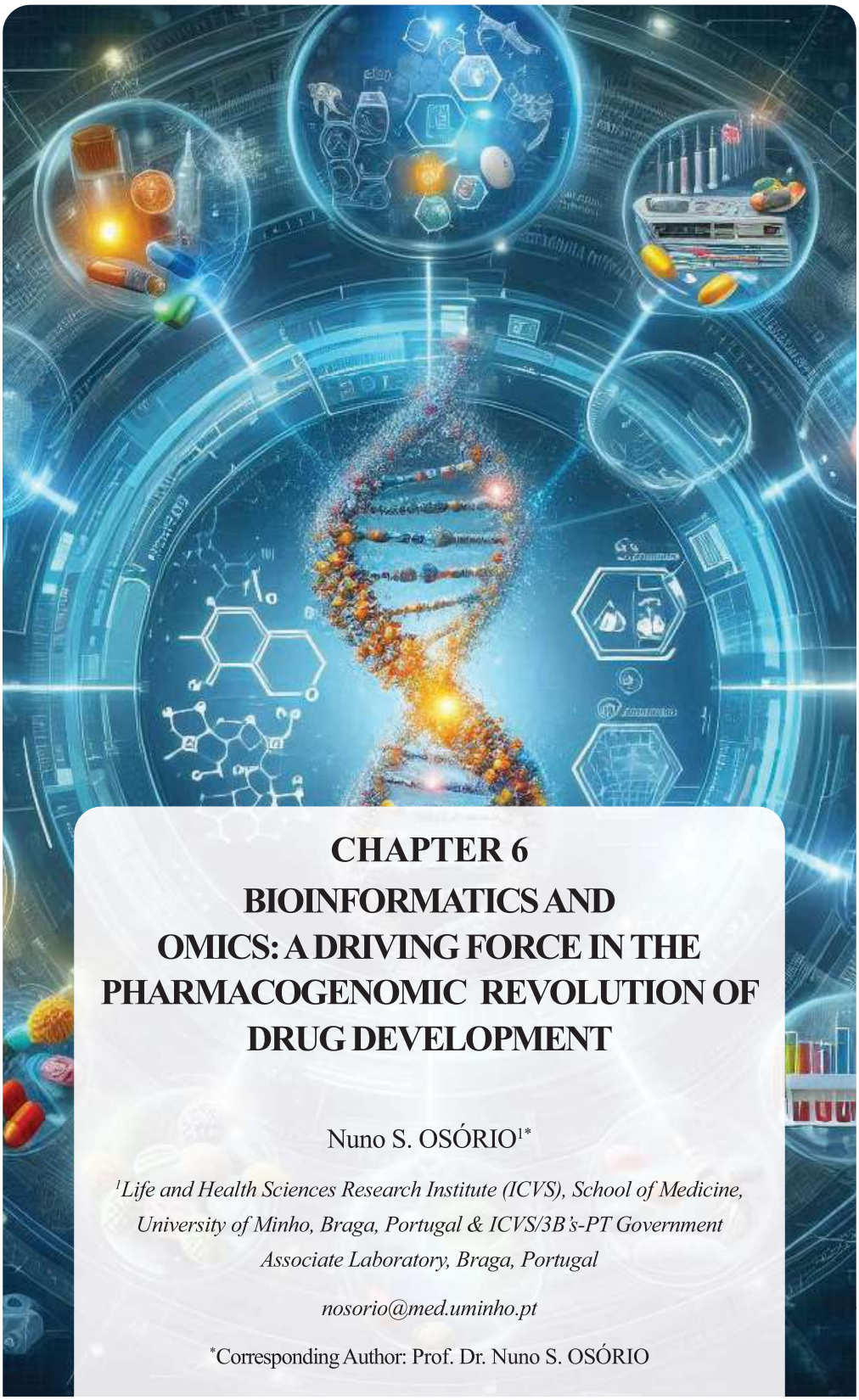

CHAPTER 6
BIOINFORMATICS AND OMICS: A DRIVING FORCE IN THE PHARMACOGENOMIC REVOLUTION OF DRUG DEVELOPMENT

Nuno S. OSÓRIO[1*]

[1]*Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal & ICVS/3B's-PT Government Associate Laboratory, Braga, Portugal*

nosorio@med.uminho.pt

[*]Corresponding Author: Prof. Dr. Nuno S. OSÓRIO


GENETIC

For video Scan the QR code.

Omics, Bioinformatics, and Pharmacogenomics Practical Training

1. INTRODUCTION

Bioinformatics and omics are interdisciplinary fields that apply large scale biological data and computational tools to several applications including the discovery and development of drugs. Drug development is a complex, costly and time-consuming process that involves several stages, such as target identification, lead optimization, preclinical testing, clinical trials and regulatory approval. This article introduces the main concepts and methods of bioinformatics and omics for drug development, and illustrates their applications and challenges. One of the key aspects of bioinformatics and omics for drug development is pharmacogenomics, which is the field concerned with understanding how genetic differences among individuals cause varied responses to the same drug and with developing drug therapies to compensate for these differences (Pirmohamed, 2001). Pharmacogenomics can help to personalize drug therapy, by identifying the genetic and environmental factors that influence drug response and adverse effects, and by selecting the optimal drug and dose for each patient (Pirmohamed, 2001). Pharmacogenomics can also help to discover new drug targets and mechanisms, by revealing the genetic variations that affect the function and expression of proteins involved in disease pathogenesis and drug action (Pirmohamed, 2001). Pharmacogenomics is a rapidly evolving field that relies on the integration of genomic, transcriptomic, proteomic, metabolomic and other omics data, as well as bioinformatic and computational tools, to achieve its goals.

According to a recent study, the average cost of developing a new drug is about $2.6 billion and the average time is about 10 years (DiMasi et al., 2016). Moreover, the success rate of drug development is very low, with only about 10% of the candidates that enter clinical trials reaching the market (Wong et al., 2019). Therefore, there is a great need for improving the efficiency and effectiveness of drug development.

A significant hurdle in drug development is the limited understanding of the molecular mechanisms underlying diseases, the interactions between drugs and their targets, and the diverse responses of individuals to the same drug treatment. Traditionally, drug discovery has relied on phenotypic screening, which is a trial-and-error method that tests the effects of compounds on disease models without knowing their targets or mechanisms of action. However, this approach is often limited by the availability and relevance of the models, the complexity and variability of the phenotypes, and the difficulty of identifying the targets and mechanisms of the active compounds (Swinney, 2013).

To overcome these limitations, a new paradigm of drug development has emerged, which is based on the concept of pharmacogenomics and relies on the use of bioinformatics and Omics data and technologies. Omics allow the comprehensive and systematic analysis of various biological molecules, such as genes, transcripts, proteins and metabolites, in a given biological system. Bioinformatics confer the computational methods and tools that enable the storage, processing, analysis and integration of omics data. The developments in these fields fuel the goals of pharmacogenomics, promoting the understanding on how genetic differences among individuals cause varied responses to the same drug, aiming to develop drug therapies to compensate for these differences.

The use of omics in biomedical research has seen a significant increase in recent years (Figure 1).

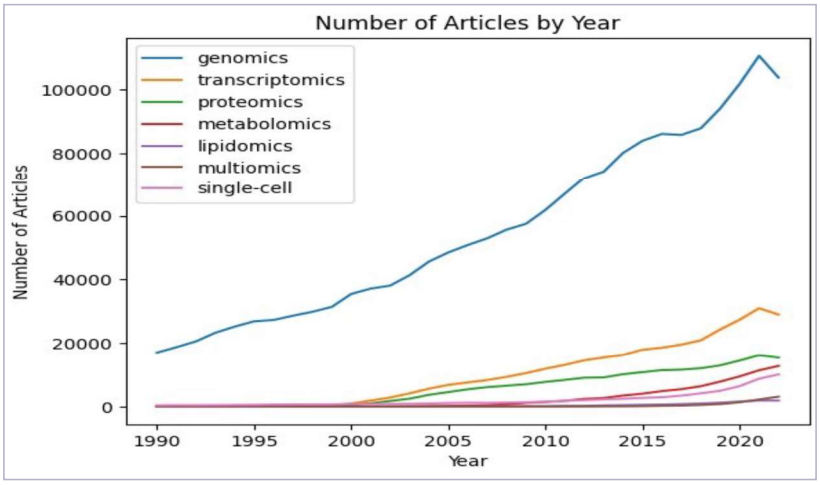

Figure 1. Yearly publication trends in various omics fields from 1990 to 2022. The data was collected from the PubMed database using the Bio.Entrez and Bio.Medline modules from the Biopython library. The search was performed using the keywords "genomics", "transcriptomics", "proteomics", "metabolomics", "lipidomics", "multiomics",and"single-cell". The number of articles published each year in each field was plotted using the matplotlib library. An interactive version of the figure is available here: https://nunososorio.github.io/omics/

Bioinformatics and omics have revolutionized drug discovery and development in several ways. First, they can help to identify novel and validated targets for drug development, by revealing the key genes, pathways and networks that are involved in disease pathogenesis and drug response. Second, they can help to discover and optimize new lead compounds, by screening large libraries of chemical or biological molecules against omics data and predicting their properties and activities. Third, they can help repurposing existing drugs for new indications, by finding new uses and targets for drugs that have already been approved or tested for other diseases. Fourth, they can help to design and conduct preclinical and clinical studies, by selecting the most appropriate models, biomarkers and endpoints for drug testing and evaluation. Fifth, they can help to personalize drug therapy, by identifying the genetic and environmental factors that influence drug response and adverse effects in different individuals and populations (Loging, 2016).

2. APPLICATIONS OF OMICS FOR DRUG DEVELOPMENT

The study of biological systems involves various omics fields, including genomics, metagenomics, transcriptomics, proteomics, metabolomics, lipidomics, and others. Each of these fields focuses on a different aspect of the system: genomics on DNA sequences, metagenomics on the genetic material of entire communities of organisms, transcriptomics on RNA sequences, proteomics on proteins, metabolomics on metabolites, and lipidomics on lipids. These diverse fields provide a comprehensive view of the biological system from various angles, allowing for a more complete understanding of its structure, function, and dynamics.

2.1. Genomics

Genomics is a specialized branch of biology that focuses on the comprehensive study of an organism's genome. The genome is the complete set of genes or genetic material present in a cell or organism. Genomics involves the sequencing and analysis of genomes through uses of high-throughput DNA sequencing and bioinformatics to assemble and analyze the function and structure of entire genomes.

The primary aim of genomics is to understand the complex genomic architecture of life by mapping out the DNA sequences of organisms. This includes understanding their structure, function, and evolutionary history. Genomics also seeks to reveal the genetic basis of different traits and diseases. This is achieved by identifying potential targets and biomarkers for therapeutics.

One of the significant challenges in genomics is the handling of large-scale and complex data from genomic samples. These data can include whole-genome sequencing, exome sequencing, and different dimensions of data necessary for genome-wide association studies. Each of these techniques provides a different perspective on the genome and can reveal different types of genetic variation.

To make sense of this vast amount of data, researchers use various techniques and tools. These include alignment (matching DNA sequences that are similar), assembly(putting together sequenced DNA to recreate the original genome), annotation (identifying the locations of genes and all of the coding regions in a genome and determining what those genes do), variant calling (identifying differences in a genomic sequence compared to a reference sequence), filtering (removing low quality or irrelevant data), and prioritization (determining which variants are most likely to have an impact on the organism).

These techniques and tools help to identify and characterize the genomic variants, such as single nucleotide polymorphisms (SNPs), insertions, deletions, copy number variations, and structural variations. These variants can then be studied to infer their functional impact and clinical significance.

In the context of drug development, genomics plays a crucial role. It enables the discovery of new genes and their products, provides insights into genetic mechanisms and disease pathways, and facilitates the study of pharmacogenomics (Table 1).

Table 1. Applications of genomics for drug development.

Application	Description
Discovering new genes and their products	Genomics can isolate and characterize the novel genes and their products from various organisms, such as bacteria, fungi, plants, animals, and evaluate their biological activities and properties, and synthesize and optimize their products for drug discovery and development (Challis, 2008; van der Lee et al., 2016).
Understanding the genetic and molecular mechanisms and pathways of diseases	Genomics can profile and compare the genomes of healthy and diseased individuals, and reveal how they are associated with various diseases, such as cancer, cardiovascular diseases, neurological diseases, and discover new drug targets and biomarkers (Yang et al., 2013).
Studying the pharmacogenomics and personalized medicine	Genomics can investigate how the genomic variation affects the response and adverse effects of drugs, such as their absorption, distribution, metabolism, excretion, and toxicity, and design and evaluate the efficacy and safety of drugs and dosages for individual patients (Chang et al., 2021).

2.2. Metagenomics

Metagenomics is a branch of genomics that aims to measure and analyze the genomic profiles of microbial communities, such as their composition, diversity, function, interaction, evolution. Metagenomics can reveal the role of microbial communities in health and disease, and identify potential targets and biomarkers for therapeutics.

One of the specific challenges of metagenomics is to obtain and process the high-throughput and complex data from microbial samples, such as soil, water, air, food, or human body, and to distinguish and characterize the individual microbes and their genes, proteins, and metabolites. To address this challenge, various techniques and tools have been developed, such as sequencing, assembly, annotation, classification, comparison. These techniques and tools can help to identify and quantify the microbial taxa and their functional genes, proteins, and metabolites, and infer the phylogenetic and metabolic relationships and networks among them.

In the realm of drug development, metagenomics plays a pivotal role by unearthing novel microbes and their derivatives, elucidating the influence of the microbiome on health and disease, and scrutinizing the evolution and propagation of microbial pathogens along with their resistance mechanisms (Table 2).

Table 2. Applications of metagenomics for drug development.

Application	Description
Discovering new microbes and their products	Metagenomics can isolate and characterize the novel microbes and their genes, proteins, and metabolites from various environmental or clinical samples, and evaluate their biological activities and properties, and synthesize and optimize their products for drug discovery and development (Jethwa et al., 2023).
Understanding the microbiome and its impact on human health and disease	Metagenomics can profile and compare the microbiome of healthy and diseased individuals, such as their gut, skin, oral, and reveal how they are associated with various diseases, such as obesity, diabetes, cancer, inflammatory bowel disease, and modulate their microbiome for disease prevention and treatment (Jorth et al., 2014).
Studying the evolution and transmission of microbial pathogens and their resistance	Metagenomics can track and monitor the emergence and dissemination of microbial pathogens and their resistance genes, such as bacteria, viruses, fungi, and their hosts, vectors, and reservoirs, and design and evaluate the efficacy and safety of drugs and vaccines for infection management and eradication (Datta et al., 2020).

2.3. Transcriptomics

Transcriptomics is a branch of genomics that aims to measure and analyze the transcriptome (the set of all RNA molecules) of biological samples, such as cells, tissues, organs. Transcriptomics can reveal the gene expression patterns and regulatory networks of cells, and how they are influenced by various factors, such as development, differentiation, disease, drug response.

One of the specific challenges of transcriptomics is to obtain and process the high-throughput and complex data from RNA samples, such as mRNA, miRNA, lncRNA, and to distinguish and characterize the individual transcripts and their isoforms, functions, interactions. To address this challenge, various techniques and tools have been developed, such as sequencing, alignment, quantification, annotation, differential expression, alternative splicing. These techniques and tools can help to identify and quantify the transcripts and their isoforms, and infer their functions and interactions with other molecules, such as DNA, proteins, and metabolites.

Transcriptomics aids drug development by identifying new drug targets and biomarkers associated with diseases, understanding the effects of drugs on cellular processes, and developing personalized medicine strategies based on gene expression and transcriptome signatures (Table 3).

Table 3. Applications of transcriptomics for drug development.

Application	Description
Discovering new drug targets and biomarkers	Transcriptomics can identify the genes and transcripts that are associated with diseases, such as cancer, cardiovascular diseases, neurological diseases, and evaluate their potential as drug targets and biomarkers for diagnosis, prognosis, and treatment (Pedrotty et al., 2012; Kaczkowski et al., 2016).
Understanding the mechanisms of action and the effects of drugs on cells	Transcriptomics can monitor the changes in gene expression and transcriptome profiles in response to drug treatment, and reveal how drugs affect the cellular processes and pathways, such as cell cycle, apoptosis, signal transduction (Cui et al., 2010).
Developingnew therapeutic strategies and personalized medicine	Transcriptomics can classify the patients and diseases based on their gene expression and transcriptome signatures, and predict their response and resistance to drugs, and optimize the drug dosage and combination for individualized treatment (Liu et al., 2022).

2.4. Proteomics

Proteomics is a branch of omics that aims to measure and analyze the proteome (the set of all proteins) of biological samples, such as their expression, modification, interaction, function, structure. Proteomics can reveal the molecular mechanisms, pathways, and networks underlying biological phenomena, such as development, differentiation, disease, and drug response.

One of the specific challenges of proteomics is to obtain and process the high-throughput and complex data from protein samples, such as cell lysates, tissue extracts, or body fluids, and to identify and quantify thousands of proteins and their post-translational modifications, such as phosphorylation, acetylation, ubiquitination. To address this challenge, various techniques and tools have been developed, such as mass spectrometry (MS), liquid chromatography (LC), gel electrophoresis (GE), protein microarrays. These techniques and tools can help to separate, detect, identify, and quantify the proteins and their modifications, and infer their interactions and functions.

Proteomics contributes to drug development by identifying novel drug targets and biomarkers, elucidating the drug's mechanisms of action and its impact on cells, tissues, and organs, and fostering the development of innovative therapeutic strategies and precision medicine (Table 4).

Table 4. Applications of proteomics for drug development.

Application	Description
Discovering new drug targets and biomarkers	Proteomics can identify and characterize the proteins and their modifications that are involved in or affected by diseases, such as cancer, neurodegenerative diseases, cardiovascular diseases, and evaluate their potential as drug targets and biomarkers for diagnosis, prognosis, and treatment (Lee et al., 2011).
Understanding the mechanisms of action and the effects of drugs on cells, tissues, and organs	Proteomics can monitor the changes in protein expression and modification in response to drug treatment, and reveal how drugs interact with their targets and modulate their functions and pathways. Proteomics can also identify the off-target and side effects of drugs, and assess their pharmacokinetics, pharmacodynamics, toxicity (Kennedy, 2002; Šleno et al., 2008).
Developingnew therapeutic strategies and personalized medicine	Proteomics can guide the design and optimization of drugs and drug delivery systems, such as peptides, antibodies, nanoparticles and enhance their specificity, stability, solubility, and bioavailability. Proteomics can also enable the selection and stratification of patients based on their protein profiles, and predict their response and resistance to drugs (Zhang et al., 2010).

2.5. Metabolomics and Lipidomics

Metabolomics and lipidomics are branches of omics that aim to measure and analyze the metabolome and the lipidome of biological samples, such as their composition, function, interaction. Metabolomics and lipidomics can reveal the molecular mechanisms, pathways, and networks underlying biological phenomena, such as development, disease, and drug response.

One of the specific challenges of metabolomics and lipidomics is the unequivocal identification of metabolites or lipids due to the large complexity of structures and nomenclatures (Bowen et al., 2010). This challenge arises from the fact that these fields deal with a vast array of chemically diverse compounds, each with its own unique structure and properties. This makes it difficult to develop a universal method for identifying all metabolites or lipids in a sample. Furthermore, the lack of standardized nomenclature for these compounds adds another layer of complexity to their identification.

To address this challenge, various techniques and tools have been developed, such as mass spectrometry, which can identify and quantify thousands of metabolites and lipids in a complex mixture (Bowen et al., 2010).

Metabolomics and lipidomics aid drug development by discovering new metabolites and lipids, understanding their impact on health and disease, and studying the effects of drugs on metabolism, including the identification of biomarkers for drug efficacy and toxicity (Table 5).

Table 5. Applications of metabolomics and lipidomics for drug development.

Application	Description
Discovering new metabolites and lipids and their products	Metabolomics and lipidomics can isolate and characterize the novel metabolites and lipids and their genes, proteins, and enzymes from various biological or environmental samples, and evaluate their biological activities and properties, and synthesize and optimize their products for drug discovery and development (Shyur et al., 2008; Stuart et al., 2020).
Understanding the metabolism and lipid metabolism and their impact on human health and disease	Metabolomics and lipidomics can profile and compare the metabolism and lipid metabolism of healthy and diseased individuals, and reveal how they are associated with various diseases, such as cancer, diabetes, cardiovascular disease and modulate their metabolism and lipid metabolism for disease prevention and treatment (Rasmiena et al., 2013; Zhao et al., 2014).
Studying the effects of drugs on metabolism and lipid metabolism and their biomarkers	Metabolomics and lipidomics can monitor the changes in metabolism and lipid metabolism in response to drug treatment, and identify the target metabolites and lipids and their pathways of drugs, and discover the biomarkers for drug efficacy and toxicity (Armitage et al., 2016; Guleria et al., 2018).

2.6. Data Acquisition and Preprocessing

Data acquisition and preprocessing are fundamental to omics bioinformatics, underpinning the success of a wide range of subsequent analyses such as differential biomolecule expression, functional enrichment, and functional network analyses. The quality of these initial stages is crucial to the robustness and reliability of downstream analyses.

Omics data, including genomics and transcriptomics, are typically generated using next-generation sequencing (NGS) platforms, which produce FASTQ files containing sequence and quality information for each read (Liu et al., 2012). Conversely, proteomics, metabolomics, and lipidomics data are often generated by mass spectrometry (MS), a technique that measures the mass-to-

charge ratio of ionized molecules in a sample (Griffiths et al., 2009). Regardless of the data type, preprocessing is an essential first step in data analysis, involving the conversion of raw data into a matrix of features and their intensities across samples. This task is performed by specialized software such as bcl2fastq for NGS data and MaxQuant (Cox et al., 2008) for MS data.

In the context of bioinformatics and omics for drug development, the first step is to acquire and preprocess the omics data. These data are usually generated by high-throughput technologies that measure the abundance or activity of various biological molecules in a given biological system. Omics data can be classified into different types according to the level of biological organization, each with its own characteristics, advantages, and limitations, and each requiring specific methods and tools for data acquisition and preprocessing.

Data acquisition involves generating omics data from biological samples using high-throughput technologies. This process includes several steps such as sample collection, preparation, processing, measurement, and storage, and can be influenced by many factors that can affect the quality and reproducibility of the omics data. Therefore, data acquisition should follow standard protocols and best practices to ensure the reliability and validity of the omics data.

Data preprocessing involves transforming the raw omics data into a format suitable for further analysis. This process includes several steps such as data cleaning, quality control, normalization, transformation, and annotation. The general goal of data preprocessing is to remove noise, artifacts, errors, and biases from the omics data, and to enhance the signal, accuracy, and comparability of the omics data. This step is essential for improving the quality and usability of the omics data, and for reducing the complexity and dimensionality of the omics data.

Data acquisition and preprocessing form the bedrock of omics bioinformatics, serving as the fundamental pillars that uphold the success of a myriad of subsequent analyses. These analyses span a broad spectrum, encompassing differential biomolecule expression, functional enrichment, and functional network analyses, among others. The robustness and reliability of these downstream analyses are inextricably tied to the quality of the initial data acquisition and preprocessing stages, underscoring their pivotal role in the field of omics and bioinformatics.

Despite the differences in focus, various omics fields share common methods for data acquisition and preprocessing. For simplicity, we've listed some examples of data acquisition and preprocessing methods and tools, categorized by different types of omics data, in Table 6.

Table 6. A summary of data acquisition and preprocessing methods and tools for genomics, transcriptomics, proteomics, and metabolomics data.

Omics	Data acquisition	Data preprocessing
Genomics	FASTQ files from NGS platforms, such as bcl2fastq, Torrent Suite.	Quality control and assessment, such as FASTQC; trimming and filtering, such as Trimmomatic (Bolger et al., 2014); alignment to reference genome, such as BWA (Li et al., 2009 a); manipulation and conversion of alignment files, such as SAMtools (Li et al., 2009 b).
Transcriptomics	FASTQ files from NGS platforms, such as bcl2fastq, Torrent Suite.	Quality control and assessment, such as FASTQC; trimming and filtering, such as Trimmomatic (Bolger et al., 2014); alignment to reference genome and transcriptome, such as STAR (Dobin et al., 2013); estimation of gene and transcript expression levels, such as RSEM (Li et al., 2011)..
Proteomics	mzML files from mass spectrometry platforms, such as msconvert (Chambers et al., 2012), ProteoWizard (Kessner et al., 2008).	Quality control and analysis, such as MaxQuant (Cox et al., 2008); downstream analysis, such as Perseus (Tyanova et al., 2016).
Metabolomics	Data-Dependent and DataIndependent Acquisition modes (DDA and DIA, respectively) are both widely used to acquire MS2 spectra in untargeted liquid chromatography tandem mass spectrometry (LC-MS/MS) metabolomics analyses (Guan et al., 2020); newer technologies such as collision cross section (CCS) data for ion mobility, high resolution mass spectra from Orbitrap, direct injection data, data independent acquisition (DIA)/ all ion fragmentation (AIF), imaging MS and multidimensional chromatography (Sud et al., 2016).	Many free data preprocessing tools, such as XCMS, MZmine, MAVEN, and MetaboAnalyst, as well as commercial software (Zhang et al., 2015); a growing number of users are adopting a workflow-based approach for their LC–MS data processing, for example XCMS Online, Metabolomic Analysis and Visualization ENgine (MAVEN), MZmine2, MetaboAnalyst and metabolomics specific Galaxy workflows—Galaxy-M and Workflow4metabolomics (Misra et al., 2016).

3. MAJOR CHALLENGES IN BIOINFORMATICS AND OMICS

Bioinformatics and omics face a multitude of significant challenges that encompass aspects such as data quality, the sheer size or dimensionality of the data, reproducibility of results, standardization of methods, and the ethical and privacy concerns associated with handling such data. These challenges underscore the complexity of the field and highlight the need for robust, innovative solutions to ensure the continued advancement and success of bioinformatics and omics research.

3.1. Data Quality

Data preprocessing includes filtering, transformation, and scaling of the data to reduce noise and improve data quality. However, this process can be challenging due to issues such as missing values, batch effects, and the need for multiple testing corrections. Missing values are a common problem in omics data, which occur when a feature is not detected or quantified in some samples due to technical or biological reasons. Missing values can affect the downstream analysis and interpretation of the data, such as differential expression, clustering, and correlation. Therefore, it is important to handle missing values appropriately, either by imputing them with reasonable values or by removing them from the analysis. Several methods and tools have been developed for missing value imputation in data, such as k-nearest neighbors (Troyanskaya et al., 2001), random forest (Lebedev et al., 2014), and Bayesian principal component analysis (Fang et al., 2018).

Normalization is another important step in omics data analysis, which aims to remove systematic biases and variations in the data that are not related to the biological conditions of interest, such as instrument settings, sample preparation, and run order. Normalization can improve the comparability and reproducibility of the data and reduce the false discovery rate in the downstream analysis. Several methods and tools have been developed for normalization of data, such as median (Bolstad et al., 2003), quantile (Bolstad et al., 2003), cyclic loess (Yang, 2002), and vsn (Huber et al., 2002).

Batch effects are another source of unwanted variation in omics data, which occur when the data are collected in different batches or groups, such as different days, operators, or instruments. Batch effects can confound the biological signal and lead to spurious results and conclusions. Therefore, it is important to identify and correct batch effects in data, either by adjusting the experimental design or by applying statistical methods. Several methods and tools have been developed for batch effect correction in omics data, such as ComBat (Johnson et al., 2007), RUV (Gagnon-Bartsch et al., 2013), and limma (Ritchie et al., 2015).

Multiple testing is another challenge in omics data analysis, which arises when performing multiple statistical tests on the same data, such as testing for differential expression of thousands of features. Multiple testing can increase the chance of false positives and inflate the type I error rate. Therefore, it is important to control the false discovery rate (FDR) in data analysis, which is the expected proportion of false positives among the rejected hypotheses. Several methods and tools have been developed for FDR control in data, such as Benjamini-Hochberg (Benjamini et al., 1995) and qvalue (Storey, 2003).

Despite these challenges, the use of common methods across different omics fields facilitates providing a more comprehensive understanding of biological systems.

3.2. Data Volume and Dimensionality

One of the major challenges in bioinformatics and omics is the high volume and dimensionality of the data, which can pose computational and statistical difficulties for data analysis and interpretation. This phenomenon is known as the curse of dimensionality (Bellman, 1961), which refers to the problems that arise when dealing with data that have many features or variables, such as noise, redundancy, sparsity, overfitting. To overcome the curse of dimensionality, various methods and tools have been developed for dealing with high data volume and dimensionality, such as dimensionality reduction.

Dimensionality reduction is a branch of machine learning that aims to reduce the number of features or variables of omics data, while preserving the essential information and structure of the data. Dimensionality reduction can enhance data visualization and interpretation, and improve the performance and efficiency of data analysis and modeling.

One of the challenges of dimensionality reduction is to find the optimal balance between the complexity and the accuracy of the reduced data, and to avoid the loss of important or relevant information and the introduction of noise or artifacts. To address this challenge, various methods and algorithms have been developed, such as linear, nonlinear, supervised, unsupervised. These methods and algorithms can help to identify the intrinsic and latent dimensions and factors of omics data, and project the data onto lower-dimensional spaces or manifolds.

Some of the examples of dimensionality reduction methods for omics data are listed in Table 7.

3.3. Reproducibility

Reproducibility is the ability to obtain the same results from the same data and methods by different researchers. It is essential for the validity and reliability of scientific research, especially in the fields of bioinformatics and omics, where complex data and algorithms are involved. However, reproducibility is often hampered by various factors, such as lack of data availability, insufficient documentation, inconsistent software versions, and human errors (Peng, 2011).

Table 7. Some of the applications of dimensionality reduction for omics data.

Application	Description
Visualization and exploration	Dimensionality reduction can project the omics data onto a lower-dimensional space, such as two or three dimensions, and reveal the structure and diversity of the data, such as clusters, outliers, trends. For example, principal component analysis (PCA) (Pearson, 1901), t-distributed stochastic neighbor embedding (t-SNE) (Maaten et al., 2008), and uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) are popular methods for visualizing and exploring omics data.
Noise reduction and feature selection	Dimensionality reduction can filter out the noise and redundancy of omics data, and retain the most informative and discriminative features or variables. For example, singular value decomposition (SVD) (Eckart et al., 1936), independent component anal- ysis (ICA) (Hyvärinen et al., 2000), and sparse coding (Olshausen et al., 1997) are methods that can decompose omics data into a set of basis vectors or components that capture the main sources of variation in the data.
Data integration and comparison	Dimensionality reduction can align and fuse omics data from different sources, platforms, or modalities, and identify the common and unique features, patterns, and associations among them. For example, canonical corre- lation analysis (CCA) (HOTELLING, 1936), mul- tiple co-inertia analysis (MCIA) (Culhane et al., 2005), and multi-view learning (Wang et al., 2014) are methods that can integrate and compare omics data of different types, such as genomic, transcriptomic, proteomic.

Several solutions have been proposed and implemented to ensure and enhance the reproducibility of bioinformatics and omics analyses. One such solution is data sharing, which involves making the raw and processed data publicly available through online repositories like GEO (Barrett et al., 2011), SRA (Leinonen et al., 2011), EGA (Lappalainen et al., 2015), and OpenAIRE.eu. This is highly effective when done with proper metadata and identifiers, adhering to principles like MIAME (Brazma et al., 2001) and FAIR (Wilkinson et al., 2016).

Another solution is code sharing, where the source code and scripts used for the analysis are made publicly available through online platforms such as GitHub (Blischak et al., 2016), Bitbucket.org, Zenodo.org or OSF.io. That should be accompanied by proper documentation and licensing, including README, LICENSE, and links to DOI and to all relevant documents.

Workflow management serves as a solution that employs various tools and frameworks to automate and standardize the analysis pipeline. Notable examples of these tools are Snakemake (Köster et al., 2012), Nextflow (Di Tommaso et al., 2017), and Galaxy (Afgan et al., 2018). These tools, when coupled with robust configuration and dependency management systems such as Conda.io, Docker.com, and Singularity (Kurtzer et al., 2017), can streamline the process and enhance the efficiency of data analysis workflows.

By adopting these solutions, bioinformatics and omics researchers can increase the transparency, reproducibility, and reusability of their analyses. This, in turn, facilitates the verification, validation, and dissemination of their findings.

3.4. Data Standardization

One of the major challenges in bioinformatics and omics is the lack of standardization of complex data objects. These are data structures and formats that can store and organize large omics data using Python or R programming languages. Complex Python and R objects can facilitate efficient and scalable data processing, analysis, and visualization, and enable interoperability and compatibility among different tools and platforms (Hu et al., 2022).

However, the lack of standardization can lead to difficulties in data integration, comparison, and reproducibility. Different tools and platforms may use different data structures and formats, which can make it challenging to combine or compare the data from different sources. Moreover, the lack of standardization can also hinder the reproducibility of the data analysis, as the results may depend on the specific data structures and formats used.

To address this challenge, several efforts have been made to develop standardized data objects for omics data. For example, the Bioconductor project has developed a series of S4 classes for storing and manipulating omics data in R, such as ExpressionSet, SummarizedExperiment, and MultiAssayExperiment (Chervitz et al., 2011). Similarly, the pandas library in Python provides DataFrame for handling tabular data (Yudin, 2021) or the xarray library as examples of advanced data manipulation tools ((Hoyer et al., 2017).

Despite these efforts, the standardization of complex data objects in bioinformatics is still a work in progress. More efforts are needed to develop and promote the use of standardized data objects, and to ensure their compatibility and interoperability with existing and future tools and platforms (Katayama et al., 2010). Some of the examples of complex objects for omics data are listed in Table 8.

Table 8. Complex data objects highly used in omics data.

Data Object	Description
Xarray	A Python package that makes working with labelled multi-dimensional arrays simple, efficient, and fun! Xarray introduces labels in the form of dimensions, coordinates and attributes on top of raw NumPy-like arrays, which allows for a more intuitive, more concise, and less error-prone developer experience.
netCDF	Network Common Data Form (netCDF) is a set of software libraries and machine independent data formats that support the creation, access, and sharing of array-oriented scientific data. It is used extensively in the atmospheric and oceanographic communities to store variables, such as temperature, pressure, wind speed, and wave height.
HDF5	Hiyerarşik Veri Biçimi (HDF5), büyük miktarda sayısal veriyi depolamak ve düzenleHierarchical Data Format (HDF5) is a set of file specifications designed to store and organize large amounts of numerical data. It's a self-describing file format for storing complex data collections, widely used in the field of high-performance computing.
Zarr	Zarr is a Python package providing an implementation of chunked, compressed, N-dimensional arrays, designed for use in parallel computing. It is well suited to the storage of large multi-dimensional arrays of data, which are often encountered in scientific data processing and machine learning.

3.5. Ethics and Privacy Challenges

One of the major challenges in bioinformatics and omics is ensuring the ethics and privacy of data. These fields often deal with sensitive information, such as genomic data, that can uniquely identify individuals. Therefore, it is crucial to handle such data ethically and ensure its privacy (Heeney et al., 2011).

However, the lack of standardization and clear guidelines can lead to difficulties in ensuring ethics and privacy. Different tools, platforms, and researchers may have different interpretations of what constitutes ethical use and adequate privacy protection. This can make it challenging to ensure that all data is handled ethically and that the privacy of individuals is protected (Mittelstadt et al., 2016).

Moreover, the rapid advancement and complexity of these fields can further complicate the issue. New technologies and methods can lead to new ethical and privacy challenges that were not previously considered. Therefore, it is important to continuously update and refine the ethical guidelines and privacy protection measures (Greenbaum et al., 2011).

To address this challenge, several efforts have been made to develop standardized guidelines for ethics and privacy in bioinformatics and omics. These guidelines provide clear instructions on how to handle data ethically and ensure its privacy. However, the implementation of these guidelines can be challenging and requires continuous effort and vigilance.

4. FUTURE DIRECTIONS

In this section, you will learn about some of the advanced and emerging applications of bioinformatics and omics for drug development, such as single-cell omics, multiomics, standardization of complex data objects, dimensionality reduction, and metagenomics.

4.1. Single-Cell and Spatial Omics

Single-cell omics is a branch of omics that aims to measure and analyze the molecular profiles of individual cells, such as their DNA, RNA, proteins, metabolites. Single-cell omics can reveal the heterogeneity and dynamics of cell populations, such as their gene expression, epigenetic modifications, signaling pathways, metabolic activities. Single-cell omics can also identify rare or novel cell types, such as stem cells, cancer cells, immune cells, and their functional roles and interactions in various biological processes and diseases (Chappell et al., 2018).

One of the most widely used techniques for single-cell omics is single-cell RNA sequencing (scRNAseq), which can measure the transcriptome (the set of all RNA molecules) of individual cells. scRNA-seq can capture the gene expression patterns and regulatory networks of cells, and reveal their differentiation trajectories and developmental stages. scRNA-seq can also be combined with other techniques, such as single-cell ATAC-seq, which can measure the chromatin accessibility (the degree of openness of DNA regions) of individual cells, or single-cell proteomics, which can measure the proteome (the set of all proteins) of individual cells. These techniques can provide complementary information and enable a more comprehensive and integrated analysis of single-cell omics data.

Some of the applications of single-cell omics for drug development include:

- Discovering new cell types and biomarkers for diagnosis, prognosis, and treatment of diseases, such as cancer, neurodegenerative diseases, autoimmune diseases. For example, scRNA-seq can identify the subtypes and heterogeneity of tumor cells and their microenvironment, and reveal their molecular signatures and drug resistance mechanisms (Ding et al., 2020; Zhang et al., 2021).

- Understanding the mechanisms of action and the effects of drugs on cells, tissues, and organs, such as their pharmacokinetics, pharmacodynamics, toxicity. For example, scRNA-seq can monitor the changes in gene expression and cell states in response to drug treatment, and identify the target cells and pathways of drugs (Gawel et al., 2019).
- Developing new therapeutic strategies and personalized medicine, such as cell-based therapies, gene therapies, immunotherapies. For example, scRNA-seq can guide the selection and engineering of cells for transplantation, gene editing, or immune modulation, and optimize the dosage and timing of drug administration (Ding et al., 2020).

4.2. Integrated Multiomics

Multiomics is a branch of omics that aims to measure and analyze multiple types of molecular profiles of biological samples, such as their genome, epigenome, transcriptome, proteome, metabolome. Multiomics can capture the complexity and interactions of multiple biological layers, and reveal how they are influenced by genetic and environmental factors, and how they affect the phenotypes and functions of cells, tissues, and organisms.

One of the challenges of multiomics is to integrate the heterogeneous and high-dimensional data from different omics platforms and sources, such as sequencing, mass spectrometry, microarrays. In this context, Python's MOFA (Multi-Omics Factor Analysis) stands out as a powerful tool for multiomics data integration. MOFA is a general framework for the integration of multi-omics data sets in a completely unsupervised manner. It identifies the main sources of variability in the data, which are represented as factors. Each factor captures a particular pattern that is shared across multiple omics data types. This allows for the disentanglement of the different biological and technical sources of variability, providing a more comprehensive understanding of the system under study (Argelaguet et al., 2018). To further address the challenges of multiomics integration, various computational methods can be applied, such as data normalization, transformation, imputation, alignment, fusion, correlation, clustering, classification, regression, network inference. These methods and tools can help to identify the common and unique features, patterns, and associations among different omics data, and infer the causal and regulatory relationships among different biological layers.

4.3. Collaborative Bioinformatics

Collaborative analysis is the process of working together with other researchers to perform and improve the analysis of bioinformatics and omics data. It is beneficial for the advancement and innovation of scientific research, especially in the fields like drug development and pharmacogenomics, where interdisciplinary and integrative approaches are required. However, collaborative analysis is often challenged by various factors, such as data heterogeneity, method diversity, communication barriers, and coordination difficulties (Schmitt et al., 2011).

In the pursuit of facilitating and improving the collaborative analysis of bioinformatics and omics data, a multitude of solutions have been proposed and implemented. One such solution is the use of version control tools and systems that are widely used in software development like Git, Subversion (SVN), and Mercurial. These systems track and manage changes made to the data and code, employing proper branching and merging strategies such as feature, release, and hotfix branches.

Another approach is through cloud computing, which utilizes services and platforms like AWS, Google Cloud, or Azure. These provide on-demand access to computing resources including storage, processing, and networking. They also ensure proper security and scalability features, such as encryption, authentication, and load balancing. in the context of bioinformatics research and data science Google Colaboratory, often referred to as "Colab", has emerged as a popular tool for collaborative analysis. Colab is a cloud service based on Jupyter Notebooks and provides an interactive environment that enables users to write and execute Python code through the browser. Colab has a free version of research and education allowing access to computing resources including GPUs and TPUs, making it a valuable tool for bioinformatics and omics data analysis.

Custom-made web applications, developed using Python and R, have also emerged as a viable solution in the field of bioinformatics and omics analyses. Tools and frameworks, such as Rstudio Shiny and Streamlit, have facilitated the creation and sharing of these interactive applications. These applications can be conveniently deployed on specialized cloud servers like streamlit.io or shinyapps.io. Remarkably, recent technological advancements, such as Shinylive, have made it possible to deploy these applications on static servers, thereby eliminating the need to run R or Python on the server side (Jia et al., 2022). This development has significantly enhanced the flexibility for bioinformaticians to deploy and share their applications. However, at this early stage of Shinylive's development, the applications may take a considerable amount of time to load, and not all modules are compatible. An example of these tools in action is the interactive version of Figure 1, which can be found here. In summary, these applications incorporate essential user interface and user experience design elements, such as widgets, plots, and tables.

By adopting these solutions, bioinformatics and omics researchers can enhance collaboration, communication, and integration of their analyses. This, in turn, accelerates the discovery and development of new knowledge and solutions.

5. CONCLUSION

Bioinformatics and omics are revolutionizing drug development, offering new insights and solutions for drug discovery and improvement. The integration of large-scale biological data and computational tools is enabling the research for new targets, biomarkers, and drugs, and the elucidation of disease mechanisms and drug effects.

One of the key solutions in this field is the development and implementation of novel and robust computational tools and techniques. These include machine learning, AI, network analysis, and data mining, which are used for analyzing and integrating large-scale and heterogeneous omics data. Another important aspect is the establishment and standardization of data quality, reproducibility, and sharing protocols and platforms. This ensures the reliability and accessibility of omics data and results, which is crucial for the advancement of the field. Moreover, the enhancement and expansion of omics databases and resources provide comprehensive and up-to-date information on genes, proteins, metabolites, pathways, diseases, drugs, and interactions. This wealth of information is invaluable for researchers and practitioners in the field. Collaboration and communication between experts from different disciplines, such as biologists, chemists, geneticists, statisticians, and computer scientists, are also essential. This fosters interdisciplinary and trans- lational research and innovation, leading to breakthroughs in drug development.

The ethical and social implications of bioinformatics and omics for drug development cannot be overlooked. Issues such as the protection of data privacy, the regulation of data ownership, and the promotion of data equity and justice need to be addressed.

In the context of pharmacogenomics, bioinformatics and omics are particularly impactful. They enable the study of how genes affect a person's response to drugs, leading to the development of personalized medicine. This has the potential to revolutionize drug development and our understanding of health and disease.

In conclusion, bioinformatics and omics for drug development is a rapidly evolving and promising field. Despite the challenges, the opportunities are immense and the future is bright. By harnessing the power of biological data and computational tools, bioinformatics and omics can pave the way for a new era of personalized medicine.

ACKNOWLEDGES

This work was supported by the project "Strategic Necessity: Molecule to Drug" with the grant number 2022-1-TR01-KA220-VET-000088373.

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A. vd. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. (2018)

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., Stegle, O. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. 14,. (2018)

Armitage, E. G., Southam, A. D. Monitoring cancer prognosis, diagnosis and treatment efficacy using metabolomics and lipidomics 12, 146. (2016)

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., Holko, M., Ayanbule, O., Yefanov, A., Soboleva, A. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. 39, D1005–D1010. (2011)

BELLMAN, R. *Adaptive Control Processes*Princeton University Press (1961)

Benjamini, Y., Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57, 289–300. (1995)

Blischak, J. D., Davenport, E. R., Wilson, G. A Quick Introduction to Version Control with Git and GitHub. (Ouellette, F., Ed.)PLOS Comput. Biol. 12, e1004668. (2016)

Bolger, A. M., Lohse, M., Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data 30, 2114–2120. (2014)

Bolstad, B. M., Irizarry, R. ., Åstrand, M., Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias 19, 185–193. (2003)

Bowen, B. P., Northen, T. R. Dealing with the unknown: Metabolomics and Metabolite Atlases. J. Am. Soc. Mass Spectrom. 21, 1471–1476. (2010)

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U. vd. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat. Genet. 29, 365–371. (2001)

Challis, G. L. Genome Mining for Novel Natural Product Discovery. J. Med. Chem. 51, 2618–2628. (2008)

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M. Y., Paulse, C., Creasy, D. vd. A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. 30, 918–920. (2012)

Chang, W. C., Tanoshima, R., Ross, C. J. D., Carleton, B. C. Challenges and Opportunities in Implementing Pharmacogenetic Testing in Clinical Settings. Annu. Rev. Pharmacol. Toxicol. 61, 65–84. (2021)

Chappell, L., Russell, A. J. C., Voet, T. Single-Cell (Multi)omics Technologies. Annu. Rev. Genomics Hum. Genet. 19, 15–41. (2018)

Chervitz, S. A., Deutsch, E. W., Field, D., Parkinson, H., Quackenbush, J., Rocca-Serra, P., Sansone, S. A., Stoeckert, C. J., Taylor, C. F., Taylor, R., Ball, C. A. Data Standards for Omics Data: The Basis of Data Sharing and Reuse, ss. 31–69 (2011)

Cox, J., Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372. (2008)

Cui, Y., Paules, R. S. Use of Transcriptomics in Understanding Mechanisms of Drug-Induced Toxicity. Pharmacogenomics 11, 573–585. (2010)

Culhane, A. C., Thioulouse, J., Perriere, G., Higgins, D. G. MADE4: an R package for multivariate analysis of gene expression data 21, 2789–2790. (2005)

Datta, S., Rajnish, K. N., Samuel, M. S., Pugazlendhi, A., Selvarajan, E. Metagenomic applications in microbial diversity, bioremediation, pollution monitoring, enzyme and drug discovery. A review. Environ. Chem. Lett. 18, 1229–1241. (2020)

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., Notredame, C. Nextflow enables reproducible computational workflows. Nat. Biotechnol. 35, 316–319. (2017)

DiMasi, J. A., Grabowski, H. G., Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. J. Health Econ. 47, 20–33. (2016)

Ding, S., Chen, X., Shen, K. Single-cell RNA sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy. Cancer Commun. 40, 329–344. (2020)

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner 29, 15–21. (2013)

Eckart, C., Young, G. The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218. (1936)

Fang, Z., Ma, T., Tang, G., Zhu, L., Yan, Q., Wang, T., Celedón, J. C., Chen, W., Tseng, G. C. Bayesian integrative model for multi-omics data with missingness. (Hancock, J., Ed.) 34, 3801–3808. (2018)

Gagnon-Bartsch, J. A., Jacob, L., Speed, T. P. Removing Unwanted Variation from High Dimensional Data with Negative Controls. Tech. Reports from Dep. Stat. Univ. California, Berkeley 1–112. (2013)

Gawel, D. R., Serra-Musach, J., Lilja, S., Aagesen, J., Arenas, A., Asking, B., Bengnér, M., Björkander, J., Biggs, S., Ernerudh, J., Hjortswang, H., Karlsson, J. E., Köpsen, M., Lee, E. J., Lentini, A., Li, X., Magnusson, M., Martínez-Enguita, D., Matussek, A. vd. A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases. Genome Med. 11, 47. (2019)

Greenbaum, D., Sboner, A., Mu, X. J., Gerstein, M. Genomics and Privacy: Implications of the New Reality of Closed Data for the Field. (Bourne, P. E., Ed.)PLoS Comput. Biol. 7, e1002278. (2011)

Griffiths, W. J., Wang, Y. Mass spectrometry: from proteomics to metabolomics and lipidomics. Chem. Soc. Rev. 38, 1882. (2009)

Guan, S., Taylor, P. P., Han, Z., Moran, M. F., Ma, B. Data Dependent–Independent Acquisition (DDIA) Proteomics. J. Proteome Res. 19, 3230–3237. (2020)

Guleria, A., Kumar, A., Kumar, U., Raj, R., Kumar, D. NMR Based Metabolomics: An Exquisite and Facile Method for Evaluating Therapeutic Efficacy and Screening Drug Toxicity. Curr. Top. Med. Chem. 18, 1827–1849. (2018)

Heeney, C., Hawkins, N., de Vries, J., Boddington, P., Kaye, J. Assessing the Privacy Risks of Data Sharing in Genomics. Public Health Genomics 14, 17–25. (2011)

HOTELLING, H. RELATIONS BETWEEN TWO SETS OF VARIATES. Biometrika 28, 321–377. (1936)

Hoyer, S., Hamman, J. xarray: N-D labeled Arrays and Datasets in Python. J. Open Res. Softw. 5, 10. (2017)

Hu, B., Canon, S., Eloe-Fadrosh, E. A., Anubhav, Babinski, M., Corilo, Y., Davenport, K., Duncan, W. D., Fagnan, K., Flynn, M., Foster, B., Hays, D., Huntemann, M., Jackson, E. K. P., Kelliher, J., Li, P. E., Lo, C. C., Mans, D., McCue, L. A. vd. Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale. Front. Bioinforma. 1,. (2022)

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression 18, S96–S104. (2002)

Hyvärinen, A., Oja, E. Independent component analysis: algorithms and applications 13, 411–430. (2000)

Jethwa, A., Bhagat, J., George, J. T., Shah, S. Metagenomics for Drug Discovery. Nov. Technol. Biosyst. Biomed. Drug Deliv.Springer Nature Singapore, Singapore , ss. 125–153 (2023)

Jia, L., Yao, W., Jiang, Y., Li, Y., Wang, Z., Li, H., Huang, F., Li, J., Chen, T., Zhang, H. Development of interactive biological web applications with R/Shiny. Brief. Bioinform. 23,. (2022)

Johnson, W. E., Li, C., Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods 8, 118–127. (2007)

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., Whiteley, M. Metatranscriptomics of the Human Oral Microbiome during Health and Disease. (Kolter, R., Ed.)MBio 5,. (2014)

Kaczkowski, B., Tanaka, Y., Kawaji, H., Sandelin, A., Andersson, R., Itoh, M., Lassmann, T., Hayashizaki, Y., Carninci, P., Forrest, A. R. R. Transcriptome Analysis of Recurrently Deregulated Genes across Multiple Cancers Identifies New Pan-Cancer Biomarkers. Cancer Res. 76, 216–226. (2016)

Katayama, T., Arakawa, K., Nakao, M., Ono, K., Aoki-Kinoshita, K. F., Yamamoto, Y., Yamaguchi, A., Kawashima, S., Chun, H. W., Aerts, J., Aranda, B., Barboza, L. H., Bonnal, R. J., Bruskiewich, R., Bryne, J. C., Fernendez, J. M., Funahashi, A., Gordon, P. M., Goto, N. vd. The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. J. Biomed. Semantics 1, 8. (2010)

Kennedy, S. The role of proteomics in toxicology: identification of biomarkers of toxicity by protein expression analysis 7, 269–290. (2002)

Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P. ProteoWizard: open source software for rapid proteomics tools development 24, 2534–2536. (2008)

Köster, J., Rahmann, S. Snakemake—a scalable bioinformatics workflow engine 28, 2520–2522. (2012)

Kurtzer, G. M., Sochat, V., Bauer, M. W. Singularity: Scientific containers for mobility of compute. (Gursoy, A., Ed.)PLoS One 12, e0177459. (2017)

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., Vaughan, B., Laurent, T., Rowland, F., Marin-Garcia, P., Barker, J., Jokinen, P., Torres, A. C., de Argila, J. R., Llobet, O. M. vd. The European Genome-phenome Archive of human data consented for biomedical research. Nat. Genet. 47, 692–695. (2015)

Lebedev, A. V., Westman, E., Van Westen, G. J. P., Kramberger, M. G., Lundervold, A., Aarsland, D., Soininen, H., Kłoszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Lovestone, S., Simmons, A. Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. NeuroImage Clin. 6, 115–125. (2014)

Lee, J. min, Han, J. J., Altwerger, G., Kohn, E. C. Proteomics and biomarkers in clinical trials for drug development. J. Proteomics 74, 2632–2641. (2011)

Leinonen, R., Sugawara, H., Shumway, M. The Sequence Read Archive. Nucleic Acids Res. 39, D19–D21. (2011)

Li, B., Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. (2011)

Li, H., Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform 25, 1754–1760. (2009a)

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. The Sequence Alignment/Map format and SAMtools 25, 2078–2079. (2009b)

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M. Comparison of Next-Generation Sequencing Systems. J. Biomed. Biotechnol. 2012, 1–11. (2012)

Liu, Z., Liu, J., Liu, X., Wang, X., Xie, Q., Zhang, X., Kong, X., He, M., Yang, Y., Deng, X., Yang, L., Qi, Y., Li, J., Liu, Y., Yuan, L., Diao, L., He, F., Li, D. CTR-DB, an omnibus for patient-derived gene expression signatures correlated with cancer drug response. Nucleic Acids Res. 50, D1184–D1199. (2022)

Loging, W. T. *Bioinformatics and Computational Biology in Drug Discovery and Development*. (Loging, W. T., Ed.) Cambridge University Press (2016)

Maaten, L. van der, Hinton, G. Visualizing Data using t-SNE Laurens. J. Mach. Learn. Res. 9, 2579–2605. (2008)

McInnes, L., Healy, J., Saul, N., Großberger, L. UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. 3, 861. (2018)

Misra, B. B., van der Hooft, J. J. J. Updates in metabolomics tools and resources: 2014–2015. Electrophoresis 37, 86–110. (2016)

Mittelstadt, B. D., Floridi, L. *The Ethics of Biomedical Big Data*. (Mittelstadt, B. D. & L. Floridi, Ed.)Law, Governance and Technology SeriesSpringer International Publishing, Cham , C. 29 (2016)

Olshausen, B. A., Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res. 37, 3311–3325. (1997)

Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos. Mag. J. Sci. 2, 559–572. (1901)

Pedrotty, D. M., Morley, M. P., Cappola, T. P. Transcriptomic Biomarkers of Cardiovascular Disease. Prog. Cardiovasc. Dis. 55, 64–69. (2012)

Peng, R. D. Reproducible Research in Computational Science. Science (80-.). 334, 1226–1227. (2011)

Pirmohamed, M. Pharmacogenetics and pharmacogenomics. Br. J. Clin. Pharmacol. 52, 345–347. (2001)

Rasmiena, A. A., Ng, T. W., Meikle, P. J. Metabolomics and ischaemic heart disease. Clin. Sci. 124, 289–306. (2013)

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47–e47. (2015)

Schmitt, C. P., Burchinal, M. Data Management Practices for Collaborative Research. Front. Psychiatry 2,. (2011)

Shyur, L. F., Yang, N. S. Metabolomics for phytomedicine research and drug development. Curr. Opin. Chem. Biol. 12, 66–71. (2008)

Sleno, L., Emili, A. Proteomic methods for drug target discovery. Curr. Opin. Chem. Biol. 12, 46–54. (2008)

Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Stat. 31,. (2003)

Stuart, K. A., Welsh, K., Walker, M. C., Edrada-Ebel, R. Metabolomic tools used in marine natural product drug discovery. Expert Opin. Drug Discov. 15, 499–522. (2020)

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K. S., Sumner, S., Subramaniam, S. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. 44, D463–D470. (2016)

Swinney, D. C. Phenotypic vs. Target-Based Drug Discovery for First-in-Class Medicines. Clin. Pharmacol. Ther. 93, 299–301. (2013)

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. B. Missing value estimation methods for DNA microarrays 17, 520–525. (2001)

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., Cox, J. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods 13, 731–740. (2016)

van der Lee, T. A. J., Medema, M. H. Computational strategies for genome-based natural product discovery and engineering in fungi. Fungal Genet. Biol. 89, 29–36. (2016)

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 11, 333–337. (2014)

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R. vd. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. (2016)

Wong, C. H., Siah, K. W., Lo, A. W. Estimation of clinical trial success rates and related parameters 20, 273–286. (2019)

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., McDermott, U., Garnett, M. J. Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 41, 955–961. (2013)

Yang, Y. H. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30, 15e – 15. (2002)

Yudin, A. Data Analysis with Pandas. Basic Python Data Manag. Financ. Mark.Apress, Berkeley, CA , ss. 93–150 (2021)

Zhang, A., Sun, H., Yan, G., Wang, P., Wang, X. Metabolomics for Biomarker Discovery: Moving to the Clinic. Biomed Res. Int. 2015, 1–6. (2015)

Zhang, C. C., Kast, J. Applications of Current Proteomics Techniques in Modern Drug Design. Curr. Comput. Aided-Drug Des. 6, 147–164. (2010)

Zhang, Y., Wang, D., Peng, M., Tang, L., Ouyang, J., Xiong, F., Guo, C., Tang, Y., Zhou, Y., Liao, Q., Wu, X., Wang, H., Yu, J., Li, Y., Li, X., Li, G., Zeng, Z., Tan, Y., Xiong, W. Single-cell RNA sequencing in cancer research. J. Exp. Clin. Cancer Res. 40, 81. (2021)

Zhao, Y. Y., Cheng, X. long, Lin, R. C. Lipidomics Applications for Discovering Biomarkers of Diseases in Clinical Chemistry, ss. 1–26 (2014)