Chapter 4
Guidelines for Bioinformatics and the Statistical Analysis of Omic Data

Surajit Bhattacharya and Heather Gordish-Dressman

4.1 Bioinformatics: An overview

The Human Genome Project [1] was initiated to find all the nucleotides that constitute a human genome and identify and map genes making up that genome. The Human Genome Project, along with the other model organism projects, have enabled us to better understand the genetic and molecular underpinnings of developmental stages and disease conditions. These projects have also produced petabytes of data, mostly sequences of nucleotides, which are not comprehensible to a biologist until and unless it is given a biological perspective. Bioinformatics is a multidisciplinary branch of science, utilizing knowledge from many other fields including physics, mathematics, computational science, and statistics, to assist in transforming these raw nucleotide sequences to a more comprehensible biological dataset (Fig. 4.1). Bioinformatics can be broadly classified into two groups, genomics and proteomics. Genomics includes the study of the genomic data pertaining to defects in transcriptional and posttranscriptional mechanisms, while proteomics pertains to changes in proteins occurring mostly during the translational and post-translational phase. In this section we discuss the bioinformatics tools and algorithms used in various kinds of genomic experiments. Bioinformatics for transcriptome analyses can be found in the transcriptome chapter.

Genomic analyses focus on deoxyribonucleic acid (DNA), which is the building block of any biological system and is the primary component of the central dogma.

S. Bhattacharya
Center for Genetic Medicine Research, Children's National Medical Center, Washington, DC, USA

H. Gordish-Dressman (✉)
Center for Translational Research, Children's National Medical Center, Washington, DC, USA

Department of Pediatrics, The George Washington University School of Medicine and Health Sciences, Washington, DC, USA
e-mail: HGordish@childrensnational.org

J. G. Burniston, Y.-W. Chen (eds.), *Omics Approaches to Understanding Muscle Biology*, Methods in Physiology, https://doi.org/10.1007/978-1-4939-9802-9_4

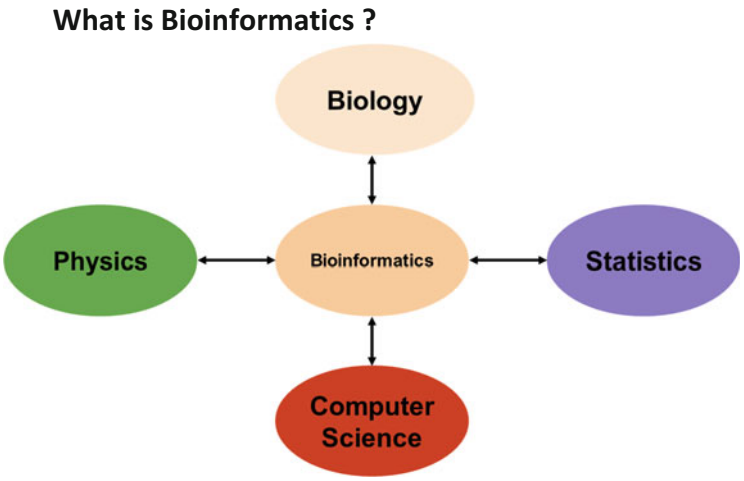

Fig. 4.1 Bioinformatics is a multidisciplinary branch of science: bioinformatics is a multidisciplinary branch of science, utilizing ideas from major branches of science such as physics, statistics, biology, and computer science

DNA is transcribed into ribonucleic acid (RNA), and messenger RNA (mRNA) directs protein translation. DNA is found in the nucleus and is composed of four major groups of nitrogen-based organic compounds (nucleotides): deoxyadenosine (A), deoxyguanosine (G), deoxythymidine (T), and deoxycytidine (C). These four nucleotides are arranged in different combinations to form single-stranded DNA. As A forms a chemical bond with T and C bonds with G, the DNA arranges itself into the shape of a double-stranded helical structure where each strand complements the other. If one strand has a nucleotide arrangement of ATTTCGATA, the second complementary strand will have an arrangement of TAAAGCTAT. This arrangement of nucleotides is called a sequence and the two strands of DNA sequences are referred to as the forward strand (read from left to right; also called 5′–3′) or the reverse strand (read from right to left; also called 3′–5′). RNA, another common biological material studied in genomics, differs slightly from DNA. Besides being made up of ribonucleic acid rather than deoxyribonucleic acid, RNA contains a unique nucleotide, uracil (U), in place of thymine.

From its isolation from pus in late 1868 [2] to the end of the Human Genome Project, the identification of the sequence of DNA has been an important aspect in understanding the functionality of the genes. This chapter is intended to be a resource for researchers interested in designing omic experiments or analyzing data from such experiments. It is organized into two major applications. The first has a focus on bioinformatics tools and techniques, and the second has a focus on statistical analyses. This chapter is intended to be a high-level introduction for researchers interested in performing their own analyses. It describes the major components of bioinformatics and statistical analysis and also directs the reader to key tools and sources of further information. Unfortunately it is beyond the scope of

this chapter to provide a comprehensive discussion of statistical theory and complex statistical models. This chapter discusses the general study design and aspects that should be taken into account before an experiment is begun. We describe some basic principles of statistical analysis and some commonly used methods, but further guidance from a statistician will be required for more complex study designs and statistical models.

4.2 Major Applications

4.2.1 Bioinformatics of Genomics Data

The first section discusses the bioinformatics tools and algorithms used in genomics. It describes typical workflows and the tools available for performing an omic experiment and underscores the importance of both the tools being used and a clear understanding of the underlying algorithm. One of the primary goals of any genomic research is to understand whether there are any changes or variants in the genome of a subject as compared to a reference genome. The reference genome is a representative genome of an organism, including humans, that is made up of multiple samples and is a representative example of a species' set of genes. The variation from this reference genome can be of two types: single nucleotide polymorphisms (SNP) and structural variants (SV). While SNPs constitute a change in a single nucleotide, SVs constitute a variation of $\geq$50 kilobases (KB) which can be an insertion, deletion, inversion, translocation, or duplication. It is these variants, or differences, which we are interested in when we perform omic experiments.

Next-generation sequencing (NGS) is the common term used for the modern-day high-throughput sequencing techniques, which allow one to define the precise order of nucleotides in the DNA of the organism sequenced. As this chapter focuses on the bioinformatics aspect of omic experiments, we will focus more on the results obtained from sequencing instruments than the sequencing techniques used. A general overview of sequencing techniques can be found in the following reviews [3, 4]. Here we briefly discuss the workflow for whole genome sequencing (WGS) analysis, obtained from a short-read sequencer, such as that produced by Illumina. A similar pipeline is used for exome sequencing. The pipeline can be divided into five parts (workflow is shown in Fig. 4.2):

1. *Preprocessing of raw sequences*: Raw output from sequencers, referred to as binary base calls (BCL), is converted to the human readable Fastq format [5]. Fastq format generally has three lines representing the sequence header, the sequence, and a quality score (*PHRED* score) [6, 7]. The PHRED score is a measure of the quality of the nucleotide identification associated with each base written in ASCII format. The PHRED score Q is defined as

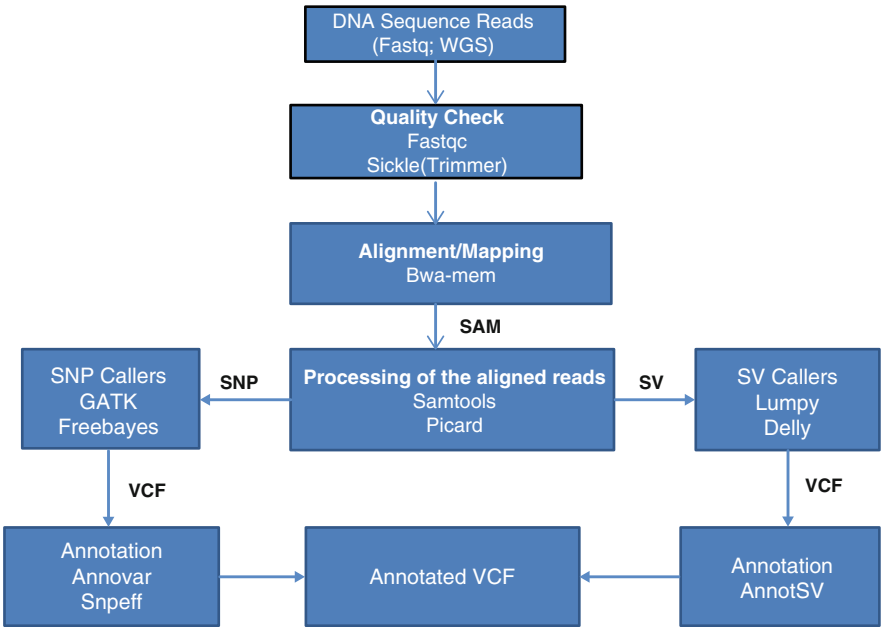

Fig. 4.2 Workflow for WGS: sequence files obtained from sequencers are converted to Fastq files, and an initial quality check using *FastQC* is performed. If the quality is low, the file can be trimmed using a trimming software such as *Sickle*. Next, alignment to the reference genome is performed using *BWA*. Aligned SAM files are processed using *Samtools* and *Picard* based on the type of variants to be identified. For single nucleotide polymorphisms (SNPs) and small insertion deletions (indels), *GATK* or *Freebayes* is chosen. For larger structural variants (SV), tools like *Delly* and Lumpy can be used. The output of all of these tools is a variant call format (VCF) file. An annotation tool for annotating the VCF file is chosen based on the variant that needs to be identified. *Annovar* and *Snpeff* are used for SNP annotation and *AnnotSV* for SV annotation. The final product is an annotated VCF file

$$Q = -10\log P$$

where P is the base-calling error probability. Thus, a larger Q value indicates a greater accuracy in identifying the base. The generally accepted PHRED score threshold is 30 which translates to the chances that the base was incorrectly identified as 1 out of 1000. The tool *FastQC* [8] can be used not only to provide a visual representation of PHRED score per sequence but to provide additional useful information such as sequence GC content, sequence length distribution, and sequence duplication distribution. If a sequence is found to be of low quality, trimming tools such as *cutadapt* [9], *sickle* [10], and *scythe* (tool available at https://github.com/vsbuffalo/scythe) can be used to remove unwanted sequences. As one example, sickle allows one to input a PHRED score threshold so that the user can define which sequences will be trimmed due to quality issues. Each of

these tools has various capabilities and the documentation for each should be consulted and understood before using.

2. *Alignment of the sequence*: Multiple tools/algorithms are available for the alignment/mapping of the whole genome sequence(s) to the reference genome. There are several important aspects to consider when choosing an alignment algorithm/software. These include the accuracy of alignment to the reference genome, the amount of computational memory used, and the time required for the alignment. One of the most important and commonly used algorithms is the Burrows-Wheeler algorithm (*BWA*) [11]. The BWA uses backward search with a compression algorithm, the Burrows-Wheeler transform (BWT) [12], to align sequence reads to the reference genome at a reduced memory usage and time [13]. The BWA software package consists of three algorithms, BWA-backtrack, BWA-SW, and BWA-MEM. BWA-backtrack is used for sequence reads that are shorter than 70 base pairs, while both BWA-SW and BWA-MEM are used for longer reads. The choice between BWA-SW and BWA-MEM is dependent on the platform used to generate the reads and the characteristics of the data. This algorithm can align 7 Gbp (giga base pairs) to the human reference genome per CPU day [13]. Output from this alignment step is a Sequence Alignment Map (SAM).
3. *Processing of the aligned files*: The output from this step needs to be further processed before it can be used for calling variants. Read group names are added to the aligned reads using *Picard* (https://broadinstitute.github.io/picard/) to identify each of the reads from different runs separately. *Samtools* [14] uses a merge function to combine different SAM files into a single merged binary aligned BAM file. Quality control processes can then be applied to the merged files using Samtools, and PCR duplicates can be removed by Picard Mark Duplicates.
4. *Identifying single nucleotide polymorphisms and structural variants*: Typically, short-read sequencers are more adept at identifying or calling small nucleotide variants and small insertion/deletions (indels) than structural variants. The Genome Analysis Toolkit (*GATK*) [15] is the standard and most commonly used pipeline for the identification of small nucleotide polymorphism and small indels from an aligned and processed BAM file. Other softwares, *Freebayes* [16] and *Heap* [17], are also available with Heap having the advantage in the detection of SNPs at low coverage.

   Although short-read sequencers are not ideal for identifying structural variants, there are multiple tools that can detect and identify structural variants from BAMs obtained by aligning short-read sequencer reads. All SV calling tools for short reads are based on five basic ideas:

   (a) *Read pair (RP)*: Discovers SVs by looking at the span and orientation of paired end reads. This method is used to identify almost all types of SVs. Pairs that are mapped far apart may denote deletion, whereas closer read pairs denote insertion. Orientation inconsistencies denote inversion or tandem

duplications [18]. Some of the tools associated to this method are *BreakDancer* [19], *PEMer* [20], and *MODIL* [21].

(b) *Read count (RC)*: This method can be also known as *read depth*. In these algorithms, a Poisson or modified Poisson distribution is assumed for the mapping depth, and deviation from that distribution denotes deletion or duplication [22]. This kind of SV is also called copy number variants (CNV). Tools associated with this method are *EXCAVATOR* [23], *CNVnator* [24], and *BIC-seq* [24].

(c) *Split read (SR)*: This method defined SV break points by identifying deviations between the sample sequence and the reference genome. A gap in the sample sequence denotes deletion, while additional sequence (above that in the reference) denotes insertion. Tools include *Socrates* [25] and *Splitread* [26].

(d) *De novo assembly (AS)*: In these group of algorithms, the short fragments are reassembled to form the original sequence, to identify the SV [27]. Tools include *Cortex* [28] and *Magnolya* [29].

(e) *Hybrid methods*: These algorithms use a combination of two or more methods discussed above. For example, *Delly* [30] and *Lumpy* [31] use read pair and split read algorithms, to detect SV.

As there are multiple methods to detect SVs, it is advisable to get a consensus result from multiple methods to confidently detect a true set of SVs in a sample. Tools like *svMerge* [32], *SURVIVOR* [33], and *Parliament* [34] combine results from multiple SV-identifying tools, to get a consensus output in variant call format (VCF).

5. *Annotation*: The output from the SV identifiers are in the form of a VCF, which includes information on the chromosome number and location of the variant along with type of SNP/SV and quality. It does not, however, contain information on the gene affected or whether or not the variant is a rare variant. This information is added by annotation tools. Annotation tools like *Annovar* [35] can perform annotation for both SVs and SNPs. Using Gene transfer format (GTF), Annovar identifies genes affected and uses databases such as *dbSNP* [36] (for SNPs) and Database of Genomic Variants [37] (*DGV*; for SVs) to identify the frequency of occurrence of the particular SV across different populations. *Clinvar* [38] database is used to identify the clinical significance of the variants (whether it causes disease or not), whereas the *CADD* [39, 40] tool is used to classify the variant based on its pathogenicity. *Snpeff* [41] and *SnpSift* [42] can be used to annotate and filter SNPs and small insertion deletion (indels; <50 bases), respectively, whereas *AnnotSV* [43] performs the same task for SVs.

4.2.2 Other Bioinformatics Analyses

Beyond expression analysis and variant discovery, NGS and microarrays can also be used to determine and evaluate the effect of other biological factors that affect a system. They include:

1. *Transcription factor binding site*: Chromatin immunoprecipitation (*ChIP*) is an immunoprecipitation technique by which the genomic region to which a transcription factor binds is extracted and then analyzed using either microarray or sequencing techniques. The basic working of microarray and sequencing remains the same for ChIP experiments, only the analysis tool pipeline changes.
   (a) *Microarray*: These experiments are known as *ChIP-on-chip*. The design of the experiments is similar to gene expression determination experiments, where the control is without the antibody to pull down the genomic region, while the experimental condition has the pulled down genomic region. Tiling Analysis Software (open-source software from Affymetrix) is used to process raw data into signals, with associated p-values for each probe. This is followed by visualization using Integrated Genome Browser (IGB), a proprietary software from BioViz. Another proprietary software, Partek, can also be used to analyze and visualize the regions to which transcription factors bind to. An open-source Bioconductor R pipeline is also available [87] which uses *Ringo* [88] for processing the raw microarray data into normalized data, *BioMart* [89] to annotate the data, and *topGO* [90] to perform the gene ontology. Ringo also has functions to visualize the genomic region.
   (b) *RNA-seq*: These experiments are known as *ChIP-seq*. Preprocessing, alignment, and processing of BAMs are done in the same fashion as RNA-seq data. FastQC is used for quality check, followed by alignment with Bowtie or BWA, followed by processing of SAM files by Samtools and Picard. To identify genomic regions to which transcription factors bind, we identify regions in the genome where maximum reads map to. These are called peaks. Peak calling is performed using tools like model-based analysis of ChIP-Seq (*MACS*) [91] and spatial clustering for identification of ChIP-enriched regions (*SICER*) [92]. Normalization and fold change calculation of transcription factor binding between two samples (control and experimental) are estimated using *edgeR*, *Gfold*, and *Deseq2*.
2. *Methylation*: Determination of methylation is done using bisulfite treatment. This converts unmethylated cytosine to uracil but leaves methylated cytosine unchanged [93]. Microarray and/or sequencing techniques can be then performed to determine the difference in methylation between an experimental condition and a control condition.
   (a) *Microarray*: The most popular platforms are Illumina 450 K and Epic (850 K) arrays. These arrays are named after the number of probes used; the 450 K has 450,000 probes of known methylation region and the 850 K is composed of

850,000 probes. There are multiple tools to analyze this data. Partek, a proprietary software, performs differential methylation and visualization of methylation regions. Open-source workflow *Champ* [94, 95], a R Bioconductor workflow, uses other Bioconductor packages like *minfi* [96] for analysis and visualization of methylation, a modified version of *combat* function from sva package [97] for batch correction, *limma* for differential expression calculation, and *plotly* [98] for visualization.

(b) *Sequencing*: The workflow can be divided into two steps.

(i) *Preprocessing and alignment-based identification*: The preprocessing and the alignment steps are similar to preprocessing step as described in the Genomics and Transcriptomics sections; the only differences are in the tools used for alignment. FastQC is used for quality check and sickle used for trimming. Most of the alignment methods use a modified version of other short read alignment algorithms like Bowtie2. Tools like *Bismark* [99], *BS-Seeker* [100], and *B-Solana* [101] use Bowtie algorithm, with methods to handle cytosine to thymine (uracil in RNA) conversion. LAST [102], another algorithm, uses score matrix to handle C-T conversion, and *Bisulfighter* [103] uses this algorithm for alignment. Another algorithm, BSMAP [104], converts the thymine to cytosine in the bisulfite reads in silico, in positions in which cytosine is present in the reference.

(ii) *Differential methylation*: The best tool to use is based on whether replicates are present in the experiment to be analyzed. For experiments in which there are no replicates available for control or experimental conditions, methylkit [102] or RnBeads [105] uses Fisher's exact tests, and *ComMet* [103, 106] (part of *Bisulfighter* methylation analysis pipeline) or *Methpipe* [107] uses hidden Markov models. For experiments having replicates, tools like *limma*, *BSmooth* [108], and *Biseq* [109] use regression models. Good reviews of these tools and methods can be found in [110, 111].

4.2.3 Platforms Used

Most of the pipelines used in the discussion involve a combination of a Linux environment and R IDEs like Rstudio. Open-source user interface (UI) option for analyzing NGS studies is limited, with Galaxy [116] being the only reliable option. Cloud platform UIs like DNAnexus and artificial intelligence RNA-seq (AIR) are proprietary and are subscription based. Cloud-based platforms such as Amazon and Microsoft Azure can also be used for processing large amounts of genomic data, but they are mostly command line based.

4.2.4 Downstream Analysis and Visualization

After the analysis to identify the most significantly expressed genes is complete, multiple downstream analyses can then be performed to give biological perspective to the genes defined. Following are the two most important downstream analyses done on transcriptomic analyzed data:

1. *Gene ontology and pathway analysis*: To better understand the biological system studied in a given experiment, one needs to understand the function and pathways of the genes that are expressed in the system. Gene ontology [117] (*GO*) classification tools can classify the function of the genes under three broad classes: cellular components (*CC*), molecular functions (*MF*), and biological processes (*BP*). This kind of analysis is called a gene enrichment analysis (GEA). GEA can also be divided into three categories:
   (a) *Singular enrichment analysis (SEA)*: A user-defined gene list derived from any high-throughput NGS/microarray experiment can be used as an input [118]. Generally, the input genes are those highly significant genes based on user-defined arbitrary threshold (e.g., multiple testing adjusted p-value < 0.05). The genes are classified as falling into the three broad categories (BP, CC, and MF) described above. Fisher's exact test [119] and its modifications (*EASE* score [118]) or a chi-square test [120] is used in this method to determine whether the genes are related to the categorization based on function. Tools such as *DAVID* [118, 121], *GOStat* [122], and *Bingo* [123] perform singular enrichment analysis.
   (b) *Gene set enrichment analysis (GSEA)*: In gene set enrichment analysis, all genes from a high-throughput genomic experiment are used as an input, thus ensuring the analysis is free from any potential bias due to arbitrary thresholds such as those used in SEA [124]. This allows even genes with small differential expression changes to be considered for further enrichment analysis. Maximum enrichment scores (MESs) are calculated from the rank order of all gene members in an annotation category. These maximum enrichment scores are analyzed and p-values calculated using either Kolmogorov-Smirnov-like statistics to compare the observed MESs to those obtained with from a random distribution or with parametric statistics which compare fold changes between experiments. *ErmineJ* [125] and *FatiScan* [126] can be used to perform this type of analysis.
   (c) *Modular enrichment analysis (MEA)*: Modular enrichment analysis combines SEA-kind enrichment analysis with network discovery algorithms, to allow term to term relationships. Kappa statistics of agreement are used for this analysis. Due to the structure of the kappa statistics, genes which do not occur in multiple neighboring terms are excluded from the analysis. Some tools that perform this type of analysis are *ADGO* [127], *DAVID*, and *GeneCodis* [128].

These gene ontology classification tools combine data from multiple sources including *KEGG* [129] for pathway analysis, *Pfam* [130] for protein domains, and *TRANSFAC* [131] for gene regulation.

2. *Correlation networks*: Another important piece of information from the obtained gene lists are whether or not there is any relationship between the statistically significant gene lists. Although we learn a lot about genes that share common functionality from the GEA, tools like *GeneMania* [132] provide the user with information on the interaction between genes, thus yielding a more comprehensive picture of the interaction pathways. GeneMania provides the users with predictions of co-expression, co-localization, and physical interactions between genes. Biological General Repository for Interaction Datasets (*Biogrid*) [133] is another database which provides the user with chemical, genetic, and protein interactions using known experimental results and is curated regularly. *STRING* [133], a protein interaction database, is used to understand the interaction between the proteins translated from the genes. One can also use weighted gene correlation network analysis (WGCNA) [134], a package in R [135], which uses expression data from microarray or RNA-seq to construct a correlation network between genes in a given experiment.

Visualization Tools like *FastQC and MultiQC* [136] are used to evaluate and visualize the quality of the sequencing from both Fastq and BAM data. Visualization of expression data can be done using Heatmap.2 function of the *gplots* [137] package in R. *GOPlot* [138], an R package, which can be used to visualize Gene Ontology data.

4.2.5 Conclusion

Although newer long-read sequencing techniques as well as optical mapping techniques are being developed to better understand variation in genes and changes in expression, no technology has yet displaced short-read sequencing and/or microarrays. With the advancement in computational hardware, as well as development of more extensive computational algorithms to better handle big data, bioinformatics tools for analysis of NGS data are rapidly changing. In these changing times, one should not only concentrate on the tools being used, but also better understand and appreciate the underlying algorithm to better evaluate his/her data. In addition, a thorough understanding of experimental design and statistical analysis is incredibly important to yield a thoughtful and meaningful analysis.

4.3 Statistical Analysis

4.3.1 Study Design Principles

Study design at its most basic level is defined as the methods for planning the collection of data in order to obtain the maximum amount of information using the least amount of resources. As such, good study design is an integral part of any experiment. There are many available resources that discuss study design in general, and an Internet search of study design fundamentals will yield hundreds of sources. However, omic experiments have certain common traits that make many typical study design elements less applicable. We often do not have large numbers of samples available or the high cost of the technologies limits the number of samples we have the resources for. We are often testing many more parameters than we have samples for, causing difficulties with our statistical analyses [139, 140]. Lastly, due to the scarcity of individual samples, we are often tempted to test complicated hypotheses that we do not have the data to truly support.

Before we begin an experiment, there are several things we should consider so that our experiment can be planned to appropriately answer our scientific question. We need to have a testable hypothesis, we need to define our replicates (how many we have and what type they are), we need to understand the sources of variability in our sample and how it relates to the population, and we need to verify that our sample size is adequate to answer our question. Lastly, we need to understand what our collected data points are representing.

4.3.1.1 Hypothesis

Having a testable hypothesis may seem an obvious step; however, not all hypotheses are well thought out or testable. Omic experiments are often used for hypothesis "generation" where we use the data we have to inform what is happening in a biological system. We then use this information to generate hypotheses that can be tested in future experiments. Because of this feature, it is mistakenly thought that omic experiments are discovery in nature and do not need an a priori defined hypothesis. While this is not entirely false, it is also not true [141]. One may have a more general question to ask such as "are RNA levels related to biceps strength?" with the intention of discovering what RNA levels change in response to strength. However, when we perform our statistical analysis, we are in fact testing a hypothesis. That hypothesis may involve assessing the relationship between RNA levels at one particular strength level, or assessing the change in RNA levels as strength changes, or assessing whether or not the relationship between RNA level and strength is different with respect to some other factors such as gender. These three hypotheses are all tested using different statistical methods and require different study designs and numbers of samples. So, we must know what hypothesis we want to test before we plan our experiment, even if our goal is only to "discover" what

RNA levels are important to us. If we don't consider a hypothesis, we may be left at the end with data that cannot answer our questions.

Imagine a case where we are interested in assessing whether exercise in men has an effect on the expression level of gene *X*. We measure gene expression in 10 men who exercise on a regular basis and in 10 men who are sedentary. If we are truly interested in whether or not exercise causes a change in expression of gene *X*, we cannot test that hypothesis with this sample. All we have from our sample design is expression levels in exercisers and expression levels in sedentary individuals. We may find significant differences in the expression of gene *X*; however, we cannot say those differences are due to exercise. They could be due to any number of differences between the two groups of men, not necessarily related to exercise. Our intentions may have been to test the hypothesis that exercise induces changes in gene expression, but our sample cannot test that hypothesis. In order to test our intended hypothesis, we would need repeated measurements of expression on the same individuals both before and after exercise. Only then could we infer that the exercise was, in part, responsible for any differences observed in expression levels.

The above example shows how it can be a mistake to design an experiment without a clear hypothesis in mind. We can spend valuable resources collecting data that does not, in the end, meet our needs.

4.3.1.2 Replicates

The understanding and choice of replicates is often a source of confusion because the term "replicates" has various definitions depending on who is defining it. At its simplest, a replicate is a measurement taken more than once. Some refer to multiple measurements taken from the same sample at the same time as replicates. Others refer to measurements taken from the same sample at different times as replicates. And yet others consider replicates to be different aliquots of the same RNA extraction or multiple arrays hybridized with the same RNA.

We can divide replicates into two types, biological and technical, which serve different purposes. Technical replicates are used to establish the variability, or experimental error, of the measurement technique. They are performed on the same biological sample so that the differences we see can be wholly attributed to the measurement technique. All measurement techniques have variability in how well they measure the parameter of interest and we need to quantify this to better allow us to measure other sources of variability we encounter. Those performing omic experiments should be familiar with this phenomenon as we must know how much noise is in the measurement to accurately assess the signal. Biological replicates, on the other hand, are broadly speaking biologically different samples. They are used to establish the biological variability which exists between organisms which should be identical, and they are typically the reason we are performing the experiment in the first place. The biological entities in our experiment represent the wider population. We can only make inferences, by applying inferential statistics, about the population from which our samples were taken. When we apply statistical tests to

our sample we are inferring what we would expect to see in the larger population. Vaux et al. [142] provides an excellent overview of the difference between technical and biological replicates and describe why one can be used in hypothesis tests and inferential statistics and the other cannot. A particularly apt description of the mistake in the blurring of technical and biological replicates is given by Bell [143].

Consider the three following experiments, each of which contains replicates. The first experiment assesses gene expression in 10 separate mice. Here we have 10 biological replicates and we expect expression to be the same in each mouse. We are measuring how much difference in expression exists in our mice. This sample of 10 mice is representative of the population of all mice and we can apply inferential statistics to infer that our conclusions apply to the population. The second experiment assesses protein expression in a cell line that has been aliquoted into 20 plate wells. Here we have 20 technical replicates; we expect protein levels to be exactly the same from the same cell line; therefore, we are measuring the variability in the measurement method. Our population here is this single cell line. The third experiment is more complicated. We have one cell line that is aliquoted into two petri dishes, one of which is treated with a potential drug. From each dish, we take three aliquots and measure gene expression of gene *Y*. Here we have three technical replicates that could be used to estimate the variability in the gene expression measurement, but we do not have any biological replicates; all measurements were made on the same sample. We may be tempted to compare gene expression between the treated samples and the untreated samples and attribute any differences we see to the treatment. However, our population remains one single cell line. Any conclusion we make from that comparison applies only to that single cell line; we cannot infer the same would be true in other cell lines.

We must be careful to understand what our replicates actually represent. Because the p-values calculated from statistical tests are based in part on the number of data points we have, we can inadvertently artificially increase our sample size and bias our conclusions by mistaking technical replicates for biological replicates.

4.3.1.3 Understanding Sources of Variability

All statistical tests assume that what we observe in our sample can be inferred in the population under study. When we understand our sample, we can be more confident that our conclusions are generalizable to the population under study. We can appropriately deal with the many sources of variability, both biological and technical, that, when ignored, can severely impact our statistical analysis and the conclusions from that analysis.

Many sources of variability are familiar to those performing omic experiments. These can include date and method of data collection or preparation, protocol used, the technician performing preparatory analyses, software used for data processing, and a host of other experiment-/platform-specific characteristics. We can account for these potential differences in our samples either statistically or technically; however, we need to have this information to do so. We can plan our experiment so that all

samples use the same preparation techniques and are performed by the same technician, or we can account for these differences through the use of covariates in our statistical tests.

Equally important are sources of variability within our study population. These may not be as familiar to those performing omic experiments but they can affect the conclusions from our experiments. When one is building a hypothesis based on a defining variable, such as the effect of a drug, both the samples receiving the drug and those not receiving the drug should be as similar to each other as possible. This becomes more difficult when the defining variable is also an inherent difference between samples, for example, assessing gene expression in individuals with a disease compared to those without. It is difficult to find affected and unaffected individuals who are similar in all other respects, making any expression differences possibly due to other factors rather than the disease. The more completely we characterize and understand the variability in our samples, the more confidence we have in the validity of our conclusions.

4.3.1.4 Adequate Sample Size

A last area of importance in study design, but one that is often the most difficult to deal with, is ensuring we have an adequate sample size for the hypothesis we are testing. All of the previously described areas of study design are important in determining what your sample size should be or what power you have given an already existing sample. Sometimes we have all of the information we need to adequately perform our sample size or power calculations; however, more often we do not. This section does not aim to discuss the theory behind sample size or power calculations or to fully describe all of the individual techniques that have been developed for specific types of experiments. It is instead intended to make the reader aware of what is necessary to complete these calculations and to impart their importance. A more thorough and easily readable discussion of sample size calculations can be found in Whitley and Ball [144]. An additional resource by Billoir et al. [145], although more technical, discusses sample size and power specifically for high-throughput experiments with a particular focus on metabolomics.

Sample size and power calculations utilize the same mathematical equations and have common elements. A checklist for the elements needed for power and/or sample size calculations include:

1. Well-defined hypothesis
2. The statistical test we will use to test our hypothesis
3. The effect size we expect to see (i.e., how much difference are we trying to detect)
4. How much variation in our outcome can we expect
5. The significance level (type I error rate) we want to test the hypothesis at
6. The power (type II error) we want/have
7. The sample size we need/have

You'll notice that the list includes both power and sample size. If one wants to determine sample size, power must be provided. If one wants to determine power, sample size must be given. One can choose to calculate the power that a sample has to test the hypothesis or calculate the number of samples needed to test the hypothesis at a given power level. We'll discuss what each of these components is and how to define them.

A well-defined hypothesis and the statistical test we will use to test it are necessary because each statistical test has its own equation to calculate sample size and/or power. The equation to calculate sample size for a *t*-test is very different than that for a chi-square test. We must know what test we will perform so that we can use the correct equation. The choice of statistical test, as described above, is determined by the hypothesis one is testing. So, our first step is to define our hypothesis and determine what test will be used. Only then can we perform the appropriate analysis to define our sample size/power.

The effect size describes how large of an effect you would define as statistically significant. In other words, how large of a difference in your dependent variable would you expect to see or define as meaningful. If you have preliminary data available, this difference may come from the difference you've observed in that data. Or you may decide that only a difference of a certain amount is meaningful; therefore, you want to have a large enough sample to detect that effect size. Imagine you are planning an experiment to test the hypothesis that expression of gene *A* is different in men and women. You may have some previous data that indicates expression in men is 10% greater than in women. This is the effect size, or difference, you would like to plan your experiment to detect. Or you may not have any previous data but consider any difference between men and women smaller than 10% to be meaningless; again 10% would be the effect size you are planning to detect. Defining the effect size can be difficult, especially if you have no preliminary data nor an idea of what would be meaningful. If this is the case, we have to improvise and make our best guess as to what the effect size will be and understand that detecting a smaller effect size requires a larger sample size. We can use literature sources of similar experiments as a guide, or we can calculate sample size or power for a range of effect sizes and hope that the effect size that ultimately exists in our sample falls within that range.

The expected variability is directly related to our effect size. It is an estimate of how much variability we expect to find in our dependent variable. Not surprisingly, the more variable our dependent variable is, the larger the sample size needed will be. Unless we have preliminary data, the expected variability can be difficult to estimate. Like effect size, we can make a best guess based on similarities in literature sources or use a range of values.

The significance level, or the type I error rate, is the level at which we want to test our hypothesis. It defines the false-positive rate that we are willing to accept and still define our conclusion as statistically significant. This value is conventionally set at 0.05 for any single statistical test; 5% is the level of uncertainty we are comfortable with. However, in omic experiments, we are rarely performing only a single statistical test and are often performing many hundreds or thousands of tests. To account

for this, we often adjust the significance level in our sample size/power calculations to account for these many tests. This phenomenon is described in more detail in the multiple testing section of this chapter.

The last elements are power and sample size. These are the values we are interested in calculating. If we want to determine what sample size is needed to test our hypothesis, we need to define the power we are comfortable with. Power, or type II error, is the rate at which we expect to conclude there is a statistically significant difference when there truly is. It is defined as 1 minus the false-negative rate (1-FN). Conventionally we perform statistical tests with 80% power; however, we are free to increase our power if warranted. Once we define these elements, we have all of the necessary information to calculate the number of samples we need to test our hypothesis at the given significance and power levels.

Sometimes we already have our sample in hand. It may be a sample of convenience that happens to be available, or we may know that we have the money and resources for a certain number of samples. In these cases, we can calculate how much power we have given our sample. This is an important step that is not only usually required in funding proposals, but one that is in our best interest to perform. Underpowered studies, those that do not have adequate power to test their hypotheses, are a common problem and often result in inconclusive results. At the end of an experiment, a nonsignificant *p*-value indicates that either there truly is no effect or the sample was not powered to detect the effect. If you know that your analysis was adequately powered, you can be confident in concluding there is no effect. If on the other hand you have an underpowered study, you're left with an inconclusive answer. Assuring that your experiment is adequately powered allows one to adequately answer the question they are interested in and assures the funding agency that the money spent on the experiment will be of use.

The table below (from [144]) shows the relationships between all of the elements that go into sample size or power calculations. As mentioned above, a smaller effect size or a dependent variable with a large level of variability requires a larger sample size. An experiment with greater power requires a larger sample size than one with lower power.

Factor	Magnitude	Impact of identification of effect	Required sample size
P-value	Small	Stringent criteria; may be difficult to achieve significance	Large
	Large	Relaxed criteria; easier to attain significance	Small
Power	Low	Identification unlikely	Small
	High	Identification more probable	Large
Effect size	Small	Difficult to identify	Large
	Large	Easy to identify	Small
Variability	Small	Easy to identify	Small
	Large	Difficult to identify	Large

4.3.1.5 Data Representation

One last topic the researcher should be aware of is the pre-data processing that has occurred prior to statistical analysis. Just as it's important to understand what hypothesis you are testing and what your replicates represent, it's also important to understand what one's data represent. Typically, omic experiments utilize data from an instrument that goes through various steps of processing and normalization before any statistical analysis is performed. Rarely is raw data collected from an instrument used in analysis. Each type of experiment, i.e., proteomic, transcriptomic, etc., has specific processing methods to yield useable data points. It is important to understand what these data points represent. Are they values in an experimental sample relative to a control sample? Are they ratios of values in treated samples to untreated samples? Before any statistical analysis is performed, the researcher needs to know exactly what the data is representing so that the appropriate hypothesis can be tested.

4.3.2 Statistical Analysis Methods

The statistical analysis methods chosen to test hypotheses depend primarily on the type of data being analyzed, the definition of the dependent variable, and the sample size. Every statistical test has underlying assumptions that must be met for the method to be appropriate and there are specific tests developed for the analysis of small sample sizes. This section will describe the most commonly used statistical tests and define the appropriate use of them.

4.3.2.1 Data Type

The first characteristic that must be defined is the type of data one is applying the statistical test to. Typically, data falls into two broad categories, continuous and categorical. These names are descriptive of the data. A continuous variable is quantitative and measured on a continuous scale and can take on any value limited only by the measurement precision. The value of a continuous variable indicates some sort of amount. For example, protein expression reported in RFUs is a continuous variable that can take on any value between a minimum and a maximum, can be measured to the precision possible by the instrument, and indicates how much of the protein is expressed. Categorical data is, on the other hand, data that describes inclusion into a category. For example, experiment batch is an example of categorical data where the data defines which category or group the data point falls into, either it is a data point that was collected in batch A or it is a data point that was collected in batch B. The value of the category is not meaningful itself, i.e., batch A is not more or less of a batch than batch B. Categorical data can be further divided into nominal (categories that have no natural order) and ordinal (categories that have

a natural order). This may seem a rather elementary description that most scientists will be familiar with, but the importance of knowing what type of data you have is sometimes forgotten. The statistical methods for the analysis of continuous data are completely different than those used for categorical data and they are, for the most part, not interchangeable. For a continuous variable, such as protein expression, it makes sense to compare mean levels between two groups. It does not make much sense, however, to compare mean levels of batch, a categorical variable, between two groups.

4.3.2.2 Dependent Versus Independent Variables

The distinction between the dependent variable and the independent variable(s) is an important element to a statistical analysis. While, mathematically, simple statistical tests between two variables, such as a Pearson correlation or a Chi-square test, do not differ based on which variable is considered dependent or independent, many statistical tests do differ. In addition, which variable is defined as the dependent variable dictates the interpretation of the statistical test.

Typically, an experiment has two classes of variables, the dependent variable that is tested and measured and the independent variables that are changed or controlled. The dependent variable is said to be "dependent" on the independent variable; as the independent variable changes, it exerts an effect on the dependent variable which is measured. We can say that the independent variable is the "input" to a statistical model which describes the change in the "output" or the dependent variable. This may seem an unnecessary complication for analysis, but this distinction defines how the results are interpreted. For an example of how interpretation is impacted, consider a simple experiment where one has a group of mice of varying ages and the interest is to evaluate if the expression of gene *ABC* changes as the mice get older. If expression of gene *ABC* is defined as the dependent variable and age as the independent variable, a correlation or linear regression would be used to assess the relationship. The conclusion, assuming a significant *p*-value, would indicate that age exerts an effect on gene expression; the gene expression level *depends* on how old the mouse is. If, on the other hand, age is defined as the dependent variable and expression as the independent variable, the same correlation or linear regression would be used. The conclusion from this analysis however, assuming a significant *p*-value, would indicate expression exerts an effect on age; the age of the mouse *depends* on what the expression of gene *ABC* is. These are two closely related conclusions, but they are saying something quite different. One is implying that the expression of gene *ABC* changes as the mice get older; the other is implying that age is dictated by the expression of gene *ABC*.

In more complicated statistical analyses where one has more than two variables, the definition of dependent/independent variables will mathematically change the statistical test. Here it is critically important to define the dependent variable appropriately so that the statistical test used is testing the hypothesis correctly. Before a choice of a statistical test is made, make sure that the dependent variable

is defined correctly. Multiple regression models, where one has a single dependent variable and two or more independent variables, treat the dependent and independent variables differently. The model estimates calculated as part of a multiple regression model take into account all independent variables in the model and calculate their cumulative effect on the dependent variable.

4.3.2.3 Parametric Versus Nonparametric

The commonly used statistical tests described here fall into two groups, those that are parametric and those that are nonparametric. There are semi-parametric models that do not fit this paradigm; however, they are beyond the scope of this discussion. Parametric tests are those that make assumptions about the population distribution from which the sample is drawn; nonparametric tests do not make this assumption. Parametric tests use the data points as measured or after a systematic transformation has been applied, whereas nonparametric tests use the rank order of the data points (which value is the smallest, which is the next smallest, and so on) and ignore the distance between the points. Each parametric statistical test requires that the distribution assumption is met. If, and only if, one can show their data is drawn from the assumed population, the statistical test is valid.

When performing a statistical test that relies on an underlying distribution, such as the common student's *t*-test, the test is only valid if your sample has been drawn from a normal distribution. Applying a *t*-test to a highly skewed dependent variable violates the assumption of the test which can lead to either a reasonable conclusion or a biased conclusion. However, the difficulty is one doesn't know which. Before applying a statistical test, look at the dependent variable graphically and determine if it meets the assumptions. If it does not, a data transformation (i.e., log, square root, reciprocal transformation) can be applied, or a nonparametric test can be used.

4.3.2.4 Common Statistical Tests

When setting out to perform a statistical analysis, answer the three following questions of your data. First, what is the dependent variable and what is/are the independent variable(s)? Second, what type of data are the dependent and independent variables? Third, can we show that the dependent variable is drawn from a specific distribution, or do we not want to make that assumption? Once these three questions are answered, the choice of the appropriate statistical test is straightforward.

The table below is intended as a guide to the most commonly used statistical tests in omic experiments and those tests likely to be available using nonstatistical software. It is not an exhaustive list and analyses that require more complicated methods or designs should be performed with the guidance of a statistician.

Underlying distribution of dependent variable	Dependent variable	Independent variable	Test	Use
Normal	Continuous	Continuous	Pearson correlation	Assess the linear relationship between two continuous variables
Normal	Continuous	Categorical with two levels	Independent *t*-test for equal variances	Compare means between two independent groups; variances equal in both groups
Normal	Continuous	Categorical with two levels	Welch *t*-test for unequal variances	Compare means between two independent groups; variances unequal
Normal	Continuous	Categorical with three or more levels	ANOVA	Compare means between three or more independent groups
Normal	Continuous	Paired categorical with two levels	Paired *t*-test	Compare means between two paired groups (often pre/post)
Normal	Continuous	Paired categorical with three or more levels	Repeated measures ANOVA	Compare three or more means between groups (often pre/post)
Normal	Continuous	Any type/any number	Linear regression	Assess the effect the independent variable(s) exert on the dependent variable; can predict what the dependent variable would be if one knows the value for the independent variable(s)
Binomial	Categorical (two levels)	Any type/any number	Logistic regression	Assess the effect the independent variable(s) exert on the dependent variable; can predict what the dependent variable would be if one knows the value for the independent variable(s)
Binomial	Categorical (three or more levels)	Any type/any number	Ordinal or nominal regression	Assess the effect the independent variable(s) exert on the dependent variable; can predict what the dependent variable would be if one knows the value for the independent variable(s)
Chi2	Categorical	Categorical	Chi square test	Tests for an association between two categorical variables

(continued)

Underlying distribution of dependent variable	Dependent variable	Independent variable	Test	Use
Chi2	Categorical	Categorical	Fisher's exact test	Tests for an association between two categorical variables; used for small sample sizes
Chi2	Categorical	Categorical (2 levels)—Paired	McNemar's test	Tests for an association between two paired categorical variables
Chi2	Categorical	Categorical (3+ levels)—paired	Cochran's Q test	Tests for an association between two paired categorical variables where one variable has three or more levels
Binomial	Single proportion	Known proportion	Binomial test	Test deviations from a theoretically expected distribution of observations into two categories
Normal	Continuous	Continuous	Kolmogorov-Smirnov one sample test	Compares the distribution of the experimental variable to a normal distribution (normality test)
Normal	Continuous	Continuous	Kolmogorov-Smirnov two-sample test	Compares two continuous variables to determine if they arise from similar distributions
N/A	Continuous	Continuous	Spearman correlation	Assesses the linear relationship between two ranked continuous variables
N/A	Continuous	Categorical with two levels	Wilcoxon rank sum	Compare ranks between two independent groups
N/A	Continuous	Categorical with three or more levels	Kruskal-Wallis	Compare ranks between three or more independent groups
N/A	Continuous	Paired categorical with two levels	Wilcoxon sign rank	Compare ranks between two paired groups (often pre/post)
N/A	Continuous	Paired categorical with three or more levels	Freidman test	Compare three or more ranks between groups (often pre/post)

I highly recommend two reference books to guide the researcher in choosing the correct statistical test to use. These two references, [146, 147], discuss statistical tests in easy to understand language devoid of the heavy use of jargon. Pett discusses the choice of statistical methods specifically for small sample sizes, a common problem in omic experiments.

4.3.2.5 Statistical Model Complexity

Statistical models range from simple, where one is testing the relationship between two variables, to extremely complex where one can have many comparison groups, many time points, nested designs, and many covariates. As described above, it is important to make sure the sample size one has is adequate to test the hypothesis of interest. It is also important to make the best use of the number of samples one has, often a small number.

Consider one wants to look at expression levels between two groups of mice, treated and untreated, and also want to see if there is a sex difference. Analyze the data using a linear regression with expression as your dependent variable and with two independent variables, sex and treatment. This will allow you test the simultaneous effect of sex and treatment, and it will allow one, through the use of a sex $\times$ treatment interaction term, to determine if expression differences between treated and untreated are different in males and females. This is a much more efficient analysis than performing a comparison in males and females separately. It allows one to perform a direct comparison between males and females.

However, a more complex statistical model can require a larger sample size to adequately test the hypothesis. If one has many treatment groups or many doses of one treatment and are interested in evaluating the change in expression over many time points, a small sample size of six mice is probably not adequate. Here we must balance the need for using our data in the most efficient manner (using a more complex model) with the need for a model that can be adequately supported by our sample size.

4.3.2.6 Multiple Testing

The issue of multiple testing and the need to account for it is not a new phenomenon; however, it has gained a new prominence with the rise of high-throughput experiments where thousands, and sometimes millions, of statistical tests are performed on the same samples. Whereas before we concerned ourselves with the testing of 10 or 20 outcomes, now we are concerned with a number of statistical tests that is many orders of magnitude larger.

Each statistical test produces a p-value which most researchers are familiar with. Using the conventional significance level of 0.05 as our decision point, a p-value ≤ 0.05 will lead the researcher to reject the null hypothesis and conclude there is a significant effect; a p-value >0.05 will lead the researcher to accept the null hypothesis and conclude there is no evidence of a significant effect. In layman's terms, the p-value defines how likely the conclusion to accept or reject the null hypothesis has occurred by chance. A p-value of 0.05 indicates that there is a 1 in 20 chance that the conclusion occurred by chance (false positive) and a 19 in 20 chance that the conclusion is true. The smaller the p-value is, the less likely it occurred by chance. This concept applies to each statistical test performed.

Consider that for a single significance test performed at the conventional 0.05 level, we have a 5% chance (or a 1 in 20 chance) that our conclusion to reject the null hypothesis occurred by chance alone (is a false positive). Because of a mathematical property of probabilities, the likelihood of this conclusion occurring by chance (a false positive) increases with the number of tests performed. So that if one performs 20 statistical tests, there is a 64% chance of at least one conclusion being a false positive. This 64% of having at least one false positive is obviously much greater than 5%. Now consider when we perform thousands of statistical tests; the likelihood that at least one of our conclusions is false approaches 100%. Unfortunately, we cannot differentiate which of our significant conclusions are in fact false positives and which are true findings.

To combat this problem, much effort has been put into methods to control the error in an analysis [148] and these methods consider two different ways to solve this problem. The first methods control the family-wise error rate (FWER), the probability of finding one or more false positives among all of the statistical tests performed. The most common method, the Bonferroni method, adjusts the overall error rate to control the probability of at least one false positive *overall* rather than for each individual statistical test. Most researchers are probably familiar with the Bonferroni correction where the *p*-value used as our cutoff for accepting or rejecting the null hypothesis is adjusted by the number of hypotheses tested. For example, if the cutoff (i.e., significance level) for a single statistical test is 0.05, then the new cutoff for 10 statistical tests is 0.05/10 or 0.005. Now, only *p*-values that are $\leq$0.005 would be evidence to reject the null hypothesis and conclude there is a significant finding. Unfortunately, the Bonferroni correction specifically and the adjustment of the FWER generally are usually considered too conservative for most omic experiments [149] for reasons that go beyond the scope of this chapter. A second methodology, control of the false discovery rate (FDR), has been developed which controls the expected proportion of false positive among all of the significant findings (i.e., where the null hypothesis has been rejected). Many different methods have been developed to control each of these error rates (FWER and FDR) and they have been refined to resolve specific limitations. A review of the commonly used methods for both control of the FWER and the FDR can be found in Chen [150].

In any omic experiment where we are performing multiple statistical tests, as in testing the expression of thousands of genes, we need to account for multiple testing. If we do not, we cannot define which, if any, of our genes yielding a significant *p*-value are true findings and which are false positives.

4.3.2.7 Limitations and Future Directions

This chapter is intended as a resource to assist in the design and analysis of omic experiments; however, it is not a comprehensive guide. It introduces a high-level view of the bioinformatic techniques commonly used and an overview of the statistical methods typically used for omic experiments, but it does not discuss in detail the underlying mathematics or probability theory. It is advised that one

consults with a statistician early in the experimental process, preferably during study design, to ensure that one is able to test the hypothesis of interest with the data collected. In addition, the guidance of a statistician is highly recommended for complex study designs as the methods used for their analysis go beyond the common methods discussed here. Some additional resources are given as references [151–161]. These resources describe statistical tests and their interpretation from a non-statistician perspective and include discussions of statistical analyses specifically for omic experiments.

Understanding the purpose of your experiment will determine what type of statistics are best suited and the inferential statistics described in this chapter may not even be needed. The purpose of inferential statistics is to make broad conclusions about a population from a smaller sample through the calculation of a *p*-value. If your purpose differs, maybe to describe a sample rather than make inferences about the broader population, the calculation of a *p*-value is unwarranted and unnecessary. If you have only technical replicates, all originating from the same biological sample, inferential statistics are unnecessary. Here any conclusions can only inform you about the population which consists of a single biological sample.

The statistical analysis of omic experiments is an evolving field where consensus on the most appropriate methods can be hard to find. As research continues, improved methods for the adjustment of multiple testing better suited to related outcomes will be developed and better methods of data integration from various sources and platforms will be defined. However, there remain several challenges in the analysis of omic data, whether it be integrating molecular and clinical data or adequately testing hypotheses so that the resulting conclusions are relevant for the population.

References

1. Hood, L., & Galas, D. (2003). The digital code of DNA. *Nature, 421*(6921), 444–448.
2. Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics, 122*(6), 565–581.
3. Levy, S. E., & Myers, R. M. (2016). Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics, 17*(1), 95–115.
4. Reis-Filho, J. S. (2009). Next-generation sequencing. *Breast Cancer Research, 11*(S3), S12.
5. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research, 38*(6), 1767–1771.
6. Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research, 8*(3), 175–185.
7. Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research, 8*(3), 186–194.
8. Andrews, S. (2010). *FastQC a quality control tool for high throughput sequence data*. Retrieved November 25, 2018 from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
9. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal, 17*(1), 10.

10. Joshi, N. A., & Fass, J. N. (2011). *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files*.
11. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754–1760.
12. Adjeroh, D., Bell, T., & Mukherjee, A. (2008). *The Burrows-Wheeler transform: Data compression, suffix arrays, and pattern matching*. New York: Springer.
13. Lam, T. W., Sung, W. K., Tam, S. L., Wong, C. K., & Yiu, S. M. (2008). Compressed indexing and local alignment of DNA. *Bioinformatics, 24*(6), 791–797.
14. Li, H., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25* (16), 2078–2079.
15. McKenna, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297–1303.
16. Garrison, E., & Marth, G. (2016). *Haplotype-based variant detection from short-read sequencing*.
17. Kobayashi, M., et al. (2017). Heap: A highly sensitive and accurate SNP detection tool for low-coverage high-throughput sequencing data. *DNA Research, 24*(4), 397–405.
18. Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology, 3*, 92.
19. Chen, K., et al. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods, 6*(9), 677–681.
20. Korbel, J. O., et al. (2009). PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology, 10*(2), R23.
21. Lee, S., Hormozdiari, F., Alkan, C., & Brudno, M. (2009). MoDIL: Detecting small indels from clone-end sequencing with mixtures of distributions. *Nature Methods, 6*(7), 473–474.
22. Magi, A., Tattini, L., Pippucci, T., Torricelli, F., & Benelli, M. (2012). Read count approach for DNA copy number variants detection. *Bioinformatics, 28*(4), 470–478.
23. Magi, A., et al. (2013). EXCAVATOR: Detecting copy number variants from whole-exome sequencing data. *Genome Biology, 14*(10), R120.
24. Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research, 21*(6), 974–984.
25. Schröder, J., et al. (2014). Socrates: Identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics, 30*(8), 1064–1072.
26. Karakoc, E., et al. (2012). Detection of structural variants and indels within exome data. *Nature Methods, 9*(2), 176–178.
27. Earl, D., et al. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research, 21*(12), 2224–2241.
28. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics, 44*(2), 226–232.
29. Nijkamp, J. F., van den Broek, M. A., Geertman, J.-M. A., Reinders, M. J. T., Daran, J.-M. G., & de Ridder, D. (2012). De novo detection of copy number variation by co-assembly. *Bioinformatics, 28*(24), 3195–3202.
30. Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics, 28*(18), i333–i339.
31. Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology, 15*(6), R84.
32. Wong, K., Keane, T. M., Stalker, J., & Adams, D. J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biology, 11*(12), R128.
33. Jeffares, D. C., et al. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications, 8*, 14061.

34. English, A. C., et al. (2015). Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genomics, 16*(1), 286.
35. Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164.
36. Sherry, S. T., et al. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research, 29*(1), 308–311.
37. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The database of genomic variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research, 42*(Database issue), D986–D992.
38. Landrum, M. J., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research, 46*(D1), D1062–D1067.
39. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2018). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research, 47*(D1), D886–D894.
40. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics, 46*(3), 310–315.
41. Cingolani, P., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly, 6*(2), 80–92.
42. Cingolani, P., et al. (2012). Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics, 3*, 35.
43. Geoffroy, V., et al. (2018). AnnotSV: An integrated tool for structural variations annotation. *Bioinformatics, 34*(20), 3572–3574.
44. Freeman, W. M., Walker, S. J., & Vrana, K. E. (1999). Quantitative RT-PCR: Pitfalls and potential. *BioTechniques, 26*(1), 112–125.
45. Bumgarner, R. (2013). Overview of DNA microarrays: Types, applications, and their future. *Current Protocols in Molecular Biology, 101*(1), 22–21.
46. Solomon, M. J., Larsen, P. L., & Varshavsky, A. (1988). Mapping protein-DNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell, 53*(6), 937–947.
47. Van Gelder, R. N., von Zastrow, M. E., Yool, A., Dement, W. C., Barchas, J. D., & Eberwine, J. H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences of the United States of America, 87*(5), 1663–1667.
48. Shalon, D., Smith, S. J., & Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research, 6*(7), 639–645.
49. Ritchie, M. E., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47.
50. Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). Affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics, 20*(3), 307–315.
51. Dunning, M. J., Smith, M. L., Ritchie, M. E., & Tavare, S. (2007). Beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics, 23*(16), 2183–2184.
52. Bolstad, B. M., Irizarry, R., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics, 19*(2), 185–193.
53. Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics, 26*(19), 2363–2367.
54. Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., & Liaw, A. (2009). gplots: Various R programming tools for plotting data. *R Packag. version 2*.
55. Student. (1908). The probable error of a mean. *Biometrika*. Retreived May 07, 2016, from http://seismo.berkeley.edu/~kirchner/eps_120/Odds_n_ends/Students_original_paper.pdf.

56. Fisher, R. A. (1919). XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh, 52*(02), 399–433.
57. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1), 289–300.
58. Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8*, 3–62.
59. Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics, 19*(R2), R227–R240.
60. Mikheyev, A. S., & Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources, 14*(6), 1097–1102.
61. Eisenstein, M. (2012). Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology, 30*(4), 295–296.
62. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*(4), R36.
63. Trapnell, C., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols, 7*(3), 562–578.
64. Trapnell, C., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28*(5), 511–515.
65. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357–359.
66. Ferragina, P., & Manzini, G. (2001). An experimental study of a compressed index. *Information Sciences, 135*(1–2), 13–28.
67. Dobin, A., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15–21.
68. Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods, 12*(4), 357–360.
69. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols, 11*(9), 1650–1667.
70. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology, 33*(3), 290–295.
71. Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., & Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature Biotechnology, 33*(3), 243–246.
72. Wang, L., Wang, S., & Li, W. (2012). RSeQC: Quality control of RNA-seq experiments. *Bioinformatics, 28*(16), 2184–2185.
73. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods, 5*(7), 621–628.
74. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics, 26*(4), 493–500.
75. Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*(1), 323.
76. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1976). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.
77. Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics, 31*(2), 166–169.
78. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics, 30*(7), 923–930.

79. Lawrence, M., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology, 9*(8), e1003118.
80. Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research, 4*, 1521.
81. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139–140.
82. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550.
83. Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, 135*(3), 370.
84. Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16*(2), 117–186.
85. Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X., & Zhang, Y. (2012). GFOLD: A generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics, 28*(21), 2782–2788.
86. Tarazona, S., et al. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research, 43*(21), e140.
87. Toedling, J., & Huber, W. (2008). Analyzing ChIP-chip data using bioconductor. *PLoS Computational Biology, 4*(11), e1000227.
88. Toedling, J., Sklyar, O., & Huber, W. (2007). Ringo – an R/bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics, 8*(1), 221.
89. Durinck, S., et al. (2005). BioMart and bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics, 21*(16), 3439–3440.
90. Alexa, A., Rahnenfuhrer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics, 22*(13), 1600–1607.
91. Zhang, Y., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology, 9*(9), R137.
92. Xu, S., Grullon, S., Ge, K., & Peng, W. (2014). Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods in Molecular Biology, 1150*, 97.
93. Hayatsu, H. (2008). Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis – a personal account. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences, 84*(8), 321–330.
94. Morris, T. J., et al. (2014). ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics, 30*(3), 428–430.
95. Tian, Y., et al. (2017). ChAMP: Updated methylation analysis pipeline for illumina BeadChips. *Bioinformatics, 33*(24), 3982–3984.
96. Aryee, M. J., et al. (2014). Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics, 30*(10), 1363–1369.
97. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics, 28*(6), 882–883.
98. Carson Sievert, P. T. I., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2018). Create interactive web graphics via 'plotly.js' [R package plotly version 4.8.0]. *Comprehensive R Archive Network (CRAN)*.
99. Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics, 27*(11), 1571–1572.
100. Chen, P.-Y., Cokus, S. J., & Pellegrini, M. (2010). BS Seeker: Precise mapping for bisulfite sequencing. *BMC Bioinformatics, 11*(1), 203.

101. Kreck, B., Marnellos, G., Richter, J., Krueger, F., Siebert, R., & Franke, A. (2012). B-SOLANA: An approach for the analysis of two-base encoding bisulfite sequencing data. *Bioinformatics, 28*(3), 428–429.
102. Frith, M. C., Mori, R., & Asai, K. (2012). A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Research, 40*(13), e100.
103. Saito, Y., Tsuji, J., & Mituyama, T. (2014). Bisulfighter: Accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Research, 42*(6), e45.
104. Xi, Y., & Li, W. (2009). BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics, 10*(1), 232.
105. Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., & Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. *Nature Methods, 11*(11), 1138–1140.
106. Saito, Y., & Mituyama, T. (2015). Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions. *BMC Genomics, 16*(Suppl 12), S3.
107. Song, Q., Decato, B., Hong, E. E., Zhou, M., & Fang, F. (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One, 8*(12), 81148.
108. Hansen, K. D., Langmead, B., & Irizarry, R. A. (2012). BSmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology, 13*(10), R83.
109. Hebestreit, K., Dugas, M., & Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics, 29*(13), 1647–1653.
110. Wreczycka, K., Gosdschan, A., Yusuf, D., Grüning, B., Assenov, Y., & Akalin, A. (2017). Strategies for analyzing bisulfite sequencing data. *Journal of Biotechnology, 261*, 105–115.
111. Tsuji, J., & Weng, Z. (2015). Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. *Briefings in Bioinformatics, 17*(6), bbv103.
112. Eberwine, J., et al. (1992). Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America, 89*(7), 3010–3014.
113. Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine, 50*(8), 96.
114. Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.
115. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology, 36*(5), 411–420.
116. Afgan, E., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research, 46*(W1), W537–W544.
117. Ashburner, M., et al. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics, 25*(1), 25–29.
118. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research, 37*(1), 1–13.
119. Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society, 85*(1), 87.
120. Ludbrook, J. (2008). Analysis of 2×2 tables of frequencies: Matching test to experimental design. *International Journal of Epidemiology, 37*(6), 1430–1435.
121. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44–57.
122. Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics, 23*(2), 257–258.
123. Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: A cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics, 21* (16), 3448–3449.

124. Subramanian, A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102*(43), 15545–15550.
125. Lee, H. K., Braynen, W., Keshav, K., & Pavlidis, P. (2005). ErmineJ: Tool for functional analysis of gene expression data sets. *BMC Bioinformatics, 6*(1), 269.
126. Al-Shahrour, F., et al. (2007). From genes to functional classes in the study of biological systems. *BMC Bioinformatics, 8*, 114.
127. Nam, D., Kim, S.-B., Kim, S.-K., Yang, S., Kim, S.-Y., & Chu, I.-S. (2006). ADGO: Analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics, 22*(18), 2249–2253.
128. Nogales-Cadenas, R., et al. (2009). GeneCodis: Interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Research, 37*(Web Server issue), W317–W322.
129. Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27–30.
130. Finn, R. D., et al. (2014). Pfam: The protein families database. *Nucleic Acids Research, 42* (Database issue), D222–D230.
131. Matys, V., et al. (2003). TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research, 31*(1), 374–378.
132. Warde-Farley, D., et al. (2010). The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research, 38* (Web Server issue), W214–W220.
133. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: A general repository for interaction datasets. *Nucleic Acids Research, 34*(Database issue), D535–D539.
134. Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology, 4*(1), Article17.
135. Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics, 9*(1), 559.
136. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics, 32*(19), 3047–3048.
137. Gregory, R., Warnes, R., Bolker, B., Bonebakker, L., Gentleman, M., Liaw, W. H. A., Lumley, T., Maechler, B., Magnusson, A., Moeller, S., Schwartz, M., & Venables, B. (2016). Various R programming tools for plotting data. *R Package Version, 2*(4), 1.
138. Walter, W., Sánchez-Cabo, F., & Ricote, M. (2015). GOplot: An R package for visually combining expression data with functional analysis. *Bioinformatics, 31*(17), 2912–2914.
139. Ghosh, D., & Poisson, L. M. (2009). "Omics" data and levels of evidence for biomarker discovery. *Genomics, 93*, 13–16.
140. Wheelock, A. M., & Wheelock, C. E. (2013). Trials and tribulations of 'omics data analysis: Assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine. *Molecular BioSystems, 9*, 2589.
141. Kraus, L. (2015). Editorial: Would you like a hypothesis with those data? Omics and the age of discovery science. *Molecular Endocrinology, 29*(11), 1531–1534.
142. Vaux, D. L., Fidler, F., & Cumming, G. (2012). Replicates and repeats—What is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO Reports, 13*(4), 291.
143. Bell, G. (2016). Comment: Replicates and repeats. *BMC Biology, 14*, 28.
144. Whitley, E., & Ball, J. (2002). Statistics review 4: Sample size calculations. *Critical Care, 6* (4), 335.
145. Billoir, E., Navratil, V., & Blaise, B. J. (2015). Sample size calculation in metabolic phenotyping studies. *Briefings in Bioinformatics, 16*(5), 813–819.
146. Urdan, T. C. (2010). *Statistics in plain English* (3rd ed.). New York: Routledge.

147. Pett, M. A. (1997). *Nonparametric statistics for health care research: Statistics for small samples and unusual distributions*. Thousand Oaks, CA: Sage.
148. Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B, 64*(Part 3), 479–498.
149. Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology, 2*, 8.
150. Chen, S. Y., Feng, Z., & Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease, 9*(6), 1725–1729.
151. Forshed, J. (2017). Experimental design in clinical 'omics biomarker discovery. *Journal of Proteome Research, 16*, 3954–3960.
152. Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H., & Walter, S. (1995). Basic statistics for clinicians: 1. Hypothesis testing. *CMAJ, 152*(1), 27–32.
153. Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H., & Walter, S. (1995). Basic statistics for clinicians: 2. Interpreting study results: Confidence intervals. *CMAJ, 152*(2), 169–173.
154. Guyatt, G., Walkter, S., Shannon, H., Cook, D., Jaeschke, R., & Heddle, N. (1995). Basic statistics for clinicians: 4. Correlation and regression. *CMAJ, 152*(4), 497–504.
155. Hanley, J. A., & Moodie, E. E. M. (2011). Sample size, precision and power calculations: A unified approach. *Journal of Biometrics and Biostatistics, 2*, 5.
156. Ioannidis, J. P. A., Tarone, R., & McLaughlin, J. K. (2011). The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology, 22*(4), 450–456.
157. Jarschke, R., Guyatt, G., Shannon, H., Walter, S., Cook, D., & Heddle, N. (1995). Basic statistics for clinicians: 3. Assessing the effects of treatment: Measures of association. *CMAJ, 152*(3), 351–357.
158. Mazzocchi, F. (2015). Could big data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Reports, 16*(10), 1250–1255.
159. Rajasundaram, D., & Selbig, J. (2016). More effort — More results: Recent advances in integrative 'omics' data analysis. *Current Opinion in Plant Biology, 30*, 57–61.
160. Senn, S., & Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics, 6*, 161–170.
161. Signe, A., Esteban, F. J., Stavreus-Evers, A., Simon, C., Giudice, L., Lessey, B. A., Horcajadas, J. A., Macklon, N. S., D'Hooghe, T., Campoy, C., Fauser, B. C., Salamonsen, L. A., & Salumets, A. (2014). Guidelines for the design, analysis and interpretation of 'omics' data: Focus on human endometrium. *Human Reproduction Update, 20*(1), 12–28.