Comparative Analysis of Microarray and Bulk RNA-seq Data for Identifying Putative Interactions of NR5A1 in Azoospermia

BY

Muhammad Waseem Abbas

Fall-2021-MSBI-00000361112

Supervised by
Dr. Rehan Zafara Paracha

in

MS Bioinformatics
September 2023

School of Interdisciplinary and Engineering Sciences (SINES)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

Annex A to NUST Letter No.
0972/102/Exams/Thesis-Cert
dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr Muhammad Waseem Abbas Registration No. 00000361112 of __SINES___ has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: ____________________

Name of Supervisor: Dr. Rehan Zafar Paracha

Date: 22/11/23

Signature of HoD with stamp: Fmalik

Dr. Fouzia Malik
HoD Sciences
Associate Professor
SINES - NUST, Sector H-1
Islamabad

Date: 22 NOV 2023

Countersign by

Signature (Dean/Principal): ____________________

Dr. Hammad M. Cheema
Principal & Dean
SINES - NUST, Sector H-12
Islamabad

Date: 27/11/2023

This thesis is dedicated to my beloved Father Manzoor Hussain, my Mother, my Brothers Hafiz Asghar Ali, Hafiz Akbar Ali, Hafiz Naseem Abbas and Hafiz Zulfiqar Ali.

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes the outcome of the work done.

Muhammad Waseem Abbas

Acknowledgement

All glory belongs to the Almighty Allah, who is the fount of all wisdom. Almighty Allah has given me the ability to do something fresh, engaging, and path-pointing that has allowed me to attain this current level of understanding. All respect is due to the Holy Prophet Hazrat Muhammad (PBUH), who is a source of wisdom and guidance.

I dedicate this thesis to my beloved Brother Hafiz Asghar Ali. This achievement would not have been possible without your consistent support, confidence in my talents, and the numerous late-night discussions that kept me motivated. You are my confidant, source of inspiration, and loyal champion. Your tireless support for me has served as a catalyst for my pursuit of knowledge and academic success.

My sincere gratitude goes out to my supervisor, Dr. Rehan Zafar Paracha, for his superb leadership, mentoring, and support during this research project. His priceless knowledge, skill, and dedication have greatly influenced my thinking and helped make this thesis a success. I want to thank the GEC members Dr. Uzma Habib and Dr. Muhammad Tariq Saeed for their advice and support during this time.

Due to the company of my valuable Parents and Brother Hafiz Akbar Ali, whose presence in my life has strengthened my academic journey, the journey has become even more meaningful. I would like to express my sincere gratitude to my beneficial friend Muhammad Azam Rasheed, whose constant backing has been a continual source of inspiration. His undying belief in my competence has inspired me to overcome obstacles and pursue excellence.

And now to my colleagues and senior Maryum Nisar. My intellectual and per-sonal development have been enriched by their presence and our shared experiences. They are the recipients of this thesis, which serves as a testament to our continuing friendship and our same goals.

Contents

Nomenclature

Acronyms / Abbreviations

APCs	Antigen Presenting Cells
BCR	B cell receptor
BRN	Biological Regulatory Network
CAR	Coxsackie-adenovirus receptor
CD4+T cells	Helper T cells
CD8+T cells	Cytotoxic T cells
CNS	Central nervous system
CRS	Cytoreductive surgery
DAMPs	Danger-associated molecular pattern signals
DC	Dendritic Cells
EGF	Epidermal growth factor
FDA	US Food and Drug Administrationa
G-CSF	Granulocyte colony stimulating factor
GM-CSF	Granulocyte monocyte colony stimulating factor
HER2+	Human epidermal growth factor receptor 2+
HN	Hemagglutinin-neuraminidase
IL	Interleukin
INF	Interferon
MAPK	Mitogen-activated protein kinases
MHC	Major Histocompatibility complex
NDV	Newcastle Disease Virus
NK cells	Natural Killer Cells
NSCLC	Non-small cell lung cancer
PAMPs	Pathogen-associated molecular patterns
ROS	Reactive Oxygen specie
SCLC	Small Cell Lung Cancer
SLAM	Signalling lymphocytic activation molecule
STAT3	Signal Transducer and Activator of Transcription 3

SVV	Seneca Valley Virus
T-VEC	Talimogene laherparepvec
TAA	Tumor associated antigens
TAM	Tumor Microenvironment
TCR	T cell receptor
TK	Thymidine kinase
TNF	Tumor Necrosis Factor
VGF	Vaccinia growth factor
VV	Vaccinia Virus
WHO	World Health Organization

List of Tables

List of Figures

ABSTRACT

Azoospermia is one of the major causes of male infertility and is described as the absence of spermatozoa in the ejaculate. Azoospermia is the cause of infertility in more than one percent of males in the general population whereas 10\%–15\% of infertile men are affected by this problem. It is classified into two types i.e. obstructive azoospermia (OA) and non-obstructive azoospermia (NOA). NOA is the most prevalent kind of azoospermia and affects approximately 60\% of azoospermic males. It is caused by spermatogenesis failure due to different factors. There is no proper treatment available for NOA, however, sperm can be retrieved in some cases for in-vitro fertilization. This process is very expensive and has a very low success rate. Treatment options are urgently needed to increase sperm production and for targeting underlying causes. Multiple studies to understand the disease mechanism and improve the sperm retrieval rate have been reported; however, they reveal a comparatively low success rate. Studies have identified a number of genes important for spermatogenesis. Among the most important genes, Nuclear receptor subfamily 5, group A, member 1 (NR5A1, also known as steroidogenic factor 1[SF-1]) is important. NR5A1 is a nuclear hormone receptor, that plays a crucial role in regulating steroid hormone biosynthesis by targeting different genes in humans. Some transcription factors, cofactors, and transcription co-activators participate with NR5A1 in regulating NR5A1 target genes. Mutations identified in the NR5A gene have been acknowledged as being causally associated with Non-Obstructive Azoospermia (NOA). Some studies support the association of NR5A1 mutations with NOA and some conflict its association with NOA. Datasets from different platforms, one NGS dataset (GSE216907), and 2 micro-array datasets (GSE45885 and GSE 45887) were used in the current study for gene expression analysis of important genes associated with NOA. In the dataset, GSE45885, 839 genes demonstrated differential expression, while GSE45887 displayed 772 differentially expressed genes, and GSE216907 exhibited 1168 genes with differential expression. The number of common differentially expressed genes in the three datasets was 16. The common DEGs were used for pathway enrichment analysis and the HedgeHog Signaling pathway was identified as important with P-value 0.04. However, the expression of NR5A1, target genes of NR5A1, and its regulating cofactors are normal in the used datasets. This study also demonstrates the interaction profile of NR5A1 with its target genes and with cofactors

at the molecular level. The protein structures of NR5A1 and interacting partners were docked using the High Ambiguity Driven protein–protein DOCKing (HADDOCK) server. The binding affinity and interaction profile of NR5A1 protein with all interacting partners were analyzed. The NR5A1 protein shows interaction with all its target proteins, cofactors, and coactivators. NR5A1 shows the strongest interaction with CTNNB1 among all interacting proteins. The identified mutations of NR5A1 were searched in the interacting residues of NR5A1 with their interacting partners. From the identified mutations in different studies, only one mutation was present in the interacting residue and is present in only 0.4\% of the azoospermic cases used in that study. This study suggests that the mutations identified are not in the interacting residues of NR5A1, and the expression profile of NR5A1 and its interacting residues is also normal in the NGS dataset used in this study. The findings that negate the link between NR5A1 mutations and NOA are supported by this study.

Chapter 1

INTRODUCTION

1.1 Male Infertility

The current definition of infertility (medical definition) is the inability to get conceived after 12 months of unprotected genitals through the fertile phase of the menstrual cycles [1]. Large population surveys have estimated that marital infertility affects 70 million people worldwide [2]. Additionally, according to the WHO, 9% of couples globally suffer from fertility issues, and 50% of these cases are due to male factors [3]. In accordance with the current US figures collected in surveys up to 12% of males are infertile [4]. Studies also demonstrate that the chance of cancer increases in men with symptoms of the medical condition. Men with septic semen parameters have a 20 times higher chance of developing testicular cancer [5]. Male infertility and risk for prostate cancer are related [6]. Many kinds of abnormalities are found in sperm causing male infertility including Oligospermia, Globozospermia, Asthenozospermia, Azoospermia, Aspermia, etc.

1.2 Azoospermia

One out of six couples experience infertility issues with male infertility as the underlying cause in 50% of the cases [7]. Azoospermia is one of the major causes of male infertility and is described as the absence of spermatozoa in the ejaculate. It is classified into two types i.e. obstructive azoospermia (OA) and non-obstructive azoospermia (NOA). Azoospermia is the cause of infertility in more than one percent of males in the general population whereas 10%–15% of infertile men are affected by this problem [8].

A 40% rate of obstructive azoospermia is found in males with azoospermia [9]. Several genitourinary tract infections that result in blockage can be the cause of OA. These include the naturally occurring bilateral absence of the vas deferens, blockage of the ejaculatory and epididymal conduits, and atresia of the seminal vessels. Treatments in pelvic and inguinal regions resulting in complete congestion, such as a bilateral vasectomy, can all lead to OA [10]. Spermatogenesis is normal in most of the cases with OA. As a result, surgical removal of the obstruction using procedures like vasovasostomy or vasoepididymostomy is mostly used as part of therapy options for OA [11].

The most common type of azoospermia is non-obstructive azoospermia and affects approximately 60% of azoospermic males. The major cause of NOA is sper-matogenesis failure due to primary testicular failure and testicular failure. Pituitary or hypothalamic dysfunction could be the cause. NOA in most cases has an idiopathic pathogenesis. Both OA and NOA are important medical disorders that need to be treated. This study focuses on addressing NOA. NOA is more challenging to manage and treat due to the unknown impact on sperm production in the testes and the larger patient population.

1.2.1 Etiology

NOA is classified on the basis of underlying causes. NOA can be due to testicular failure (primary testicular failure), pre-testicular failure (secondary testicular failure), and idiopathic. Testicular failure affects 10% of total infertile men and is characterized by increased luteinizing hormone (LH) and follicle-stimulating hor-mone (FSH) levels, and smaller testis [11]. Pre-testicular failure (epigenetic hypogonadotropic hypogonadism) is characterized by decreased LH and FHS, and the small size of the testis [11]. The one with the inconclusive picture of testicular failure is also known as idiopathic. Idiopathic testicular failure is characterized by an increased level of FSH with a normal testicle size, normal FSH level with a small testicle size, or normal FSH with a normal testicle size. For example, testis maturation arrest in some cases is characterized by normal LH, FSH, and testicles and is caused due to some genetic abnormalities. [12]. Abnormal spermatozoa synthesis because of chro-mosomal dysfunctions as in Klinefelter syndrome or Y chromosome micro-deletions of sub-regions AZFa, AZFb, or AZFc also leads to NOA [13].

1.2.2 Pathophysiology

Pre-testicular and testicular NOA are the two types depending on the causing factors as shown in Figure 1.1. Pretesticular failure can be by birth or acquired and is due to the following factors:

- Pituitary Tumors: Pituitary tumors such as prolactinomas may be the causing factor of NOA. Prolactinomas cause excessive production of prolactin, a hormone in females responsible for milk production. [14]. Gonadotropin-releasing (GnRH) hormone is produced less when prolactin is produced, which in turn reduces the LH and FSH production from the pituitary glands. LH and FSH are essential for stimulating the testicles to produce testosterone and spermatozoa, resulting in impaired sperm production leading to azoospermia [15].
- Kallmann Syndrome: This is a condition in which failure to release GnRH hormone causes a defect at the hypothalamus level. It results from the failure of GnRH-releasing neurons to migrate.

Primary testicular failure presenting 10% of infertile men, with increased LH and FSH levels is caused by many factors. Mumps is a respiratory disorder caused by the mumps virus that has a linkage to the testis. The mumps virus harms seminiferous tubules and the interstitium of the testis, resulting in the lack of spermatogenic cells [16]. Primary testicular failure due to ganodotoxic drugs or treatments like chemotherapy or radiotherapy because almost 5-% of males are affected with cancer during their lifetime [17]. In some cases, genetic defects such as Klinefelter syndrome, are characterized by reduced testosterone levels, tight testicles, diminished penis size, and reduced hair growth on the body.

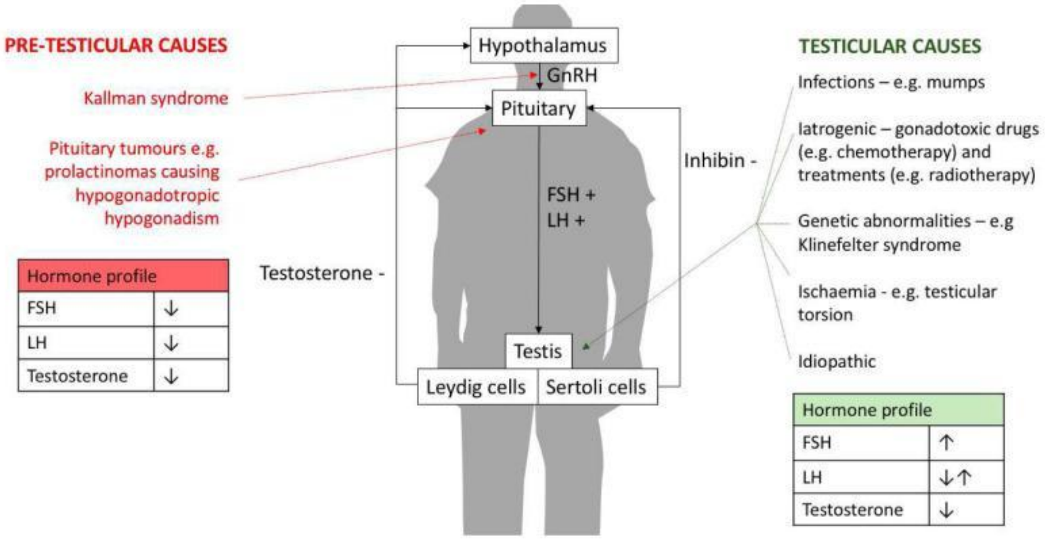


Figure 1.1. Pathophysiology of NOA Testicular and Pre-testicular origins of NOA. The figure is adapted from [18].

1.2.3 Epidemiology

Almost 15% of all couples experience infertility. Approximately 10 to 15% of infertile men and about 1% of all males suffer from azoospermia [19]. The majority of patients of azoospermia (around 600000) in the US have NOA [20]. In comparison with the non-azoospermic population, males with azoospermia have a higher risk of developing cancer. On average, 5% to 8% of testicular cancer patients will experience azoospermia. [21]. Azoospermia is mostly treated in medical centers and not reported in developing countries because of expensive or inaccessible therapy. Therefore, the exact ratio and cases of azoospermia are unknown [22].

1.2.4 Risk Factors

Obstructive azoospermia is due to blockage and in most cases, spermatogenesis occurs normally. However, non-obstructive azoospermia has an idiopathic pathogenesis with more cases between the age of 23-35 years. The risk factors associated with non-obstructive azoospermia are varied [23] as shown in Figure 1.2. Using chemical substances for treating conditions like cancer (Chemotherapy). Radiation therapy used for tumor shrinkage increases the risk of developing azoospermia. Testicular injury, heavy metal exposure, surgery in the reproductive area performed with the wrong technique, and exposure to high temperature for a long time. The use of recreational drugs such as some narcotics also increases the risk for azoospermia. Sometimes infections such as mumps also result in azoospermia.

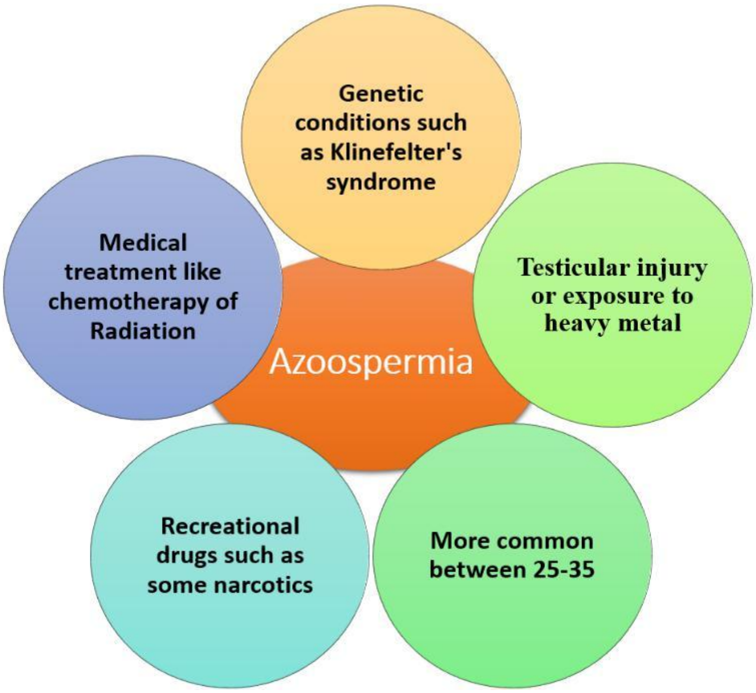

Genetic conditions such as Klinefelter's syndrome
Medical treatment like chemotherapy of Radiation
Testicular injury or exposure to heavy metal
Azoospermia
Recreational drugs such as some narcotics
More common between 25-35

Figure 1.2. Risk Factors associated with NOA Different risk factors associated with nonobstructive azoospermia.

1.2.5 Symptoms

Azoospermia is a condition that is diagnosed when a couple is struggling with infertility. A number of symptoms are associated with NOA varying with the causing factor. The main symptom of the disease is the absence of sperm in the ejaculate. Men with NOA caused due to hormonal imbalance may experience conditions like reduced body and facial hair. Mood swings due to hormonal imbalance and erectile dysfunction can also be the sign of NOA. Testis size is also affected in NOA causing discomfort or swelling around the testicles.

1.2.6 Diagnosis

There are a number of symptoms that indicate azoospermia, but the true causes are revealed by the medical tests and therapies. Semen analysis is the basic approach for diagnosing azoospermia. Patient's medical history (medical treatments or pro-cedures), and blood tests for confirming the percentages of LH and FSH hormones. Testicular biopsy also contributes to diagnosing azoospermia.

In men with NOA without a previous infertility record, karyotype and Y chromosome microdeletion (YCMD) testing are recommended [24]. Karyotype analysis is capable of identifying both structural and functional chromosomal abnormalities that are affecting up to 19% of NOA cases. Among karyotype abnormalities, Klinefelter syndrome is the most prevalent (47, XXY; occurs in 1/600 men). An increasing number of X chromosomes is associated with decreased spermatogenesis. In 10-20% of azoospermic men, abnormalities in AZF regions (AZFa, AZFb, AZFc) are identified using sequence-tagged sites (STS) PCR amplification. For epigenetic NOA patients, the most prevalent variation is Kallmann syndrome. Kallmann syndrome results in insufficient LH and FSH hormone production due to a decrease in GnRH. For diagnosing such cases specific genetic testing based on the mode of inheritance of KAL1, FGFR1, and other genes is required.

1.3 Methodologies used in Current Research

The increasing cases of azoospermia, make it more important to use new techniques and technologies that help early diagnosis and treatment of NOA. Advanced high throughput sequencing (HTS) techniques play an important role in finding out the potential causal agents. Analyzing these technologies in parallel and comparing the results can help in finding potential therapeutic targets. A couple of these tech-niques include

1.3.1 Microarray Analysis

microarray technology is capable of immediately detecting a wide range of different compounds in a sample. Therefore, the use of microarray technology has become very common in high-throughput applications. Large-scale genetic testing, gene expression profiling, comparative genomic hybridization, and resequencing are a few examples of the numerous uses of microarray technology. The development of microarray technology involved an extensive combination of many different scien-tific and technological fields. The fields include optics, microfabrication, chemistry, microfluidics, enzymology, and mechanics. RNA microarrays developed in the late 1990s are effective instruments for analyzing gene expression. microarrays can be sub-categorized as follows:

1. Spotted microarray in contrast to oligonucleotide arrays, which generate probes directly on the array, use specific sequences called probes (to detect expression) that are printed on a glass or plastic slide.

2. As compared to one color microarray in which one type of sample is hybridized on the array, two-color microarrays with two types of sample are hybridized sequentially on the microarray.

With the passage of time, microarrays have significantly improved, from initial arrays. The initial arrays were capable of holding a few hundred or thousand expressed sequence tags (ESTs). The latest microarrays are now capable of holding millions of probes covering the entire genome, including exons, introns, miRNAs, long coding RNAs, and other transcriptome variants [25]. A quickly evolving topic, microarray data evaluation has been applied to solve a number of issues. It is used for finding differentially expressed genes, developing prognostic or diagnostic predictors, and detecting data clusters. Data exploration, quality assurance, normalization, statistical analysis, and examination of biological significance or pathways are the steps involved in the analysis process. A number of tools are available for analyzing microarray data, ranging from free software to paid products.

1.3.2 Next generation sequencing

High throughput sequencing (HTS), also referred to as next-generation sequencing (NGS), is used to sequence DNA and RNA as well as find variations and mutations. With the use of NGS, a significant number of genes, maybe even an entire genome, can be easily sequenced. This technology combines the benefits of many sequencing chemistries, platforms, and bioinformatics techniques. By utilizing this combination, NGS makes it possible to sequence many DNA or RNA sequences in parallel, irrespective of their length, or even complete genomes, in a relatively short amount of time. After Sanger sequencing, it represents an innovative development in sequencing technology. NGS includes a number of essential steps in the sequencing procedure. The steps for NGS, as shown in Figure 1.3, include samples for sequencing, culture growth, DNA extraction, DNA quality control, library preparation, pooling and loading, template generation and sequencing, followed by bioinformatics anal-ysis, and annotation and interpretation of variations and mutations. Numerous uses for the sequence modifications and mutations discovered using NGS include disease diagnosis, prognosis, therapeutic discoveries, and patient follow-up [26].

1.3.3 Protein-protein interactions

PPIs (protein-protein interactions) are crucial for cell function at the molecu-lar level. For discovering more about residue interactions, binding area, and structural flexibility, understanding the molecular processes of PPIs is essential. The growing PDB library serves as proof of the advancements in the analysis of single protein structures. The structural description of protein complexes is still difficult, among

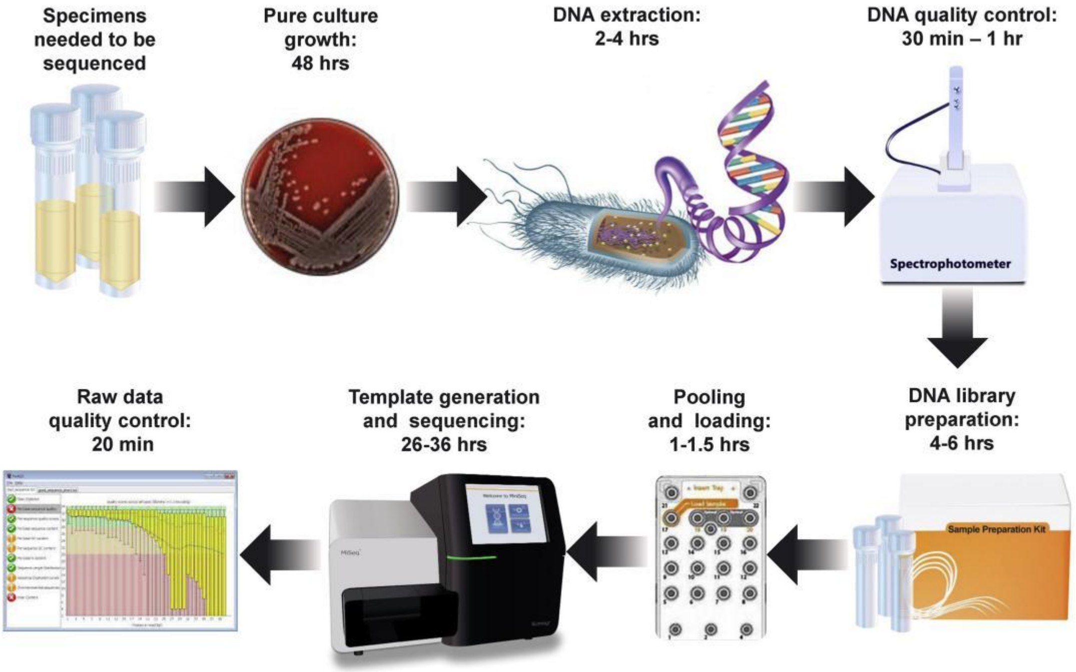

Specimens needed to be sequenced
Pure culture growth: 48 hrs
DNA extraction: 2-4 hrs
DNA quality control: 30 min – 1 hr
Spectrophotometer
DNA library preparation: 4-6 hrs
Sample Preparation Kit
Pooling and loading: 1-1.5 hrs
Template generation and sequencing: 26-36 hrs
Raw data quality control: 20 min

Figure 1.3. NGS Workflow A generalized NGS workflow starting from specimen leading towards bioinformatics analysis including different techniques. The workflow is adapted from [27].

other things. It is demonstrated by the fact that if the molecular complex has a significant molecular weight, it is either challenging or impossible to collect and evaluate the essential data using NMR spectroscopy. Presently, in silico protein-protein docking is the only technique available for the comprehensive study of large protein complexes. This method uses the unbound (free-form) protein structures, obtained experimentally or by comparative modeling, to estimate the most likely protein alignments in a complex. The following list shows the list of software packages that have been developed for predicting protein-protein interactions. The majority of these packages are based on their energetic and/or geometrical characteristics.

- ClusPro
- GRAMM-X
- HDOCK

- DOCKSCORE
- HawkDock
- ZDOCK

HADDOCK (High Ambiguity-Driven DOCKing), a recently created tool, uses a novel way to dock the provided proteins based on nuclear magnetic resonance (NMR) (and non-NMR) empirical information. If experimental data are not available, HADDOCK can also use an ab initio technique in addition to data-driven dockings [28].

Chapter 2

LITERATURE REVIEW

Spermatogenesis failure, which results in no sperm in the ejaculate, is NOA. The subsequent part describes a brief review of previous NOA research papers. Sev-eral studies have been performed for microarray and NGS dataset analysis of NOA patients. These studies were focused on understanding the expression profiles of dif-ferentially expressed genes and finding the disease bio-markers. The purpose of this chapter is to provide a summary of the important findings from earlier research studies on NOA.

2.1 Spermatogenesis and NOA

The association between spermatogenesis and NOA could provide the answer to one of the most basic queries about NOA. Spermatogenesis itself is a very com-plex process in which spermatogonia (male germ cells) are developed into mature sperm cells in different stages. Normal spermatogenesis is carried out with the help of different cells i.e. Leydig cells and Sertoli cells. Spermatogenesis takes place in-side seminiferous tubules and involves a pattern of mitotic, meiotic, and post-meiotic divisions. Following are the steps of spermatogenesis [29]:

- Spermatogonial Phase: Spermatogonial stem cells, which are located on the outermost layer of the tubules that contain seminiferous tissue, divide during mitosis to form spermatogonia. These spermatogonia will later develop into primary spermatocytes in some cases.
- Meiotic phase: Consists of the two phases of cell division known as meiosis I and meiosis II, which is when primary spermatocytes go through this process.

Secondary spermatocytes originate during meiosis I, while haploid spherical spermatids are generated during meiosis II.

- Post-meiotic phase: Round spermatids go through morphological changes to become elongated spermatids. The shape of the cell modifies during this phase, and the size of the cell decreases. Spermatids also go through spermiogenesis, which comprises the production of the acrosome, flagellum (tail), and nucleus condensation.
- Sperm maturation: Elongated spermatids (also known as immature sperm cells), transit from the adluminal region of the seminiferous tubules toward the lumen, undergoing more maturation along the path. In this procedure, extra cytoplasm is eliminated, and the acrosome continues to grow.
- Sperm Release: When sperm reach their full maturity, they are eventually re-leased into the epididymis, where they complete their development, become motile (able to swim), and fertilize an egg.

NOA is due to spermatogenesis failure or disruption that can occur at various stages during spermatogenesis.

- Spermatogonial Phase: Spermatogonial stem cell abnormalities or defects in their development might cause insufficient generation of spermatocytes, which can result in NOA.
- Meiotic phase: NOA can result from meiotic arrest, which occurs when spermatocytes are not able to move through meiosis I or meiosis II. This may be due to epigenetic mutations of other factors having an impact on meiotic division.
- Post-Meiotic Phase: Sperm that are deformed or ineffective can be produced as a result of deficiencies in spermatid enlargement and spermiogenesis, which contribute to NOA [30].

2.2 Existing Bio-markers

Semen analysis or testicular biopsy can be performed that acts as a bio-marker for NOA. The study of results of a testicular biopsy clarifies symptoms of NOA. NOA is a condition caused due to many factors that interfere with the levels of hormones such as LH, FSH, Inhibin B, and Anti-Müllerian Hormone (AMH). The profiles of these hormones of NOA are a potential bio-marker. For example, a reduced level of inhibin B is an indication of spermatogenesis failure and NOA [31]. Different studies have been performed using bioinformatics tools to find out molecular bio-markers associated with NOA. Previous studies show that a number of epigenetic mutations and abnormalities have been associated with NOA. These mutations or abnormalities may be involved in spermatogenesis, testicular growth, hormone regulation, and other processes.

Different genes play significant roles in the extremely complex process of spermatogenesis. Mutations in these genes may also cause spermatogenesis failure leading to NOA. A number of genes are very important in this regard including SYCP3, STAG3, SPATA20, SPACA4, NR5A1, TEX11, etc [32]. The meiotic arrest is a condition in which spermatogenesis is stopped during the meiosis step. This step involves the process of cell division that results in the generation of mature sperm. Mutations in the SYCP3 and SYCP2 genes have been identified in this condition [33]. Variations in these genes are found to be an effective risk for spermatogenesis failure leading to NOA.

2.3 Microarray associated findings for NOA

The molecular biology and genomics fields use the efficient method of microarray analysis to evaluate the expression level of thousands of genes at once. It offers information on how certain genes are expressed in various conditions or tissues. In scientific areas like gene expression profiling, bio-marker development, disease categorization, and pathway analysis, this method has been extensively used. Many studies have been performed to deeply study the gene expression levels in NOA using microarray.

Jie Lian and his colleagues in 2009 performed a study in which miRNA ex-pression profiles were looked at in the testes of individuals with non-obstructive azoospermia (NOA) and healthy, using microarray technologies. In NOA patients, 154 miRNAs showed differential lower expression, whereas 19 showed elevated ex-pression, indicating altered microRNA expression. RT-PCR studies on a few specific miRNAs, such as miR-302a, miR-491-3p, miR-520d-3p, and miR-383, evidenced these findings. Numerous miRNA clusters, including those with the potential to cause cancer, were lowly expressed in NOA patients. [34].

Genomic integrity maintenance depends on DNA repair mechanisms. An-other was carried out to assess the relationship between NOA and the DNA re-pair genes (322 genes). The relationship between the DNA repair genes RAD23B, OBFC2A, PMS1, UBE2V1, ERCC5, SMUG1, RFC4, PMS2L5, MMS19, SHFM1, INO80, PMS2L1, CHEK2, TRIP13, and POLD4 has been revealed by this work. The expression of RAD23B, OBFC2A, PMS1, UBE2V1, ERCC5, SMUG1, RFC4, PMS2L5, MMS19, SHFM1, and INO80 was elevated compared in six human sam-ples with various NOA. The expression profiles of PMS2L1, CHEK2, TRIP13, and POLD4 were down-regulated [35].

2.4 High throughput sequencing associated finding for NOA

Genetic research and application have been revolutionized by the use of the intense Next-generation sequencing (NGS), a sequencing methodology. Millions of DNA fragments may be rapidly and simultaneously sequenced with NGS, enabling genetic data analysis at a scale and depth that have never been achievable previously. Our knowledge of the genetic causes of NOA has considerably improved because of NGS. Several studies have been performed to date for the identification of genetic variants, mutations, and genes linked to NOA. Next-generation sequencing has provided new insights into the fundamental causes of NOA.

Govindkumar et al., 2019, conducted a study in which NGS profiles of 8 NOA were studied and identified 19 genes FAM71F1, CAPN11, BTG4, OAZ3, AKAP4, CHRNB3, CCDC83, PDHA2, PDCL2, ADAM29, SPATA3, SPERT, UBQLN3, SPAC A4, FBXO39, GGN, H1FNT, ZCCHC13 and POU5F2 therapeutic target genes for NOA [36].

In China, a study was conducted in 2019 in which NGS was performed on 34 NOA patients. This research screened and found low-frequency mutations of the genes involved with azoospermia. This data was utilized to create a database of single nucleotide variations (SNVs) linked to NOA. [37].

M Cerván Martín and his colleagues performed a study on 715 men with different types of male infertility disorders of which 505 were suffering from NOA. This study concluded that PIN1 gene polymorphism plays a crucial role in the de-velopment of single-cell-only syndrome, responsible for most of the cases of male infertility [38].

2.5 Genes important for Normal Spermatogenesis

Spermatogenesis is a very complex process consisting of different stages. Normal spermatogenesis is carried out with the help of several genes playing their role in different stages. Some of the important genes with their role in spermatogenesis are described below.

(A) Anti-Mullerian Hormone

Anti-Mullerian hormone (AMH) also known as Mullerian inhibiting substance or factor plays an important role in spermatogenesis. AMH is a glycoprotein that is secreted by Sertoli cells and is responsible for Mullerian duct regression in male embryos. The role of AMH is not confirmed in adult males [39]. Sertoli cells formed during embryonic development determine the quantity of germ cells in adults. In studies on factors affecting male fertility, AMH has recently attained more attention [40].

(B) Wilms’ tumour 1

Wilms’ tumour 1 (WT1) is a transcriptional regulator and is involved in several processes in vertebrates. WT1 is important for the development of some organs includ-ing kidneys, gonads, and heart [41]. Previous studies suggest that Wt1 is necessary in mice during the early stage of gonad growth and development [42]. Additionally, immunohistochemistry studies demonstrated that WT1 protein is strongly expressed in the Sertoli cells associated with early spermatogonia [43]. This suggests expression pattern of WT1 has a key role in normal spermatogenesis.

(C) Nuclear receptor subfamily 5 group A member 1

Nuclear receptor subfamily 5 group A member 1 (NR5A1) is a nuclear hor-mone receptor, that plays a crucial role in regulating steroid hormone biosynthesis by targeting different genes in humans [44]. NR5A1 targets several genes including AMH, STAR, MC2R, etc [44]. Some transcription cofactors such as SOX9 [45], WT1 [46], GATA4, and transcription co-activators including CTNNB1 [47] probably in-teract with NR5A1 and help in regulating NR5A1 target genes as shown in figure 2.1.

2.6 Association of NR5A1 with Spermatogenesis

Several factors including varicoceles, hormonal imbalance, testicular trauma, anatomical abnormalities of reproductive systems, chromosomal abnormalities, and Y chromosome microdeletions have been associated with male infertility. Thousands of genes taking place in spermatogenesis, testicular development, and endocrine regulation of testicular function are considered as the etiology of the disease. At least 15%

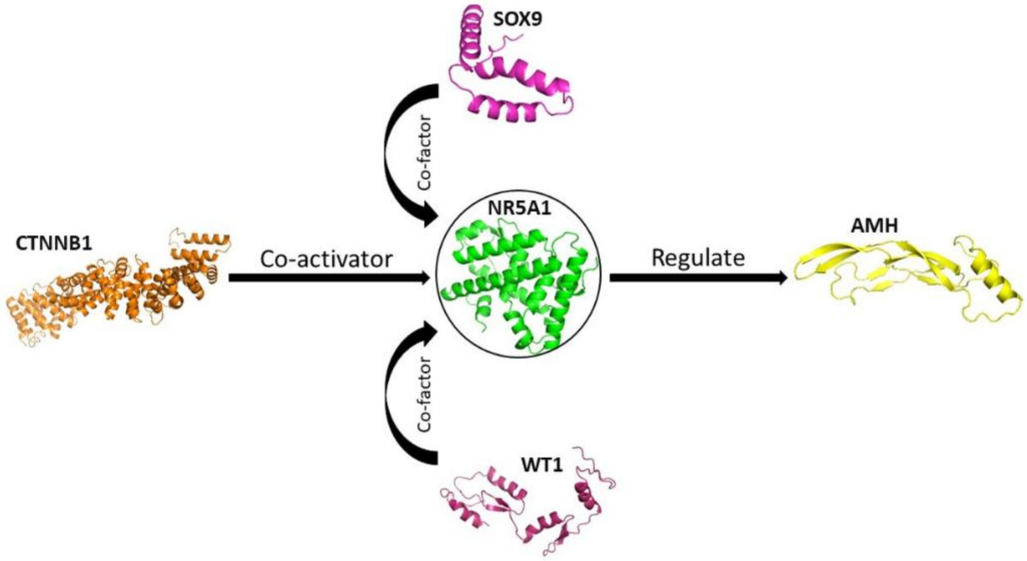


Figure 2.1. Target gene and co-factors of NR5A1

of infertile men are presented with defects in such genes [48].NR5A1 is among the genes that proved to be associated with male infertility by biological and functional evidence and is replicated in numerous independent studies [49]. NR5A1(Nuclear receptor subfamily 5 group A member 1, NM_004959.5) is located on chromosome 9q33, spanning about 30 kb long, and consisting of 7 exons (1 non-coding exon followed by 6 coding exons) [50]. The steroidogenic factor 1 (SF1) protein, encoded by the NR5A1gene, plays a pivotal role in steroidogenesis, sexual and adrenal development, and reproduction [51]. It is expressed in Sertoli and Leydig cells of the developing testis and Sertoli cells of the prepubertal and adult testis. Attempts to identify mutations of the NR5A1gene revealed several point mutations, which impair its function, leading to severe spermatogenic failure and male infertility.

2.7 Mutational Analysis of NR5A1

Currently, more than 188 different mutations in NR5A1 have been described, and they are scattered throughout all the protein domains [52]. These are found in a wide range of infertile phenotypes, including 46 XY disorders of sex development(DSD), cryptorchidism, non-obstructive azoospermia, and oligospermia patients.

Moreover, case-control association studies between polymorphisms and different types of male infertility have also been conducted in diverse populations, generat-ing different outcomes [53].

Researchers have different opinions on the association of NR5A1 mutations with spermatogenic failure leading to azoospermia. NR5A1 mutations associated with spermatogenic failure were first identified in 2010. It has been shown that missense mutations are present in the hinge region and proximal ligand-binding domain of NR5A1. These mutations may lead to the complete absence of spermatogenesis or a progressive decline in the quality and quantity of spermatozoa leading to azoosper-mia [54]. Another study in 2015 found that missense mutations in NR5A1 lead to azoospermia or severe oligozoospermia in about 1% of Caucasian men [55]. In 2018 a study conducted in India in which NR5A1 was sequenced in 502 infertile men versus 427 fertile men as controls. This study concluded that NR5A1 mutations are not associated with male infertility in Indian men [56]. Later on in 2021, a study was conducted in Vietnam that aimed to identify the single nucleotide polymorphism (SNP) associated with male infertility in the NR5A1 gene in a Vietnamese cohort of 202 infertile men and 199 healthy controls. However, no association was established between NR5A1 rs1110061 and male infertility [53].

2.8 Study Rationale

NOA is a very complex disease because it is caused by spermatogenesis failure. Spermatogenesis failure itself is a very complicated process involving differ-ent stages. There may be many reasons causing spermatogenesis failure leading to NOA, representing that several genes have an important role in this. Previous studies have identified several genes that are up-regulated, down-regulated, or even deleted in some cases. Mutations in spermatogenesis-associated genes are also reported to cause azoospermia. A key obstacle in finding a successful treatment for NOA is the variation in the expression pattern of these transcripts, which varies depending on the underlying etiology.

2.9 Proposed Solution

The identification of the most important genes for spermatogenesis and their expression profile in different datasets of NGS and microarray. Functional annotation of significant genes and their reported mutations from literature. Network analysis of the most important genes for identifying their interacting partners. PPI analysis of interacting genes to produce a better understanding at the molecular level and to observe the impact of mutated residues in interactions.

2.10 Objectives

- Using bioinformatics analysis, including microarray and Next-generation se-quencing, for identifying expression patterns of important genes, through exten-sive analysis using various samples of NOA patients, to establish their relevant correlations with NOA.
- Using literature for identifying genes and their mutations associated with azoosper-mia, and output expression files from different technologies to correlate expres-sion levels.
- Network analysis of selected genes to identify the known interacting partners, because proteins always work in groups. Performing PPIs of interacting part-ners to deeply understand the interaction at the molecular level and the impact of identified mutations in interaction.

Chapter 3

MATERIALS AND METHODS

This section of the thesis looks into how several high-throughput methods, such as microarray and RNA-Seq, are utilized to detect differentially expressed genes. Finding a relation between over-expressed and under-expressed genes to pinpoint the interacting residues is a further objective. This may aid in producing meaningful therapeutic results.

3.1 Data Collection

The analysis began with the collection of NOA-related datasets. Numerous freely accessible sources were explored to find datasets for microarray and RNA-seq research. Microarray and RNA-seq datasets were accessed by GEO https://www. ncbi.nlm.nih.gov/geoand Array Express https://www.ebi.ac.uk/arrayexpress/. While selecting datasets, the following parameters were taken into account.

1. The dataset was searched by entering the term "Azoospermia" in the search bar.
2. The dataset comprised "Homo-sapiens" and the samples were not taken from cell lines.
3. The dataset included samples from both the control and disease groups.
4. Dataset containing any medication, antibody, or small molecule should be elim-inated.

From the aforementioned databases, two microarray datasets and one mRNA-seq dataset were obtained. Local or Pakistani datasets were not accessible in open repositories. The diversity of the research sample area is demonstrated by the fact that the datasets selected are from different parts of the world. Summary of the datasets that were selected and presented in Tables 3.1 and 3.2.

Table 3.1. Microarray Datasets

Accession No.	Platform	No. of Samples	Region
GSE45885	Affymetrix	31	Norway
GSE45885	Affymetrix	20	Norway

Table 3.2. Next generation sequencing Dataset

Accession No.	Platform	No. of Samples	Region
GSE216907	Illumina Hiseq 2000	10	India

3.2 MicroArray Data Analysis

The initial step is to analyze the datasets using microarrays to identify differ-entially expressed genes. Secondary datasets were gathered from above mentioned freely accessible web repositories. For microarray analysis, the recently published maEndToEnd pipeline using R language is employed [57]. Figure 3.1 depicts the microarray procedure.

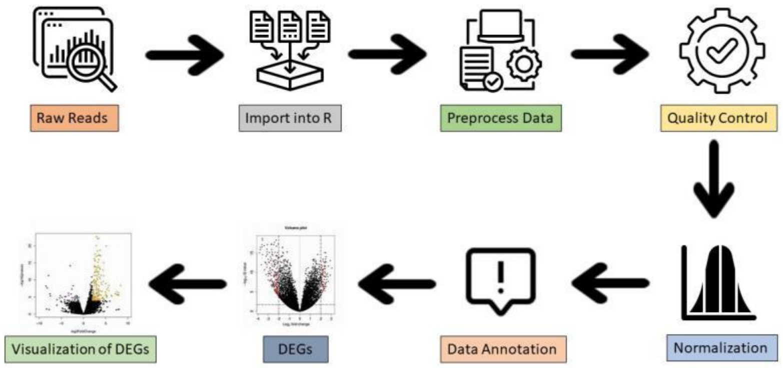


Figure 3.1. Generalized methodology of Microarray analysis

3.2.1 R Software

R serves as a programming language and an entirely free statistical analy-sis software. R Studio is a free and open-source R development environment. R Studios' most recent version, 4.2.1, was used for analysis. Bioconductor is a popu-lar bioinformatics package that offers tools for the evaluation and interpretation of high-throughput genomics data. BiocManager was employed to install the very latest version of Bioconductor, 3.1.7.

3.2.2 Uploading Raw Data

The data was gathered with ArrayExpress the package of Bioconductor using getAE and getGSE methods. IDF, ADF, SDRF, and CEL files were among those downloaded files. CEL files contain values of gene expression, whereas IDF and SDRF files contain additional details. The IDF file provides the title, description, submitter contact information, and protocols for the experiment. The SDRF file contains critical information about the experimental samples, such as their experimental group(s).

3.2.3 Storing Data in ExpressionSet

Downloaded files provide information about samples, experiments, and expression values. Once the data is imported, an Expressionset is created using Bioconductor's biobase function, and these data are saved in the ExpressionSet. ExpressionSet's data consists of assayData, metaData, and experimentData. Sample identifiers are displayed in columns, while microarray probes are displayed in rows, in AssayData. Metadata is made up of two types of data: featureData and phenoData. ExperimentData contains the description of the experiment. In the phenoData database, samples are grouped in columns and their descriptions are in rows. Both AssayData and FeatureData contain similar row information as well as freely assignable columns. The metadata also includes gene annotation for the features.

3.2.4 Quality Control

From raw data, once the ExpressionSet is generated, the quality of raw data is assessed using plots that include boxplot, PCA (Principal Component Analysis) plots, and RLE (Relative Log Expression) plots. The function (log2) of the biobase package was used for turning data into a logarithmic scale for checking the quality of the data. Quality control is an essential component in upholding data standards. Outliers were detected using the package array quality metrics, and once they were detected, normalization and summarization were carried out to remove biases. It produced a number of charts and an HTML report.

(A) Principle Component Analysis

Principal component analysis is a statistical approach used to reduce the size of a large dataset [58]. The PCA plot clearly differentiates between groups. The dataset includes samples (NOA and normal samples). Along the x and y axes, the PCA image depicts the relationship between Principal Component 1 and Principal Component 2 based on the raw log2 data. Each point represents a sample, with the color indicating either the sample is healthy or has NOA.

(B) Boxplot

An example of a graph that shows the intensity value distribution of a sample of data is a boxplot. It also identifies any outliers that might have an impact on the mean of the entire set of data. Samples are shown on the horizontal axis, while scaled intensity levels derived from log2 transformed data are shown on the y-axis [59]. Each box represents a single sample. If the intensity value distribution of each sample differs from the others, the data must be normalized for sample comparison.

(C) Relative Log Expression

Relative log expression (RLE) is the median of each gene's log intensity over all arrays. The rma function of the oligo package does RLE in addition to the other operations. The intensity distribution around the median for each sample is shown on the RLE plot. Each sample is represented by a box in the plot. The vertical axis shows the scaled log2 converted intensity values, and the horizontal axis shows the samples.

3.2.5 Data Preprocessing

Preprocessing of the data comprises of the following steps:

- Background adjustment
- Calibration
- Summarization
- Annotation

(A) Background Adjustment and Calibration

The intensity of every feature on microarray chips is calculated using a computer scanner. These scanners, which were subjected to various sorts of noise both within and between arrays, used a multitude of programs to measure fluorescence intensity. It was required to account for probe intensities produced by non-specific hybridization due to the noise this induced in the data. To obtain ambient strength for each feature on the microarray chip and to get rid of these disruptions, background correction is required. It is called calibration to normalize each feature's intensity value across the array so that they can be compared to one another.

(B) Summarization

On the microarray chip, each transcript is represented by several probes cre-ated especially for the Affymetrix technology. To determine an accurate and single-intensity value for each gene, several probe results must be combined, calibrated, and normalized into a single-intensity measurement. After summarization, each transcript or gene will have only one intensity value.

(C) Robust Multichip Average

There are many software available for pre-processing microarray data, albeit the methods change depending on the stage. A versatile algorithm in the oligo package is called Robust Multichip Average (RMA). Oligo implements the microarray background correction, calibration, and summary in a single step. Data is calibrated using quantile normalization, and deconvolution to correct for background. The RMA algorithm of the oligo package is used to summarize the data.

3.2.6 Quality Evaluation of Calibrated Data

After pre-processing, calibrated data was examined to confirm data quality once again. PCA plot and heat map were used to evaluate the quality. Following that, the calibrated data PCA plot and the raw data PCA plot were compared.

(A) HeatMap

A heatmap is used to group samples depending on the phenotype. Additionally, it determines the separation between each sample and shows the outcomes according to that separation. Each cell in the heatmap represents the level of a gene’s expression in a specific sample or situation [60]. The samples in this plot can be divided into two categories, such as NOA and normal.

3.2.7 Filtering Low Intensity Features

Microarray data contain representations of the overwhelming majority of the probes in the background intensity range. There is not much variance shown by these probes. As a result, exhibits low intensity and minimal fluctuation. They could con-sequently be categorized as genes with differential expression because their intensity value is so low that it rarely even falls within the detection range. To extract the genuine differentially expressed genes, Limma proposes deleting these data. The row medians of expression data were calculated, and a histogram was created to weed out low expression levels.

3.2.8 Annotation of Filtered Features

The next stage is to annotate or label genes with existing information, such as gene names, gene symbols, and so on. AnnotationDbi from Bioconductor is used to an-notate transcripts in expression data. Annotation databases such as "hugene10sttranscri ptcluster.db" are available from Bioconductor for each platform.

3.2.9 Fitting Linear model on Data

To analyze differentially expressed features between samples with NOA and normal samples, a linear model was fitted to the expression data for each gene. Limma software is used for model fitting. The goal of limma is to find similarities between the two groups. Limma acquires knowledge of variance across genes using Empirical Bayes and other techniques and allows analysis for small numbers of arrays through t-statistics. We generate design and contrast matrices for the variables of interest before fitting the linear model to the data. To determine whether the design matrix was effectively constructed, with the appropriate normal or diseased assigned to each sample. The rows of the design matrix contain information on the samples, and the columns indicate the variables used in linear models. The numbers 0 and 1 represent the assignment of samples to variables. A linear model with adequate contrasts for the test hypothesis was fitted to each gene using the design matrix. In our instance, we utilized the limma function makeContrasts to generate a contrast matrix of "NOA-Normal" data by comparing NOA and Normal samples. The data was then fitted with linear models using contrasts. Using the fit() method to locate genes with significant differential expression. A table with the gene symbol, gene name, log2FC, original p.value, changed p.value, test statistics (t), and B statistics is generated as output with this process.

3.2.10 Filtering out Deferentially expressed Genes

It is necessary to filter the list of DEGs created one step prior to excluding genes whose expression varies significantly between NOA and normal. Genes with log2FC greater than 1 and less than -1 with p-vales less than 0.5 were filtered as DEGs. The expression of a gene is considered to be "over-expressed" if its log2FC is greater than 1 and "under-expressed" if its log2FC is less than -1.

3.2.11 Graphical Representation of DEGs

For visualization DEGs, volcano, and enhanced volcano plots were used. These plots clearly show over and under-expressed genes and even genes that are not sig-nificant or with low log2FC values. Each dot on the plot represents a separate gene. These graphs display the -log10 p-values for the genes along with their log2 fold changes on the horizontal axis.

3.3 Next Generation Sequencing (RNA Seq)

Data from NGS are analyzed using a web server named Galaxy (https://usegalaxy.org/). Many tools and procedures for evaluating NGS data are available through the open-source web server Galaxy. Throughout the analysis, using an online server helps to save lots of memory and processing time. For the analysis of NGS data, we made use of a pipeline that was published in Nature Methods in 2016 [35]. The appendix portion contains the source code. Figure 3.2 shows the general NGS procedure.

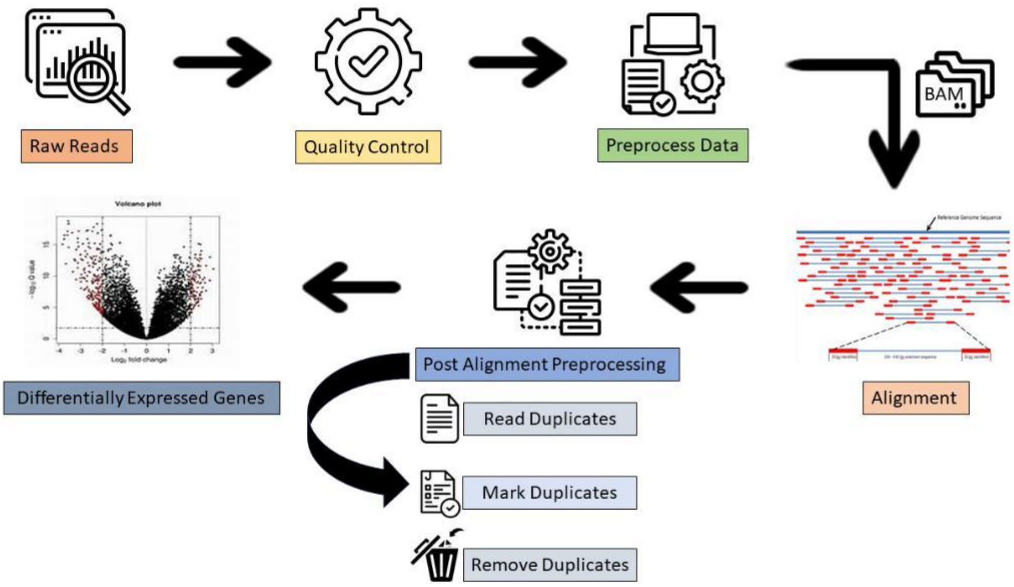


Figure 3.2. Generalized methodology of NGS analysis

3.3.1 Galaxy Pipeline

Galaxy has a straightforward yet effective user interface that handles all tool updates automatically. It can be accessed by the URL https://usegalaxy.eu/login. On high-throughput data, like RNAseq [61], it is used to perform complex-level analysis. For these jobs, Galaxy offered a variety of tools, including FastQC, RseQC, HISAT2, and StringTie. Users of the Galaxy interface have access to 250 GB of data storage for high-throughput analysis.

3.3.2 Importing Raw Reads in Galaxy

Data is first imported into Galaxy from other databases, such as the European Nucleotide Archive (ENA). Numerous options are available to import data from other databases in Galaxy. Data was imported via the EBI SRA option in Galaxy. One can access data in Galaxy by entering the samples’ accession numbers through the EBI

SRA. In our case, the downloaded data also contains the Fasta-formatted raw readings of the mRNA from each sample.

3.3.3 Preprocessing Raw Data

Many biological and experimental imperfections may be present in the data obtained from high-throughput sequencing. These artifacts include adapter contami-nation, GC content, over-representative regions, total base count, etc. These artifacts must be removed before continuing with the next step since they can bias our find-ings or result in erroneous positive results. The following steps are taken during the pre-processing of raw data to eliminate these biases.

(A) Quality Assessment of Raw Data

The reliability of the raw data must first be assessed. FastQC, a tool, was used to evaluate the quality of the raw data. FastQC offers a thorough quality check report for data that addresses many topics, including:

1. Basic data statistics, such as file type, file name, total sequencing, encoding, and GC content.
2. Per base sequence quality is described by a quality score, with a score of >28 indicating good quality.
3. A quality score map for each sequence describes the distribution of scores across all sequences.
4. Per base sequence content.
5. GC content per sequence describes how many GCs are there in a sequence.
6. Based on the per base N content.
7. Sequence length distribution, which is the distribution of the length of the se-quence across all sequences.

8. Sequence duplication level, i.e. how many reads in this data are duplicated.
9. Over-represented sequences, which are single sequences that are over-represented due to a biological effect or contamination.
10. Adapter content, i.e. whether or not adapter sequences employed in the experi-ment for hybridization are present.

(B) Preprocessing of Raw Reads

The FastQC report’s artifacts list identifies the places that require attention, such as adapter sequences, overrepresented regions, low quality, etc. FastP, another technique, was used to eliminate these noises. FastP is a pre-processing program that may be used in one step and offers a variety of choices, such as trimming low-quality bases, adapter removal, analysis of over-represented regions, etc.

(C) Quality Check after Preprocessing

FastQC reports on pre-processed measurements were used to assess data qual-ity. Because pre-processing enhanced data quality, these measurements were sub-jected to extra analysis.

3.3.4 Alignment to Reference Genome

To determine where raw readings are located, transcripts are aligned with a reference genome in a process known as alignment. The alignment is done using HISAT2 (Galaxy Version 2.2.1+galaxy1), which is a quick and accurate tool. The most recent version of the human reference genome in Galaxy, "Human CHM13 2.0 (T2T Consortium Jan.2022)," is the reference genome utilized for the read alignment. The information for the paired-end data strand was selected as "Forward(FR)". The number of transcripts that are aligned to the reference genome is listed in the report of alignment that HISAT2 prints after alignment. After alignment to the reference genome, raw fasta reads are transformed into BAM (Binary Alignment Map) file format.

3.3.5 Identification and Removal of Duplicate Reads

The final result can be affected by duplicate reads that may be due to experimental or biological artifacts. To examine the differential expression of genes, biological duplicates are required. To avoid false positive results, duplicates created by the PCR (Polymerase Chain Reaction) must be removed. Duplicates are found using MarkDuplicates (Galaxy Version 2.18.2.3), and they are removed using RmDup (Galaxy Version 2.0.1). Default values are used for both tools. MarkDuplicates uses a flag that RmDup can recognize to mark the repeated readings. RmDup recognizes these duplicates and removes them from the aligned BAM file.

3.3.6 Transcript Assembly and Quantification

StringTie (Galaxy Version 2.2.1+galaxy0) was used to assemble and quantify RNA sequence reads in BAM files. StringTie is an efficient and fast aligner. It has also the option of de Novo transcript assembly. We utilized the following StringTie input options:

1. Does the input contain long reads? No
2. Enter the strand information Forward(FR).
3. Use a reference file to assist with assembly. Use GTF or GFF3 as a reference.
4. The file CHM13-T2T-v2.0.gff3 is a reference file.
5. Only use reference transcripts? Yes
6. Differential expression output files? Ballgown
7. Availability of output coverage file? Absolutely

8. Output of the gene abundance estimation file?

Yes String tie generates the following output:

- A gtf file with assembled transcripts
- Gene abundances in tabular format
- The ballgown requires five files as input, which it utilizes to estimate differential expression.

3.3.7 Differentially Expressed Genes

The following five files are produced by StringTie and Tablemaker that can be imported into Ballgown:

1. e_data.ctab: This file contains exon-level expression measurements of the data.One row is dedicated per exon while the columns contain e_id, chr, start, end, etc. The file also contains the following quantification information for each sample:
    - rcount: Number of overlapping reads in exon.
    - ucount: Number of uniquely mapped and overlapped exon.
    - mrcount: Number of reads that overlap exon after multi-mapping correc-tion.
    - cov: Per-base read coverage average.
    - cov sd: Standard Deviation of read coverage.
    - mcov: Per-base average of read coverage in multi-mapped reads.
    - mcov sd:Multi-map-corrected per-base coverage standard deviation.
2. i_data.ctab: This file contains intron expression levels. Each row represents a single intron, and columns comprise i_id, chr, strand, start, end, and so on.

3. t_data.ctab:Transcript concentrations are present in this file. Each row represents a transcript, and the columns contain the following information:

    - t_id: Transcript ID
    - t_name: Cufflinks-generated Transcript ID
    - num_exon: the number of exons in the transcript;
    - length: the length of the transcript;
    - gene_id: the ID of the gene relating to the specific transcript;
    - gene_name: the transcript's HUGO gene name.

4. FPKM: Cufflinks estimates FPKM values for the transcript.

5. e2t.ctab: A table having two columns, e_id and t_id, that link exons to tran-scripts. The file should have the same ids as the e_data.ctab and t_data.ctab tables.

6. i2t.ctab: A table containing two columns, i_id and t_id, that connect introns to their respective transcripts. The file should have the same ids as the i_data.ctab and t_data.ctab tables.

    For statistical analysis, Differential expression analysis, and visualization of assembled transcripts, Ballgown which is an R-programming Bioconductor tool, is used. Stringtie's output files are imported, and the following information is necessary so that DEGs have to be obtained.

    - Phenotypic data: Information about the samples.
    - Expression data: Information on the size of the intron, exon, or both. The sample contains transcripts and genes.
    - Genomic data: Information regarding gene and transcript coordinates as well as exons.

The following step involves seeing distributed FPKM data that have been stan-dardized for library size. A linear model is employed to calculate differential expression.

3.3.8 Graphical Representation of DEGs

For visualization DEGs, volcano, and enhanced volcano plots were used, which clearly show over and under-expressed genes and even genes that are not sig-nificant or with low log2FC values. Each dot on the plot represents a separate gene. These graphs display the -log10 p-values for the genes along with their log2 fold changes on the horizontal axis.

3.4 Network Analysis

After differential gene expression analysis, common differentially genes between all the datasets were found. The common differentially genes were used for pathway enrichment analysis. The pathway with a significant P-value Hedgehog Sig-naling Pathway in our study with a p-value of 0.04 was used for network analysis to identify the most important genes in the pathway. Cytoscape is an open-source platform for visualizing, analyzing, and modeling complex networks was used for network analysis. Centrality measures, such as betweenness, and closeness centrality, were used to identify the most important nodes in the network.

(A) Betweenness Centrality

Betweenness centrality quantifies the importance of a node in a network based on its position as a bridge or intermediary between other nodes. Nodes with high betweenness centrality lie on many of the shortest paths between pairs of nodes in the network. Betweenness centrality is often used to identify nodes that are critical for maintaining network connectivity or for controlling the flow of information in various applications, including social networks, transportation networks, and biological networks.

(B) Closeness Centrality

Closeness centrality is a measure that assesses how quickly information or influence can spread from a node to all other nodes in the network. It quantifies how "close" a node is to all other nodes in terms of geodesic distance, where geodesic distance is the shortest path length between nodes. Closeness centrality is used to identify nodes that can rapidly disseminate information or influence throughout the network. It’s particularly relevant in situations where quick communication or the efficient transfer of resources is essential.

3.5 Protein Selection

Protein was selected based on interactions performed using STRING database, a biological database, and web resource for known protein-protein interactions freely available at https://string-db.org/. The following steps were used for protein selection:

1. Searched for important genes on the basis of their association with azoospermia and spermatogenesis from literature and expression profiles in the used datasets.
2. Mutations in identified genes leading to azoospermia were also searched in literature and used for further analysis. The expression profiles of selected genes were then searched in the used datasets.
3. Interaction analysis of the selected genes was performed using the STRING database and only genes with known interactions were kept for further analysis.

The selected genes were used for further analysis i.e. protein-protein interaction and later on genes were filtered on the basis of protein structural availability.

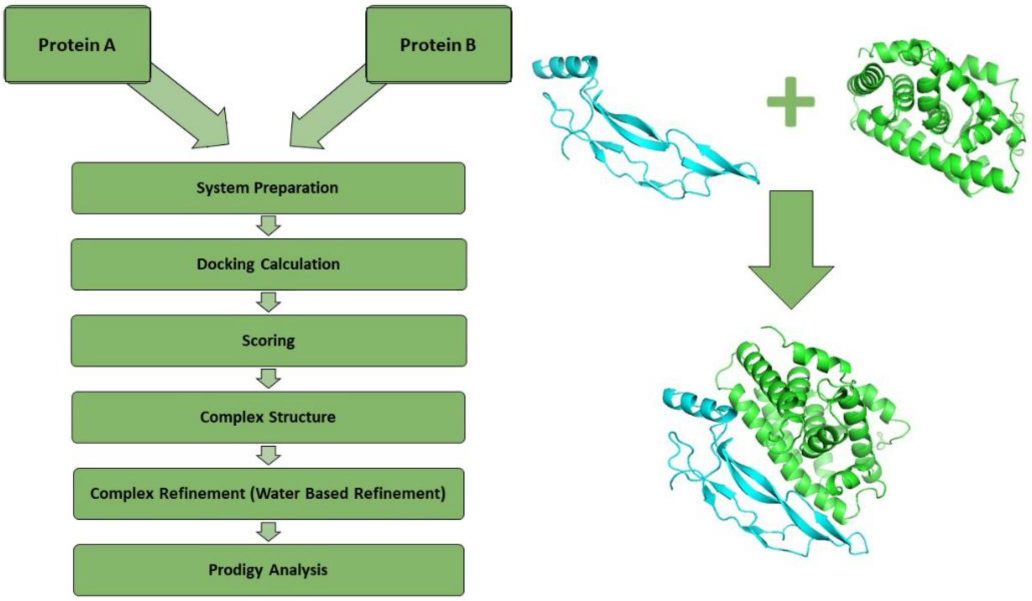


Figure 3.3. Generalized Workflow of protein-protein interaction

Selected genes were looked up in PROTEIN DATA BANK (PDB) freely available at https://www.rcsb.org/. The model with high resolution and no mutations was selected for for remaining steps. Uniprot for used for functional analysis domain study. Position-specific Blastp was used to search the sequence, and the percentage of identity was verified. Here prprotein models with percentage identity above 99% were used as it is, models with identity in range 75%-98% were modeled using SWISS-MODEL, models with similarity percentage range between 55%-74% were modeled using trRosetta and proteins which showed similarity less than 55% were filtered.

3.6 Protein-protein Interactions

Predictions about the binding affinities of the proteins and interacting residues are possible from protein-protein interaction. Despite the fact that there are other PPI interaction web servers available, studies continually place HADDOCK as one of the finest PPI programs in terms of quality, usability, and effectiveness. The docking workflow employed in research is shown in Figure 3.3.

(A) Prediction of Protein Binding Areas

Protein interface residues were computer-generated using SPPIDER web server available at https://sppider.cchmc.org/. It appears to provide more accurate interpretations of the interface residues that can be used for docking than High Ambi-guity Driven biomolecular DOCKing (HADDOCK).

(B) Docking Analysis

After preparing the required files and predicting interacting residues, protein-protein interaction is performed with the help of HADDOCK. Two protein structures were given as input along with the list of interacting residues. The resulting protein-protein complexes are analyzed on the basis of different factors and scoring functions including:

- HADDOCK SCORE: The projected protein-protein complex's overall quality is indicated by the HADDOCK score, an aggregate value. It is used to rank and choose the most advantageous docking poses and is derived from several energy components and phrases. The lower the HADDOCK the more biologically relevant and energetically favourable will be the interaction.
- Cluster Size: HADDOCK groups the resulting positions based on their struc-tural similarity after completing docking calculations. The number of poses that make up a certain cluster is referred to as the cluster size. This knowledge aids in locating the most prevalent and possibly stable binding mechanisms.
- RMSD from the Overall Lowest-Energy Structure: Root Mean Square Devia-tion, or RMSD, is a metric used to compare or contrast the structural similari-ties or differences between two protein conformations. The average deviation (measured in angstroms) between each position and the cluster's lowest-energy stance is given by this word in the context of HADDOCK. It aids in evaluating a cluster's structural diversity.

- Van der Waals energy: Describes how strongly atoms in a docked complex in-teract with one another. It contributes to the total binding energy and represents the attractive and repulsive forces between unbounded atoms. Complex stability is influenced by favorable van der Waals interactions.
- Electrostatic Energy: Electrostatic energy explains the electrostatic interactions between charged atoms or groups in the complex, such as attracting (such as Coulombic) and repulsive forces. It adds to the overall binding energy in a manner similar to van der Waals energy.
- Desolvation energy: During docking, when molecules are in close range to one another, this phrase refers to the energy needed to dislodge water molecules from the binding interface. It consists of both positive and negative contribu-tions, the latter of which carries desolvation liabilities.
- Restraints Violation Energy: HADDOCK uses practical or bioinformatics-derived information to direct the docking computations, such as NMR-derived distance restraints or other interaction data. The degree to which the generated postures adhere to these limitations is measured by the restraints violation energy. Better interaction is indicated by lower values.
- Buried Surface Area: During complex formation, the interacting molecules that become inaccessible to the solvent are referred to as buried surface area. A more stable interaction is represented by a greater buried surface area.
- Z-Score: The Z-Score is a statistical metric used to compare a given docking score to a variation in scores for random or loose configurations. Higher Z-scores indicate more significant and advantageous binding, and they are used to determine the importance of the HADDOCK score.

When employing the HADDOCK algorithm to study protein-protein interactions, the above-mentioned numerous energy components and metrics are essential for assessing and choosing the most biologically pertinent docking poses and complexes.

(C) Complex Refinement

In order to increase the precision of protein-protein and protein-ligand complex predictions, HADDOCK offers the option for water-based refinement. Haddock's top cluster of protein complexes was submitted to the refinement website for water-based refinements. In HADDOCK, the full refinement stage, which occurs later in the docking process, is when water-based refinement often enters the picture. Incorporating water into the computations seeks to more accurately mimic the biological setting and increase the predictability of protein-protein interactions. A more explicit or implicit model of the solvent (water) is provided in water-based refining. This indi-cates that the calculations account for the interactions between and among the water molecules and the proteins. The refinement process involves the following steps:

- Input files (Protein Complex)
- Access to HADDOCK Refinement Server
- Upload input files
- Solvent selection (water)
- Job Submission
- Result download and Analysis

(D) Prodigy Analysis

For the advancement of treatments and the comprehension of biological processes and illnesses, it is essential to know the structural properties of protein-protein interactions. A vital part of this is the accurate prediction of the binding strength for a protein-protein complex [62]. Haddock provides a webserver known as Pro-tein Binding Energy Prediction (PRODIGY) Prodigy webserver is used for finding binding affinity and binding residues between protein-protein complexes given the 3D structure based on different parameters. The best protein-protein complex from the refinement outputs is submitted as input with default parameters. The parameters used for analyzing the prodigy results include:

- Binding affinity assumption (G) expressed in kcal mol-1: This is a prediction of how the Gibbs free energy (G) will change when the two proteins attach to one another. It is measured in kilocalories per mole (kcal/mol) units. The thermodynamic stability of protein-protein complexes is frequently evaluated using the constant G.
- Kd (M) at °C: The equilibrium constant for the breakdown of a complex into its individual proteins is represented by the dissociation constant (Kd). It is measured in molarity units (M) and frequently calculated using the temperature and G value. Stronger binding is indicated by a lower Kd.
- Interacting Contacts (ICs): The amount of interactions or relationships be-tween the two proteins within a given distance threshold (5.5 angstroms) is presumably revealed by this. It can further be categorized based on the proper-ties of interacting residues such as charged-charged, and charged-polar.
- Non-interacting surface (NIS): As the name indicates it represents the per-centage of polar and apolar non-interacting residues.

Chapter 4

RESULTS

The results of the methods used to identify variations in gene expression are presented in this portion of the thesis. Additionally, it analyzes the connection between up-regulated and down-regulated genes, shows the interacting residues between af-fected proteins, and assesses the likelihood of substantial therapeutic improvements. Results include microarray analysis, RNA-Seq analysis, network analysis, and protein-protein interaction analysis.

4.1 Microarray Analysis

After data is extracted from Array Express and GEO, processing is performed using appropriate tools. The analysis includes pre-processing of the data, normaliza-tion, background adjustment, and detection of DEGs. All the datasets are of Affymetrix microarrays.

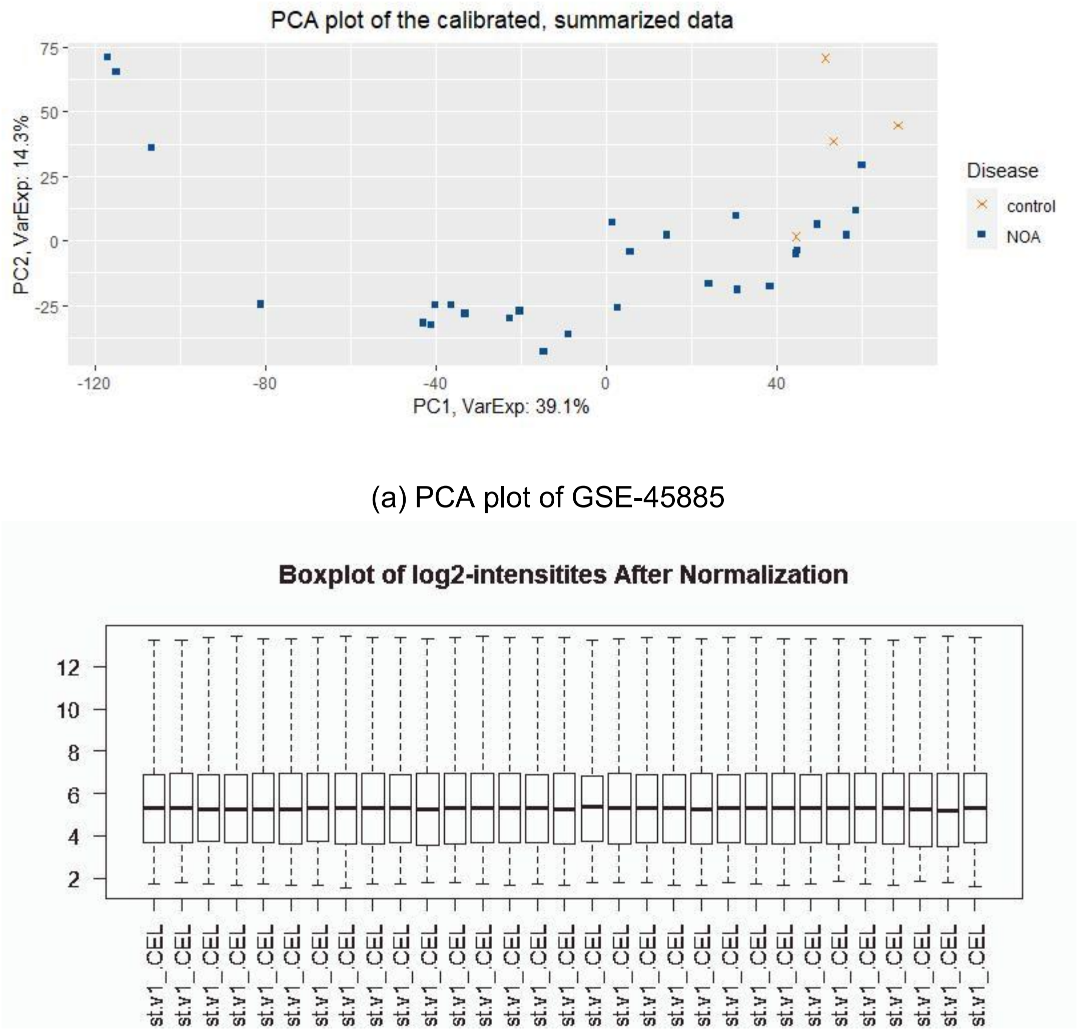
4.1.1 GSE-45885

(A) Pre-Processing of Raw Data

Pre-processing of microarray data generated the normalized, calibrated, summarized, and annotated data. In this normalized data, samples in the data are compared with each other. The variation between normal and disease is represented by the PCA plot, Figure 4.1a shows NOA samples as orange crosses and control and normal samples as blue boxes, where PC1 exhibits 39.1% variation and PC2 exhibits 14.3% variation. Figure 4.1b shows a normalized boxplot of the dataset with sam-ples on the y-axis and log2 intensity values on the x-axis. Boxplot of the normalized data plotted to find out whether the data normalized correctly or not. Each sample in the box plot is represented by each box whereas, bars on both sides of the boxes show upper and lower quantiles. As the median values of the samples coincide with each other, it shows data is normalized and samples are compared with each other. Relative Log Expression is another quality check procedure, through which median log-transformed values are plotted in order to check inter-sample median distribution. In the RLE plot represented in Figure 4.2a, the x-axis describes samples and the y-axis describes log2 expression values. The median of each sample is around 0 and shows the median coincide with each other. Extended lines at each end of the box describe the distribution of data. Both the histogram and bar plot describe the distribution of median intensities and p-values respectively. The X-axis of the histogram as shown in Figure 4.2b shows the median intensity and the y-axis represents the density of probes or genes at a given intensity. There are a few bars on the left side with low intensity, we used a 4 cutoff threshold which is standard to remove these values from the data. The bar plot in Figure 4.3 shows the frequency of the p-value in which the x-axis represents p-values and the y-axis shows a number of genes. Most of the genes are significant with a low p-value. Heatmap as shown in Figure4.4a is used to illustrate sample clustering with each other. Both phenotypes cluster into two different clusters representing that samples of normal have a high correlation with each other.

(B) Differential Expressed Genes and Visualization

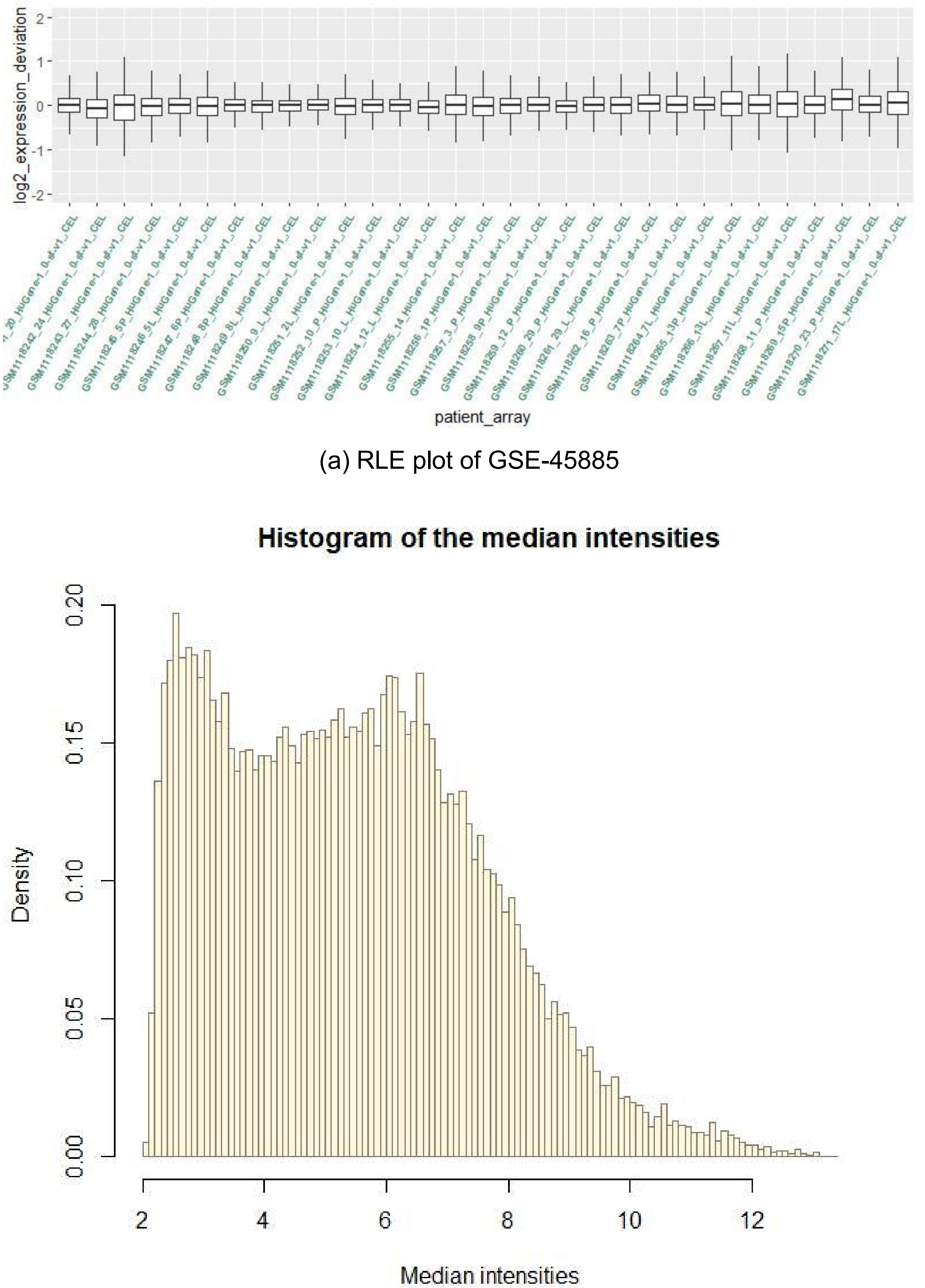
After DEGS analysis, volcano and enhanced volcano plots are used for visualization. The threshold for filtering genes was p-value less than 0.05 and log2FC ±1. Gene distribution according to log2FC on the horizontal axis and p-value (-log10) on the vertical axis is displayed in Figure 4.4b. Each dot stands for a unique gene. Genes present in the grey area are non-significant. Blue-dotted genes are filtered through just the p-value threshold. Same as blue-dotted genes, green-dotted genes only passed the LogFC threshold. Genes represented with red dots are significant genes with low p-values i.e. less than 0.05 and log2 fold change greater than ±1. These genes passed through both cutoff values. Highlighted genes less than 1 log2FC are down-regulated genes and genes with greater than 1 log2FC are up-regulated genes. A total of 883 differentially expressed genes passed out through both thresholds. The top 10 Filtered DEGs are listed in Table 4.1.



(a) PCA plot of GSE-45885

(b) Boxplot of GSE-45885

Figure 4.1. PCA and Boxplot of Dataset GSE-45885



(a) RLE plot of GSE-45885

(b) Histogram of GSE-45885

Figure 4.2. Quality Assessment of Dataset GSE-45885

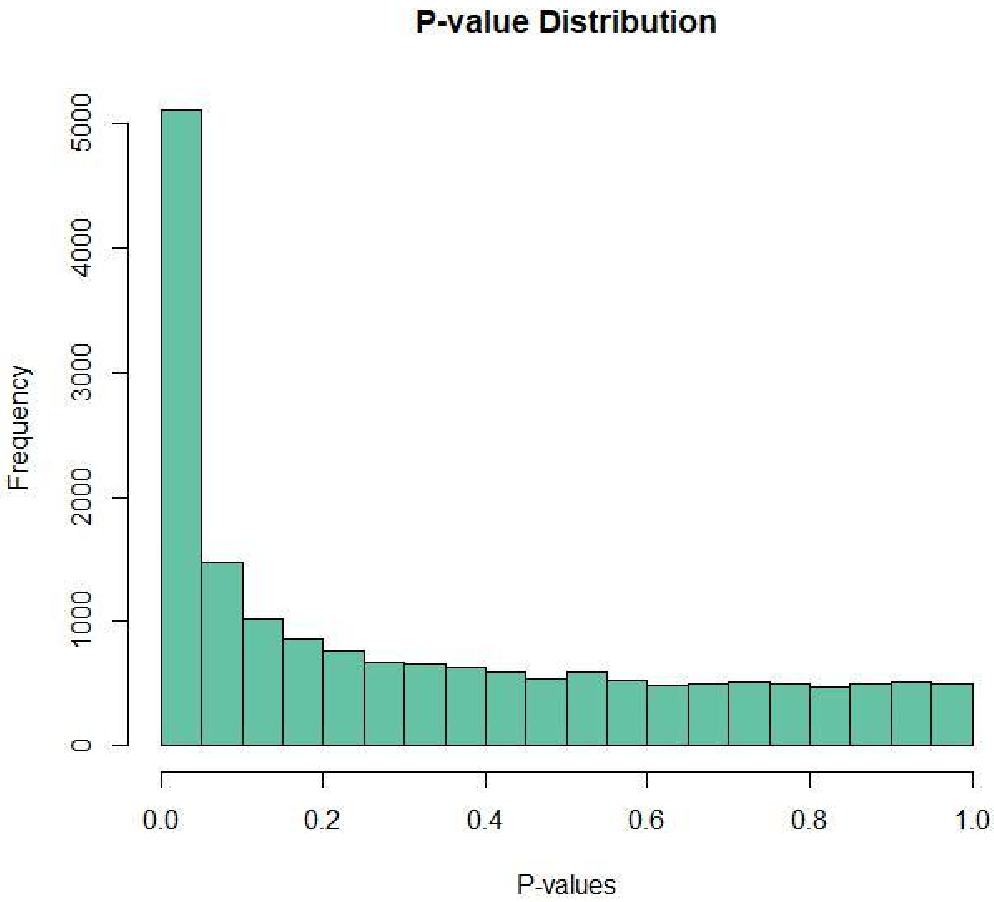
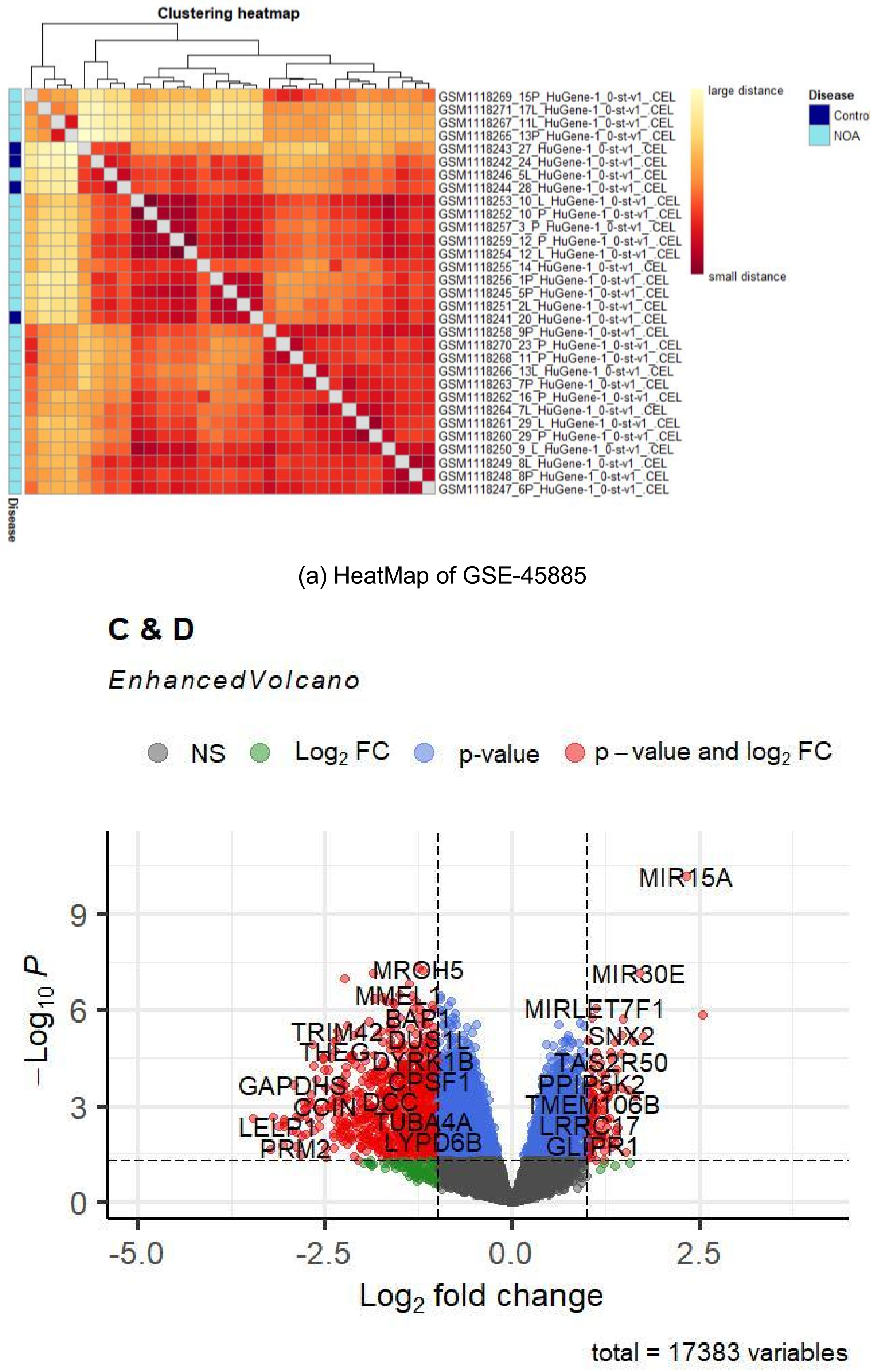


Figure 4.3. P-value Distribution Graph of Dataset GSE-45885



(a) HeatMap of GSE-45885

(b) Enhanced Volcano plot of GSE-45885

Figure 4.4. Visualization of Dataset GSE-45885

Table 4.1. Top 10 DEGs of Dataset GSE45885

PROBE ID	SYMBOL	GENE NAME	log2FC	P-value	adj. P-value
8087881	MIRLET7G	microRNA let-7g	1.810552564	3.93E-12	9.16E-08
7971661	MIR15A	microRNA 15a	2.321978085	6.59E-11	7.67E-07
8153273	MROH5	maestro heat like repeat family member 5 (gene/pseudogene)	-1.244966667	4.86E-08	0.00021734
8048350	PLCD4	phospholipase C delta 4	-1.190788111	5.94E-08	0.00021734
7900488	MIR30E	microRNA 30e	1.706107373	7.00E-08	0.00021734
7906527	ATP1A4	ATPase Na+/K+ transporting subunit alpha 4	-1.849361041	7.47E-08	0.00021734
8043639	FER1L5	fer-1 like family member 5	-2.235849415	1.01E-07	0.000260554
8014298	HEATR9	HEAT repeat containing 9	-1.375539888	1.51E-07	0.000350827
7911767	MMEL1	membrane metalloendopeptidase like 1	-1.499232822	3.28E-07	0.000549306
7973629	REC8	REC8 meiotic recombination protein	-1.337024926	3.63E-07	0.000549306

si
s

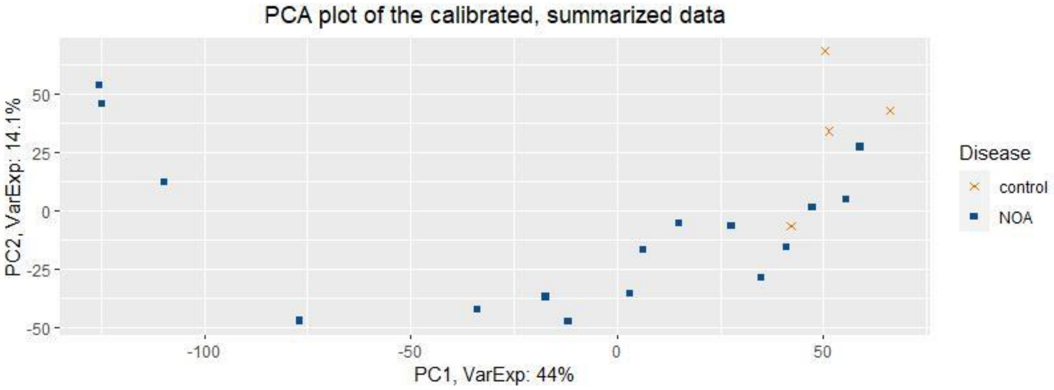
4.1.2 GSE-45887

(A) Pre-Processing of Raw Data

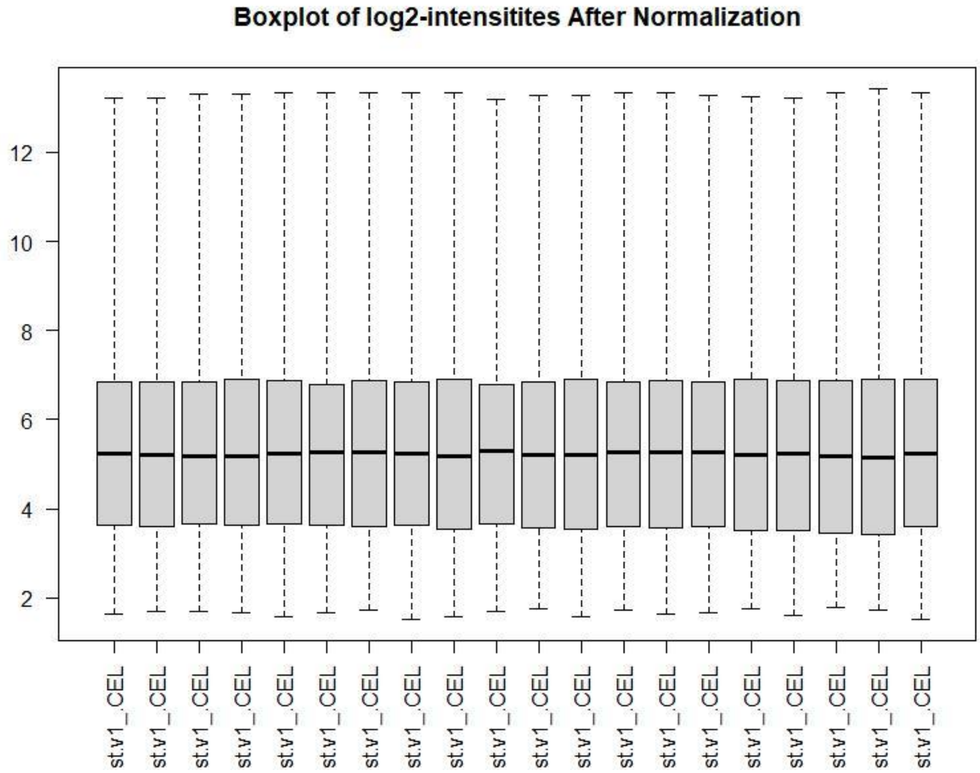
Pre-processing of microarray data generated the normalized, calibrated, summarized, and annotated data. In this normalized data, samples in the data are compared with each other. The variation between normal and disease is represented by the PCA plot, Figure 4.5a shows NOA samples as orange crosses and control and normal samples as blue boxes, where PC1 exhibits 44% variation and PC2 exhibits 14.1% variation. Figure 4.5b shows a normalized boxplot of the dataset with sam-ples on the y-axis and log2 intensity values on the x-axis. Boxplot of the normalized data plotted to find out whether the data normalized correctly or not. Each sample in the box plot is represented by each box whereas, bars on both sides of the boxes show upper and lower quantiles. As the median values of the samples coincide with each other, it shows data is normalized and samples are compared with each other. Relative Log Expression is another quality check procedure, through which median log-transformed values are plotted in order to check inter-sample median distribution. In the RLE plot represented in Figure 4.6a, the x-axis describes samples and the y-axis describes log2 expression values. The median of each sample is around 0 and shows the median coincide with each other. Extended lines at each end of the box describe the distribution of data. Both the histogram and bar plot describe the distribution of median intensities and p-values respectively. The X-axis of the histogram as shown in Figure 4.6b shows the median intensity and the y-axis represents the density of probes or genes at a given intensity. There are a few bars on the left side with low intensity, we used a 4 cutoff threshold which is standard to remove these values from the data. The bar plot in Figure 4.7 shows the frequency of the p-value in which the x-axis represents p-values and the y-axis shows a number of genes. Most of the genes are significant with a low p-value. Heatmap as shown in Figure 4.8a is used to illustrate sample clustering with each other. Both phenotypes cluster into two different clusters representing that samples of normal have a high correlation with each other.

(B) Differential Expressed Genes and Visualization

After DEGS analysis, volcano and enhanced volcano plots are used for visualization. The threshold for filtering genes was p-value less than 0.05 and log2FC ±1. Gene distribution according to log2FC on the horizontal axis and p-value (-log10) on the vertical axis is displayed in Figure 4.8b. Each dot stands for a unique gene. Genes present in the grey area are non-significant. Blue-dotted genes are filtered through just the p-value threshold. Same as blue-dotted genes, green-dotted genes only passed the LogFC threshold. Genes represented with red dots are significant genes with low p-values i.e. less than 0.05 and log2 fold change greater than ±1. These genes passed through both cutoff values. Highlighted genes less than 1 log2FC are down-regulated genes and genes with greater than 1 log2FC are up-regulated genes. A total of 772 differentially expressed genes passed out through both thresholds. The top 10 Filtered DEGs are listed in Table 4.2.

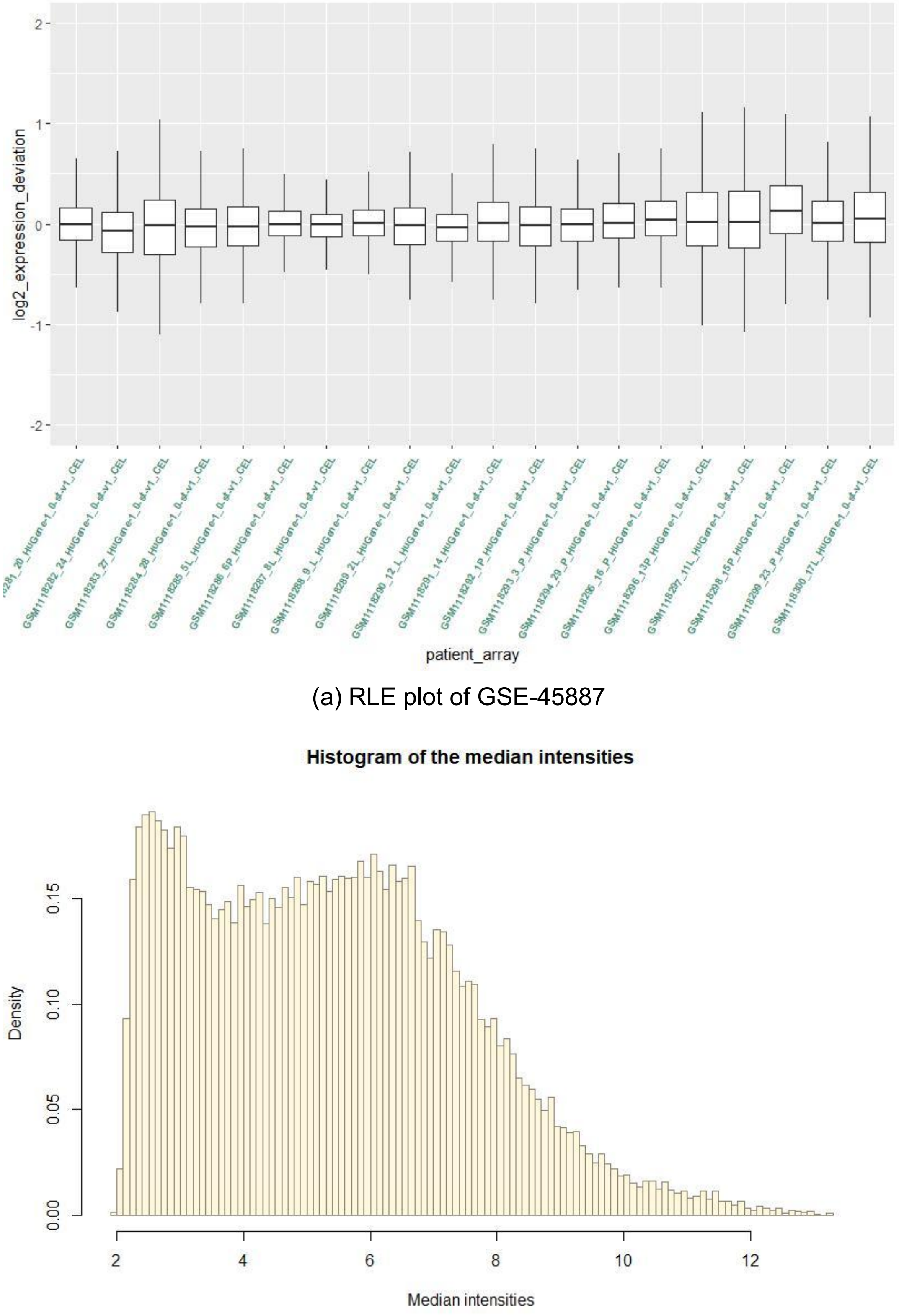


(a) PCA plot of GSE-45887



(b) Boxplot of GSE-45887

Figure 4.5. PCA and Boxplot of Dataset GSE-45887



(a) RLE plot of GSE-45887

(b) Histogram of GSE-45887

Figure 4.6. Quality Assessment of Dataset GSE-45887

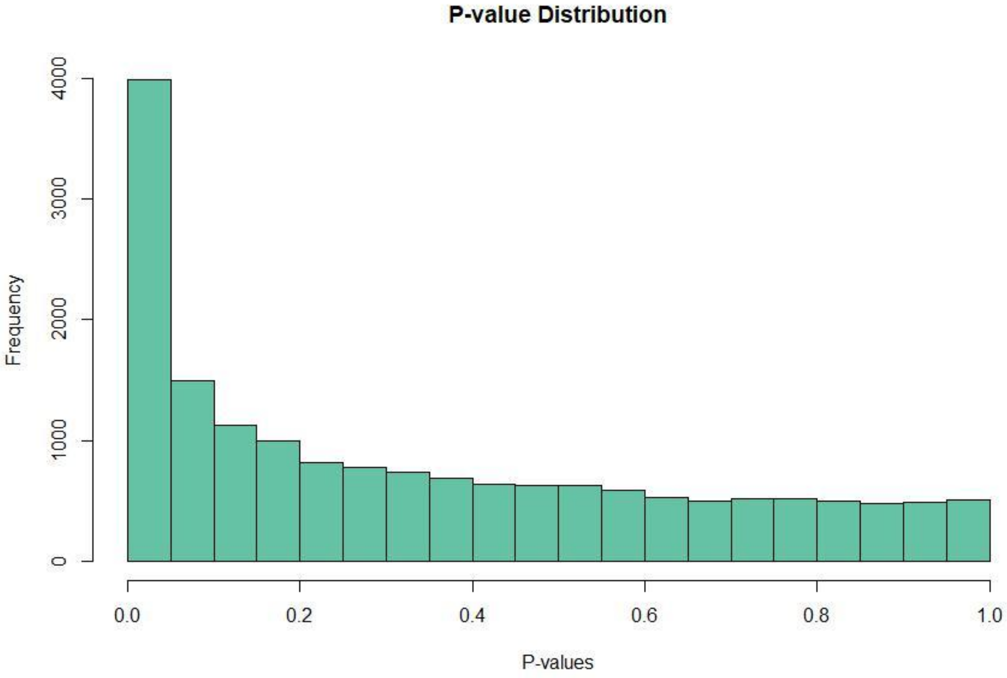
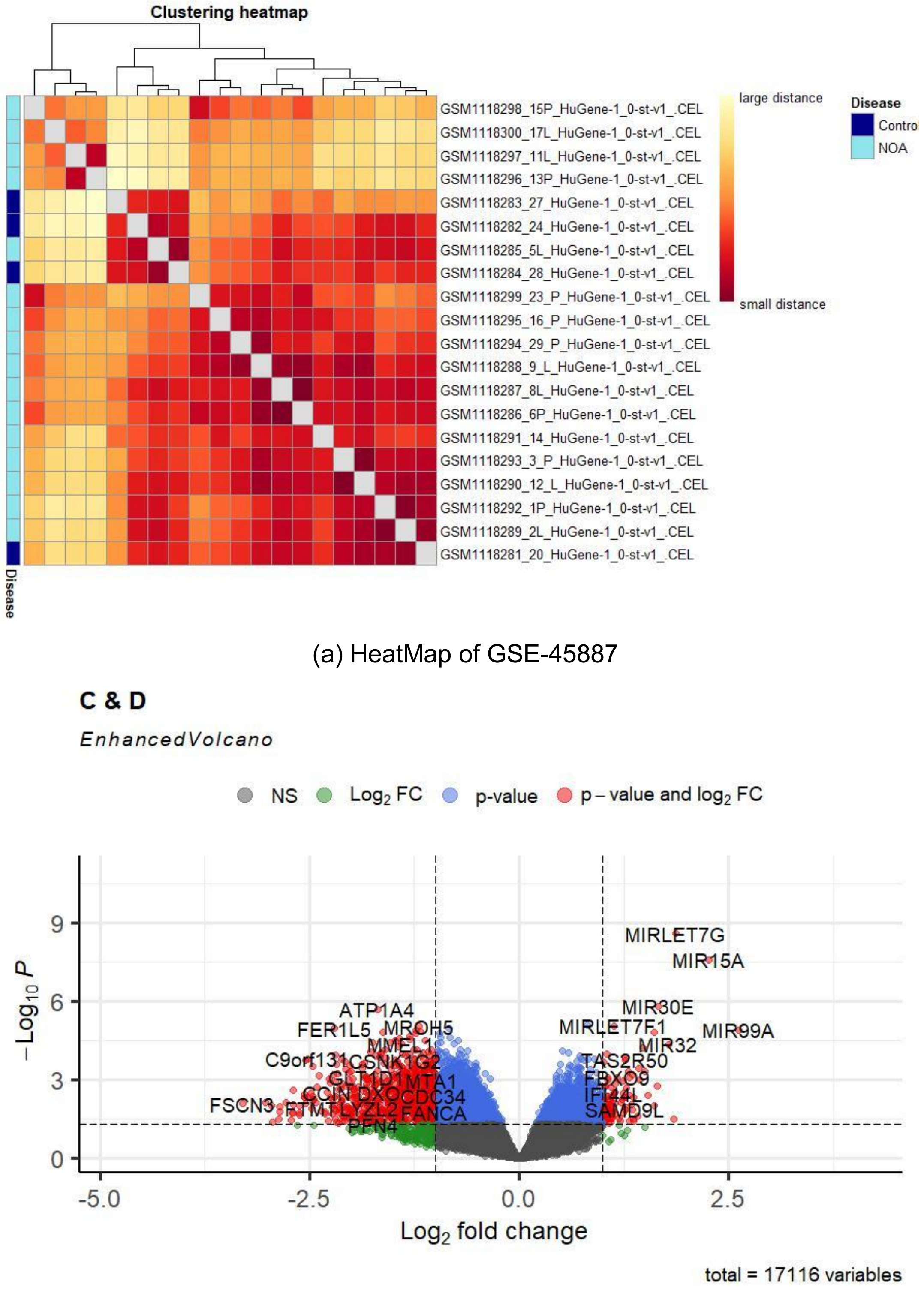


Figure 4.7. P-value Distribution Graph of Dataset GSE-45887



(a) HeatMap of GSE-45887

(b) Enhanced Volcano plot of GSE-45887

Figure 4.8. Visualization of Dataset GSE-45887

Table 4.2. Top 10 DEGs of Dataset GSE45887

PROBEID	SYMBOL	GENENAME	log2FC	P.Value	adj.P.Val
8087881	MIRLET7G	microRNA let-7g	1.873159046	2.57E-09	5.82E-05
7971661	MIR15A	microRNA 15a	2.26066658	2.63E-08	0.000297782
7900488	MIR30E	microRNA 30e	1.664081679	1.55E-06	0.010210371
7906527	ATP1A4	ATPase Na+/K+ transporting subunit alpha 4	-1.691674474	2.03E-06	0.010210371
8156521	MIRLET7F1	microRNA let-7f-1	1.130795034	8.74E-06	0.017797315
8153273	MROH5	maestro heat like repeat family member 5 (gene/pseudogene)	-1.199332484	9.06E-06	0.017797315
8043639	FER1L5	fer-1 like family member 5	-2.202215685	1.06E-05	0.017797315
8067942	MIR99A	microRNA 99a	2.625447662	1.26E-05	0.017797315
8014298	HEATR9	HEAT repeat containing 9	-1.236875639	1.29E-05	0.017797315
8048350	PLCD4	phospholipase C delta 4	-1.186656337	1.40E-05	0.017797315

si
s

4.2 Next Generation Sequencing

After the retrieval of data from GEO, analysis is performed with the help of an appropriate tool in the Galaxy server to identify DEGs. The analysis includes pre-processing of the data, transcript analysis, and identification of DEGs.

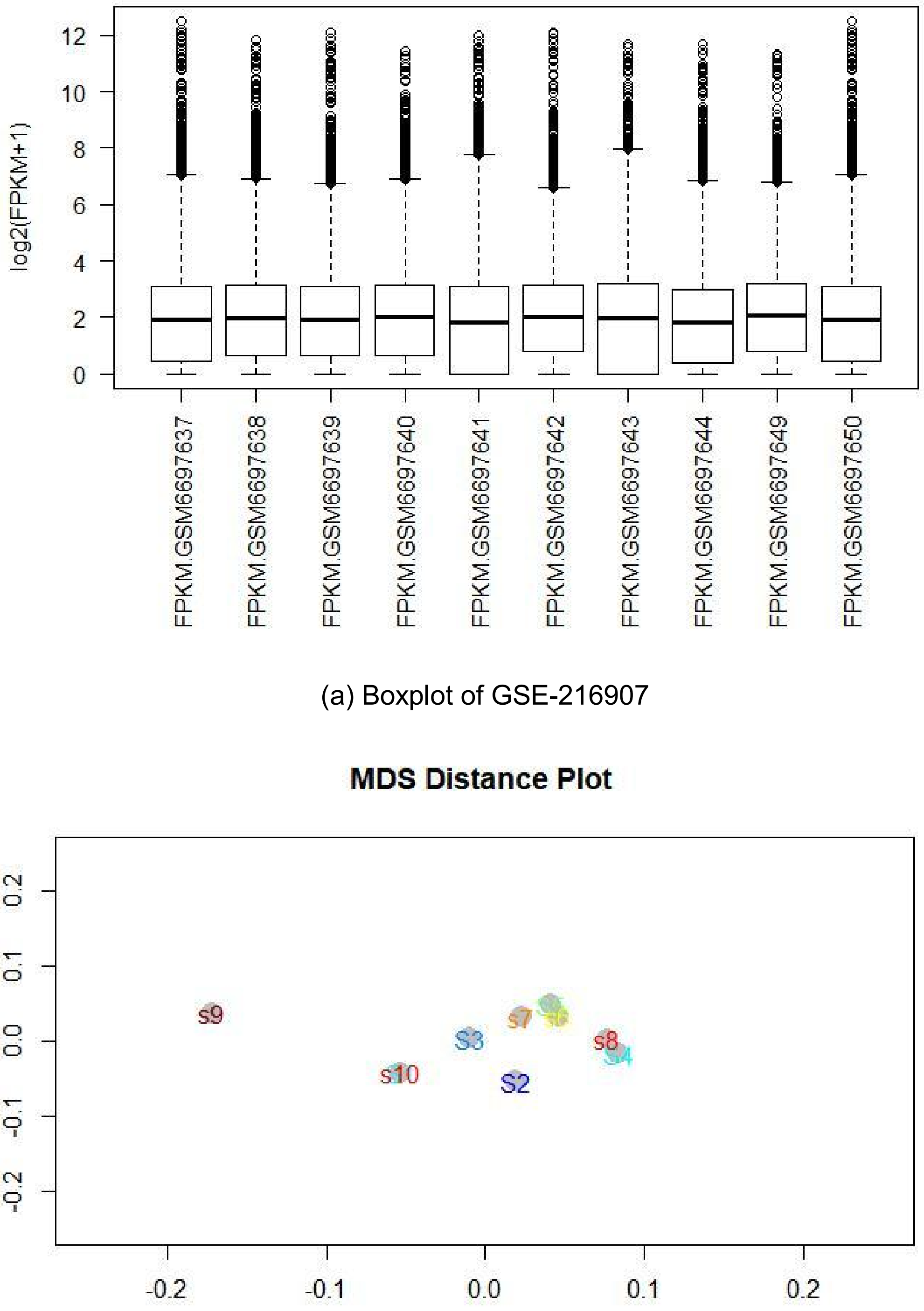
4.2.1 GSE-216907

(A) Pre-processing of Transcript Data

There are 10 samples total in this collection, 8 of which are from NOA pa-tients and 2 from healthy individuals. The analysis of the raw data was done in Galaxy, and global read trimming was done at 13 on the front. After pre-processing the raw data, the transcript data was analyzed using Ballgown in R. The median distribution of FPKM values for each sample is displayed in Figure 4.9a as a box plot of log2 transformed FPKM (fragments per kilobase of exon model per million reads mapped). This shows that the median of each sample is around 2. The whiskers depict the distribution of the data, while the dots stand in for any outliers that were found. A vi-sualization graph used in particular for assessing the similarity or dissimilarity across samples based on their genetic or molecular profiles is the Multi-Dimensional Scaling (MDS) plot. MDS as shown in Figure 4.9b enables to visualization of relationships between samples in a lower-dimensional environment and aids in the reduction of the dimensionality of complicated data. Transcript length distribution is shown in Figure 4.10. The horizontal axis shows transcript length in base pairs, and the vertical axis shows frequency. Right-handed long tail and positively skewed distribution are seen. There are a maximum of 100bp-long transcripts available. The frequency distribution of genes' log2 fold change values is shown in Figure 4.11a. Values for the fold change are shown on the horizontal axis, while values for the frequency of the differential expression are shown on the vertical axis. The physiologically important genes are filtered out using a cutoff value of ± 1.

(B) Visualization of DEGs

After DEGS analysis, volcano and enhanced volcano plots are used for visualization. The threshold for filtering genes was p-value less than 0.05 and log2FC ±1. Gene distribution according to log2FC on the horizontal axis and p-value (-log10) on the vertical axis is displayed in Figure 4.11b. Each dot stands for a unique gene. Genes present in the grey area are non-significant. Blue-dotted genes are filtered through just the p-value threshold. Same as blue-dotted genes, green-dotted genes only passed the LogFC threshold. Genes represented with red dots are significant genes with low p-values i.e. less than 0.05 and log2 fold change greater than ±1. These genes passed through both cutoff values. Highlighted genes less than 1 log2FC are down-regulated genes and genes with greater than 1 log2FC are up-regulated genes. Top 10 Filtered DEGs are listed in Table 4.3.



(a) Boxplot of GSE-216907

(b) MDS plot of GSE-45887

Figure 4.9. Quality Assessment of Dataset GSE-216907

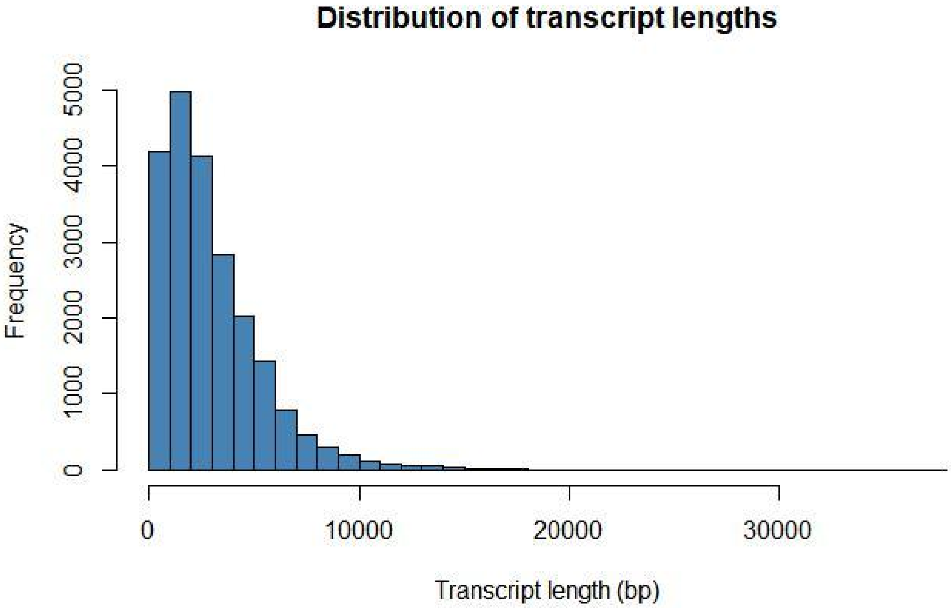
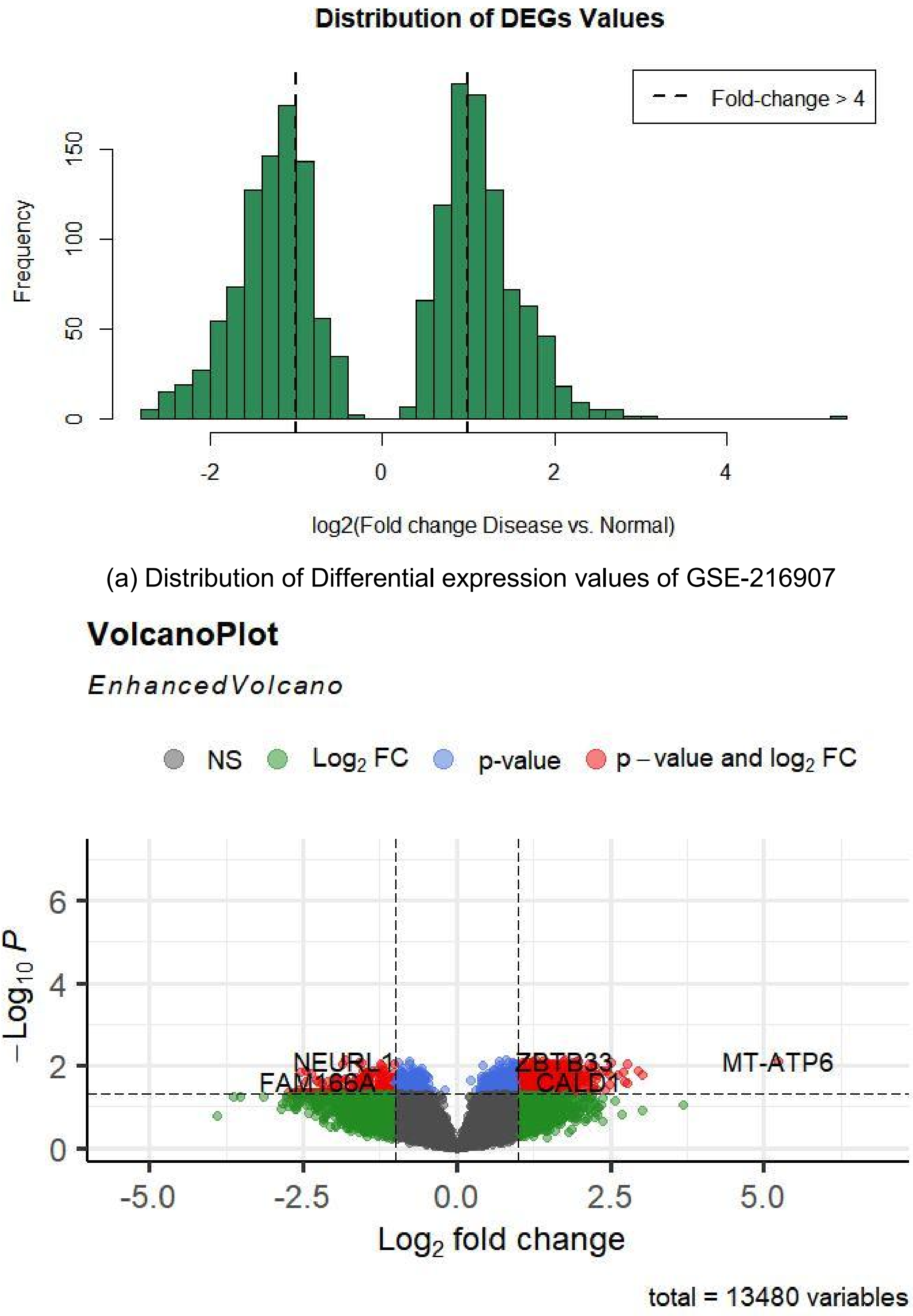


Figure 4.10. Distribution of transcript length of GSE-216907

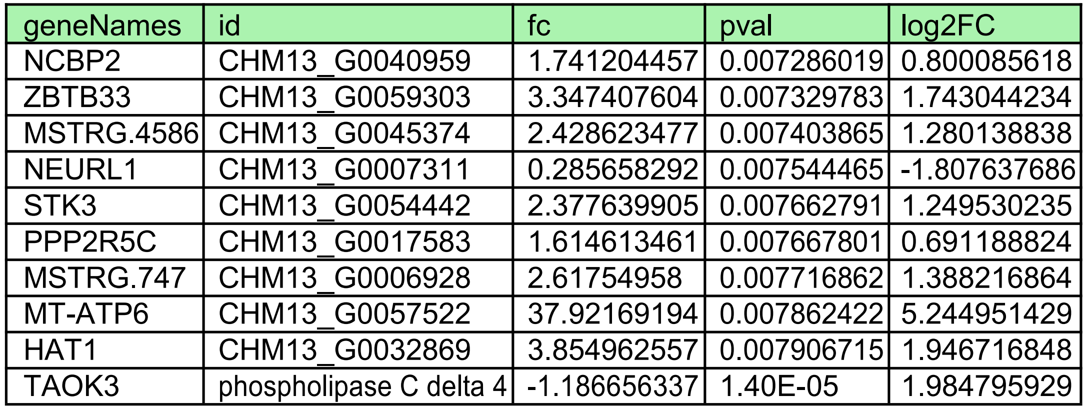


(a) Distribution of Differential expression values of GSE-216907

(b) Distribution of Differential expression values of GSE-216907

Figure 4.11. Visualizing DEGs of Dataset GSE-216907

Table 4.3. Top 10 DEGs of Dataset GSE216907



geneNames	id	fc	pval	log2FC
NCBP2	CHM13_G0040959	1.741204457	0.007286019	0.800085618
ZBTB33	CHM13_G0059303	3.347407604	0.007329783	1.743044234
MSTRG.4586	CHM13_G0045374	2.428623477	0.007403865	1.280138838
NEURL1	CHM13_G0007311	0.285658292	0.007544465	-1.807637686
STK3	CHM13_G0054442	2.377639905	0.007662791	1.249530235
PPP2R5C	CHM13_G0017583	1.614613461	0.007667801	0.691188824
MSTRG.747	CHM13_G0006928	2.61754958	0.007716862	1.388216864
MT-ATP6	CHM13_G0057522	37.92169194	0.007862422	5.244951429
HAT1	CHM13_G0032869	3.854962557	0.007906715	1.946716848
TAOK3	phospholipase C delta 4	-1.186656337	1.40E-05	1.984795929

in
g

4.3 Network Analysis

Network analysis was performed using the pathway of 16 common differen-tially expressed genes using Cytoscape. The significant nodes were identified on the basis of betweenness centrality and closeness centrality. As shown in Figure 4.12 the most important nodes identified on the basis of betweenness centrality include SMO, COS2, and PTCH1 proteins.

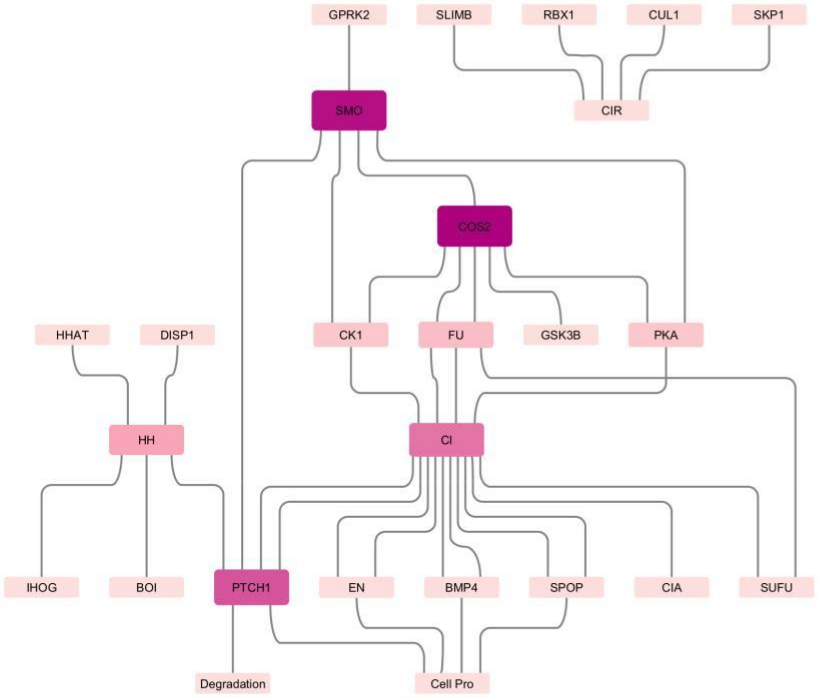


Figure 4.12. Important Nodes identified on basis of Betweenness Centrality

On the other hand, the most important nodes identified using closeness centrality include EN complex, BMP4, and SPOP proteins as shown in Figure 4.13.

4.4 Protein Selection

The proteins selected on the basis of their association with azoospermia from the literature review and their expression values are presented in Table 4.4.

Network analysis performed on selected proteins using STRING database, a biological database, and web resource for known protein-protein interactions freely

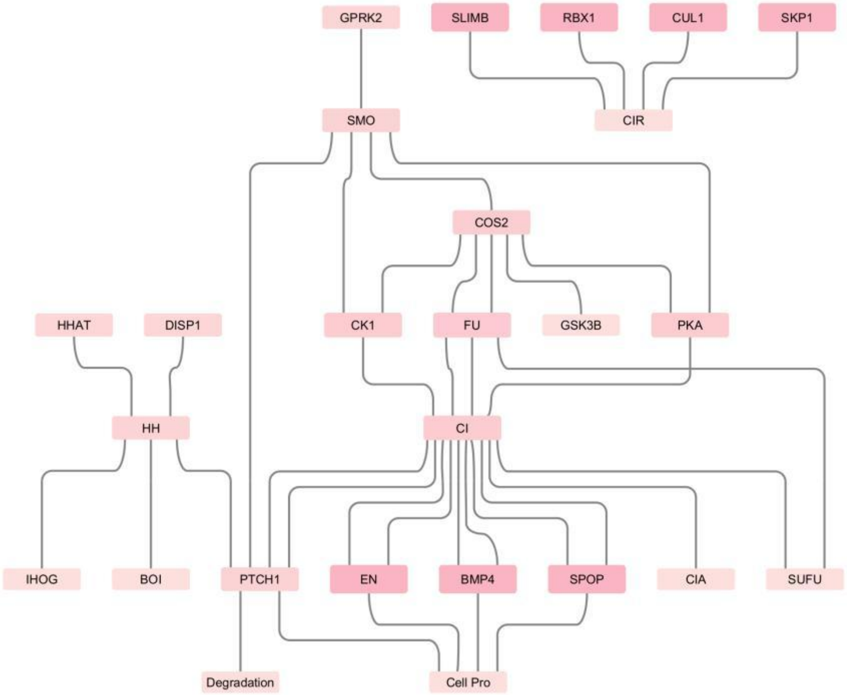
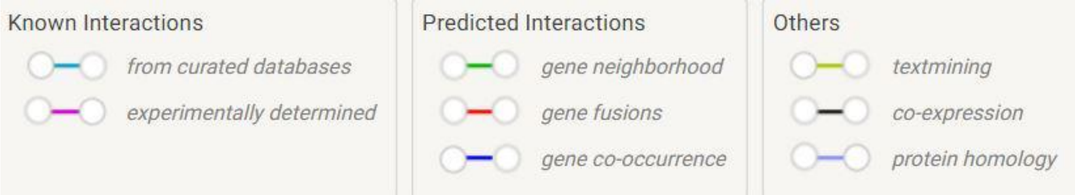


Figure 4.13. Important Nodes identified on basis of Closeness Centrality

Table 4.4. List of Genes for PPI Analysis

Gene-ID	Gene-Name	log2FC	Expression
CCNG1	Cyclin G1	1.845241283	Up-regulated
SYCP3	Synaptonemal Complex Protein 3	0.293490419	Normal
TEX11	Testis Expressed 11	0.368793257	Normal
SYCE1	synaptonemal complex (SC) central element 1	0.112801709	Normal
NR5A1	Nuclear Receptor Subfamily 5 Group A Member 1	0.08869329	Normal
ZMYND15	zinc finger MYND-type containing 15	-1.901059618	Down-regulated
UBQLN3	ubiquilin 3	-2.760846572	Down-regulated
THEG	Testicular Haploid Expressed Gene	-2.004005676	Down-regulated
STPG1	Sperm Tail PG-Rich Repeat Containing 1	-1.245574217	Down-regulated
SPEM1	spermatid maturation 1	-2.204479872	Down-regulated
SPATA32	Spermatogenesis Associated 32	-2.234732231	Down-regulated

available at https://string-db.org/. The network analysis results of CCNG1, SYCP3, TEX11, SYCE1, and NR5A1 generated using STRING are shown in Figure 4.15. Nodes represent the query proteins in red color and edges represent protein-protein associations with different colors for different types of interactions (known interac-tions, predicted, and others) as presented in Figure4.14.



Known Interactions

- from curated databases
- experimentally determined

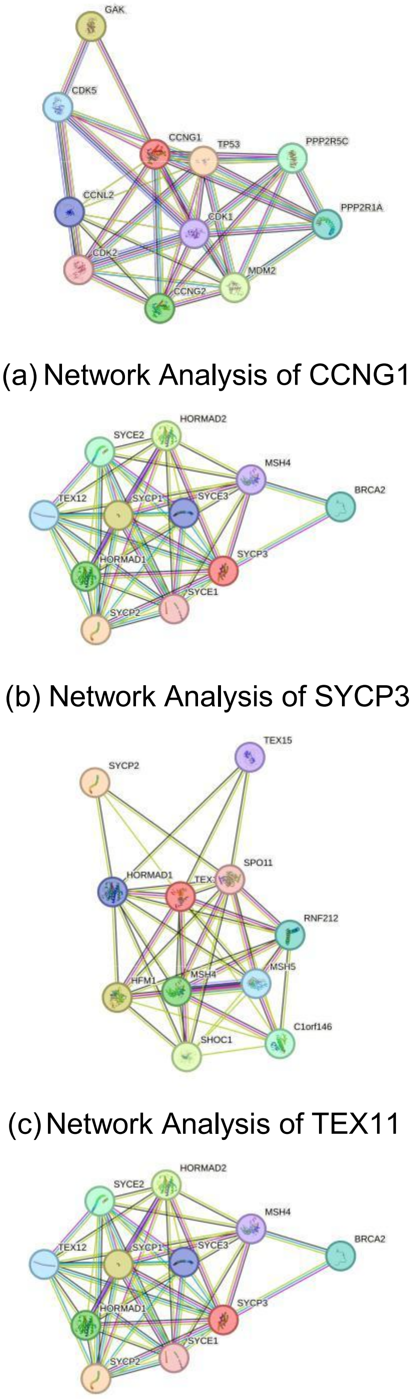
Predicted Interactions

- gene neighborhood
- gene fusions
- gene co-occurrence

Others

- textmining
- co-expression
- protein homology

Figure 4.14. Edge colors for different Interaction types



(a) Network Analysis of CCNG1

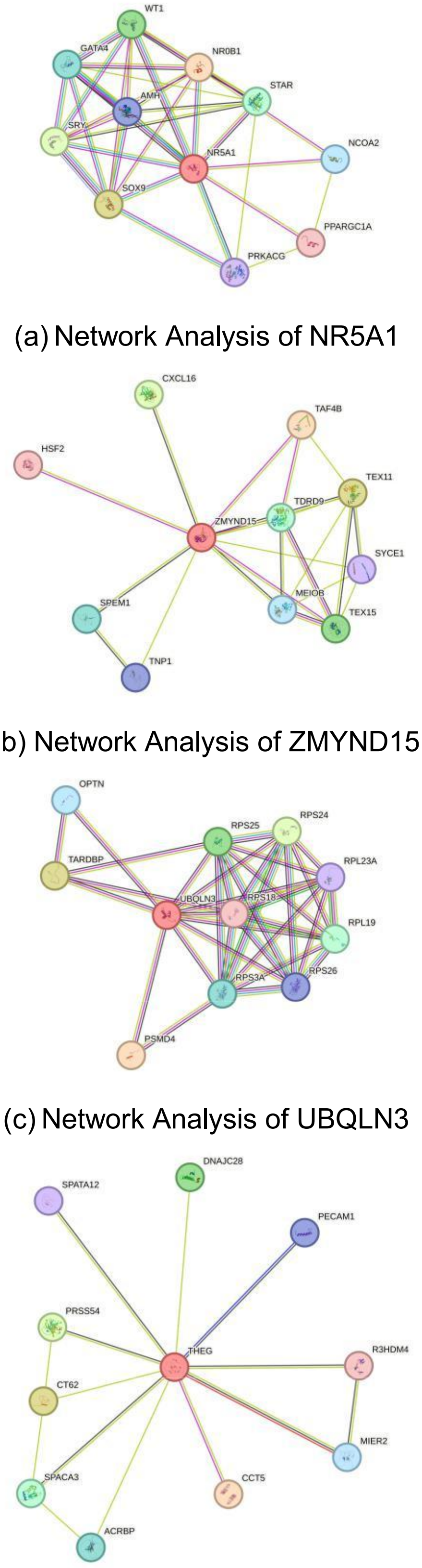
(b) Network Analysis of SYCP3

(c) Network Analysis of TEX11

(d) Network Analysis of SYCE1

Figure 4.15. Network analysis results of CCNG1, SYCP3, TEX11 and SYCE1 from STRING Database

The network analysis results of NR5A1, ZMYND15, UBQLN3, and THEG generated using STRING are shown in Figure 4.16. Nodes represent the query pro-teins in red color and edges represent protein-protein associations with different colors for different types of interactions (known interactions, predicted, and others) as pre-sented in Figure 4.14.



(a) Network Analysis of NR5A1

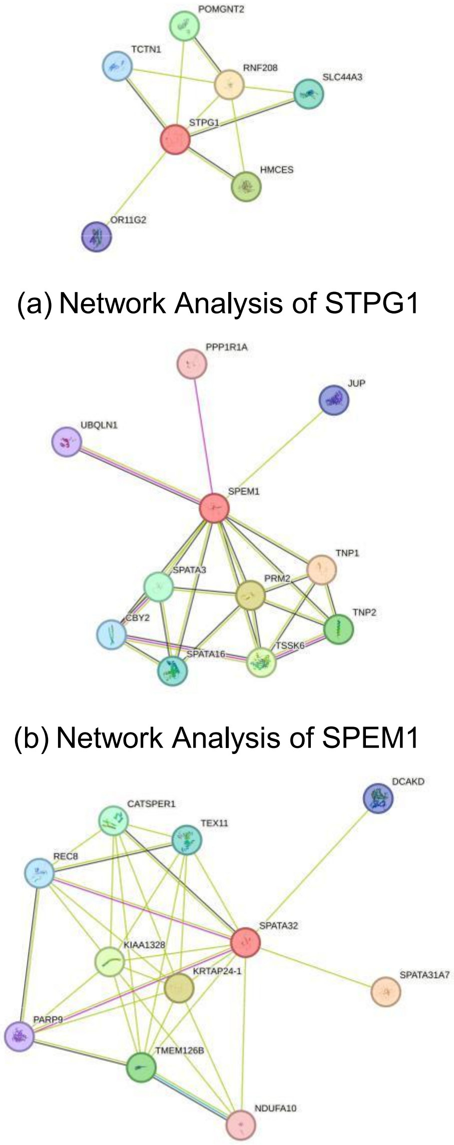
(b) Network Analysis of ZMYND15

(c) Network Analysis of UBQLN3

(d) Network Analysis of THEG

Figure 4.16. Network analysis results of NR5A1, ZMYND15, UBQLN3 and THEG from STRING Database

The network analysis results of STPG1, SPEM1, and SPATA32 generated using STRING are shown in Figure 4.17. Nodes represent the query proteins in red color and edges represent protein-protein associations with different colors for different types of interactions (known interactions, predicted, and others) as presented in Figure 4.14.



(a) Network Analysis of STPG1

(b) Network Analysis of SPEM1

(c) Network Analysis of SPATA32

Figure 4.17. Network analysis results of STPG1, SPEM1, and SPATA32 from STRING Database

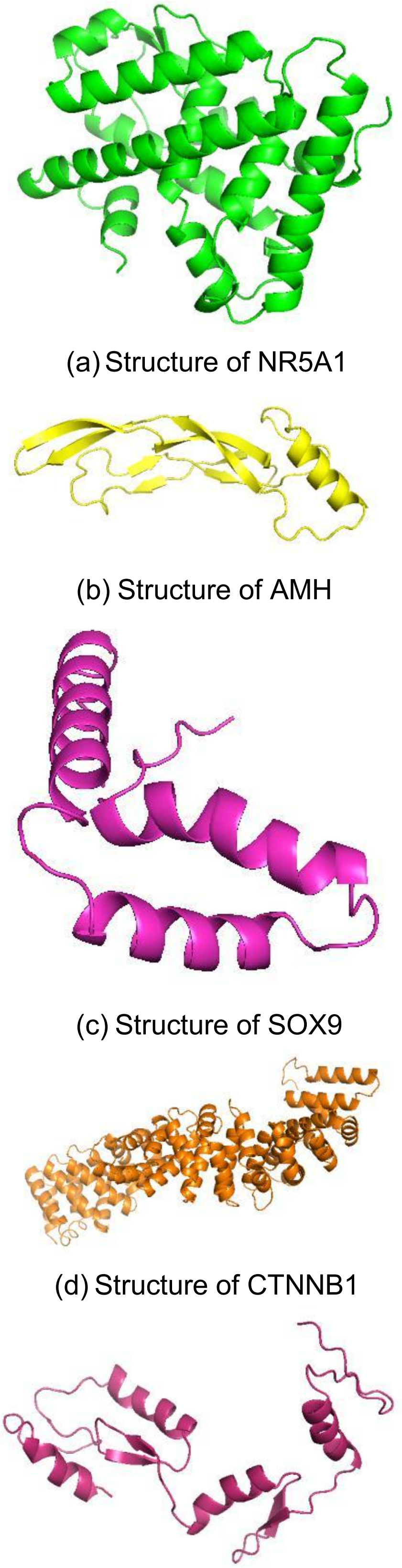
4.5 Protein-protein Interaction

Predictions about the binding affinities of the proteins and interacting residues are possible from protein-protein interaction. Despite the fact that there are other PPI interaction web servers available, studies continually place HADDOCK as one of the finest PPI programs in terms of quality, usability, and effectiveness.

The proteins with known interactions were aligned with another requirement for performing interactions as described above and only a few genes meeting the criteria were selected as displayed in Table 4.5. NR5A1 is associated with normal spermatogenesis and mutations in NR5A1 have been reported to be associated with Azoospermia. There is a conflict between the association of NR5A1 mutations with azoospermia. NR5A1 regulates its target protein AMH with the help of its transcrip-tional co-factors SOX9, WT1, and co-activator CTNNB1. The structures of NR5A1, AMH, SOX9, CTNNB1, and WT1 are presented in Figure 4.18.

Table 4.5. Proteins filtered for HADDOCK Interaction protein-protein Analysis

Gene ID	Gene Name	log2FC	Expression
NR5A1	Nuclear Receptor Subfamily 5 Group A Member 1	0.08869329	Normal
CTNNB1	Catenin Beta 1	0.332850661	Normal
AMH	Anti-Müllerian hormone	-0.88427359	Normal
SOX9	SUMO-conjugating enzyme UBC9	-1.425957221	Normal
WT1	Wilms' tumour suppressor gene 1	-0.735163824	Normal



(a) Structure of NR5A1

(b) Structure of AMH

(c) Structure of SOX9

(d) Structure of CTNNB1

(e) Structure of WT1

Figure 4.18. Protein structures of NR5A1, AMH, SOX9, CTNNB1 and WT1

4.5.1 Protein Binding Areas

The list of interacting residues for each of the selected proteins was generated using the SPPIDER (Species and Proteins Profile-based Infrared Database and Enhanced Retrieval) webserver. The PDB file for each protein was given input and SPPIDER generated the list of interacting residues for further analysis.

4.5.2 Docking Analysis

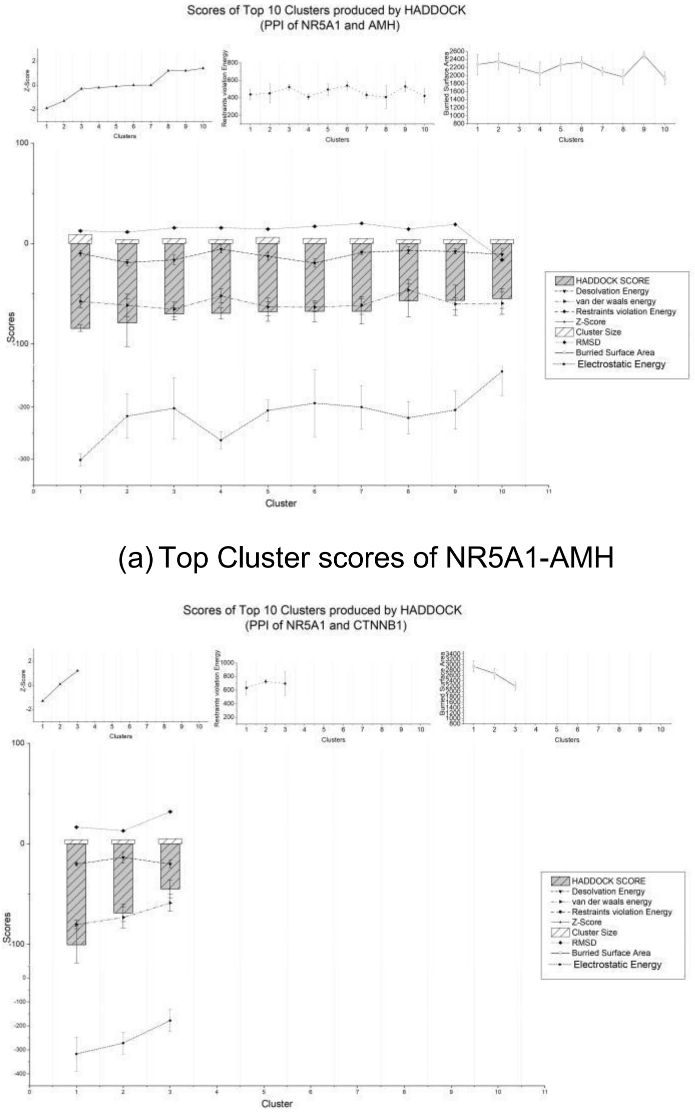
After preparing the required files and predicting interacting residues, protein-protein interaction was performed with the help of HADDOCK. Two protein struc-tures were given as input along with the list of interacting residues. The resulting protein-protein complexes are analyzed on the basis of different factors as discussed in Section 3.5(B) of Chapter 3.

When employing the HADDOCK algorithm to study protein-protein interac-tions, the above-mentioned numerous energy components and metrics are essential for assessing and choosing the most biologically pertinent docking poses and com-plexes. The evaluation parameter scores for the top clusters of all complexes are given in Table 4.6.

Table 4.6. The statistics of the top and most reliable clusters according to HADDOCK before Refinement.

Protein-Protein Complex	NR5A1-AMH	NR5A1-SOX9	NR5A1-CTNNB1	NR5A1-WT1
HADDOCK score	-84.5 +/- 3.1	-103.4 +/- 16.4	-100.5 +/- 18.1	-100.4 +/- 24.8
Cluster size	9	5	20	6
RMSD from the overall lowest-energy structure	12.5 +/- 0.7	1.6 +/- 1.3	16.5 +/- 0.1	9.9 +/- 0.2
Van der Waals energy	-56.7 +/- 6.4	-72.7 +/- 9.8	-80.3 +/- 4.2	-61.4 +/- 9.5
Electrostatic energy	-308.3 +/- 11.3	-217.0 +/- 36.9	-317.2 +/- 71.1	-278.9 +/- 56.6
Desolvation energy	-10.0 +/- 2.8	-19.9 +/- 2.1	-20.1 +/- 2.0	-17.4 +/- 1.8
Restraints violation energy	438.6 +/- 60.3	326.0 +/- 112.0	633.2 +/- 98.2	342.5 +/- 94.2
Buried Surface Area	2271.4 +/- 258.7	2044.4 +/- 257.0	2936.1 +/- 208.0	2212.2 +/- 229.9
Z-Score	-1.9	-1.3	-1.3	-2.1

The chart of the PPI scores of the top clusters of most probable complexes of NR5A1 with AMH and CTNNB1 respectively interacting partners is shown in Figure 4.19.

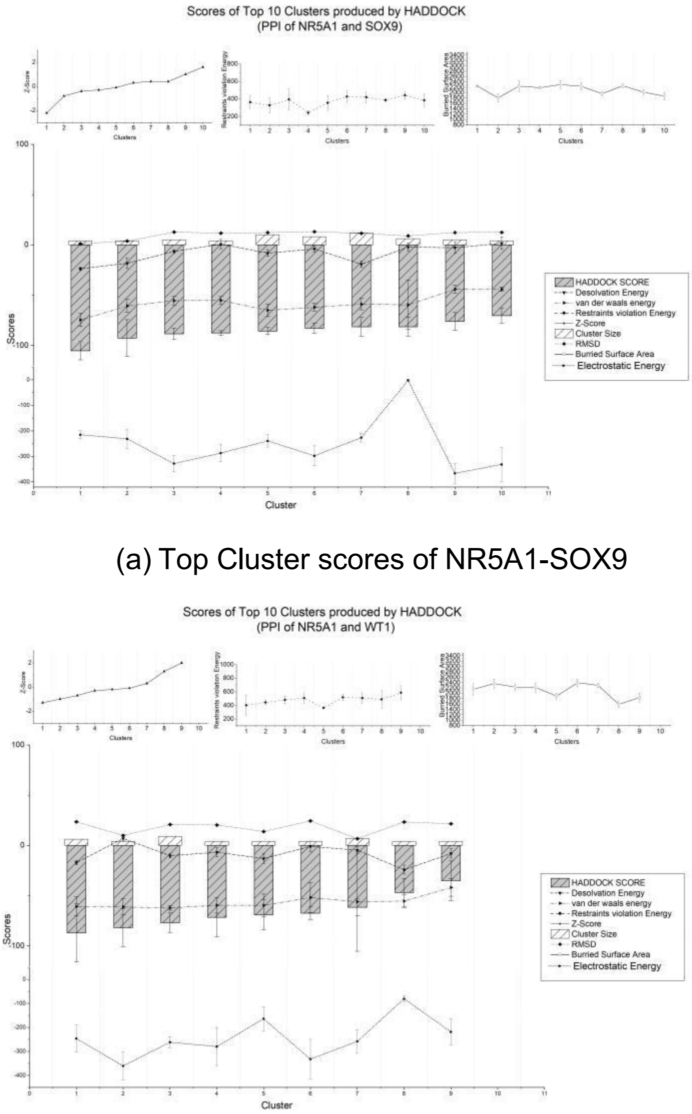


(a) Top Cluster scores of NR5A1-AMH

(b) Top 10 Cluster scores of NR5A1-CTNNB1

Figure 4.19. Top Cluster scores of NR5A1 with AMH and CTNNB1

The chart of the PPI scores of the top clusters of most probable complexes of NR5A1 with SOX9 and WT1 respectively interacting partners is shown in Figure 4.20.



(a) Top Cluster scores of NR5A1-SOX9

(b) Top Cluster scores of NR5A1-WT1

Figure 4.20. Top Cluster scores of NR5A1 with SOX9 and WT1

4.5.3 Complex Refinement

HADDOCK provides the option for water-based refinement to improve the accuracy of protein-protein complex predictions. Refinement of the protein-protein complex is performed to more accurately mimic the biological setting and increase the predictability of the interaction. A more explicit or implicit model of the solvent (water) is provided in water-based refining. This indicates that the calculations account for the interactions between and among the water molecules and the proteins. HADDOCK recommended using the first complex from the top cluster as the most credible input for refinement analysis. The evaluation is performed on the basis of the same parameters as for prior complexes. The underlying Table 4.7 displays the statistical parameters for analyzing the refinement results.

Table 4.7. The statistics of the most reliable clusters according to HADDOCK after Refinement.

Protein-Protein Complex	NR5A1-AMH	NR5A1-SOX9	NR5A1-CTNNB1	NR5A1-WT1
HADDOCK score	-138.7 +/- 4.1	-161.2 +/- 3.3	-209.7 +/- 3.8	-117.8 +/- 6.8
Cluster size	20	20	20	20
RMSD from the overall lowest-energy structure	2.6 +/- 0.0	0.5 +/- 0.3	0.5 +/- 0.3	0.6 +/- 0.3
Van der Waals energy	-68.8 +/- 1.7	-82.5 +/- 3.0	-94.2 +/- 6.5	-48.8 +/- 1.8
Electrostatic energy	-296.8 +/- 24.3	-248.4 +/- 27.8	-549.3 +/- 53.4	-353.5 +/- 17.6
Desolvation energy	-10.5 +/- 4.2	-29.1 +/- 3.0	-5.7 +/- 1.7	1.7 +/- 3.5
Restraints violation energy	0.0 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0
Buried Surface Area	2164.0 +/- 52.6	2301.5 +/- 47.0	3300.6 +/- 64.6	2147.0 +/- 60.1
Z-Score	0.0	0.0	0.0	0.0

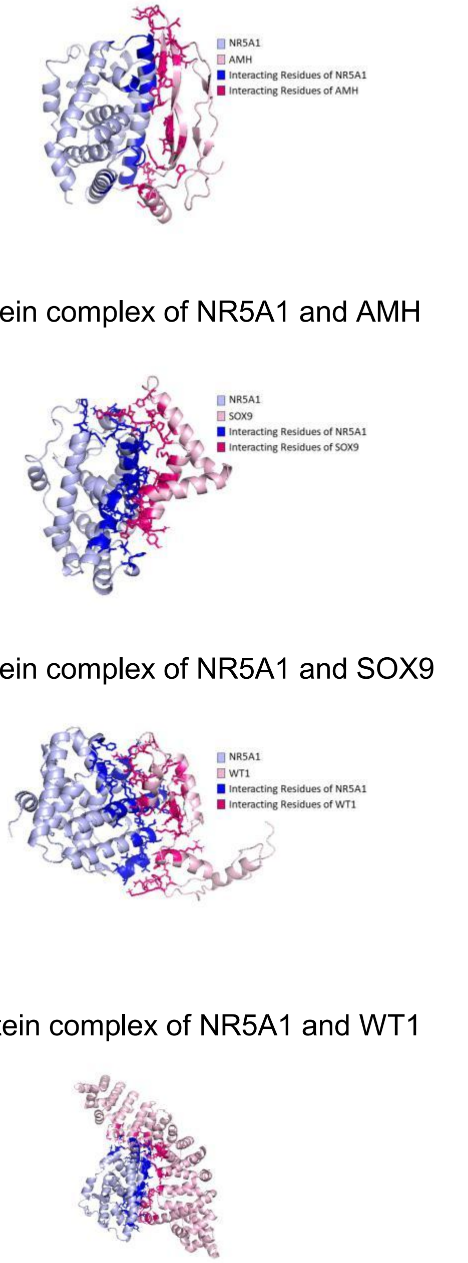
4.5.4 Prodigy Analysis

Haddock Prodigy webserver is used for finding binding affinity and binding residues between the complex proteins of the protein complex based on different parameters. The best protein-protein complex from the refinement outputs is submitted as input with default parameters. The parameters used for analyzing the prodigy results are discussed in section 3.5(D) of chapter 3. The underlying Table 4.8 displays the statistical results for prodigy analysis of the protein-protein complexes:

Table 4.8. Prodigy statistical parameters for all complexes

Protein-protein complex	NR5A1-AMH	NR5A1-SOX9	NR5A1-CTNNB1	NR5A1-WT1
G (kcal mol-1)	-9.7	-10.3	-11.6	-9.7
Kd (M) at °C	1.4e-07	5.4e-08	6.7e-09	1.5e-07
ICs charged-charged	12	14	14	11
ICs charged-polar	13	12	22	11
ICs charged-apolar	21	26	31	16
ICs polar-polar	1	1	6	6
ICs polar-apolar	13	11	17	16
ICs apolar-apolar	16	13	20	8
NIS charged	22.62	26.24	25.7	27.1
NIS apolar	48.52	44.11	42.31	40.81

The results of the HADDOCK prodigy analysis were downloaded, Pymol tool was used to see the 3D structures of the protein complexes in different forms. The prodigy results of NR5A1 with other proteins are presented in Figure 4.21 with light blue and light pink colors representing the two proteins and dark regions representing their interacting areas.


NR5A1
AMH
Interacting Residues of NR5A1
Interacting Residues of AMH
NR5A1
SOX9
Interacting Residues of NR5A1
Interacting Residues of SOX9
NR5A1
WT1
Interacting Residues of NR5A1
Interacting Residues of WT1

(a) Protein complex of NR5A1 and AMH

(b) Protein complex of NR5A1 and SOX9

(c) Protein complex of NR5A1 and WT1

(d) Protein complex of NR5A1 and CTNNB1

Figure 4.21. Protein Complex o NR5A1 with other Proteins

The information regarding the association of identified mutation in NR5A1 with the interacting residues is presented in table 4.9. Three mutations reported in one study were found to be present in the interacting residue positions of NR5A1. These mutations were associated with spermatogenesis failure due to different factors. Only one mutation (identified as p.Asp257Asn) out of these, present in 0.4% (4 out of 270) samples was associated with azoospermia. The mutations identified in other studies were not in the interacting region of NR5A1 with its interacting partners used in this study.

Table 4.9. Mutations at Interacting Residues

Mutations Identified	Reference	NR5A1-AMH	NR5A1-SOX9	NR5A1-CTNNB1	NR5A1-WT1
p.Gly123Ala (c.368G>C)	Bashamboo et al. (2010)	No	No	No	No
p.Pro129Leu(c.386C>T)	Bashamboo et al. (2010)	No	No	No	No
p.Pro131Leu(c.392C>T)	Bashamboo et al. (2010)	No	No	No	No
p.Arg191Cys (c.571C>T)	Bashamboo et al. (2010)	No	No	No	No
p.Gly212Ser (c.634G>A)	Bashamboo et al. (2010)	No	No	No	No
p.Asp238Asn(c.712G>)	Bashamboo et al. (2010)	No	No	No	No
p.Pro97Thr(c.467C>A)	Zare-Abdollahi D, Safari S, Mirfakhraie R, et al.	No	No	No	No
p.Glu237Lys(c.709G>A)	Zare-Abdollahi D, Safari S, Mirfakhraie R, et al.	No	No	No	No
P.Gly146Ala(c.437G>C)	Andrologia vol. 50,3 (2018)	No	No	No	No
p.Thr75Thr(c.225G>C)	Ropke et al. (2013)	No	No	No	No
p.Pro125Pro(c.375G>A)	Ropke et al. (2013)	No	No	No	No
p.Gly146Ala(c.437G>C)	Ropke et al. (2013)	No	No	No	No
p.Gly165Arg(c.493G>C)	Ropke et al. (2013)	No	No	No	No
p.Pro210Pro(c.630G>A)	Ropke et al. (2013)	No	No	No	No
p.Val240Val(c.720G>A)	Ropke et al. (2013)	No	No	No	No
p.Asp257Asn(c.769G>A)	Ropke et al. (2013)	No	Yes	No	No
p.Ile323Thr(c.968T>C)	Ropke et al. (2013)	No	No	No	No
p.Cys422Cys(c.1266C>T)	Ropke et al. (2013)	No	No	No	No
p.Lys440Lys(c.1320G>A)	Ropke et al. (2013)	Yes	Yes	No	No

The list of interacting residues (the residues between a distance of 5Å upon protein-protein complex formation) for NR5A1 in all the complexes is provided in Table A.1 in the Appendix portion. The highlighted residues with green and pink colors are the two common interacting residues in all the complexes.

Chapter 5

DISCUSSION

Azoospermia is one of the major causes of male infertility and is described as the absence of spermatozoa in the ejaculate. It is classified into two types i.e. obstructive azoospermia (OA) and non-obstructive azoospermia (NOA). Azoospermia is the cause of infertility in more than one percent of males in the general population whereas 10%–15% of infertile men are affected by this problem [8]. NOA is the most prevalent kind of azoospermia and affects approximately 60% of azoospermic males. It is caused by spermatogenesis failure due to different factors. There is no proper treatment available for NOA, however, sperm can be retrieved in some cases for in-vitro fertilization. This process is very expensive and has a very low success rate. Treatment options are urgently needed to increase sperm production and for targeting underlying causes.

This work aims to demonstrate relevant associations between NR5A1 mu-tations, spermatogenesis failure, and NOA. Differential expression analysis (DEA) was performed on datasets generated from different platforms like Microarray and NGS. The purpose of DEA was to analyze gene expression patterns of patients with spermatogenesis failure causing NOA. The selected genes were also searched in the literature. Network analysis of selected genes was performed and interactions were verified from the literature. The protein structures of selected proteins were docked using the HADDOCK server. The binding affinity and interaction profile of all inter-acting partners were analyzed.

Three datasets were selected for the analysis. These three datasets included one RNA-seq dataset and two microarray datasets. Using logFC ±1 and 0.05 as the p-value, DEGs were marked in microarray and mRNA seq datasets.

Since RNA-seq is more precise than microarrays, it has been selected as the best method for determining gene expression. After identifying genes based on their association with azoospermia, target identification for protein-protein interactions was performed. Different criteria were set for selecting the list of proteins for PPI. The first was selecting significant genes on the basis of network analysis using the STRING database. Only proteins exhibiting known interactions were selected. The interactions of selected proteins were verified from the literature. The second criterion was the availability of 3D structures of selected proteins. In order to find the 3D structures of selected proteins, the UniProt database was used, and particular entries with zero mutations were chosen for all proteins. The fasta sequence was then used to find proteins with similar amino acid sequences in PSI-Blast. The protein that aligned most optimally was selected. The proteins with greater than 98% were used as it is. The protein structures with percentage identity between 75% to 97% were modeled on SWISS-MODEL (a fully automated protein homology modeling web server). The protein structures with percentage identity between 65% to 74% were modeled on trRosetta (algorithm for predicting fast and accurate protein structures). The remaining proteins with known interaction but a similarity percentage less than 65% were filtered.

None of the selected genes could pass the first criterion as protein A for interacting however were present as protein B except NR5A1. From the literature, it was verified that NR5A1 is one of the most important genes among therapeutic target genes of NOA. Different studies identified mutations in NR5A1 associated with azoospermia. STRING networking database was used and NCOA2, NROB1, AMH, SOX9, GATA4, and WT1 were selected on the basis of their expression level in our datasets and known interactions with NR5A1. NR5A1 also exhibited known interaction with one of its co-activator genes CTNNB1. The proteins meeting the second criterion were NR5A1, AMH, SOX9, WT1, and CTNNB1.

NR5A1 is a nuclear hormone receptor, that plays a crucial role in regulating steroid hormone biosynthesis by targeting different genes in humans. Some transcription cofactors and transcription co-activators participate with NR5A1 in regulating NR5A1 target genes. Previous studies indicate that AMH is one of the NR5A1-mediated target genes. SOX9 and WT1 are identified as transcription cofac-tors interacting with NR5A1 to help in regulating NR5A1 target genes. In addition to this CTNNB1 is also identified as one of the transcription co-activators that helps in regulating NR5A1-mediated target genes. Researchers have different opinions on the association of NR5A1 mutations with NOA. Some studies say that mutations in NR5A1 are the cause of NOA and some say that mutations in NR5A1 are not associated with NOA. We performed PPI of NR5A2 with interacting partners and searched for if the identified mutation are present in the interacting residues or not. The protein structures of NR5A1 and interacting partners were docked using the HADDOCK web server. The HADDOCK results were analyzed and the first complex from the top cluster was used for performing refinement analysis. After refinement, Prodigy analysis was performed to find the binding affinity and binding residues of the protein-protein complexes using Prodigy webserver. The binding affinity and in-teraction profile of NR5A1 protein with all interacting partners were analyzed. The interaction residues were also checked for identified mutations. The current study pro-posed that the NR5A1 protein shows interaction with all its target proteins, cofactors, and coactivators. NR5A1 shows the strongest interaction with CTNNB1 among all interacting proteins. Only one mutation from the identified mutations was present in the interacting residues present only 0.4% of the azoospermic cases used in that study.

This study suggests that the mutations identified are not in the interacting residues of NR5A1, and the expression profile of NR5A1 and its interacting residues is also normal in the NGS dataset used in this study. This study supports the studies that contradict the association of NR5A1 mutations with NOA.

REFERENCES

[1] F. Zegers-Hochschild, G. D. Adamson, S. Dyer, C. Racowsky, J. De Mouzon, R. Sokol, L. Rienzi, A. Sunde, L. Schmidt, I. D. Cooke, et al., "The international glossary on infertility and fertility care, 2017," Human reproduction, vol. 32, no. 9, pp. 1786–1801, 2017.

[2] J. Fainberg and J. A. Kashanian, "Recent advances in understanding and man-aging male infertility," F1000Research, vol. 8, 2019.

[3] J. Boivin, L. Bunting, J. A. Collins, and K. G. Nygren, "International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care," Human reproduction, vol. 22, no. 6, pp. 1506–1512, 2007.

[4] D. M. De Kretser, "Male infertility," The lancet, vol. 349, no. 9054, pp. 787–790, 1997.

[5] J. D. Raman, C. F. Nobert, and M. Goldstein, "Increased incidence of testicular cancer in men presenting with infertility and abnormal semen analysis," The Journal of urology, vol. 174, no. 5, pp. 1819–1822, 2005.

[6] T. J. Walsh, M. Schembri, P. J. Turek, J. M. Chan, P. R. Carroll, J. F. Smith, M. L. Eisenberg, S. K. Van Den Eeden, and M. S. Croughan, "Increased risk of high-grade prostate cancer among infertile men," Cancer, vol. 116, no. 9, pp. 2140–2147, 2010.

[7] A. Agarwal, A. Mulgund, A. Hamada, and M. R. Chyatte, "A unique view on male infertility around the globe," Reproductive biology and endocrinology, vol. 13, no. 1, pp. 1–9, 2015.

[8] N. Aziz, "The importance of semen analysis in the context of azoospermia," Clinics, vol. 68, pp. 35–38, 2013.

[9] J. P. Jarow, M. A. Espeland, and L. I. Lipshultz, "Evaluation of the azoospermic patient," The Journal of urology, vol. 142, no. 1, pp. 62–65, 1989.

[10] K. Jarvi, K. Lo, A. Fischer, J. Grantmyre, A. Zini, V. Chow, and V. Mak, "Cua guideline: The workup of azoospermic males," Canadian Urological Associa-tion Journal, vol. 4, no. 3, p. 163, 2010.

[11] M. Wosnitzer, M. Goldstein, and M. P. Hardy, "Review of azoospermia," Sper-matogenesis, vol. 4, no. 1, p. e28218, 2014.

[12] J. Gonsalves, F. Sun, P. N. Schlegel, P. J. Turek, C. V. Hopps, C. Greene, R. H. Martin, and R. A. R. Pera, "Defective recombination in infertile men," Human molecular genetics, vol. 13, no. 22, pp. 2875–2883, 2004.

[13] V. N. Peña, T. P. Kohn, and A. S. Herati, "Genetic mutations contributing to non-obstructive azoospermia," Best practice & research Clinical endocrinology & metabolism, vol. 34, no. 6, p. 101479, 2020.

[14] P. J. Burrows, C. G. Schrepferman, and L. I. Lipshultz, “Comprehensive office evaluation in the new millennium,” Urologic Clinics, vol. 29, no. 4, pp. 873–894, 2002.

[15] S. C. Esteves, R. Miyaoka, and A. Agarwal, “An update on the clinical assess-ment of the infertile male,” Clinics, vol. 66, no. 4, pp. 691–700, 2011.

[16] S. Zhang, Y. An, J. Li, J. Guo, G. Zhou, J. Li, and Y. Xu, “Relation between the testicular sperm assay and sex hormone level in patients with azoospermia induced by mumps,” International journal of clinical and experimental medicine, vol. 8, no. 11, p. 21669, 2015.

[17] C. Kang, N. Punjani, and P. N. Schlegel, “Reproductive chances of men with azoospermia due to spermatogenic dysfunction,” Journal of Clinical Medicine, vol. 10, no. 7, p. 1400, 2021.

[18] T. Tharakan, R. Luo, C. N. Jayasena, and S. Minhas, “Non-obstructive azoospermia: current and future perspectives,” Faculty Reviews, vol. 10, 2021.

[19] M. Cocuzza, C. Alvarenga, and R. Pagani, “The epidemiology and etiology of azoospermia,” Clinics, vol. 68, pp. 15–26, 2013.

[20] A. Chandra, C. E. Copen, and E. H. Stephen, Infertility service use in the United States: data from the National Survey of Family Growth, 1982-2010. No. 73, US Department of Health and Human Services, Centers for Disease Control and . . . , 2014.

[21] M. L. Eisenberg, P. Betts, D. Herder, D. J. Lamb, and L. I. Lipshultz, “Increased risk of cancer among azoospermic men,” Fertility and sterility, vol. 100, no. 3, ap.681–685, 2013.

[22] W. Cates, T. M. Farley, and P. J. Rowe, “Worldwide patterns of infertility: is africa different?,” The Lancet, vol. 326, no. 8455, pp. 596–598, 1985.

[23] S. C. Esteves, “Clinical management of infertile men with nonobstructive azoospermia,” Asian journal of andrology, vol. 17, no. 3, p. 459, 2015.

[24] M. I. B. P. P. C. of the American Urological Association, P. C. of the American Society for Reproductive Medicine, et al., “Report on evaluation of the azoospermic male,” Fertility and sterility, vol. 82, pp. 131–136, 2004.

[25] D. Knapen, L. Vergauwen, K. Laukens, and R. Blust, “Best practices for hy-bridization design in two-colour microarray analysis,” Trends in biotechnology, vol. 27, no. 7, pp. 406–414, 2009.

[26] D. Qin, “Next-generation sequencing and its clinical application,” Cancer biol-ogy & medicine, vol. 16, no. 1, p. 4, 2019.

[27] J. Besser, H. A. Carleton, P. Gerner-Smidt, R. L. Lindsey, and E. Trees, “Next-generation sequencing technologies and their application to the study and control of bacterial infections,” Clinical microbiology and infection, vol. 24, no. 4, ap.335–341, 2018.

[28] S. J. De Vries, M. Van Dijk, and A. M. Bonvin, “The haddock web server for data-driven biomolecular docking,” Nature protocols, vol. 5, no. 5, pp. 883–897, 2010.

[29] D. M. de Kretser, K. L. Loveland, A. Meinhardt, D. Simorangkir, and N. Wre-ford, “Spermatogenesis,” Human reproduction, vol. 13, no. suppl_1, pp. 1–8, 1998.

[30] A. Massart, W. Lissens, H. Tournaye, and K. Stouffs, “Genetic causes of spermatogenic failure,” Asian journal of andrology, vol. 14, no. 1, p. 40, 2012.

[31] D. Adamopoulos and E. Koukkou, “‘value of fsh and inhibin-b measurements in the diagnosis of azoospermia’–a clinician’s overview,” International journal of andrology, vol. 33, no. 1, pp. e109–e113, 2010.

[32] J. Xu, Y. Sun, Y. Zhang, N. Ou, H. Bai, J. Zhao, S. Xu, J. Luo, S. Han, P. Li, et al., “A homozygous frameshift variant in sycp2 caused meiotic arrest and non-obstructive azoospermia,” Clinical Genetics, 2023.

[33] J. Martínez, S. Bonache, A. Carvajal, L. Bassas, and S. Larriba, “Mutations of sycp3 are rare in infertile spanish men with meiotic arrest,” Fertility and sterility, vol. 88, no. 4, pp. 988–989, 2007.

[34] J. Lian, X. Zhang, H. Tian, N. Liang, Y. Wang, C. Liang, X. Li, and F. Sun, “Altered microrna expression in patients with non-obstructive azoospermia,” Re-productive biology and endocrinology, vol. 7, pp. 1–10, 2009.

[35] D. Hashemi Karoii, H. Azizi, and T. Skutella, “Microarray and in silico anal-ysis of dna repair genes between human testis of patients with nonobstructive azoospermia and normal cells,” Cell Biochemistry and Function, vol. 40, no. 8, pp. 865–879, 2022.

[36] B. Govindkumar, B. Kavyashree, A. K. Bajpai, S. Davuluri, K. Shruthi, S. Vasan, M. Madhusudhan, S. Chandrasekhar Darshan, C. Neelima, B. Vashishtkumar, et al., “Listing candidate diagnostic markers and transcriptomic exploration of the molecular basis of a type of male infertility (non-obstructive azoospermia), via next generation sequencing methods,” bioRxiv, p. 778670, 2019.

[37] X. Liu, Q. Xi, L. Li, Q. Wang, Y. Jiang, H. Zhang, R. Liu, and R. Wang, “Targeted next-generation sequencing identifies novel sequence variations of genes associ-ated with nonobstructive azoospermia in the han population of northeast china,” Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, vol. 25, p. 5801, 2019.

[38] M. Cerván Martín, S. González-Muñoz, L. Bossini-Castillo, A. Guzmán-Jime’nez, N. Garrido, S. Luján, A. Clavero, S. Azoonomic, A. Barros, S. Seixas, et al., “P-536 common variation in the pin1 locus increases the genetic risk to suffer from sertoli cell only syndrome,” Human Reproduction, vol. 37, no. Sup-plement_1, pp. deac107–494, 2022.

[39] P. Fenichel, R. Rey, S. Poggioli, M. Donzeau, D. Chevallier, and G. Pointis, “Anti-mullerian hormone as a seminal marker for spermatogenesis in non-obstructive azoospermia,” Human reproduction, vol. 14, no. 8, pp. 2020–2024, 1999.

[40] H.-Y. Xu, H.-X. Zhang, Z. Xiao, J. Qiao, and R. Li, “Regulation of anti-müllerian hormone (amh) in males and the associations of serum amh with the disorders of male fertility,” Asian journal of andrology, vol. 21, no. 2, p. 109, 2019.

[41] S. Chamindrani Mendis-Handagama and H. Siril Ariyaratne, “Differentiation of the adult leydig cell population in the postnatal testis,” Biology of reproduction, vol. 65, no. 3, pp. 660–671, 2001.

[42] J. Klattig, R. Sierig, D. Kruspe, M. Makki, and C. Englert, “Wt1-mediated gene regulation in early urogenital ridge development,” Sexual Development, vol. 1, no. 4, pp. 238–254, 2007.

[43] Q.-S. Zheng, X.-N. Wang, Q. Wen, Y. Zhang, S.-R. Chen, J. Zhang, X.-X. Li, R.-N. Sha, Z.-Y. Hu, F. Gao, et al., “Wt1 deficiency causes undifferentiated spermatogonia accumulation and meiotic progression disruption in neonatal mice,” Reproduction, vol. 147, no. 1, pp. 45–52, 2014.

[44] D. Zare-Abdollahi, S. Safari, R. Mirfakhraie, A. Movafagh, M. Bastami, P. Az-imzadeh, N. Salsabili, W. Ebrahimizadeh, S. Salami, and M. Omrani, “Muta-tional screening of the nr5a1 in azoospermia,” Andrologia, vol. 47, no. 4, pp. 395– 401, 2015.

[45] M. J. Wilson, P. Jeyasuria, K. L. Parker, and P. Koopman, “The transcription factors steroidogenic factor-1 and sox9 regulate expression of vanin-1 during mouse testis development,” Journal of Biological Chemistry, vol. 280, no. 7, pp. 5917–5923, 2005.

[46] B. Gurates, A. Amsterdam, M. Tamura, S. Yang, J. Zhou, Z. Fang, S. Amin, S. Sebastian, and S. E. Bulun, “Wt1 and dax-1 regulate sf-1-mediated human p450arom gene expression in gonadal cells,” Molecular and cellular endocrinol-ogy, vol. 208, no. 1-2, pp. 61–75, 2003.

[47] B. K. Jordan, J. H.-C. Shen, R. Olaso, H. A. Ingraham, and E. Vilain, “Wnt4 overexpression disrupts normal testicular vasculature and inhibits testosterone synthesis by repressing steroidogenic factor 1/β -catenin synergy,” Proceedings of the National Academy of Sciences, vol. 100, no. 19, pp. 10866–10871, 2003.

[48] P. Asero, A. Calogero, R. Condorelli, L. Mongioi’, E. Vicari, F. Lanzafame, R. Crisci, and S. La Vignera, “Relevance of genetic investigation in male infertility,” Journal of endocrinological investigation, vol. 37, pp. 415–427, 2014.

[49] F. Tüttelmann, C. Ruckert, and A. Röpke, “Disorders of spermatogenesis: per-spectives for novel genetic diagnostics after 20 years of unchanged routine,” Medizinische Genetik, vol. 30, no. 1, pp. 12–20, 2018.

[50] K. Oba, T. Yanase, M. Nomura, K.-i. Morohashi, R. Takayanagi, and H. Nawata, “Structural characterization of humanad4bp (sf-1) gene,” Biochemical and bio-physical research communications, vol. 226, no. 1, pp. 261–267, 1996.

[51] R. Cannarella, R. A. Condorelli, Y. Duca, S. La Vignera, and A. E. Calogero, “New insights into the genetics of spermatogenic failure: a review of the litera-ture,” Human genetics, vol. 138, pp. 125–140, 2019.

[52] H. Fabbri-Scallet, L. M. de Sousa, A. T. Maciel-Guerra, G. Guerra-Júnior, and M. P. de Mello, “Mutation update for the nr5a1 gene involved in dsd and infer-tility,” Human mutation, vol. 41, no. 1, pp. 58–68, 2020.

[53] B. M. Duc, L. T. L. Anh, N. Van Hai, and N. T. Duong, “Association study of nr5a1 rs1110061 with infertile male in 401 vietnamese individuals,” Vietnam Journal of Biotechnology, vol. 19, no. 4, pp. 625–631, 2021.

[54] A. Bashamboo, B. Ferraz-de Souza, D. Lourenço, L. Lin, N. J. Sebire, D. Mont-jean, J. Bignon-Topalovic, J. Mandelbaum, J.-P. Siffroi, S. Christin-Maitre, et al., “Human male infertility associated with mutations in nr5a1 encoding steroido-genic factor 1,” The American Journal of Human Genetics, vol. 87, no. 4, pp. 505– 512, 2010.

[55] A. Röpke, A.-C. Tewes, J. Gromoll, S. Kliesch, P. Wieacker, and F. Tüttelmann, “Comprehensive sequence analysis of the nr5a1 gene encoding steroidogenic fac-tor 1 in a large group of infertile males,” European Journal of Human Genetics, vol. 21, no. 9, pp. 1012–1015, 2013.

[56] D. Sudhakar, S. Nizamuddin, G. Manisha, J. Devi, N. Gupta, B. Chakravarthy, M. Deenadayal, L. Singh, and K. Thangaraj, “Nr 5a1 mutations are not associ-ated with male infertility in indian men,” Andrologia, vol. 50, no. 3, p. e12931, 2018.

[57] B. Klaus and S. Reisenauer, “An end to end workflow for differential gene ex-pression using affymetrix microarrays,” F1000Research, vol. 5, 2016.

[58] S. M. Holland, “Principal components analysis (pca),” Department of Geology, University of Georgia, Athens, GA, vol. 30602, p. 2501, 2008.

[59] N. C. Schwertman, M. A. Owens, and R. Adnan, “A simple more general box-plot method for identifying outliers,” Computational statistics & data analysis, vol. 47, no. 1, pp. 165–174, 2004.

[60] A. Pryke, S. Mostaghim, and A. Nazemi, “Heatmap visualization of population based multi objective algorithms,” in Evolutionary Multi-Criterion Optimiza-tion: 4th International Conference, EMO 2007, Matsushima, Japan, March 5-8, 2007. Proceedings 4, pp. 361–375, Springer, 2007.

[61] D. Blankenberg, G. V. Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, “Galaxy: a web-based genome analysis tool for experimentalists,” Current protocols in molecular biology, vol. 89, no. 1, pp. 19– 10, 2010.

[62] L. C. Xue, J. P. Rodrigues, P. L. Kastritis, A. M. Bonvin, and A. Vangone, "Prodigy: a web server for predicting the binding affinity of protein–protein complexes," Bioinformatics, vol. 32, no. 23, pp. 3676–3678, 2016.

Appendix A

Interacting Residues of NR5A1 in all Complex

Table A.1. Interacting Residues of NR5A1 in All Complexes

NR5A1 -AMH	Residue Number	NR5A1 -SOX9	Residue Number	NR5A1 -CTNNB1	Residue Number	NR5A1 -WT1	Residue Number
GLY	413	HIS	439	ASP	414	GLN	357
ASP	380	GLU	435	LEU	402	SER	346
ALA	458	TYR	438	GLY	413	ASP	380
ASN	444	TYR	438	THR	296	GLN	417
LYS	382	LEU	456	ASP	414	SER	342
LEU	442	PHE	383	LEU	402	LEU	343
LEU	442	ASP	257	ALA	399	SER	346
ARG	427	LEU	442	ASP	414	SER	346
LEU	420	LYS	459	ARG	427	LEU	379
GLU	445	MET	455	LEU	421	GLN	432
LYS	434	LEU	442	LYS	434	LEU	381
GLN	417	ALA	458	GLN	417	GLN	417
PHE	416	LYS	459	LEU	402	SER	378
TYR	438	MET	431	GLU	225	LEU	421
PHE	383	GLU	445	LYS	396	SER	342
ASN	444	TYR	436	ALA	458	LEU	379
LYS	459	PRO	259	GLU	395	LYS	382
PHE	383	MET	431	LEU	456	GLN	417
LEU	420	LYS	440	ALA	399	LEU	381
GLY	443	GLU	435	GLU	225	GLN	432
GLN	457	GLU	445	MET	431	GLN	432
HIS	439	ASP	380	ARG	427	SER	342
GLY	413	MET	431	GLU	395	LEU	343
GLN	394	GLN	299	ARG	427	LYS	382
HIS	441	LYS	440	ASP	403	LEU	421
ASP	414	GLN	457	LEU	406	HIS	439
HIS	439	ARG	427	LEU	379	GLN	417
ASP	414	ALA	458	GLN	418	LEU	349
GLN	457	HIS	439	VAL	424	GLU	353
GLN	418	HIS	439	LYS	396	ASP	380
GLN	418	GLY	443	CYS	412	LEU	442
PHE	383	LEU	343	LYS	459	LEU	421
TYR	438	LEU	456	LYS	391	LEU	421
ARG	448	GLY	341	LEU	456	ASN	398
GLY	443	GLN	432	GLY	413	GLN	357

LYS	440	GLN	457	ARG	427	GLN	418
ASP	414	HIS	439	ASN	222	LEU	420
LEU	442	TYR	438	GLN	417	SER	342
LEU	442	LEU	343	GLU	435	GLY	413
SER	378	GLY	341	LEU	406	ASN	398
HIS	439	GLU	435	MET	431	LEU	420
GLN	417	THR	338	LYS	459	GLN	417
LYS	382	ALA	340	ARG	427	LEU	381
GLN	417	ALA	340	HIS	411	SER	346
LEU	442	LYS	459	GLU	225	LEU	343
LEU	381	LEU	442	LYS	391	GLU	425
PHE	416	ASN	300	MET	431	VAL	424
LYS	440	MET	431	LYS	434	GLU	435
GLN	417	GLU	435	LYS	459	VAL	424
ASP	380	TYR	436	LYS	459	LYS	382
HIS	441	LYS	434	TYR	438	ARG	350
LEU	442	GLU	445	LEU	421	LEU	421
TYR	438	PHE	262	GLU	425	GLU	435
TYR	438	LYS	434	GLU	435	PHE	416
LYS	382	LYS	459	ALA	399	TYR	438
GLU	395	LEU	379	ALA	428	VAL	424
ASN	398	GLU	435	ALA	399	LEU	421
GLN	457	TYR	438	ASN	300	ASN	398
GLN	417	GLU	445	LYS	396	GLU	425
LEU	421	MET	431	PHE	383	GLN	418
TYR	438	GLY	341	VAL	424	ASP	380
GLY	443	ARG	427	ARG	350	GLU	425
ASN	444	ASN	300	ALA	400	GLN	394
LEU	456	MET	431	ASN	222	GLU	435
LYS	391	LEU	456	ALA	458	GLN	417
GLN	457	THR	338	ASN	398	ASP	380
LEU	442	GLU	445	VAL	424	ARG	350
LYS	459	GLN	339	GLN	417	LEU	354
LEU	421	GLN	339	GLU	395		
VAL	424	SER	378	ARG	427		
TYR	438	ASN	444	LEU	421		
LEU	421	HIS	439	LYS	434		
LEU	442	HIS	439	CYS	412		
LEU	379	LEU	343	GLU	395		
PHE	383	TRP	302	TYR	404		
LEU	442	TYR	438	LEU	406		
		GLN	432	VAL	424		
				ASP	403		
				GLY	413		

				GLN	457		
				GLY	413		
				LEU	456		
				ALA	393		
				ASP	403		
				ALA	399		
				GLN	457		
				LEU	402		
				ARG	427		
				GLU	435		
				MET	431		
				GLN	299		
				ASP	392		
				GLU	395		
				GLN	417		
				ASN	222		
				LEU	401		
				ASN	300		
				ASP	414		
				LYS	391		
				GLU	395		
				TYR	438		
				GLN	418		
				PRO	224		
				LYS	396		
				MET	431		
				MET	455		
				PHE	416		
				ALA	428		
				LYS	434		
				PHE	383		