Biotechnological Advances in Methods for Functional Analysis of Genes

Juliette Rose Ongus
Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI). P.O. Box 62000, 00200 Nairobi

Abstract
This review was written in an attempt to present the reader with a decent spectrum of the available methods for gene function analysis that have been applied in the past decade. Knowledge emanating from the functional analysis of genes has applications in the fields of genetics and genomics, medical diagnostics, the pharmaceutical industry and in plant and animal biotechnology. DNA sequencing provides the primary data for the functional analysis of genes by determining the sequence order of nucleic acid residues of a DNA molecule. Computational tools are then used for characterization of genes, prediction of function, establishing structural and physiochemical properties of proteins, phylogenetic analyses, and performing simulations of the cellular interactions of biomolecules. Gene expression is done to analyse promoter activity, detect RNA transcript levels, monitor protein expression and post-translational modification. The most common purpose of a gene expression study is to find statistically differentially expressed genes. One way to understand the function of a gene is to observe a biological system that lacks that gene. Several techniques have been developed to alter a gene sequence to result in an inactivated gene. The emergence of genome-editing technologies has provided new tools for introducing sequence-specific modifications into genomes.

Keywords: Gene function, Mutation, Gene expression, Genome editing

1. INTRODUCTION

Genes are made up of DNA. Genes are the basic physical and functional unit of heredity. In cells, a gene is a portion of DNA that contains both coding sequences that determine what the gene does, and non-coding sequences that determine when the gene is active (Griffiths *et al.* 2015). Their function is to hold all the information required to make and regulate the expression of all the different proteins in cells. Genes act as instructions to the production of the proteins in the organisms' cell and control what protein is made within the cell, which can affect the organism's phenotype or outward appearance (Griffiths *et al.* 2015). In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. There are both the RNA genes and protein-coding genes in the human genome (International Human Genome Sequencing Consortium, 2001). The functions of more than 50% of the discovered genes are unknown, which means that scientists must continue to work to find out what these genes do.

Knowledge emanating from the functional analysis of genes has applications in the fields of genetics and genomics, medical diagnostics, the pharmaceutical industry and in plant and animal biotechnology (Sitnicka *et al.* 2010). Functional genomics can be initiated once large-scale sequence data is made available (Griffiths *et al.* 2015). Understanding the function of a particular gene is a multistep process. The primary activity in functional genomics is the identification of coding sequences within a genome. This is supported by bioinformatics tools, which are used to predict genes. This is then followed with the analysis of gene products to measure gene and protein expression patterns that are in turn liked to phenotypes, which offer visual or quantifiable observations leading to the recognition of the functions fulfilled by the genes (Sitnicka *et al.* 2010). The final step in functional analysis involves system perturbation where the gene in question is inactivated.

Over the past few decades, several methods for the analysis of gene function have been developed. These developments have been accelerated by advancements in automation, computerization and molecular biology and biotechnology techniques (Sitnicka *et al.* 2010). The aim of this article is to present biotechnological developments made so far in the methods for the functional analysis of genes. The function of genes can therefore either be predicted by use of computational methods or demonstrated experimentally using wet laboratory procedures.

2. DNA SEQUENCING TECHNOLOGIES AND THEIR APPLICATION IN THE ANALYSIS OF GENE FUNCTION

DNA is the information store that ultimately dictates the structure of every gene product. DNA sequencing provides the primary data for the functional analysis of genes. Determining the sequence order of nucleic acid residues of a DNA molecule is the most comprehensive way of obtaining genetic information about any living organism (Berglund *et al.* 2011; Heather & Chain 2016). Coupled with bioinformatics applications it is possible to analyse the nucleic acid sequence data to infer or predict the function of genes. The term DNA sequencing refers to applications used to determine the order of the nucleotide bases adenine, guanine, cytosine and thymine in a molecule of DNA. There are two major types of sequencing applications, *de novo* sequencing and resequencing. In *de novo* sequencing, the genome of an organism is sequenced for the first time. In contrast, in resequencing, a reference sequence is already available in the database (Berglund *et al.* 2011). Advances in sequencing technologies have enabled population genetics based on the complete genomic sequences of a large number of individuals (Berglund *et al.* 2011).

Sanger sequencing, which is based on DNA chain termination with a small concentration of radio- or fluorescently-labeled di-deoxy nucleotide triphosphate (dNTPs) molecules followed by size separation by gel electrophoresis, is the gold standard for sequencing technology in that it provides a high degree of accuracy, long-read capabilities, and the flexibility to support a diverse range of applications in many research areas (Dewey *et al.* 2012). Sanger sequencing is suited for sequencing short segments of DNA and confirming of Next Generation Sequencing (NGS) output (Dewey *et al.* 2012). Sanger Sequencers generate long reads, which enable the identification of protein-coding regions in infectious disease and are applicable for metagenomic analyses (Rothberg & Leamon 2008).

The scientific community demanded an increase the throughput of DNA sequencing. Therefore, newer technologies that allow rapid sequencing of large amounts of DNA and have the capability to generate high-throughput data have to be continuously developed. From 2005, a “second-generation” of sequencing technologies (referred to as Next generation sequencing (NGS) technologies) became available. These have provided unprecedented opportunities for high-throughput functional genomic research. The technologies include the 454 Sequencer (Roche), Illumina Genome Analyzer and SOLiD system (Applied Biosystems). These technologies are applied in whole-genome sequencing, targeted resequencing, discovery of transcription factor binding sites, and noncoding RNA expression profiling (Morozova & Marra 2008). The advantage of the NGS technologies is that they are able to generate higher volumes of sequence data with a fast turnaround time at a much lower cost that is achieved by Sanger sequencing (Berglund *et al.* 2011). This in turn has revolutionized genomics and has led to a significant increase in the number of genome sequencing projects. NGS is best for examining hundreds of genes at a time or sequencing samples with a low amount of starting material.

Of the NGS platforms that are currently commercially available, the 454 sequencer was the first to reach the market (Rothberg & Leamon 2008). It utilises pyrosequencing, which is based on the detection of light emitted by secondary reactions initiated by the release of pyrophosphate whenever a nucleotide is incorporated (Rothberg & Leamon 2008). The advantages of the 454 sequencer include long reads and short run time. The major disadvantage is that it has the highest cost per base of any of the NGS systems. (Rothberg & Leamon 2008). The second technology brought to market after the 454 was the Genome Analyzer conceived by Solexa (Cambridge, UK) and commercialized by Illumina (Hayward, CA, USA) (Rothberg & Leamon 2008). The Solexa system operates via a sequencing-by-synthesis process that incorporates base-specific fluorescently labelled “end-blocked nucleotides,” which do not allow further DNA polymerization into immobilized template strands. The Illumina platform has a higher yield of data than the 454 sequencer. The Illumina platform is widely used for a variety of applications, including human whole genome and exome variant discovery and transcriptome sequencing (RNAseq) (Berglund *et al.* 2011). The SOLiD (Applied Biosystems by Life Technologies) sequencing process differs from the 454 and Illumina methods in that it relies upon sequencing by ligation in which the sequence of a DNA template is read by competitive ligation of 2-base probes to the nascent DNA strand (Rothberg & Leamon 2008). Advantages include high throughput and Inherent error correction, both of which make the platform suitable for human whole genome and exome variant discovery (Berglund *et al.* 2011). The short read-lengths of both Illumina and SOLiD, coupled with the decreased sequencing costs afforded by the high-read densities make these two technologies ideal for applications such as sequence-based expression analysis and promoter binding site studies (Rothberg & Leamon 2008).

Unlike NGS platforms, which produces short reads a few hundred base-pairs long, “third” generation technologies have been used to produce highly accurate de novo assemblies with unprecedented lengths of sequence reads with over 10,000 bp reads or map over 100,000 bp molecules (Bleidorn 2016; Zhou *et al.* 2016). Increased read lengths can be used to address long-standing problems in de novo genome assembly. Third generation sequencing instruments negate the requirement for DNA amplification. By foregoing this step, these technologies avoid PCR-introduced error and amplification bias, and may be superior for high-throughput sequencing applications, such as transcriptome sequencing (“RNAseq”), that depend on accurate quantification of relative DNA or RNA fragment abundance. (Heather & Chain 2016). The first of these single-molecule sequencing technologies is the Helicos Heliscope (Helicos BioSciences). The Helicos chemistry worked in the same manner that Illumina does, but without any bridge amplification (Heather & Chain 2016). It produced relatively short reads. The Pacific Biosciences offers a platform (referred to as ‘‘PacBio sequencing”) for single-molecule, real-time sequencing with longer read lengths (Rhoads & Au 2015). The highly contiguous de novo assemblies can close gaps in reference assemblies and characterize structural variation in individual genomes. With longer reads, it becomes possible to sequence through extended repetitive regions and detect mutations (Rhoads & Au 2015). PacBio sequencing provides information for the detection of base modifications, such as methylation. Nanopore DNA strand sequencing has emerged as a competitive, portable technology. Nanopore sequencing detects base-specific changes in ionic flux as DNA traverses small pores in solid surfaces that are placed in an electric field. As the DNA passes through the pore, a sensor detects ionic current changes caused by differences in the shifting nucleotide sequences (Deamer *et al.* 2016; Bayley 2015). Reads exceeding 150 kb have been achieved. The Oxford Nanopore MinION (a portable sequencing device the size of a cell phone) is the dominant platform currently available (Jain *et al.* 2016). Nanopore technology can detect modifications on individual nucleotides. The technology can achieve read lengths of more than 50 kb with high read accuracies (Jain *et al.* 2016).

3. PREDICTION OF GENE FUNCTION USING COMPUTATIONAL METHODS

3.1 Genomic resources and bioinformatics tools for gene function analysis

Experimental approaches for the analysis of gene function cannot scale up to accommodate the vast amount of sequence data available due to its inherent difficulty (which is expensive, time consuming and tedious) (Radivojac *et al.* 2013). Previously, in the absence of direct experimental demonstration, homology-based protein function prediction was used as the gold standard for *in silico* analysis and prediction of protein function (Grant 2011). Computational tools are therefore routinely used for characterization of genes, prediction of function, establishing structural and physiochemical properties of proteins, phylogenetic analyses, and performing simulations of the cellular interactions of biomolecules (Mehmood *et al.* 2014). Although these tools cannot generate information as reliable as experimentation, they can still facilitate informed decision for conducting costly experimentation (Mehmood *et al.* 2014). For computational predictions to be reliable, it is crucial that their accuracy be high (Radivojac *et al.* 2013). Among the methods available are methods tools for primary sequence analyses (such as gene identification and sequence analyses from primary databases), Predicting Protein Structure and Function using Protein Sequence Databases (Mehmood *et al.* 2014). More tools for analysing genomes, proteomes, predicting structures, rational drug designing and molecular simulations are still being developed. As sequencing technologies advance, biologists are slowly drowning in their data and new tools are required to perform the "downstream" analyses. The availability of high-throughput experimental data of genomic sequences from thousands of species has created new opportunities for function prediction (Radivojac ***et al.* 2013).**

3.2 Predicting gene function through homology

Similarity-based studies are the most widely used approach for function prediction (Grant 2011). The communal availability of multiple complete genomes sequences of diverse life forms, which are stored in databases for comparative analysis, provides a new perspective to genome analysis and allows for comparison with homologous sequences for relationships between genes (Koonin 2005). An analysis showing similarity with the recognised genes is helpful in the initial determination of a probable gene product and its function (Sitnicka *et al.* 2010). This method of predicting through homology relies on the assumption that if a newly sequenced gene is highly sequence-similar to an already characterized and published gene, the function of the new gene is probably similar to the experimentally verified function of the annotated homolog and that is used as the basis to infer the function of the sequence under investigation (Grant 2011; Sitnicka *et al.* 2010).

Homologs are genes sharing a common origin. There are subcategories of homologs: Orthologs are genes related via speciation (vertical descent). They originate from a single ancestral gene in their common ancestor. Orthologs occur in various species (which may also prove that they have occurred in a common ancestor) and fulfil the same or comparable functions (Koonin 2005). Xenologs are homologous genes, which are acquired by organisms through horizontal gene transfer (Poptsova & Gogarten 2007). Paralogs are genes related via duplication. They occur in various organisms or may occur in the same individual, but due to changes in structure, they have separate roles (Koonin 2005). In this case the duplication leads to divergence, that is, a division of functions. Human myoglobin and hemoglobin are examples of two paralogs responsible for the storage of oxygen in skeletal muscles and transport of oxygen between cells and pulmonary alveoli respectively (Sitnicka *et al.* 2010).

The most commonly used database with full sequences of genomes is the National Center for Biotechnology Information (NCBI) server (Ghosh & Febin 2016). The use of the BLAST algorithm on the NCBI server for homology studies is limited when the similarity between the studied sequences is low (20-30%) (Gowri & Sandhya 2006). Proteins may differ significantly at the amino acid level but can however, assume a similar structure and fulfil similar functions. Studies on evolution show that the structure of proteins is better preserved than their sequence (Ginalski *et al.* 2003). Due to this, studies on protein structure are important when determining their functions (Sitnicka *et al.* 2010).

The greatest limitation in homology-based methods is presence of uncharacterized sequences in databases (Sitnicka *et al.* 2010). Several non-homology based approaches for protein function prediction that are based on sequence features, structure, evolution, biochemical and genetic knowledge have emerged (Grant

2011).

3.3 Characterization of the proteome by open reading frame (ORF) analysis

Double stranded DNA encodes six different reading frames due to its triplet code. Three frames on one DNA strand in 5′ to 3′ direction and further three frames on the antisense strand. Gene-prediction software analyse genomic DNA sequences by examining each of the six reading frames and searches for Protein-coding segments encoded in Open Reading Frames (ORFs) delimited by the translational start codon AUG and ending with a stop codon (Grifiths *et al.* 2015; Mir *et al.* 2012). Candidates for genes are identified by ORFs of at least 100 codons. Most ORFs are completely novel, not corresponding to any familiar gene with alleles producing identifiable phenotypes. The ORFs can be analysed for function initially by using the computer to search databases to look for full or partial homology to known genes characterized in other organisms. A provisional proteome gene distribution can be deduced from such analysis (Grifiths *et al.* 2015). The genome length influences the number of ORFs it carries and the probability to observe very long ORFs. A larger genome will harbour more ORFs (Mir *et al.* 2012).

3.4 Computational Approaches for Functional Prediction and Characterisation of Noncoding RNAs

Only a small fraction of the genomes of large multicellular eukaryotes code for proteins. The rest is mostly comprised of non-protein coding DNA. The same can be said of the majority of the human transcriptome, which is defined as non-coding RNA (ncRNA) (Veneziano *et al.* 2015). The discovery of transfer RNA and ribosomal RNA in the 1950s highlighted the presence of non-coding RNAs (ncRNAs) with biological roles. A non-coding RNA (ncRNA) is a functional RNA molecule that is transcribed from DNA but not translated into proteins (Palazzo & Lee 2015). The best candidates for novel functional ncRNAs arise from only a minute fraction of the genome. A vast majority of ncRNAs has yet to be characterized thoroughly (Palazzo & Lee 2015; Signal *et al.* 2016). ncRNAs can have numerous molecular functions, including modulating transcriptional patterns, regulating protein activities, serving structural or organizational roles, altering RNA processing events, and serving as precursors to small RNAs (Wilusz *et al.* 2009). Those ncRNAs that appear to be involved in epigenetic processes can be divided into two main groups; the short ncRNAs (<30 nts) and the long ncRNAs (>200 nts). The three major classes of short non-coding RNAs are microRNAs (miRNAs), short interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs). Whole genome annotation studies have revealed that a much larger fraction of large genomes is transcribed than initially known (Mackowiak *et al.* 2015). Current computational approaches for ncRNA analysis are based on deep sequencing using next-generation sequencing (NGS) Output. A ncRNA bioinformatics analysis system has three essential components: a data analysis platform for ncRNA detection, classification and expression analysis representing the core of the system; a database for annotation information storage and for the analysis of results (Veneziano *et al.* 2015).

4. EXPERIMENTAL TECHNIQUES AND APPLICATIONS FOR ANALYSIS OF GENE FUNCTION

A gene is a locus (or region) of DNA, which is the molecular unit of heredity. Even though nearly every cell in an organism's body contains the same set of genes, only a fraction of these genes are used in any given cell at any given time. Cells in all living organisms are continually activating or deactivating genes (Griffiths *et al.* 2015). Every gene consists of functional components, each involved in a different facet of the process of gene expression. The two main functional components of a gene are: the *promoter region* and the *coding region*. The promoter region (with or without *cis*-acting elements called enhancers) controls when and in what tissue a gene is expressed. The coding region is the component of a gene that determines the amino acid sequence of the protein encoded by the gene (Griffiths *et al.* 2015). The main function of genes, therefore, is to control the synthesis of proteins in an organisms' cell. When a particular protein is required by the cell, the gene coding for that protein is activated. Genes specify the structure of proteins, which in turn are responsible for the associated phenotypic structure and function of each cell in the body. Genes are thus responsible for all inherited traits (Berg *et al.* 2015).

Genes encode proteins and proteins direct cell function. According to the central dogma of molecular biology, information flows from DNA to RNA and finally to proteins (Berg *et al.* 2015). Gene expression is the process by which genes are transcribed and translated to yield functional gene products — functional RNA species or protein products. Gene expression is a highly regulated mechanism that controls the function and adaptability of all living cells. Each step in the flow of information provides the cell with a potential control point for self-regulating its functions by adjusting the amount and type of proteins it synthesizes (Griffiths *et al.* 2015). At any given time, the amount of a particular protein in a cell reflects the balance between that protein's anabolism and catabolism biochemical pathways (Berg *et al.* 2015). Gene expression is dynamic, and the same gene may act in different ways under different circumstances. The most common purpose of a gene expression study is to find statistically differentially expressed genes (Sweeney *et al.* 2017), which are determined by comparing sample-level gene expression data between cases and controls. The study of gene function in the developmental process is determined by comparing the expression patterns between species. A particular emphasis is put on genes, which exhibit either rapid changes in expression, which have been linked to numerous interesting developmental differences, or deep conservation of expression patterns, which have been show to underlie developmental similarities on unexpectedly large evolutionary scales (Roux *et al.* 2015). The study of gene regulation provides insights into normal cellular processes, such as differentiation, and abnormal or pathological processes.

The field of gene expression analysis has undergone major advances. Several techniques now exist for studying and quantifying gene expression and its regulation: the analysis of promoter activity, detection of RNA transcript levels, monitoring protein expression and post-translational modification and gene inactivation methods.

4.1 The analysis of promoter activity

A promoter is linked to the coding region of a gene and regulates the gene's transcription, either by activating or suppressing its expression (Griffiths *et al.* 2015). *In silico* screening is used to predict promoter regions of genes. However, the analysis of promoter activity is achieved by the expression of promoter/reporter genes fusions in host cells. Instead of directly measuring the level of target gene mRNA, the promoter region of the gene of interest can be cloned in front of a detectable reporter gene such as luciferase, β-galactosidase or β-glucuronidase and measure the reporter gene expression as a reflection of the expression of the gene of interest (Fu & Xiao 2006). A reporter gene is a gene whose phenotype can easily be detected or measured quantitatively (Griffiths *et al.* 2015). Promoter activity is thus measured as the rate of transcription of the downstream reporter gene. Promoter/reporter constructs are tested in cell lines that most closely approximate the tissue that you are interested in. Gene expression is in part regulated by transcription factors that bind specific sequence motifs in genomic DNA to either upregulate or downregulate transcription. Gel shift assays are used to study protein-DNA or protein-RNA interactions. This is an electrophoretic mobility shift assay; a powerful technique to resolve nucleic acid-protein complexes formed with transcription factors in nuclear extracts (Parés-Matos 2013). DNA or RNA fragments that are tightly associated with proteins (such as transcription factors) migrate more slowly in an agarose or polyacrylamide gel (showing a positional shift). Identifying the associated sequences provides insight into gene regulation (Parés-Matos 2013).

4.2 Detection of RNA transcript levels

mRNA is the intermediary between DNA and protein in the course of gene expression (Berg *et al.* 2015). By determining which mRNA transcripts are present in a cell, it is possible to determine which genes are expressed in that cell at different stages of development and under different environmental conditions. The amount of mRNA produced correlates with the amount of protein eventually synthesized and measuring the amount of a particular mRNA produced by a given cell or tissue is often easier than measuring the amount of the final protein. Molecular characterization of any gene usually includes the analysis of temporal and spatial distribution of RNA expression. A number of widely used procedures exist for detecting and determining the abundance of a particular mRNA in a total or poly(A) RNA sample. Subsequently, this information can be used to help determine what circumstances trigger expression of various genes.

The study of the expression patterns of a few genes at a time is done on a small scale by applying techniques such as quantitative RT-PCR or in situ hybridization (Roux *et al.* 2015). Northern blot or serial analysis of gene expression (SAGE) make it possible to identify which genes are turned on and which are turned off within cells. Northern blotting is a technique where levels of mRNA are directly quantified by electrophoresis and immobilized on a membrane followed by incubation with specific probes. The RNA-probe complexes can be detected using a variety of different chemistries or radionuclide labelling.

mRNA levels can be quantified by reverse transcription of the RNA to cDNA followed by quantitative PCR (qPCR) on the cDNA. Expression levels can be measured relative to other genes (relative quantification) or against a standard (absolute quantification). Real-time PCR is the gold standard in nucleic acid quantification because of its accuracy and sensitivity. Real-time PCR can be used to quantify mRNA or miRNA expression following conversion to cDNA or to quantitate genomic DNA directly to investigate transcriptional activity.

Transcriptomics is the study of the complete set of RNA transcripts (transcriptome) that are encoded by the genome of a specific cell or organism, at a specific time or under a specific set of conditions or specific circumstances, using high-throughput methods, such as microarray analysis (Jenkinson *et al.* 2016). Comparison of transcriptomes allows the identification of genes that are differentially expressed in distinct cell populations, or in response to different environmental stimuli (Evans 2015). The study of RNA expression patterns on a genome-wide scale can be achieved using Microarrays and RNA sequencing (RNA-seq) (Roux *et al.* 2015), which is helping researchers discover novel RNA forms and variants. New technologies promise to reveal even more about RNA and make RNA-based assays common. The volume of transcriptomics data, whether from microarrays or RNA-seq experiments, is increasing exponentially in public databases (Roux *et al.* 2015). All microarray and RNA-seq datasets come from public repositories. While RNA-seq and microarrays provide genome-wide information, they often lack the fine resolution of in situ hybridizations, which are mostly small-scale (Roux *et al.* 2015).

DNA microarrays also known as biochip or DNA chip is an array of oligonucleotide probes bound to a chip surface to enable simultaneous gene expression profiling of many genes. Labelled cDNA from a sample is hybridized to complementary probe sequences on the chip, and strongly associated complexes are identified optically (Griffiths *et al.* 2015). The microarrays are used to determine expression levels across a large number of genes or to perform genotyping across different regions of a genome. High throughput transcriptomics became possible with microarrays, which detect nucleic acids in a sample by hybridization to probes on microchips. Microarrays allowed the first large-scale comparative studies of the evolution of gene expression between species (Roux *et al.* 2015). Microarrays are particularly useful for analysing large mammalian transcriptomes. DNA chips can also be used to detect mutations (Griffiths *et al.* 2015). However, microarrays detect only known sequences, so they can't be used for discovery.

Transcriptomics has expanded dramatically in the past few years because of developments in RNA sequencing (RNA-seq) (Conesa *et al.* 2016). Use of RNA-seq has exploded because of next generation sequencing (NGS), which can yield readouts of billions of bases a day from a single instrument. RNA-seq can qualitatively and quantitatively investigate any RNA type including messenger RNAs (mRNAs), microRNAs, small interfering RNAs, and long noncoding RNAs. RNA-seq analysis of RNA isoforms, which are transcribed from the same gene but have different structures, for example because of alternative splicing, are explaining how limited genomes produce complex phenotypes (Conesa *et al.* 2016). RNA-seq aids scientists working on unusual model organisms, who use the method to assemble de novo transcriptomes for organisms without sequenced genomes. Most researchers, however, are interested in differential gene expression, changes in the levels of protein-coding mRNAs in experimental samples versus controls. RNA-seq is now allowing major progresses in describing gene expression variation between species. This technique has a larger dynamic range than microarrays, and can also be used to study differences in exon usage and alternative splicing (Gallego *et al.* 2012). Importantly, it allows the study of non-model species in the absence of a sequenced genome (Grabherr *et al*. 2011; Perry *et al.* 2012), or when the genome sequence is of poor quality. The advantages of RNA-seq allow more straightforward direct comparisons of expression levels between species, and interesting insights have been provided by the first evolutionary studies using this technology (Roux *et al.* 2015). Comparative RNA-seq can be used in functional genomics (Roux *et al.* 2015).

4.3 Proteomic analysis

Proteomic analysis (proteomics) refers to the systematic identification and quantification of the proteome (the complete complement of proteins) of a biological system (cell, tissue, organ, biological fluid, or organism) at a specific point in time (Ortea *et al.* 2016). Studying proteins generates insight on how proteins affect cell processes. The biggest challenge inherent in proteomics lies in the proteome's degree of complexity compared to the genome. For example, one gene can encode more than one protein (International Human Genome Sequencing Consortium, 2001), the proteome is dynamic and is constantly changing according to different stimuli (Larance & Lamond 2015; Ortea *et al.* 2016), proteins are post-translationally modified and exist in a wide range of concentrations in the body.

Proteins can be organized in four structural levels: primary (The amino acid sequence), secondary (Local folding of the amino acid sequence into α helices and β sheets), tertiary (3D conformation of the entire amino acid sequence) and quaternary (Interaction between multiple small peptides or protein subunits to create a large unit) (Berg *et al.* 2015). Each level of protein structure is essential to the finished molecule's function. The primary sequence of the amino acid chain determines where secondary structures will form, as well as the overall shape of the final 3D conformation. The 3D conformation of each small peptide or subunit determines the final structure and function of a protein conglomerate (Griffiths *et al.* 2015).

There are different subdivisions of proteomics, including: structural proteomics (analysis of protein structure) (Manjasetty *et al.* 2012), expression proteomics (analysis differential expression of proteins) (Chernobrovkin *et al.* 2015) and interaction proteomics (characterization of protein complexes) (Völkel *et al.* 2010). Several tools are available for conducting proteomics analysis.

Mass spectrometry is the technique most often used for proteomic analysis (Aebersold 2003; Maarten *et al.* 2013). It allows scientists to detect and quantify proteins in a complex biological matrix. Mass spectrometry (MS) measures the mass-to-charge ratio of ions to identify and quantify molecules in simple and complex mixtures (Aebersold 2003). The development of high-throughput and quantitative MS proteomics workflows within the last two decades has expanded the scope of what we know about protein structure, function, modification and global protein dynamics (Larance & Lamond 2015). In proteomics research, mass spectrometry is used to determine protein structure, function, folding and interactions, identify a protein from the mass of its peptide fragments, detect specific post-translational modifications throughout complex biological mixtures, quantitate (relative or absolute) proteins in a given sample and to monitor enzyme reactions, chemical modifications and protein digestion (Schmidt *et al.* 2014).

Western blotting is used to quantify the relative expression levels for specific proteins by electrophoretically separating extracted cell proteins, transferring them to a membrane, and then probing the bound proteins with antibodies (targeted to antigens of interest) that are subsequently detected using various chemistries or radiolabelling (Griffiths *et al.* 2015). By using a western blot, researchers are able to identify specific proteins from a complex mixture of proteins extracted from cells. The Western blotting technique uses three elements to accomplish this task: SDS-PAGE separation of the proteins by size, transfer of the separated proteins from the gel to a solid support, and visualizing the target protein using a proper primary and secondary antibody (Mahmood & Yang 2012).

Two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) is a technique for separating a complex mixture of proteins in two dimensions and then staining to detect differences at the whole-proteome level (Ortea *et al.* 2016). It is capable of resolving thousands of proteins in a single run. In the first dimension, proteins are separated based on differences in isoelectric point (pI). In the second dimension, they are separated according to molecular weight. Following separation, 2-D electrophoresis gels are stained for protein visualization and analysis (Mayer *et al.* 2015). In combination with computer-assisted image evaluation systems for comprehensive qualitative and quantitative examination of proteomes, this electrophoresis technique allows cataloguing of proteins and comparison of data among groups of researchers. 2-D PAGE is well suited for the analysis of posttranslational protein modifications. It is particularly useful for low-abundance proteins (Mayer *et al.* 2015).

Immunoassays are techniques that exploit the sensitivity and specificity of antibody-antigen interactions for detection of target analytes in biological samples. Immunoassays are important for protein detection and quantification (Tak For Yu *et al.* 2015). Proteins are quantitated in solution using antibodies that are bound to color-coded beads or immobilized to a surface, which is subsequently probed with an antibody suspension and is typically detected using a chromogenic or fluorogenic reporter. Immunoassays can be used to determine levels of protein phosphorylation and other post-translational modifications by detecting these attachments using antibodies that are specific for them (Chen *et al.* 2015).

Proteomics has both a physical laboratory component and a computational component. Bioinformatics in protein analysis is applied for database searches, sequence comparisons and structural predictions (Oliva *et al.* 2012), analysis of protein post-translational modifications, protein-protein interactions (Abellan 2013) and computational methods for mass spectrometry-based proteomics (Li & Tang 2016).

4.4 Gene inactivation methods

One way to understand the function of a gene is to observe a biological system that lacks that gene. The ability to manipulate the expression levels of specific genes into their respective final protein products has been essential for understanding the functions of specific genes in the study of biological processes (Guo *et al.* 2014; Yu & Yuan 2010). Scientists are now able to modulate the expression of genes of interest and precisely modify the genomic sequences in virtually any organisms (Guo *et al.* 2014). Several techniques have been developed to alter a gene sequence to result in an inactivated gene, or one in which the expression is inactivated at a chosen time during development to study the loss of function of the gene. It is therefore possible to identify altered gene expression that may underpin a particular disease condition. Gene inactivation events can have varying effects on phenotypes (Balasubramanian *et al.* 2011). Traditionally, gene manipulation focused on introducing a foreign target gene into host cells and tissues through recombinant DNA technology and gene transfer techniques to express the gene of interest (Yu & Yuan 2010). In the past few years, there has been a revolution in the approaches scientists use to inactivate gene expression, such as the development of highly efficient gene knockdown with ribonucleic acid interference (RNAi) delivery systems and the groundbreaking genome editing technologies of zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR) (Swamy *et al.* 2016; Guo *et al.* 2014).

4.4.1 Gene knockdown by *in vivo* RNA regulation

RNA medicine is the therapeutic targeting of mRNA using synthetic RNA molecules. Two major methods are employed: the use of antisense RNAs and RNA-interference (RNAi) mediated by double-stranded RNA (Chery 2016).

4.4.1.1 Translation blockage by antisense RNA hybridization to target mRNA

Antisense RNAs are small, highly structured single-stranded molecules that act through sequence complementarity to bind and inhibit targeted specific mRNA (sense RNA) RNA function for example, inhibition of translation to protein. Antisense RNA technology works through many mechanisms depending, in part, on the region in the RNA sequence that is targeted (Chery 2016).

Antisense RNA technology was used to synthesize a complementary mRNA to Polygalacturonase (PG) gene and inhibits the synthesis of PG enzyme thus delaying over ripening and rotting of tomatoes (García-Gago *et al.* 2009). Antisense RNA technology has applications also in the fields of gene therapy, cancer therapy and therapies of several other disorders (Evers *et al.* 2015). Research is also on going on the development of antisense antiviral drugs and other RNA therapeutics (Chery 2016).

Protein production is controlled by ribosome binding to the messenger RNA (mRNA) (Eriksen *et al.* 2017). Weaker ribosome binding sites can result in a decrease in protein yield. Antisense RNA molecules can be utilized to control gene regulation. Naturally occurring antisense RNAs have been isolated in a various microbes, including the E. coli (Thomason *et al.* 2010). Riboswitches are elements in bacterial commonly found in the 5-untranslated region (UTR) of mRNAs that exert their regulatory control over the transcript in a cis-fashion by directly binding a small molecule ligand to regulate expression of the downstream coding sequence(s) without a requirement for regulatory proteins (Garst *et al.* 2011). The regulatory signal is an effector molecule that binds the nascent RNA transcript, causing a change in the RNA structure (Fuchs *et al.* 2007). The structural rearrangements can sequester the ribosome-binding site to regulate at the level of translation initiation. Inhibition of translation initiation in bacteria is achieved when ligands sequester the Shine-Dalgarno (SD) sequence of the mRNA through alternative base pairing, resulting in occlusion of the ribosomal binding site (Rinaldi *et al.* 2016). RNA molecules can be engineered to be regulatory molecules for base-paring and ligand-induced conformational change for gene regulation as riboswitches (Wittmann & Suess 2012).

4.4.1.2 mRNA degradation by RNA interference (RNAi)

RNA interference (RNAi) (also known as *post-transcriptional gene silencing* (PTGS)) is an endogenous biological RNA-dependent gene regulatory mechanism by which noncoding double-stranded RNA (dsRNA) molecules induce gene silencing by targeting complementary mRNA for degradation (Kelly & Hurlstone 2011; Agrawal *et al.* 2003) thus suppressing the synthesis of protein. In the absence of this protein one can look for clues on the function of this protein. RNAi is a conserved biological response to double-stranded RNA. RNA in cells naturally exits as a single-stranded nucleic acid molecule (unlike DNA which is double -stranded) (Chery 2016).

The RNAi pathway is initiated when dsRNA enters the cytoplasm. Endogenous triggers of RNAi pathway include double-stranded RNA (dsRNA) of viral origin, aberrant transcripts from repetitive sequences in the genome such as transposons and unique endogenous small RNAs including microRNA (miRNA) endogenous small interfering RNAs (endo-siRNAs), and Piwi-interacting RNAs (piRNAs) (Okamura & Lai 2008; Czech *et al.* 2008). RNAi is a naturally occurring pathway thought to have evolved in plants and animals over millions of years as a form of innate immunity defence against viruses, suggesting an important role in pathogen resistance (Meng *et al.* 2013).

RNAi can be triggered experimentally by exogenous introduction of dsRNA or using DNA-based vectors, which express short hairpin RNA (shRNA) in the cytoplasm that are processed by Dicer into siRNAs (Swamy *et al.* 2016; McGinnis 2010). RNAi can also be induced directly by transfecting cells with siRNAs with dinucleotide 3' overhangs (Fitzgerald *et al.* 2017).

A simplified model for the RNAi pathway is based on two steps, each involving ribonuclease enzyme. In the first step, the trigger RNA (either dsRNA or miRNA primary transcript) is processed into a small interfering RNA (siRNA) by the RNase II enzyme, Dicer and RNaseIII endonuclease, Drosha (which cleave the long dsRNA into short double-stranded fragments of 20–25 base pairs (bp) small interfering RNAs (siRNAs)) (McGinnis 2010; Kelly & Hurlstone 2011; Li & Patel 2016). In the second step, siRNAs are loaded and assembled into the effector complex RNA-induced silencing complex (RISC). The siRNA is unwound during RISC assembly and the antisense strand of the siRNA duplex becomes part of a multi-protein complex RISC and then hybridizes with the mRNA target as it guides the RISC to bind to the target mRNA molecules. The RISC cleaves the mRNA, leading to specific gene silencing (Kelly and Hurlstone, 2011; Agrawal *et al.* 2003) as a result of nucleolytic degradation of the targeted mRNA by the RNase H enzyme Argonaute (Slicer) (Pompey *et al.* 2014). Argonautes are the key effectors of RNA interference (RNAi) pathways (Kaya *et al.* 2016). If the siRNA/mRNA duplex contains mismatches the mRNA is not cleaved (Swamy *et al.* 2016).

RNAi is a specific, potent, and highly successful approach for loss-of-function studies in virtually all eukaryotic organisms (McGinnis 2010; Pompey *et al.* 2014; Koch *et al.* 2016). RNAi technology takes advantage of the cell's natural machinery, facilitated by short interfering RNA molecules, to effectively knock down expression of a gene of interest (Swamy *et al.* 2016). RNAi technology is precise (ability to target and silence individual genes, even among a family of closely related genes) (McGinnis 2010), fast (thousands of genes can be rapidly targeted using RNAi), stable (traits based on RNAi have been shown to be stable for at least five generations) (Brown *et al.* 2003), flexible (RNAi effective for different species and phenotypes) and controllable (genes can be turned off completely or just have their effects 'turned down') (McGinnis 2010). The high degrees of efficiency and specificity of RNAi make it is one of the most important technological breakthroughs in functional genomics by allowing us to directly observe the phenotypes resulting from the systematic loss-of-function of genes (McGinnis 2010). With RNAi technology it is possible to efficiently block the expression of a specific gene and evaluate its response to changes the environment. RNAi technology can be used to assess the functions of thousands of genes within the genome that potentially participate in disease phenotypes.

4.4.2 Gene knockout by mutagenesis

Gene targeting in mouse embryonic stem cells is an established technique for creating animal models for human disease or to study gene function at a whole animal level (van Deursen 2002). Gene targeting in mouse embryonic stem cells has become the 'gold standard' for determining gene function in mammals. Gene targeting is the process of disrupting or mutating a specific genetic locus with the intention of making knockout individuals (Gerlai 2016). Gene function can be investigated by systematically looking for any mutant phenotype that might provide clues about the function of the gene (Griffiths *et al.* 2015).

Gene targeting by homologous recombination enables the exchange of genetic information between genomic and exogenous DNA molecules via crossing-over events (Gerlai 2016). These exchanges are guided by flanking homologous sequences that direct the cell’s own enzymatic machinery. Homologous recombination provides a tool for targeted defined modifications of genes of interest, for the purpose of exploring gene function (Reh & Vasquez 2014). Gene targeting with homologous recombination created a revolution in the analysis of the function of genes by allowing unprecedented precision with which one could manipulate genes and study the effect of this manipulation (Gerlai 2016). The targeting construct is usually a plasmid that contains two long stretches of genomic DNA, called homology arms, which are designed to match as closely as possible the genomic DNA of the embryonic stem cell line being targeted (Reh & Vasquez 2014). These arms drive the homologous recombination event that results in insertion of the construct into the desired locus. This process is underway in the fully sequenced genomes.

Genome editing is a genetic approach used to directly manipulate an organism's genome by inserting, replacing, or removing DNA sequences. Permanent change in DNA leads to the loss of function of a gene (Gaj *et al.* 2016). Recently developed techniques using Engineered nucleases, such as zinc-finger nucleases (ZFNs) (Urnov et al., 2010), transcription activator-like effector nucleases (TALENs) (Gerlai 2016) and most recently, the bacterial clustered regularly interspaced short palindromic repeat (CRISPR)/Cas (CRISPR-associated) system have the potential to target genes directly in embryos, without the need to use embryonic stem cells (Guo *et al.* 2014; Gerlai 2016).

In some cases, knocked-out ORFs show no phenotypic effects due to the compensatory effects of other genes in the genome. More than half of the predicted ORFs may fall into this category (Grifiths *et al.* 2015). When this happens, the phenomenon of compensation undermines our ability to answer the question originally thought of as the main goal of gene targeting (Gerlai 2016).

4.4.2.1 Gene knockout by transposon-mediated insertional mutagenesis

Transposable genetic elements (also referred to as transposons or "jumping genes") are DNA sequences that can change their position within a genome, sometimes creating or reversing mutations or altering the cell's genome size (Griffiths *et al.* 2015). Transposons are found in almost all organisms (both prokaryotes and eukaryotes) (Griffiths *et al.* 2015). They occur in large numbers, for example, they make up approximately 50% of the human genome (International Human Genome Sequencing Consortium, 2001) and up to 90% of the maize genome (SanMiguel *et al.* 1996). They have an enzyme, transposase (encoded by the transposon itself), which they require for excision and insertion (Silva *et al.* 2011). DNA transposons are flanked at both ends by terminal inverted repeats (ITRs) and a single open reading frame that encodes a transposase. The inverted repeats are complements of each other (Griffiths *et al.* 2015). Retrotransposons have long terminal repeats (LTRs) on both ends (Finnegan 2012). Transposition occurs using one of the following mechanisms: In cut-and-paste transposition, an element is cut out of one site in a chromosome and pasted into a new site. In replicative transposition, an element is replicated, and one copy is inserted at a new site; one copy also remains at the original site. In retrotransposition, an element’s RNA is used as a template to synthesize DNA molecules, which are inserted into new chromosomal sites (DeNicola *et al.* 2015; Griffiths *et al.* 2015). Transposons are non-targeted gene transfer vehicles (Silva *et al.* 2011). A transposon can either disrupt gene function when it integrates into the open reading frame or it can activate expression when inserted upstream of a gene as the promoter to drive the expression of downstream sequences (DeNicola *et al.* 2015). Transposon-mediated insertional mutagenesis provides a method for near-random mutation. This approach is particularly useful for organisms that are relatively refractory to genetic manipulation (Lin *et al.* 2014).

4.4.2.2 Targeted Genome editing using engineered nucleases

The emergence of genome-editing technologies has provided new tools for introducing sequence-specific modifications into genomes. Engineered nucleases enable the manipulation and targeted alteration of any genomic sequence in a wide range of cell types and organisms (Gaj *et al.* 2016; Joung & Sander 2013). The core technologies now most commonly used include: homing endonucleases or meganucleases, zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated protein 9 (CRISPR-Cas9). This process is most often used to achieve gene knockout via insertions and/or deletions.

Targeted genome editing relies on the use of engineered nucleases, that are linked to a customizable sequence-specific DNA-binding domain which is fused to a nuclease that cleaves DNA to induce double-strand breaks (DSB s) at specific sites, which are then repaired by mechanisms that can be exploited to create sequence alterations (Joung & Sander 2013). The Common feature of all Nuclease-mediated genome editing is the generation of DNA double-strand breaks (DSBs). The result of DSBs is the activation of cellular DNA repair pathways, which facilitate the introduction of site-specific genomic modifications (Gaj *et al.* 2016). The DSBs are then repaired either by non-homologous end joining (NHEJ) or homologous recombination (HR). Non-homologous end-joining of DNA at the double stranded break is a highly error prone pathway that often leads to the generation of random base insertions or deletions of nucleotides leading to a shift in the reading frame, which is expected to have a major disruptive effect on the structure and thus the function of the protein translated from this mutant gene upon re-joining (Gerlai 2016). Homologous recombination is an error-free mechanism for DSB repair. It uses a homologous DNA sequence as template.

Nuclease-mediated genome editing enables genetic studies that were previously difficult or impossible to perform (Joung & Sander 2013). The resultant loss-of-function mutations could be used to create somatic cell-based models of disease. Alternatively, precise insertions can be introduced into endogenous genes (Joung & Sander 2013).

One major concern associated with all targeted nucleases is the observation of off-target mutations (Gaj *et al.* 2016). These can be reduced with improving the specificity of the tools for the target sequences. Targeted nucleases have been used in a technique known as gene drive to confer particular phenotypes in hosts, which are inherited by their progeny (Gaj *et al.* 2016). Gene drives have been applied in the population control of malaria mosquitos *Anopheles stephensi* (Gantz *et al.* 2015) and *Anopheles gambiae* (Hammond *et al.* 2016). Concerns have arisen about the potential societal and environmental impact of this technology (Esvelt *et al.* 2014; Akbari *et al.* 2015), owing to the ease with which CRISPR-Cas9 can be programmed (Gantz & Bier 2015). Debate has ignited on finding avenues to minimize the risk of gene-edited organisms escaping from the laboratory (DiCarlo *et al.* 2015).

4.4.2.2.1 Meganucleases for targeted genome engineering

Meganucleases, also termed homing endonucleases, are rare-cutting highly specific DNA cleaving enzymes that are encoded within the genome of nearly all forms of microbial life as well as in eukaryotic mitochondria and chloroplasts (Stoddard 2014). They are highly specific endonucleases which recognize and cleave the exon-exon junction sequence wherein their intron resides, thus giving rise to the moniker “homing endonuclease” (Silva *et al.* 2011). These enzymes recognize and cleave long DNA sequences (typically 18–30 base pairs) generating double-strand DNA breaks (DSBs) (Stoddard 2011). The binding and cleavage domains in homing endonucleases are not modular. This overlap in form and function make their repurposing challenging, and limits their utility for more routine applications of genome editing (Gaj *et al.* 2016).

4.4.2.2.2 Targeted gene knockout by zinc finger nucleases

Zinc-finger nucleases (ZFNs) are artificial restriction enzymes which function as a heterodimer in which each subunit consists of two functional domains, which are generated by fusing a zinc finger (DNA-binding domain) to a nuclease (DNA-cleaving domain comprised of a Fok I restriction endonuclease) (Urnov *et al.* 2010). Dimerization of the ZFN proteins is mediated by the FokI cleavage domain. When the DNA-binding and DNA-cleaving domains are fused together, a highly specific pair of ‘genomic scissors’ is created. The FokI domains must dimerize for activity, thus increasing target specificity by ensuring that two proximal DNA-binding events must occur to achieve a double-stranded break. The resulting cleavage event is what enables genome editing to happen. After a break is created, the cell seeks to repair it (Carroll *et al.* 2011). Zinc finger domains can be engineered to target specific desired DNA sequences and this enables zinc-finger nucleases to target unique sequences within complex genomes. Each ZFN is composed of three or four zinc-finger domains, with each individual domain made of 30 amino acid residues. Each individual zinc fingers domain typically recognizes and interacts with DNA triplets (Gaj *et al.* 2016). The difficulty associated with constructing zinc-finger arrays that can effectively recognize all DNA triplets has hindered their widespread adoption of the ZFN technology (Gaj *et al.* 2016).

4.4.2.2.3 Genome editing with Transcription activator-like effector nucleases (TALENs)

Transcription Activator-Like Effector (TALE) proteins are bacterial effectors. A TALE binding domain consists of a tandem array of repeated segments each consisting of 34 amino acids (Gerlai 2016). The amino acid sequence of these repeats is mostly the same EXCEPT for the amino acids at positions 12 and 13. Each repeat contacts DNA via the amino acid residues at positions 12 and 13, known as the repeat variable di-residues (Gaj *et al.* 2016). The specificity of each individual TALE repeat is determined by the identities of two hypervariable residues (Joung & Sander 2013). Each of these variable pair of amino acids in a TALE binds to a specific nucleotide in DNA (Gaj *et al.* 2016). TALE genes can be mutated to generate sequence-specific DNA binding proteins. TALEs are typically assembled to recognize between 12- to 20-bps of DNA, with more bases typically leading to higher genome-editing specificity. TALEs can be fused to nucleases to form Transcription Activator-Like Effector Nucleases (TALENs) for targeted double-stranded breaks in DNA. TALENs are similar in design to ZFNs having a non-specific FokI nuclease domain linked to a customizable DNA-binding domain (Joung & Sander 2013). Like ZFNs, dimerization of TALEN proteins is mediated by the FokI cleavage domain, which cuts within a 12- to 19-bp spacer sequence that separates each TALE binding site (Gaj *et al.* 2016). TALENs are artificial restriction enzymes able to cut DNA only where they encounter a specific sequence of nucleotides. This can be repaired by non-homologous end joining (NHEJ) or homologous recombination (HR). TALENs are easier to design and researchers can create specific tools by using simple protein-DNA codes (Joung & Sander 2013).

4.4.2.2.4 CRISPR/Cas9-Based Genome Editing

The Clustered, Regularly Interspaced, Short Palindromic Repeat (CRISPR) system is a form of adaptive immunity found in bacteria (Marraffini 2016), which acts against DNA from invading viruses and plasmids using RNA-guided DNA cleavage by Cas proteins (Gaj *et al.* 2016). In nature, the Cas9 endonuclease forms a complex with two RNA molecules, CRISPR RNA (crRNA) and transactivating crRNA (tracrRNA), which guide the CRISPR-associated endonuclease (Cas9) to recognize and cleave a site (Marraffini 2016).

CRISPR/ Cas9 system used in biotechnology consists of two components: a “guide” RNA (gRNA) and a non-specific CRISPR-associated endonuclease (Cas9). The guide RNA, is a short synthetic chimeric molecule, composed of essentials functional portions of crRNA and tracrRNA which are required for Cas9-binding and a user-defined 20 nucleotide “spacer” or “targeting” sequence that is complementary to the genomic target to be modified (Doudna & Charpentier 2014). Thus, one can change the genomic target of Cas9 by simply changing the targeting sequence present in the gRNA.

The Cas9 target sequence consists of a 20-bp DNA sequence complementary to the gRNA and the trinucleotide (5′-NGG-3′) protospacer adjacent motif (PAM) recognized by Cas9 (Doudna & Charpentier 2014). The end result is a double stranded break induced by Cas9, which is resolved either by non-homologous end joining (NHEJ) or homologous recombination (HR). The advantage of the CRISPR system over the TALEN method is that it is comparably simpler to perform because target site recognition is mediated entirely by the gRNA (Gaj *et al.* 2016). CRISPR-Cas9 has emerged as the most flexible and user-friendly platform for genome editing by eliminating the need for engineering new proteins to recognize each new target site, and thus its use is rapidly spreading across molecular biology laboratories (Gerlai 2016).

5. CONCLUSION

This review was written in an attempt to present the reader with a decent spectrum of the available methods for gene function analysis that have been applied in the past decade. Genetic engineering provides powerful tools for the study of gene function in both cells and organisms. This review has revealed several methods for the functional analysis of genes. Ultimately the choice of method to be applied will rely on several factors such as; the research budget, the expected throughput, the size of genome to be analysed, the application sought, the objectives to be met and the acceptable safety levels demanded among other reasons. A simple database searching for homology could predict a gene's function, reporter genes can demonstrate when and where a gene is expressed, microarrays can monitor the expression of thousands of known genes at once, targeted mutations can reveal gene function, cells and animals containing mutated genes can be made and gene targeting makes it possible to produce transgenic animals.

6. REFERENCES

Abellan, I.A. (2013). Bioinformatics Approaches to Protein Interaction and Complexes: Application to Pathogen-Host Epitope Mimicry and to Fe-S Cluster Biogenesis Model. PhD Thesis. University of Barcelona.

Aebersold, R. (2003). Quantitative Proteome Analysis: Methods and Applications. *Journal of Infectious Diseases*. 187(2): S315-S320. DOI: https://doi.org/10.1086/374756

Agrawal, N., Dasaradhi, P.V.N., Mohmmed, A., Malhotra, P., Bhatnagar, R.K., & Mukherjee, S.K. (2003). RNA Interference: Biology, Mechanism, and Applications. *Microbiology and Molecular Biology Reviews*. 67(4): 657–685. http://doi.org/10.1128/MMBR.67.4.657-685.2003

Akbari, O.S., Bellen, H.J., Bier, E., Bullock, S.L., Burt, A., Church, G.M., … Wildonger, J. (2015). Safeguarding gene drive experiments in the laboratory: Multiple strategies are needed to ensure safe gene drive experiments. *Science (New York, N.Y.)*. *349*(6251): 927–929. http://doi.org/10.1126/science.aac7932

Altelaar, A.F., Munoz J. & Heck, A.J. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*. 14: 35-48. doi:10.1038/nrg3356

Balasubramanian, S., Habegger, L., Frankish, A., MacArthur, D.G., Harte, R., Tyler-Smith, C., … Gerstein, M. (2011). Gene inactivation and its implications for annotation in the era of personal genomics. *Genes & Development*. *25*(1): 1–10. http://doi.org/10.1101/gad.1968411

Bayley, H. (2015). Nanopore sequencing: from imagination to reality. *Clinical Chemistry*. 61(1): 25–31. http://doi.org/10.1373/clinchem.2014.223016

Berg, J.M., Tymoczko, J.L., Gatto Jr. G.J. & Stryer, L. (2015). Biochemistry 8th Edition. W. H. Freeman. ISBN-10: 1464126100. ISBN-13: 978-1464126109.

Berglund, E.C., Kiialainen, A., & Syvänen, A.-C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*. 2: 23. http://doi.org/10.1186/2041-2223-2-23

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*. 14(1): 1-8.

Brown, A. E., Bugeon, L., Crisanti, A., & Catteruccia, F. (2003). Stable and heritable gene silencing in the malaria vector *Anopheles stephensi*. *Nucleic Acids Research*. *31*(15): e85.

Carroll, D. (2011). Genome Engineering With Zinc-Finger Nucleases. *Genetics*. 188(4): 773–782. http://doi.org/10.1534/genetics.111.131433

Chen, J.-Q., Wakefield, L.M., & Goldstein, D.J. (2015). Capillary nano-immunoassays: advancing quantitative proteomics analysis, biomarker assessment, and molecular diagnostics. *Journal of Translational Medicine*. 13: 182. http://doi.org/10.1186/s12967-015-0537-6

Chernobrovkin, A., Marin-Vicente, C., Visa, N., & Zubarev, R.A. (2015). Functional Identification of Target by Expression Proteomics (FITExP) reveals protein targets and highlights mechanisms of action of small molecule drugs. *Scientific Reports*. 5: 11176. http://doi.org/10.1038/srep11176

Chery, J. (2016). RNA therapeutics: RNAi and antisense mechanisms and clinical applications. Postdoc Journal : *A Journal of Postdoctoral Research and Postdoctoral Affairs*. 4(7): 35–50.

Clark, K.J., Voytas, D.F., & Ekker, S.C. (2011). A TALE of Two Nucleases: Gene Targeting for the Masses? *Zebrafish*. *8*(3): 147–149. http://doi.org/10.1089/zeb.2011.9993

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*. 17: 13. http://doi.org/10.1186/s13059-016-0881-8

Czech, B., Malone, C.D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., … Brennecke, J. (2008). An endogenous small interfering RNA pathway in *Drosophila*. *Nature*. 453(7196): 798–802. http://doi.org/10.1038/nature07007

Deamer, D., Akeson, M. & Branton, D. (2016). Three decades of nanopore sequencing. Natuer Biotechnology. 34(5): 518-24. doi: 10.1038/nbt.3423.

DeNicola, G.M., Karreth, F.A., Adams, D.J., & Wong, C.C. (2015). The utility of transposon mutagenesis for cancer studies in the era of genome editing. *Genome Biology*. 16: 229. http://doi.org/10.1186/s13059-015-0794-y

Dewey, F.E., Pan, S., Wheeler, M.T., Quake, S.R., & Ashley, E.A. (2012). DNA sequencing: Clinical applications of new DNA sequencing technologies. *Circulation*. 125(7): 931–944. http://doi.org/10.1161/CIRCULATIONAHA.110.972828

DiCarlo, J.E., Chavez, A., Dietz, S.L., Esvelt, K.M., & Church, G.M. (2015). Safeguarding CRISPR-Cas9 gene drives in yeast. *Nature Biotechnology*. 33(12): 1250–1255. http://doi.org/10.1038/nbt.3412

Doudna, J.A. & Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*. 346(6213): 1258096. DOI: 10.1126/science.1258096

Eriksen, M., Sneppen, K., Pedersen, S., & Mitarai, N. (2017). Occlusion of the Ribosome Binding Site Connects the Translational Initiation Frequency, mRNA Stability and Premature Transcription Termination. *Frontiers in Microbiology*. 8: 362. http://doi.org/10.3389/fmicb.2017.00362

Esvelt, K.M., Mali, P., Braff, J.L., Moosburner, M., Yaung, S.J., & Church, G.M. (2013). Orthogonal Cas9 Proteins for RNA-Guided Gene Regulation and Editing. *Nature Methods*. 10(11): 1116–1121. http://doi.org/10.1038/nmeth.2681

Evans, T.G. (2015). Considerations for the use of transcriptomics in identifying the 'genes that matter' for environmental adaptation. *Journal of Experimental Biology*. 218: 1925-1935; doi: 10.1242/jeb.114306

Evers, M.M., Toonen, L.J., & van Roon-Mom, W.M. (2015). Antisense oligonucleotides in therapy for eurodegenerative disorders. *Advanced Drug Delivery Reviews*. 87: 90–103.

Finnegan, D.J. (2012). Retrotransposons. *Current Biology*. 22(11): R432–R437.

Fitzgerald, K., White, S., Borodovsky, A., Bettencourt, B.R., Strahs, A., Clausen, V., Wijngaard, P., Horton, J.D., Taubel, J., Brooks, A., Fernando, C., Kauffman, R.S., Kallend, D., Vaishnaw, A. and Simon, A. (2017). A Highly Durable RNAi Therapeutic Inhibitor of PCSK9. *New England Journal of Medicine*. 376: 41-51. DOI: 10.1056/NEJMoa1609243

Fu, Y. & Xiao, W. (2006). Study of Transcriptional Regulation Using a Reporter Gene Assay. *Methods in Molecular Biology*. 313: 257-264

Fuchs, R.T., Grundy, F.J., & Henkin, T.M. (2007). S-adenosylmethionine directly inhibits binding of 30S ribosomal subunits to the SMK box translational riboswitch RNA. *Proceedings of the National Academy of Sciences of the United States of America*. 104(12): 4876–4880.

http://doi.org/10.1073/pnas.0609956104

Gaj, T., Sirk, S.J., Shui, S.L. & Liu, J. (2016). Genome-Editing Technologies: Principles and Applications. *Cold Spring Harbor Perspectives in Biology*. 8(12): a023754. doi: 10.1101/cshperspect.a023754.

Gallego, R.I., Ruvinsky, I. & Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*. 13: 505–516.

Gantz, V.M., & Bier, E. (2015). The mutagenic chain reaction: a method for converting heterozygous to homozygous mutations. *Science (New York, N.Y.)*. 348(6233): 442–444. http://doi.org/10.1126/science.aaa5945

Gantz, V.M., Jasinskiene, N., Tatarenkova. O., Fazekas, A., Macias, V.M., Bier, E. & James, A.A. (2015). Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. PNAS. 112: E6736–E6743.

García-Gago, J.A., Posé, S., Muñoz-Blanco, J., Quesada, M.A., & Mercado, J.A. (2009). The polygalacturonase FaPG1 gene plays a key role in strawberry fruit softening. *Plant Signaling & Behavior*. 4(8): 766–768.

Garst, A.D., Edwards, A.L., & Batey, R.T. (2011). Riboswitches: Structures and mechanisms. *Cold Spring Harbor Perspectives in Biology*. 3(6): a003533. http://doi.org/10.1101/cshperspect.a003533

Ghosh, S. & Febin, P.D.J. (2016). Non-canonical pathway network modelling and ubiquitination site prediction through homology modelling of NF-κB. *Gene*. 581(1): 48-56. doi: 10.1016/j.gene.2016.01.025.

Ginalski, K., Pas, J., Wyrwicz, L.S., Grotthuss, M. von, Bujnicki, J.M., & Rychlewski, L. (2003). ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Research*. *31*(13): 3804–3807.

Gowri, V.S. & Sandhya, S. (2006). Recent trends in remote homology detection: an Indian Medley. *Bioinformation*. 3: 94-96

Grabherr, M.G., Haas, B.J., Yassour, M., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*. 29: 644–652.

Grant, M. A. (2011). Integrating computational protein function prediction into drug discovery initiatives. *Drug Development Research*. *72*(1): 4–16. http://doi.org/10.1002/ddr.20397

Griffiths, A.J.F., Wessler, S.R., Carroll, S.B. & Doebley, J. (2015). An Introduction to Genetic Analysis. 11th edition. New York: W. H. Freeman; ISBN-10: 1464109486. ISBN-13: 978-1464109485.

Gerlai, R. (2016). Gene Targeting Using Homologous Recombination in Embryonic Stem Cells: The Future for Behavior Genetics? *Frontiers in Genetics*. 7: 43. http://doi.org/10.3389/fgene.2016.00043

Guo, C-A., O'Neill, L.M. & Ntambi, J.M. (2014). Gene Inactivation Strategies: An Update. John Wiley & Sons, Ltd. DOI: 10.1002/9780470015902.a0021020.pub2

Hammond, A., Galizi, R., Kyrou, K., Simoni, A., Siniscalchi, C., Katsanos, D., … Nolan, T. (2016). A CRISPR-Cas9 Gene Drive System Targeting Female Reproduction in the Malaria Mosquito vector *Anopheles gambiae*. *Nature Biotechnology*. 34(1): 78–83. http://doi.org/10.1038/nbt.3439

Heather, J.M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*. 107(1): 1–8. http://doi.org/10.1016/j.ygeno.2015.11.003

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*. 409: 860 – 921.

Jain, M., Olsen, H.E. Paten, B. & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics Community. *Genome Biology*. 17: 239

Jenkinson, C.P., Göring, H.H.H., Arya, R., Blangero, J., Duggirala, R., & DeFronzo, R.A. (2016). Transcriptomics in type 2 diabetes: Bridging the gap between genotype and phenotype. *Genomics Data*. *8*: 25–36. http://doi.org/10.1016/j.gdata.2015.12.001

Joung, J.K., & Sander, J.D. (2013). TALENs: a widely applicable technology for targeted genome editing. Nature Reviews. *Molecular Cell Biology*. 14(1): 49–55. http://doi.org/10.1038/nrm3486

Kaya, E., Doxzen, K.W., Knoll, K.R., Wilson, R.C., Strutt, S.C., Kranzusch, P.J., & Doudna, J. A. (2016). A bacterial Argonaute with noncanonical guide RNA specificity. *Proceedings of the National Academy of Sciences of the United States of America*. 113(15), 4057–4062. http://doi.org/10.1073/pnas.1524385113

Kelly, A. & Hurlstone, A.F. (2011). The use of RNAi technologies for gene knockdown in zebrafish. *Briefings in Functional Genomics*. 10(4): 189-96. doi: 10.1093/bfgp/elr014.

Koch, A., Biedenkopf, D., Furch, A., Weber, L., Rossbach, O., Abdellatef, E., … Kogel, K.-H. (2016). An RNAi-Based Control of *Fusarium graminearum* Infections Through Spraying of Long dsRNAs Involves a Plant Passage and Is Controlled by the Fungal Silencing Machinery. *PLoS Pathogens*. *12*(10): e1005901. http://doi.org/10.1371/journal.ppat.1005901

Koonin, E.V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Reviews Genetics*. 39: 309–38.

Larance, M. & Lamond A.I. (2015). Multidimensional proteomics for cell biology. *Nature Reviews Molecular Cell Biology*. 16: 269–280 doi:10.1038/nrm3970

Lee, H.B., Sundberg, B.N., Sigafoos, A.N., & Clark, K.J. (2016). Genome Engineering with TALE and CRISPR Systems in Neuroscience. *Frontiers in Genetics*. 7: 47. http://doi.org/10.3389/fgene.2016.00047

Li, S. & Patel, D.J. (2016). Drosha and Dicer: Slicers cut from the same cloth. Cell Research. 26: 511–512. doi:10.1038/cr.2016.19;

Li, S. & Tang, H. (2016). Computational Methods in Mass Spectrometry-Based Proteomics. In "Translational Biomedical Informatics". pp 63-89. ISBN: 978-981-10-1502-1

Lin, T., Troy, E.B., Hu, L.T., Gao, L., & Norris, S.J. (2014). Transposon mutagenesis as an approach to improved understanding of Borrelia pathogenesis and biology. *Frontiers in Cellular and Infection Microbiology*. 4: 63. http://doi.org/10.3389/fcimb.2014.00063

Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., … Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*. 16: 179. http://doi.org/10.1186/s13059-015-0742-x

Mahmood, T., & Yang, P.-C. (2012). Western Blot: Technique, Theory, and Trouble Shooting. *North American Journal of Medical Sciences*. 4(9): 429–434. http://doi.org/10.4103/1947-2714.100998

Manjasetty, B.A., Büssow, K., Panjikar, S., & Turnbull, A.P. (2012). Current methods in structural proteomics and its applications in biological sciences. *3 Biotech*. 2(2): 89–113. http://doi.org/10.1007/s13205-011-0037-1

Marraffini, L. (2016). Crispr-Cas, The Prokaryotic Adaptive Immune System. *The FASEB Journal*. 30(1): Supplement 107.1

Mayer, K., Albrecht, S. & Schaller, A. (2015). Targeted Analysis of Protein Phosphorylation by 2D Electrophoresis. *Methods in Molecular Biology*. 1306:167-76. doi: 10.1007/978-1-4939-2648-0_13.

McGinnis, K.M. (2010). RNAi for functional genomics in plants. *Briefings in Functional Genomics*. 9(2): 111-117.

Mehmood, M.A., Sehar, U. & Ahmad, N. (2014) Use of Bioinformatics Tools in Different Spheres of Life Sciences. *Journal of Data Mining in Genomics and Proteomics*. 5: 158. doi:10.4172/2153-0602.1000158

Meng, Z., Zhang, X., Wu, J., Pei, R., Xu, Y., Yang, D., … Lu, M. (2013). RNAi Induces Innate Immunity through Multiple Cellular Signaling Pathways. *PLoS ONE*. 8(5): e64708. http://doi.org/10.1371/journal.pone.0064708

Mir, K., Neuhaus, K., Scherer, S., Bossert, M., & Schober, S. (2012). Predicting Statistical Properties of Open Reading Frames in Bacterial Genomes. *PLoS ONE*. 7(9): e45103. http://doi.org/10.1371/journal.pone.0045103

Morozova, O. & Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*. 92: 255–264.

Okamura, K. & Lai, E.C. (2008). Endogenous small interfering RNAs in animals. *Nature reviews Molecular cell biology*. 9(9): 673-678. doi:10.1038/nrm2479.

Oliva, B., Planas-Iglesias, J., Bonet, J., Marín-López, M.A., Feliu. E. & Gursoy, A. (2012). Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence. Edited by Weibo Cai and Hao Hong, ISBN 978-953-51-0397-4, DOI: 10.5772/37856

Ortea, I., O'Connor, G. & Maquet, A. (2016). Review on proteomics for food authentication. *Journal of Proteomics*. 147: 212–225.

Palazzo, A.F., & Lee, E.S. (2015). Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics*. 6: 2. http://doi.org/10.3389/fgene.2015.00002

Parés-Matos, EI. (2013). Electrophoretic mobility-shift and super-shift assays for studies and characterization of protein-DNA complexes. *Methods in Molecular Biology*. 977:159-67. doi: 10.1007/978-1-62703-284-1_12.

Perry, G.H., Melsted, P., Marioni, J.C., Wang, Y., Bainer, R., Pickrell, J.K., … Gilad, Y. (2012). Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research*. 22(4): 602–610. http://doi.org/10.1101/gr.130468.111

Pompey, J.M., Morf, L., & Singh, U. (2014). RNAi Pathway Genes Are Resistant to Small RNA Mediated Gene Silencing in the Protozoan Parasite *Entamoeba histolytica*. *PLoS ONE*. 9(9): e106477. http://doi.org/10.1371/journal.pone.0106477

Poptsova, M.S., & Gogarten, J.P. (2007). The power of phylogenetic approaches to detect horizontally transferred genes. *BMC Evolutionary Biology*. 7: 45. http://doi.org/10.1186/1471-2148-7-45

Radivojac, P., Clark, W.T., Ronnen Oron, T., Schnoes, A.M., Wittkop, T., Sokolov, A., … Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*. 10(3): 221–227. http://doi.org/10.1038/nmeth.2340

Reh, W.A., & Vasquez, K.M. (2014). Gene Targeting by Homologous Recombination. In: eLS. John Wiley & Sons Ltd, Chichester. http://www.els.net [doi: 10.1002/9780470015902.a0005988.pub2]

Rhoads, A. & Au, K.F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*. 13(5): 278–289. http://doi.org/10.1016/j.gpb.2015.08.002

Rinaldi, A.J., Lund, P.E., Blanco, M.R., & Walter, N.G. (2016). The Shine-Dalgarno sequence of riboswitch-regulated single mRNAs shows ligand-dependent accessibility bursts. *Nature Communications*. 7: 8976. http://doi.org/10.1038/ncomms9976

Rothberg, J.M. & Leamon, J.H. (2008). The development and impact of 454 sequencing. *Nature Biotechnology*. 26(10): 1117-24. doi: 10.1038/nbt1485.

Roux, J., Rosikiewicz, M. & Robinson-Rechavi, M. (2015). What to compare and how: Comparative transcriptomics for Evo-Devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. 324B: 372–382.

SanMiguel, P., Tikhonov, A., Jin. Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. & Bennetzen, J.L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science*. 274: 765–768.

Signal, B., Gloss, B.S. & Dinger, M.E. (2016). Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs. *Trends in Genetics*. 32 (10): 620–637. DOI: http://dx.doi.org/10.1016/j.tig.2016.08.004

Sitnicka, D., Figurska, K. & Orzechowski, S. (2015). Functional Analysis of Genes. *Advances in Cell Biology*. 2(1): 1–16

Schmidt, A., Forne, I. & Imhof, A. (2014). Bioinformatic analysis of proteomics data. *BMC Systems Biology*. 8(2): S3

Stoddard, B.L. (2011). Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure (London, England : 1993)*. 19(1): 7–15. http://doi.org/10.1016/j.str.2010.12.003

Stoddard, B.L. (2014). Homing endonucleases from mobile group I introns: discovery to genome engineering. *Mobile DNA*. 5: 7. http://doi.org/10.1186/1759-8753-5-7

Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Duchateau, P., & Pâques, F. (2011). Meganucleases and Other Tools for Targeted Genome Engineering: Perspectives and Challenges for Gene Therapy. *Current Gene Therapy*. 11(1): 11–27. http://doi.org/10.2174/156652311794520111

Swamy, M.N., Wu, H. & Shankar, P. (2016). Recent advances in RNAi-based strategies for therapy and prevention of HIV-1/AIDS. *Advanced Drug Delivery Reviews*. 103: 174-86. doi: 10.1016/j.addr.2016.03.005.

Sweeney, T.E., Haynes, W.A., Vallania, F., Ioannidis, J.P., & Khatri, P. (2017). Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Research*. *45*(1): e1. http://doi.org/10.1093/nar/gkw797

Tak For Yu, Z., Guan, H., Ki Cheung, M., McHugh, W.M., Cornell, T.T., Shanley, T. P., … Fu, J. (2015). Rapid, automated, parallel quantitative immunoassays using highly integrated microfluidics and AlphaLISA. *Scientific Reports*. *5*: 11339. http://doi.org/10.1038/srep11339

Thomason, M.K., & Storz, G. (2010). Bacterial antisense RNAs: How many are there and what are they doing? *Annual Review of Genetics*. 44: 167–188. http://doi.org/10.1146/annurev-genet-102209-163523

Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S. & Gregory, P.D. (2010). Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics*. 11: 636–646. doi:10.1038/nrg2842

van Deursen, J. (2002). Gene Targeting in Mouse Embryonic Stem Cells. *Methods in Molecular Biology*. 209: 145-158. DOI: 10.1385/1-59259-340-2:145

Veneziano, D., Nigita, G., & Ferro, A. (2015). Computational Approaches for the Analysis of ncRNA through Deep Sequencing Techniques. *Frontiers in Bioengineering and Biotechnology*. 3: 77. http://doi.org/10.3389/fbioe.2015.00077

Völkel, P., Le Faou P. & Angrand, P.O. (2010). Interaction proteomics: characterization of protein complexes using tandem affinity purification-mass spectrometry. *Biochemical Society Transactions*. 38(4): 883-7. doi: 10.1042/BST0380883.

Wilusz, J.E., Sunwoo, H., & Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes & Development*. *23*(13): 1494–1504. http://doi.org/10.1101/gad.1800909

Wittmann, A. & Suess B. (2012). Engineered riboswitches: Expanding researchers' toolbox with synthetic RNA regulators. *FEBS Letters*. 586: 2076–2083.

Yu, Y. & Yuan, J.X-J. (2010). Approaches for Manipulation of Gene Expression. In "Textbook of Pulmonary Vascular Disease". Print ISBN: 978-0-387-87428-9. Online ISBN: 978-0-387-87429-6 pp 557-566

Zhou, X., Peris, D., Kominek, J., Kurtzman, C.P., Hittinger, C.T., & Rokas, A. (2016). *In Silico* Whole Genome Sequencer and Analyzer (iWGS): a Computational Pipeline to Guide the Design and Analysis of *de novo* Genome Sequencing Studies. *G3: Genes|Genomes|Genetics*. *6*(11): 3655–3662. http://doi.org/10.1534/g3.116.034249