difficult to identify groups of points with similar characteristics compared with random events, as such larger sample sizes are required to distinguish meaningful relationships. Another issue resulting from high dimensionality is multi-collinearity. Since the number of features is high, the information they carry can be correlated and become redundant; some variables might be defined as a linear combination of others which makes the data interpretation more difficult [96]. Finally, the nature of omics datasets complicates the visualization of the data. In this section, we will discuss in a first instance different strategies to explore such complex datasets and secondly focus on methods that attempt to diminish the problem of the curse of dimensionality: the dimensionality reduction methods.

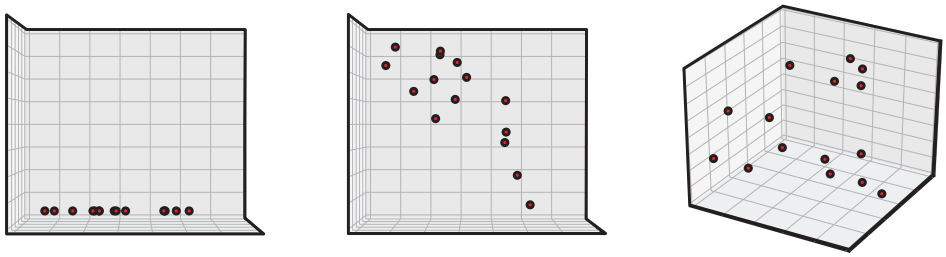

FIGURE 1.10: **Illustration of data sparsity.** Figure from [98]. The figure represents how the data occupy the available space when going from a one-dimensional space to two and three-dimensional spaces (from left to right panels).

1.4.1 Supervised and unsupervised methods

Different approaches exist to analyze high dimensional data like omics data. In the case where specific biological hypotheses need to be tested, confirmatory data analyses based on inference models can be used. It can also happen that there are

no predefined hypotheses and that the goal is to "let the data talk", in that case, exploratory data analyses (EDA) will be more adapted [99]. A broad panel of statistical methods exists to assist both approaches. Among them, a large proportion can be grouped in the popular category of machine learning methods. The term machine learning (ML) was used for the first time by Arthur Samuel around 1950 and defined a group of computer algorithms able to learn without being explicitly programmed to learn. Depending on the definition of learning, different classes of ML methods have been established. In 1997, Tom Mitchell proposed a formal definition of algorithms learning saying that "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [100]. This definition matches a class of ML methods, the supervised learning methods, used for classification and regression tasks. A common example is the identification of spam emails, where labelling emails in the spam or non-spam categories would be the task T, learning from a set of labelled emails would be the experience, and the proportion of correctly classified emails would be the performance measure P. However, ML algorithms that simply learn from the input dataset without predefined ground truth (labelled data) also exist and are part of the unsupervised ML methods. Those methods learn underlying structures in the data; hence algorithms like clustering or dimensionality reduction methods such as Principal Component Analysis (PCA), which was developed even before ML, are often included in the unsupervised learning category. In the next paragraphs, both supervised and unsupervised learning are described (See Figure 1.11).

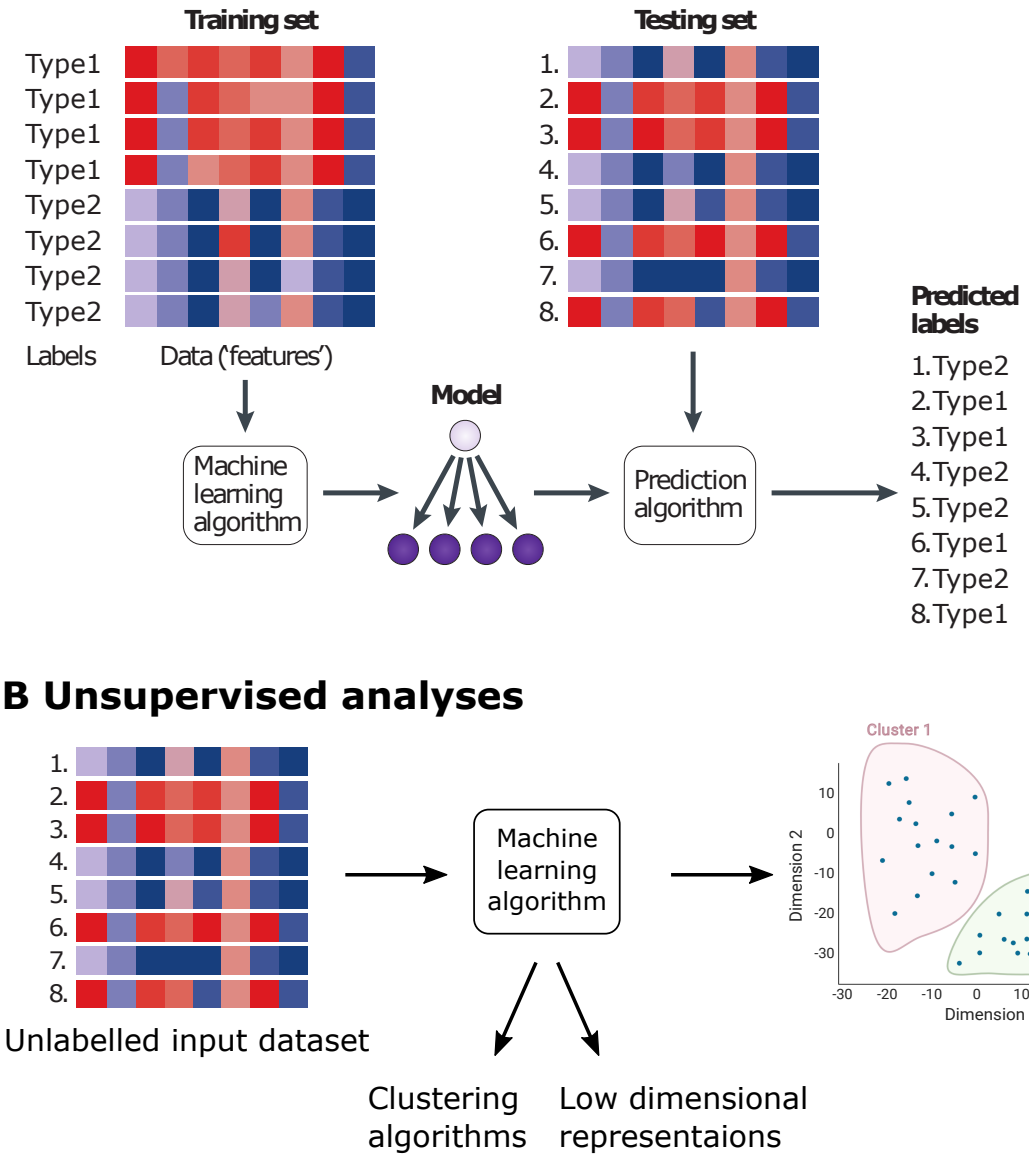

FIGURE 1.11: **Machine learning methods: supervised vs non-supervised methods.** A) Supervised methods: a model is trained on several variables, features, to recognize predefined labels. The trained model is then applied to an unlabelled dataset for prediction purposes. B) Unsupervised methods: a model learns structures underlying a dataset that has not been labelled. Those methods are divided into two main categories: clustering methods to identify subgroups of samples and dimensionality reduction methods to explore the data in lower dimensions and highlight specific structures. Figure adapted from [101].

Supervised analyses

The goal of supervised methods is to predict the value of an outcome based on a set of features given as inputs. Depending on the type of outcome, supervised

analyses can be further divided into two main categories: classification or regression problems. In classification problems, the outcome is categorical, *e.g.* a binary variable distinguishing a diseased or healthy status or a multi-classes variable like cancer subtypes. In regression problems, the objective is to predict a continuous variable. Note that some regression models, like logistic regressions, where the outcome variable is discrete, can be used though to perform classification. The main steps of supervised analyses consist in: i) defining the labels of each sample in the dataset, ii) train the model to classify the samples in the correct category, and iii) use the generated model on a dataset containing independent and unknown instances (Figure 1.11A). Several types of supervised methods exist and have to be chosen with regard to the nature of the data. The simplest supervised models are regression models. While the most common regression algorithms model linear relationships, other methods like Support Vector Machines (SVM) or neural networks can adapt to non-linear data. Another parameter that determines the type of methods to use is the data type; some methods deal only with numerical features while others like decision trees are more flexible. Figure 1.12 describes a method based on decision trees, the random forest algorithm.

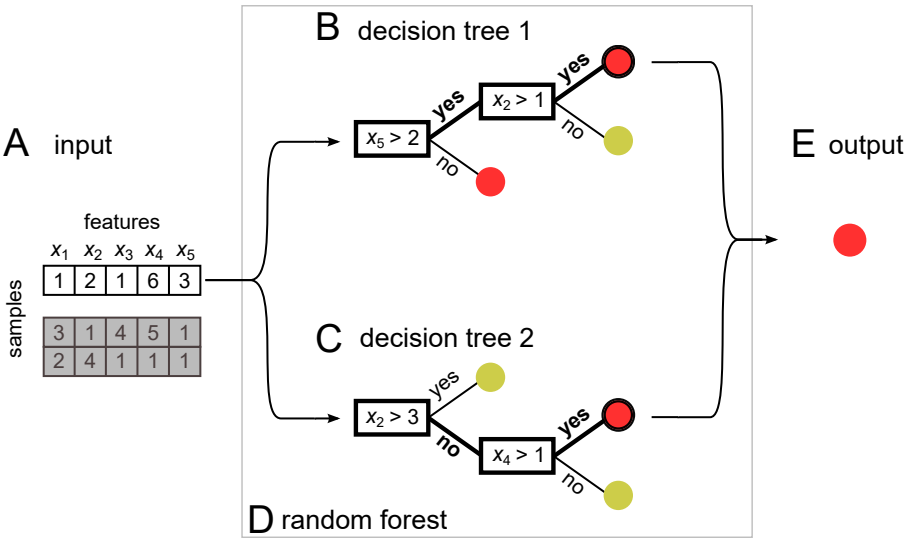

FIGURE 1.12: **The random forest method.** Figure from [102]. A labelled dataset (A) is taken as input and processed by multiple decision trees (B and C) built using random selections of features and samples. The decision trees form a random forest (D). Each tree classifies the input samples and the votes given by the different trees are then combined to provide the final predictions. The label with the most votes being chosen (here red label).

Regardless of the method used, the model and its results have to generalize to other datasets. In order to assess generalizability, the ML algorithm has to be trained

on a training dataset, and a testing dataset containing independent samples has to be used to validate the results. Two main errors underlying the generalization issue exist: bias and variance [97]. The first scenario occurs when the model is underfitting the data, *i.e.* the model has a poor performance even on the training data for example because of a model that is not complex enough (See Figure 1.13 left panel). When the model is underfitting the data, it is as well unable to generalize to other datasets. In the second case, when the number of features is too large or the number of samples small, the chances to encounter features that can perfectly discriminate two output categories or perfectly predict an outcome increase. The model, in that case, performs correctly on the training dataset but fails to generalize to other datasets and is qualified as high variance model. Such performance discrepancy indicates that the model overfits (See Figure 1.13 right panel). Note that in high dimensional data, overfitting and data sparsity, resulting from the $n << p$ problem mentioned at the beginning of this section, can be linked. Indeed, in such data, since the number of samples in the training dataset is fixed and limited, the entire input space is not covered. Thus the machine learning algorithm has not faced all possible configurations during the learning phase and the ability of the model to generalize can be diminished.

FIGURE 1.13: **High bias and high variance models.** Created with BioRender.com.

One method that can be used to detect as well as overcome overfitting is cross-validation. The method consists in randomly splitting the dataset in k folds and iteratively training the model on $k-1$ folds while reserving the remaining kth fold for testing (See Figure 1.14). The overall performance of the model can be assessed by averaging the performances in the testing folds from each iteration. As a result, while none of the samples is used simultaneously in the training and testing group, the entire dataset is used for training as well as is used in the testing phase. Hence, cross-validation can also be beneficial in studies with low sample sizes. One extreme case of cross-validation is the leave-one-out analysis, where $k = 1$. Each sample is set aside from the training set and predicted at each iteration.

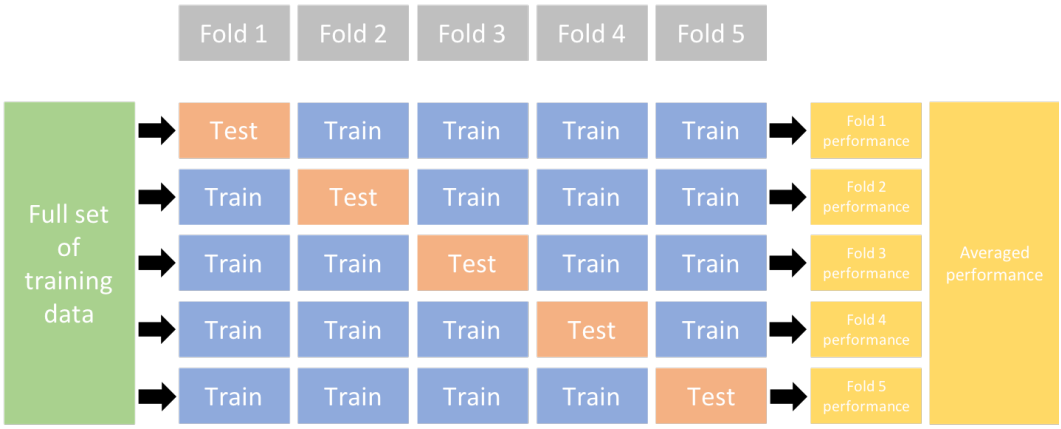

FIGURE 1.14: ***K*-fold cross-validation.** Figure from [103]. The figure illustrates 5-fold cross-validation. Five rounds are thus represented. In each of them, 4 folds are used to train the model and the model is tested on the remaining fold. The performances resulting from the test phase in each round are then averaged to estimate the overall performance of the model and its ability to generalize.

In addition, to find a compromise between bias and variance, parameter tuning and algorithm optimization might be required. Note that a third dataset, referred to as the validation dataset, can be introduced for the optimization step. In this setting, multiple models (*e.g.* one algorithm with different sets of parameters or different algorithms) learn on the training set, and their performances are evaluated on the validation dataset. The model with the best performance can then be applied on the testing dataset.

Unsupervised analyses

Unsupervised algorithms are hypothesis-free methods and can be associated to exploratory analyses [104]. The goal of such methods is usually to identify and extract useful properties of the data [105]. In contrast to the supervised methods, each element of the dataset is not labelled, no predefined groups are given to the algorithms. Thus, it is not possible to compare the algorithm output with a predefined truth and the data do not need to be split in training and testing datasets (Figure 1.11B). Since there is thus no feedback on the performance of the unsupervised model, often the validation of the results is required.

As for the supervised analyses, there are several unsupervised algorithms. A commonly used category of unsupervised methods that can unveil structure in the data is the group of clustering algorithms (*e.g. k*-means clustering, hierarchical clustering, density-based clustering). Those methods aim at grouping elements together

based on common patterns observed in the set of features. In the field of cancer, clustering algorithms can be used, for example, to identify new subtypes of cancers based on molecular data. The second most commonly used unsupervised method is the group of dimensionality reduction methods. In the next paragraph, more details about such methods are provided.

1.4.2 Dimensionality reduction methods

The goal of dimensionality reduction (DR) methods is to transform a high dimensional dataset into a low dimensional representation of the data while preserving as much as possible its initial structure. More specifically, if three clusters exist in the studied dataset, a lower dimensional representation of the same data should also reveal the initial three clusters. DR methods are part of the feature extraction techniques which aim at finding latent structures in the data. Those methods allow to summarize and transform a large number of features in a smaller number of variables, which mitigates the curse of dimensionality and is valuable for data visualization. Note that these methods are different from feature selection methods, which make a selection of the most important features in the initial dataset [106]. Mainly two families of DR methods exist: matrix factorization methods (*e.g.* PCA, PLS, ICA, NMF) or neighbour graphs approaches (*e.g.* t-SNE and UMAP).

Matrix factorization methods examples

Omics datasets, after pre-processing, often result in data matrices. For example, in the case of RNA-Seq, after aligning the reads to a reference genome (See Figure 1.8), reads counting is performed and generates a matrix in which rows represent the genes (the features) and columns the read counts for each sample (the observations). Matrix factorization consists in decomposing an initial matrix in two smaller matrices (Figure 1.15). This decomposition leads to the generation of new variables, in smaller numbers.

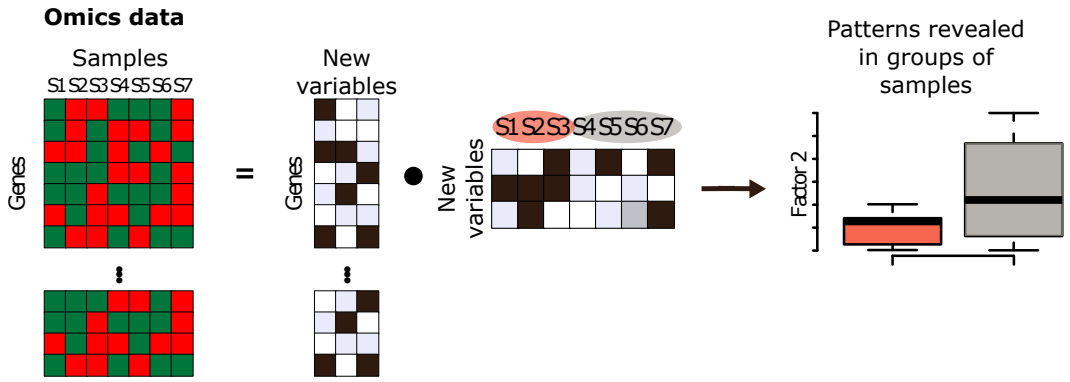

FIGURE 1.15: **Matrix factorization methods**. The input matrix is decomposed, under specific constraints, in two smaller matrices defined by new variables that can be used to reveal structures and patterns in the data.

A classical matrix factorization method is Principal Component Analysis (PCA). The goal of PCA is to project the data to a lower dimensional space while maximizing the variance in the data within this lower dimensional space. In PCA, the new variables correspond to a linear combination of the initial features. The matrix factorization results in the loading and score matrices. In the first matrix, the columns correspond to the new variables, called principal components and the rows indicate the contribution of each feature to the latent variables. The principal components are orthogonal; they correspond to the directions of maximal variance and are ranked by the importance of variance explained, *i.e.* the first principal component captures most of the variation in the dataset. The second matrix contains the coordinates of the samples in the projected space. While PCA maximizes the variance in the data, similar methods use other criteria. For example, Independent Component Analysis (ICA), which is a method attempting at disentangling independent signals that are linearly mixed, maximizes the independence between the new variables. Other methods have in addition specific constraints [107]. Non-negative Matrix factorization (NMF), for example, enforces the decomposed matrices to be positive; this method has enabled the extraction of *de novo* mutational signatures from whole genome sequencing data [108]. One limitation of those methods is that they are linear models. In the following paragraphs, two non-linear methods based on neighbour graphs are presented.

Neighbor graphs methods examples

The principle of DR methods based on neighbor graphs models is to use neighbors distances and similarities to represent the structure of the data in high dimensions and then to embed this representation in a lower dimensional space.

A method called t-Distributed Stochastic Neighbor Embedding (t-SNE) [109] has been widely used in the past years to perform DR. The t-SNE method can be seen as a neighbor graph based algorithm [110] in a sense that similarity scores based on Euclidean distances between neighbors are computed to embed the high dimensional structure in a two-dimensional space. Samples positions in the two-dimensional space are randomly initialized and are then moved iteratively so that the pair-wise samples similarities match the ones in the original space. t-SNE has limitations though. Firstly, the method can be computationally intensive when applied to huge datasets. Also, the interpretation of the t-SNE representation must be performed with caution. Indeed, the method retains local structures but has limited ability to maintain global structure [110].

Recently, a novel method called Uniform Manifold Approximation and Projection (UMAP) [110] was developed and is more and more replacing the t-SNE method. UMAP is based on topological theory. The algorithm builds what is called a simplicial complex which is a representation of the data as a weighted graph (See Figure 1.16), the weights corresponding to the likelihood that there is a connection between two points [111, 112].

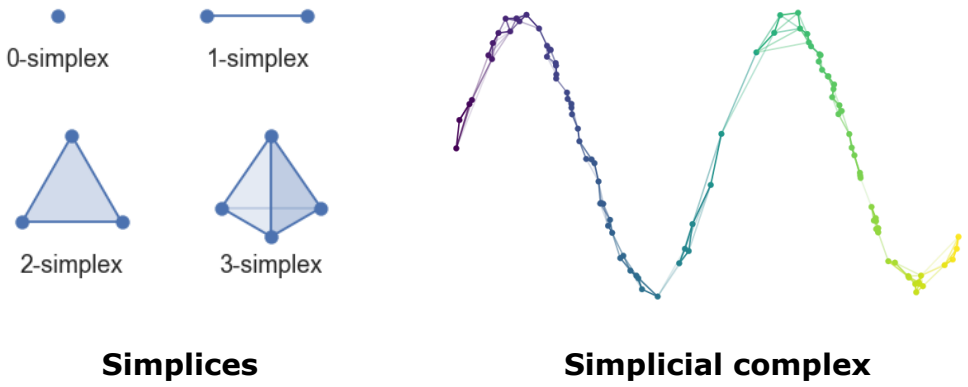

FIGURE 1.16: **UMAP topological representation**. A) The building blocks of a simplicial complex, the simplices. B) An example of a simplicial complex. Figures from [111].

As mentioned at the beginning of Section 1.4, in high dimensional spaces data sparsity increases. To connect all the points in the simplicial complex, UMAP varies the radius in which the search of neighbors is performed by fixing the number of

neighbors to consider around each point [111, 112]. This number of neighbors influences how the data structure is preserved, low and high values favoring local and global structures, respectively. Once the graphical representation of the high dimensional data is constructed, a low dimensional representation of the data is optimized so that it is as close as possible to the high dimensional representation. One of the advantages of UMAP over t-SNE is that the method better maintains the global structure of the data. Also, UMAP is computationally more efficient [110]. Note that UMAP can be applied on a lower dimensional dataset resulting, for example, from a DR method like PCA.

1.4.3 Multi-omics data integration

The methods previously described consider as input a single dataset. DR methods processing multiple matrices also exist and can be used to integrate multi-omics datasets. Such integration raises, though, multiple challenges. Firstly, the data to integrate are heterogeneous. The nature of the collected data is different, hence their statistical properties can vary. Also, it can happen that all the omics datasets are not available for each sample included in the analysis for technical reasons or due to quality issues. Hence, distinct patterns of missing data can occur in each omic dataset. Besides, integrating multiple datasets amplifies the curse of dimensionality issues already encountered in each dataset individually.

In 2018, a method called Multi-Omics Factor Analysis (MOFA) was developed to integrate multi-omics data while considering the previously mentioned challenges [113]. MOFA is an unsupervised analysis based on matrix factorization (See Section 1.4), and can be seen as an extension of PCA to multi-omics data, called modalities or also views. It is a factor analysis method which reduces the dimensions of the data to a smaller number of unobserved factors, called the latent factors. These factors differ from the PCA components. The latter are linear combinations of the initial features, while in factor analyses the initial features are expressed as linear combinations of the latent factors, plus a residual noise term. To enable multi-omics data (modalities) integration, MOFA supports different noise models depending on the nature of the data (continuous, counts or binary data). Based on this model, MOFA identifies different sources of variations across multiple omics data. MOFA presents though several limitations. The model does not capture non-linear relationships and assume features independence [113]. Also, additional features accounting for samples structure, such as groups of samples, batches or samples conditions, were not available in the initial version of MOFA but have been recently introduced in a second version, MOFA+ [114]. In this framework, the MOFA dimensionality reduction

is performed with regards to additional samples information (*e.g.* batch or cluster information) to identify sources of variations shared between groups or exclusive to one of them.

Other integrative methods can take into consideration samples structure. For example, the Partial Least Squares (PLS) method, which is a matrix factorization method, attempts to relate two matrices: a response matrix and a matrix gathering explanatory variables. The advantage of this method is that it ensures that the new variables resulting from the dimensionality reduction explain the response data. In that sense, the PLS method can be considered as a supervised DR framework. While PCA maximizes the variance of the components, PLS maximizes the covariance between the latent components of the response and explanatory datasets [115, 106]. When the response data is a categorical variable, a variant of PLS called PLS discriminant analysis (PLS-DA) can be used to perform classification tasks, *e.g.* samples groups prediction. In 2017, Lê Cao team published the mixOmics framework implementing multivariate analyses tools, including the PLS methods previously described [116]. The mixOmics tools also include the Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) method, which is a multivariate dimension reduction method that can be used for supervised multi-omics data integration [117]. DIABLO maximizes the correlation between the features of the different omics datasets, one of this dataset corresponding to the labelled samples. Hence, the method extracts what the authors call multi-omics signatures that are discriminant and can be used for prediction in a supervised framework.

Bibliography

[1] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 68.6 (Nov. 2018), pp. 394–424. ISSN: 1542-4863. DOI: 10.3322/caac.21492 (cit. on p. 7).

[2] Gilles R. Dagenais et al. "Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study". In: *The Lancet* 395.10226 (Mar. 2020), pp. 785–794. ISSN: 1474547X. DOI: 10.1016/S0140-6736(19)32007-0 (cit. on p. 7).

[3] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. "The cancer genome". In: *Nature* 458.7239 (Apr. 2009), pp. 719–724. ISSN: 00280836. DOI: 10.1038/nature07943 (cit. on pp. 7, 12–14).

[4] David S. Wishart. "Is Cancer a Genetic Disease or a Metabolic Disease?" In: *EBioMedicine* 2.6 (June 2015), pp. 478–479. ISSN: 23523964. DOI: 10.1016/j.ebiom.2015.05.022 (cit. on p. 7).

[5] Julia Eggert. "Biology of cancer". In: *Cancer Basics (Second Edition)*. Oncology Nursing Society, 2017, p. 816. ISBN: 9781935864929 (cit. on pp. 7, 13).

[6] Andreas Luch. "Nature and nurture - Lessons from chemical carcinogenesis". In: *Nature Reviews Cancer* 5.2 (Feb. 2005), pp. 113–125. ISSN: 1474175X. DOI: 10.1038/nrc1546 (cit. on p. 7).

[7] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation." In: *Cell* 144.5 (Mar. 2011), pp. 646–74. ISSN: 1097-4172. DOI: 10.1016/j.cell.2011.02.013 (cit. on p. 7).

[8] Oswald T. Avery, Colin M. Macleod, and Maclyn McCarty. "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III". In: *Journal of Experimental Medicine* 79.2 (Feb. 1944), pp. 137–158. ISSN: 15409538. DOI: 10.1084/jem.79.2.137 (cit. on p. 8).

[9] Thomas Hunt Morgan et al. *The Mechanism of Mendelian heredity*. New York : H. Holt and company, 1915 (cit. on p. 8).

[10] J. D. Watson and F. H.C. Crick. "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid". In: *Nature* 171.4356 (1953), pp. 737–738. ISSN: 00280836. DOI: 10.1038/171737a0 (cit. on p. 8).

[11] Robert A. Weinberg. *The biology of cancer (second edition)*. Ed. by Garland Science. 2014, p. 960. ISBN: 9780815345282 (cit. on pp. 10, 14, 15).

[12] Xiaotu Ma et al. "DNA methylation data analysis and its application to cancer research". In: *Epigenomics* 5.3 (2013), pp. 301–316. DOI: 10.2217/epi.13.26 (cit. on p. 10).

[13] Bert Vogelstein et al. "Cancer genome landscapes". In: *Science* 340.6127 (Mar. 2013), pp. 1546–1558. ISSN: 10959203. DOI: 10.1126/science.1235122 (cit. on p. 13).

[14] Zbyslaw Sondka et al. "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers". In: *Nature Reviews Cancer* 18.11 (Nov. 2018), pp. 696–705. ISSN: 14741768. DOI: 10.1038/s41568-018-0060-1 (cit. on p. 13).

[15] *Cancer Gene Census*. 1999. URL: https://cancer.sanger.ac.uk/census. Online. Accessed October 2020 (cit. on p. 13).

[16] E. Premkumar Reddy et al. "A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene". In: *Nature* 300.5888 (1982), pp. 149–152. ISSN: 00280836. DOI: 10.1038/300149a0 (cit. on p. 13).

[17] A. G. Knudson. "Mutation and cancer: statistical study of retinoblastoma." In: *Proceedings of the National Academy of Sciences of the United States of America* 68.4 (1971), pp. 820–823. ISSN: 00278424. DOI: 10.1073/pnas.68.4.820 (cit. on p. 13).

[18] Francisco Martínez-Jiménez et al. "A compendium of mutational cancer driver genes". In: *Nature Reviews Cancer* (Aug. 2020), pp. 1–18. ISSN: 14741768. DOI: 10.1038/s41568-020-0290-x (cit. on p. 13).

[19] L A Loeb. "Mutator phenotype may be required for multistage carcinogenesis." In: *Cancer research* 51.12 (June 1991), pp. 3075–9. ISSN: 0008-5472 (cit. on p. 14).

[20] Cristian Tomasetti and Bert Vogelstein. "Variation in cancer risk among tissues can be explained by the number of stem cell divisions". In: *Science* 347.6217 (Jan. 2015), pp. 78–81. ISSN: 10959203. DOI: 10.1126/science.1260825 (cit. on p. 14).

[21] Cristian Tomasetti, Lu Li, and Bert Vogelstein. "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention". In: *Science* (2017). DOI: doi:10.1126/science.aaf9011 (cit. on p. 14).

[22] *Lung Source: Globocan 2018 Number of new cases in 2018, both sexes, all ages*. 2018. URL: http://gco.iarc.fr/today. Online. Accessed October 2020 (cit. on p. 14).

[23] Julian Peto. "Cancer epidemiology in the last century and the next decade". In: *Nature* 411.6835 (May 2001), pp. 390–395. ISSN: 00280836. DOI: 10.1038/35077256 (cit. on p. 14).

[24] Richard Doll and A. Bradford Hill. "Smoking and carcinoma of the lung preliminary report". In: *British Medical Journal* 2.4682 (1950), pp. 739–748. ISSN: 00071447. DOI: 10.1136/bmj.2.4682.739 (cit. on p. 15).

[25] Lawrence A. Loeb and Curtis C. Harris. "Advances in chemical carcinogenesis: A historical review and prospective". In: *Cancer Research* 68.17 (Sept. 2008), pp. 6863–6872. ISSN: 00085472. DOI: 10.1158/0008-5472.CAN-08-2852 (cit. on p. 15).

[26] Helmut K. Seitz and Felix Stickel. "Acetaldehyde as an underestimated risk factor for cancer development: Role of genetics in ethanol metabolism". In: *Genes and Nutrition* 5.2 (June 2010), pp. 121–128. ISSN: 15558932. DOI: 10.1007/s12263-009-0154-1 (cit. on p. 15).

[27] Ludmil B. Alexandrov et al. "Signatures of mutational processes in human cancer". In: *Nature* 500.7463 (Aug. 2013), pp. 415–421. ISSN: 14764687. DOI: 10.1038/nature12477 (cit. on pp. 15, 22, 29).

[28] Serena Nik-Zainal et al. "The genome as a record of environmental exposure". In: *Mutagenesis* 30 (2015), pp. 763–770. DOI: 10.1093/mutage/gev073 (cit. on pp. 15, 29).

[29] Ludmil B. Alexandrov and Michael R. Stratton. "Mutational signatures: The patterns of somatic mutations hidden in cancer genomes". In: *Current Opinion in Genetics and Development* 24.1 (2014), pp. 52–60. ISSN: 0959437X. DOI: 10.1016/j.gde.2013.11.014 (cit. on p. 15).

[30] *COSMIC: Signatures of Mutational Processes in Human Cancer*. URL: https://cancer.sanger.ac.uk/cosmic/signatures_v2.tt. Online. Accessed October 2020 (cit. on p. 15).

[31] Ludmil B. Alexandrov et al. "The repertoire of mutational signatures in human cancer". In: *Nature* 578.7793 (Feb. 2020), pp. 94–101. ISSN: 14764687. DOI: 10.1038/s41586-020-1943-3 (cit. on pp. 16, 24, 29).

[32] Christopher Paul Wild. "Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology". In: *Cancer Epidemiology Biomarkers and Prevention* 14.8 (Aug. 2005), pp. 1847–1850. ISSN: 10559965. DOI: 10.1158/1055-9965.EPI-05-0456 (cit. on p. 16).

[33] Leslie Roberts. "Controversial From the Start". In: *Science* 291.5507 (Feb. 2001), pp. 1–1188. ISSN: 0036-8075. DOI: 10.1126/science.291.5507.1182a (cit. on p. 16).

[34] Thomas Laframboise. "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". In: *Nucleic Acids Research* 37.13 (2009), pp. 4181–4193. DOI: 10.1093/nar/gkp552 (cit. on p. 18).

[35] Chuanhua Xing et al. "Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to detect risk variants for human complex diseases". In: *European Journal of Human Genetics* 24.7 (July 2016), pp. 1029–1034. ISSN: 14765438. DOI: 10.1038/ejhg.2015.244 (cit. on p. 18).

[36] Rameen Beroukhim et al. "Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays". In: *PLoS Computational Biology* 2.5 (2006), pp. 323–332. ISSN: 15537358. DOI: 10.1371/journal.pcbi.0020041 (cit. on p. 18).

[37] Amit Dutt and Rameen Beroukhim. "Single nucleotide polymorphism array analysis of cancer". In: *Current Opinion in Oncology* 19.1 (2007), pp. 43–49. ISSN: 10408746. DOI: 10.1097/CCO.0b013e328011a8c1 (cit. on p. 18).

[38] Xueying Mao, Bryan D. Young, and Yong-Jie Lu. "The Application of Single Nucleotide Polymorphism Microarrays in Cancer Research". In: *Current Genomics* 8.4 (July 2007), pp. 219–228. ISSN: 13892029. DOI: 10.2174/138920207781386924 (cit. on p. 18).

[39] John W. Belmont et al. "The international HapMap project". In: *Nature* 426.6968 (Dec. 2003), pp. 789–796. ISSN: 00280836. DOI: 10.1038/nature02168 (cit. on p. 19).

[40] Kara Rogers. *International HapMap Project*. URL: https://www.britannica.com/event/International-HapMap-Project. Online. Accessed October 2020 (cit. on p. 19).

[41] Adi L. Tarca, Roberto Romero, and Sorin Draghici. "Analysis of microarray experiments of gene expression profiling". In: *American Journal of Obstetrics and Gynecology* 195.2 (Aug. 2006), pp. 373–388. ISSN: 00029378. DOI: 10.1016/j.ajog.2006.07.001 (cit. on p. 19).

[42] Interrogate single CpG sites. *Infinium Methylation Assay*. URL: https://emea.illumina.com/science/technology/microarray/infinium-methylation-assay.html. Online. Accessed October 2020 (cit. on pp. 19, 20).

[43] KA Wetterstrand. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. Online. Accessed October 2020 (cit. on pp. 21, 23).

[44] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: A revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1 (Jan. 2009), pp. 57–63. ISSN: 14710056. DOI: 10.1038/nrg2484 (cit. on p. 22).

[45] Terrence S. Furey. "ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions". In: *Nature Reviews Genetics* 13.12 (Dec. 2012), pp. 840–852. ISSN: 14710056. DOI: 10.1038/nrg3306 (cit. on p. 22).

[46] Feng Yan et al. "From reads to insight: A hitchhiker's guide to ATAC-seq data analysis". In: *Genome Biology* 21.1 (Feb. 2020), pp. 1–16. ISSN: 1474760X. DOI: 10.1186/s13059-020-1929-3 (cit. on p. 22).

[47] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental and Molecular Medicine* 50.8 (Aug. 2018), p. 96. ISSN: 20926413. DOI: 10.1038/s12276-018-0071-8 (cit. on p. 23).

[48] Francesca Finotello et al. "Next-generation computational tools for interrogating cancer immunity". In: *Nature Reviews Genetics* 20.12 (Dec. 2019), pp. 724–746. ISSN: 14710064. DOI: 10.1038/s41576-019-0166-7 (cit. on p. 23).

[49] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge". In: *Wspolczesna Onkologia, Contemporary oncology* 19.1A (2015), A68–A77. ISSN: 14282526. DOI: 10.5114/wo.2014.47136 (cit. on p. 23).

[50] John N. Weinstein et al. "The cancer genome atlas pan-cancer analysis project". In: *Nature Genetics* 45.10 (Oct. 2013), pp. 1113–1120. ISSN: 15461718. DOI: 10.1038/ng.2764 (cit. on pp. 24, 25).

[51] NCI Genomic Data Commons. *PanCanAtlas Publications*. URL: https://gdc.cancer.gov/about-data/publications/pancanatlas. Online. Accessed October 2020 (cit. on p. 24).

[52] Thomas J. Hudson et al. "International network of cancer genome projects". In: *Nature* 464.7291 (Apr. 2010), pp. 993–998. ISSN: 00280836. DOI: 10.1038/nature08987 (cit. on pp. 24, 25).

[53] Marcin Cieslik and Arul M. Chinnaiyan. "Global genomics project unravels cancer's complexity at unprecedented scale". In: *Nature* 578.7793 (Feb. 2020), pp. 39–40. ISSN: 14764687. DOI: 10.1038/d41586-020-00213-2 (cit. on p. 24).

[54] Peter J. Campbell et al. "Pan-cancer analysis of whole genomes". In: *Nature* 578.7793 (Feb. 2020), pp. 82–93. ISSN: 14764687. DOI: 10.1038/s41586-020-1969-6 (cit. on p. 24).

[55] Yilong Li et al. "Patterns of somatic structural variation in human cancer genomes". In: *Nature* 578.7793 (Feb. 2020), pp. 112–121. ISSN: 14764687. DOI: 10.1038/s41586-019-1913-9 (cit. on p. 24).

[56] Orli G. Bahcall. "UK Biobank — a new era in genomic medicine". In: *Nature Reviews Genetics* 19.12 (Dec. 2018), p. 737. ISSN: 14710064. DOI: 10.1038/s41576-018-0065-3 (cit. on p. 25).

[57] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (Oct. 2018), pp. 203–209. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0579-z (cit. on p. 25).

[58] Annalisa Buniello et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D1005–D1012. ISSN: 13624962. DOI: 10.1093/nar/gky1120 (cit. on p. 25).

[59] *TCGA Timeline Milestones was originally published by the National Cancer Institute.* URL: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/timeline. Online. Accessed October 2020 (cit. on p. 26).

[60] Mark A. Jensen et al. "The NCI Genomic Data Commons as an engine for precision medicine". In: *Blood* 130.4 (July 2017), pp. 453–459. ISSN: 15280020. DOI: 10.1182/blood-2017-03-735654 (cit. on p. 26).

[61] Galen F. Gao et al. "Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data". In: *Cell Systems* 9.1 (July 2019), pp. 24–34. ISSN: 2405-4712. DOI: 10.1016/J.CELS.2019.06.006 (cit. on p. 26).

[62] NCI Genomic Data Commons. *PanCanAtlas Publications*. URL: https://gdc.cancer.gov/about-data/publications/pancanatlas. Online. Accessed October 2020 (cit. on p. 26).

[63] Vésteinn Thorsson et al. "The Immune Landscape of Cancer." In: *Immunity* 48.4 (Apr. 2018), pp. 812–830. ISSN: 1097-4180. DOI: 10.1016/j.immuni.2018.03.023 (cit. on p. 26).

[64] Theo A. Knijnenburg et al. "Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas". In: *Cell Reports* 23.1 (2018), pp. 239–254. ISSN: 22111247. DOI: 10.1016/j.celrep.2018.03.076 (cit. on p. 26).

[65] Jessica W. Lau et al. "The cancer genomics cloud: Collaborative, reproducible, and democratized - A new paradigm in large-scale computational research". In: *Cancer Research* 77.21 (Nov. 2017), e3–e6. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-17-0387 (cit. on p. 26).

[66] Sheila M. Reynolds et al. "The ISB cancer genomics cloud: A flexible cloud-based platform for cancer genomics research". In: *Cancer Research* 77.21 (Nov. 2017), e7–e10. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-17-0617 (cit. on p. 26).

[67] Sergei Yakneen et al. "Butler enables rapid cloud-based analysis of thousands of human genomes". In: *Nature Biotechnology* 38.3 (Mar. 2020), pp. 288–292. ISSN: 15461696. DOI: 10.1038/s41587-019-0360-3 (cit. on p. 26).

[68] W.D. Travis. "Advances in neuroendocrine lung tumors". In: *Annals of Oncology* 21 (Oct. 2010), pp. vii65–vii71. ISSN: 09237534. DOI: 10.1093/annonc/mdq380 (cit. on p. 27).

[69] Jules L. Derks et al. "New Insights into the Molecular Characteristics of Pulmonary Carcinoids and Large Cell Neuroendocrine Carcinomas, and the Impact on Their Clinical Management". In: *Journal of Thoracic Oncology* 13.6 (June 2018), pp. 752–766. ISSN: 15561380. DOI: 10.1016/j.jtho.2018.02.002 (cit. on pp. 27, 28, 30).

[70] Michele Simbolo et al. "Exploring the molecular and biological background of lung neuroendocrine tumours." In: *Journal of thoracic disease* 11.Suppl 9 (May 2019), S1194–S1198. ISSN: 2072-1439. DOI: 10.21037/jtd.2019.03.66 (cit. on p. 27).

[71] Cancer.Net. *Lung Cancer - Non-Small Cell: Statistics*. URL: https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics. Online. Accessed October 2020 (cit. on p. 27).

[72] Katerina Politi and Roy S. Herbst. "Lung cancer in the era of precision medicine". In: *Clinical Cancer Research* 21.10 (May 2015), pp. 2213–2220. ISSN: 15573265. DOI: 10.1158/1078-0432.CCR-14-2748 (cit. on pp. 27, 29).

[73] William D Travis et al. "The 2015 World Health Organization Classification of Lung Tumors". In: *Journal of Thoracic Oncology* 10 (2015), pp. 1243–1260. DOI: 10.1097/JTO.0000000000000630 (cit. on p. 27).

[74] Inigo Martincorena and Peter J Campbell. "Somatic mutation in cancer and normal cells (Erratum)". In: *Science* 351.6277 (Mar. 2016), aaf5401–aaf5401. ISSN: 0036-8075. DOI: 10.1126/science.aaf5401 (cit. on p. 28).

[75] Viviane Teixeira Loiola de Alencar, Maria Nirvana Formiga, and Vladmir Cláudio Cordeiro de Lima. "Inherited lung cancer: A review". In: *ecancermedicalscience* 14 (Jan. 2020). ISSN: 17546605. DOI: 10.3332/ECANCER.2020.1008 (cit. on p. 28).

[76] C. I. Amos, W. Xu, and M. R. Spitz. "Is There a Genetic Basis for Lung Cancer Susceptibility?" In: *Chemoprevention of Cancer*. Vol. 151. 1999, pp. 3–12. DOI: 10.1007/978-3-642-59945-3{_}1 (cit. on p. 28).

[77] Yohan Bosse and Christopher I. Amos. "A decade of GWAS results in lung cancer". In: *Cancer Epidemiology Biomarkers and Prevention* 27.4 (Apr. 2018), pp. 363–379. ISSN: 10559965. DOI: 10.1158/1055-9965.EPI-16-0794 (cit. on p. 28).

[78] Thorgeir E. Thorgeirsson et al. "A variant associated with nicotine dependence, lung cancer and peripheral arterial disease". In: *Nature* 452.7187 (Apr. 2008), pp. 638–642. ISSN: 14764687. DOI: 10.1038/nature06846 (cit. on p. 28).

[79] James D McKay et al. "Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes". In: *Nature Genetics* 49.7 (July 2017), pp. 1126–1132. ISSN: 1061-4036. DOI: 10.1038/ng.3892 (cit. on p. 28).

[80] Michael S. Lawrence et al. "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457 (2013), pp. 214–218. ISSN: 00280836. DOI: 10.1038/nature12213 (cit. on p. 29).

[81] Kenichi Yoshida et al. "Tobacco smoking and somatic mutations in human bronchial epithelium". In: *Nature* 578.7794 (Feb. 2020), pp. 266–272. ISSN: 14764687. DOI: 10.1038/s41586-020-1961-1 (cit. on p. 29).

[82] Eric A. Collisson et al. "Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network". In: *Nature* 511.7511 (July 2014), pp. 543–550. ISSN: 14764687. DOI: 10.1038/nature13385 (cit. on p. 29).

[83] Neal I. Lindeman et al. "Updated Molecular Testing Guideline for the Selection of Lung Cancer Patients for Treatment With Targeted Tyrosine Kinase Inhibitors: Guideline From the College of American Pathologists, the International Association for the Study of Lung Cancer, and the ". In: *Journal of Thoracic Oncology* 13.3 (Mar. 2018), pp. 323–358. ISSN: 15561380. DOI: 10.1016/j.jtho.2017.12.001 (cit. on p. 29).

[84] The Cancer Genome Atlas Research Network. "Comprehensive genomic characterization of squamous cell lung cancers". In: *Nature* 489.7417 (Sept. 2012), pp. 519–525. ISSN: 0028-0836. DOI: 10.1038/nature11404 (cit. on p. 29).

[85] Martin Peifer et al. "Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer". In: *Nature Genetics* 44.10 (Oct. 2012), pp. 1104–1110. ISSN: 1061-4036. DOI: 10.1038/ng.2396 (cit. on p. 29).

[86] Julie George et al. "Comprehensive genomic profiles of small cell lung cancer". In: *Nature* 524.7563 (Aug. 2015), pp. 47–53. ISSN: 0028-0836. DOI: 10.1038/nature14664 (cit. on p. 29).

[87] Lynnette Fernandez-Cuesta and Matthieu Foll. "Molecular studies of lung neuroendocrine neoplasms uncover new concepts and entities". In: *Translational Lung Cancer Research* 8.S4 (Dec. 2019), S430–S434. ISSN: 22186751. DOI: 10.21037/tlcr.2019.11.08 (cit. on pp. 29, 30).

[88] Lynnette Fernandez-Cuesta et al. "Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer". In: *EBioMedicine* 10 (Aug. 2016), pp. 117–123. ISSN: 23523964. DOI: 10.1016/j.ebiom.2016.06.032 (cit. on p. 29).

[89] Joshua D. Cohen et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test". In: *Science* 359.6378 (Feb. 2018), pp. 926–930. ISSN: 10959203. DOI: 10.1126/science.aar3247 (cit. on p. 30).

[90] Tiffany M Delhomme et al. "Needlestack: an ultra-sensitive variant caller for multi-sample next generation sequencing data". In: *NAR Genomics and Bioinformatics* 2.2 (June 2020). ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa021 (cit. on p. 30).

[91] Lynnette Fernandez-Cuesta et al. "Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids." In: *Nature communications* 5 (Mar. 2014), p. 3518. ISSN: 2041-1723. DOI: 10.1038/ncomms4518 (cit. on p. 30).

[92] Julie George et al. "Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors". In: *Nature Communications* 9.1 (Dec. 2018), p. 1048. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03099-x (cit. on p. 30).

[93] Natasha Rekhtman et al. "Next-Generation Sequencing of Pulmonary Large Cell Neuroendocrine Carcinoma Reveals Small Cell Carcinoma–like and Non–Small Cell Carcinoma–like Subsets". In: *Clinical Cancer Research* 22.14 (July 2016), pp. 3618–3629. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-15-2946 (cit. on p. 30).

[94] Michele Simbolo et al. "Gene Expression Profiling of Lung Atypical Carcinoids and Large Cell Neuroendocrine Carcinomas Identifies Three Transcriptomic Subtypes with Specific Genomic Alterations". In: *Journal of Thoracic Oncology* 14.9 (Sept. 2019), pp. 1651–1661. ISSN: 15560864. DOI: 10.1016/j.jtho.2019.05.003 (cit. on p. 30).

[95] Charles M. Rudin et al. "Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data". In: *Nature Reviews Cancer* 19.5 (May 2019), pp. 289–297. ISSN: 14741768. DOI: 10.1038/s41568-019-0133-9 (cit. on p. 30).

[96] Naomi Altman and Martin Krzywinski. "The curse(s) of dimensionality". In: *Nature Methods* 15.6 (June 2018), pp. 399–400. ISSN: 1548-7091. DOI: 10.1038/s41592-018-0019-x (cit. on pp. 30, 31).

[97] Pedro Domingos. "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10 (2012). DOI: 10.1145/2347736.2347755 (cit. on pp. 31, 35).

[98] Tom Ronan, Zhijie Qi, and Kristen M. Naegle. "Avoiding common pitfalls when clustering biological data". In: *Science Signaling* 9.432 (June 2016), re6–re6. ISSN: 19379145. DOI: 10.1126/scisignal.aad1932 (cit. on p. 31).

[99] Susan Holmes and Wolfgang Huber. *Modern Statistics for Modern Biology*. Ed. by Cambridge University Press. 2019, p. 402 (cit. on p. 32).

[100] Tom M. Mitchell. *Machine learning*. McGraw-Hill Science/Engineering/Math, 1997, p. 432 (cit. on p. 32).

[101] Maxwell W. Libbrecht and William Stafford Noble. "Machine learning applications in genetics and genomics". In: *Nature Reviews Genetics* 16.6 (June 2015), pp. 321–332. ISSN: 1471-0056. DOI: 10.1038/nrg3920 (cit. on p. 33).

[102] Danielle Denisko and Michael M. Hoffman. "Classification and interaction in random forests". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.8 (Feb. 2018), pp. 1690–1692. ISSN: 10916490. DOI: 10.1073/pnas.1800256115 (cit. on p. 34).

[103] Bradley Boehmke and Greenwell Brandon. *Hands-On Machine Learning with R*. URL: https://bradleyboehmke.github.io/HOML/process.html. Online. Accessed October 2020 (cit. on p. 36).

[104] Nikolay Oskolkov. *Unsupervised OMICs Integration*. 2019. URL: https://towardsdatascience.com/unsupervised-omics-integration-688bf8fa49bf. Online. Accessed October 2020 (cit. on p. 36).

[105] Gökcen Eraslan et al. "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* (Apr. 2019), p. 1. ISSN: 1471-0056. DOI: 10.1038/s41576-019-0122-6 (cit. on p. 36).

[106] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Ed. by Springer. 2017, p. 745 (cit. on pp. 37, 41).

[107] Genevieve L Stein-O'brien et al. "Enter the Matrix: Factorization Uncovers Knowledge from Omics Determining the Dimensions of Biology from Omics Data". In: *Trends in Genetics* 34 (2018), pp. 790–805. DOI: 10.1016/j.tig.2018.07.003 (cit. on p. 38).

[108] Ludmil B. Alexandrov et al. "Deciphering Signatures of Mutational Processes Operative in Human Cancer". In: *Cell Reports* 3.1 (Jan. 2013), pp. 246–259. ISSN: 22111247. DOI: 10.1016/j.celrep.2012.12.008 (cit. on p. 38).

[109] Laurens Van Der Maaten and Geoffrey Hinton. *Visualizing Data using t-SNE*. Tech. rep. 2008, pp. 2579–2605 (cit. on p. 39).

[110] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *ArXiv* (Feb. 2018) (cit. on pp. 39, 40).

[111] Leland McInnes. *How UMAP Works — umap 0.4 documentation*. 2018. URL: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html. Online. Accessed October 2020 (cit. on pp. 39, 40).

[112] Andy Coenen and Adam Pearce. *Understanding UMAP*. URL: https://pair-code.github.io/understanding-umap/. Online. Accessed October 2020 (cit. on pp. 39, 40).

[113] Ricard Argelaguet et al. "Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets." In: *Molecular systems biology* 14.6 (June 2018), e8124. ISSN: 1744-4292. DOI: 10.15252/msb.20178124 (cit. on p. 40).

[114] Ricard Argelaguet et al. "MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome Biology* 21.1 (May 2020), p. 111. ISSN: 1474760X. DOI: 10.1186/s13059-020-02015-1 (cit. on p. 40).

[115] Kim-Ahn Lê Cao. *Webinar mixOmics: PLS methods*. 2020 (cit. on p. 41).

[116] Florian Rohart et al. "mixOmics: An R package for 'omics feature selection and multiple data integration". In: *PLOS Computational Biology* 13.11 (Nov. 2017). Ed. by Dina Schneidman, e1005752. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005752 (cit. on p. 41).

[117] Amrit Singh et al. "DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays". In: *Bioinformatics* 35.17 (2019), pp. 3055–3062. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty1054 (cit. on p. 41).

Acronyms

ATAC-seq Assay of Transposase Accessible Chromatin sequencing. 22

cDNA complementary DNA. 19

CGC Cancer Genomics Cloud. 26

ChiP-Seq Chromatin immunoprecipitation Sequencing. 22

CML chronic myelogenous leukemia. 12

COSMIC Catalogue Of Somatic Mutations In Cancer. 29

CpG cytosine–phosphate–guanine. 10, 19, 22

ctDNA circulating tumor DNA. 29

dbGAP Database of Genotypes And Phenotypes. 25

DIABLO Data Integration Analysis for Biomarker discovery using Latent cOmponents. 41

DNA Deoxyribonucleic acid. 8–17, 20, 22, 29, 30

DR dimensionality reduction. 37, 39–41

EDA exploratory data analyses. 32

EGA European-Genome Phenome Archive. 25

EGFR Epidermal Growth Factor Receptor. 29

GDC Genomic Data Common. 26

GEO Gene Expression Omnibus. 25

GWAS Genome-Wide Association Studies. 18, 25, 28

HapMap Haplotype Map project. 19

HGP Human Genome Project. 16

IARC International Agency for Research on Cancer. 7

ICA Independent Component Analysis. 38

ICGC International Cancer Genome Consortium. 24–26

indels insertions or deletions. 12, 16

ISB-CGC ISB Cancer Genomics Cloud. 26

LCNEC Large Cell Neuroendocrine Carcinoma. 28, 30

LD linkage disequilibrium. 19

LOH loss of heterozygosity. 13, 18

LUAD Lung Adenocarcinomas. 27–29

LUSC Lung Squamous Cell Carcinomas. 27–29

miRNAs micro RNAs. 10, 23

ML machine learning. 32, 34

MOFA Multi-Omics Factor Analysis. 40

MRI Magnetic Resonance Imaging. 25

mRNAs messenger RNAs. 10, 17, 19, 22

NCI National Cancer Institute. 26

NGS Next Generation Sequencing. 20, 21, 23, 28

NIH National Institutes of Health. 23, 26

NMF Non-negative Matrix factorization. 38

NSCLC Non Small Cell Lung Cancer. 27, 29

PCA Principal Component Analysis. 32, 38, 40, 41

PCAWG Pan-Cancer Analysis of Whole Genomes. 24, 26

PCR Polymerase Chain Reaction. 21