Chapter I GENERAL INTRODUCTION.

**CONTENTS AND STRUCTURE**

**A. BACKGROUND.**
I-A.1. **Gene expression and promoters**: Knowledge of the gene expression control is essential for understanding the molecular mechanisms of cellular and physiological events.
I-A.2. **Mammalian promoters** are complex and many have been characterized mainly based on the TFBSs, CpG islands and the epigenetic modifications.
I-A.3. **The DNA microarray technique** has been the most widely used method for gene expression profiling, and has generated a huge amount of data.
I-A.4. **RNA sequencing** seems to be a more reliable method of gene expression profiling.
I-A.5. **Mammalian testis** forms an ideal tissue for studying gene expression regulation.
I-A.6. **Non obstructive azoospermia**: There is a need to study more about NOA in terms of diagnostics, therapeutics and so on.

**B. CURRENT WORK AND OBJECTIVES.**
I-B.1. **Gene expression control**: Establishing gene expression patterns and understanding the mechanism of expression control remain challenging tasks.
I-B.2. **Promoter analysis and predictions** also need to be improved.
I-B.3. **Better methods**: The current work is based on a few postulations related to the methods of gene expression data analysis.
I-B.4. **Objectives**: The current work intends to improve the compilation of available gene expression data and their meta-analysis, derive a reliable set of co-expressed gene clusters and efficiently analyze the corresponding promoter clusters.
A. BACKGROUND.

I-A.1. Gene expression and promoters: Knowledge of the gene expression control is essential for understanding the molecular mechanisms of cellular and physiological events.

Most of the cellular functions are executed at the molecular level by proteins. DNA is used by all cells to store the information for expression of proteins. This bio-molecular information is stored in the form of a sequence of four nucleotides, viz., Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). DNA propagates this sequence information and hence forms the genetic material. The double stranded nature of the DNA molecule and the complementary pairing of nitrogenous bases (A-T and G-C) makes it easy to replicate. The molecular machinery inside a cell uses specific regions of DNA, the genes, to produce proteins via two separate steps: transcription and translation. In the first step (transcription), the nucleotide sequence in one of the strands of each gene is used as a template and a RNA molecule is produced with a sequence of nucleotides complementary to that of the template. Triplets of nucleotides present in RNA molecule are then used to code for amino acids, which are polymerized into a protein, during the second step (translation). These fundamental molecular processes have been well established and collectively referred to as the central dogma of molecular biology [Crick, 1970]. While the sequence of nucleotides in the DNA is same in all cells in a multicellular organism, the way it is used varies in different regions in the body [Brown, 2002a; Alberts et al., 2002]. Different genes are turned on or shut off in varying combinations across the cell types/tissues. In fact, such differential expression of genes results in a gradual differentiation of homogenous embryonic (totipotent) stem cells into various cell types/tissues, during development [Tata, 1993; Gilbert, 2000; Lodish et al., 2000; Alonso et al., 2012] and forms the basis of normal differential functions of tissues in normal conditions. Thus, the 'genome' a representative DNA set of an organism, is a constant, the 'transcriptomes' and 'proteomes' are the representative sets of RNAs and proteins - respectively, form the functional units of genomes and vary across locations and conditions within the organism. In fact, Brown [2002] views the proteome as, "on one hand, the culmination of genome expression and, on the other hand, the starting point for the biochemical activities that constitute cellular life".
Regulation of gene expression is complex in eukaryotes, particularly in mammals. Well-orchestrated regulation of gene expression is important for normal functioning of multicellular organisms. The control of gene expression is implied by the routine phrase: ‘DNA makes RNA makes protein’. Control mechanisms exist before, during and after each of the steps. The transcription initiation is controlled by many ways as described below [Figure I-1]. After transcription, the RNA is processed in various ways, particularly in mammals, and the translational process is regulated at different stages involving initiation, elongation, termination and post-translational modification [Brown, 2002a; Alberts et al., 2002]. These control mechanisms permit alteration of the composition of the transcriptome and proteome frequently and thereby allow the cell to adjust its biochemical capabilities in response to changes in the extracellular environment and to signals received from other cells. Such regulatory events underlie the functioning of individual cells as well as the processes of tissue differentiation and development. Changes in regulatory mechanisms can obviously lead to abnormal tissue functions and diseases.

Transcriptional regulation plays a crucial role in differential expression of genes across normal and abnormal conditions in mammals. Although gene expression can be regulated at different levels, regulation at transcription initiation is very crucial [Brown, 2002a; Alberts et al., 2002]. Initiation of the transcription of a gene involves binding of specific proteins, including the RNA polymerase, around the transcription start site (TSS) and culminates with the RNA polymerase leaving the promoter and beginning of the synthesis of RNA molecule. These are the first few steps in the expression of any gene [Brown, 2002a; Alberts et al., 2002]. A deviation in the gene expression patterns can result in diseases. Many disorders have been well studied in terms of such deviation in
expression profiles [Examples: William et al., 2003; Cookson et al., 2009; Ryten et al., 2009; Handunnetthi et al., 2010; Raspe et al., 2012; Emmert et al., 2012;].

**Varying lengths of promoters are known to be important in controlling transcription.**

Promoter is a DNA sequence that influences transcription of the neighboring gene [Lewin, 2003]. In the recent literature, the promoter has been used to refer to the regulatory region of DNA around the TSS and this region promotes the transcriptional regulation of the corresponding gene by permitting interactions with transcription factors and the RNA polymerases. The promoters targeted by RNA polymerase II are different compared to the promoters of non-protein coding genes [Figure I-2]. However, different lengths of the promoters, from -5000 bp to -500 bp upstream and +1 to +500 bp of protein-coding genes have been considered for analysis by scientists, as described later in this thesis.

Promoters can be classified based on their relative distance from the transcription start site [http://www.experiencefestival.com/a/Promoter_-_Promoter_elements/id/5396671] or based on their chromatin architecture [Cairns, 2009]. Based on their relative distance from the TSS, they are classified into three types

1. **Core promoter:** It is the minimum DNA region required for the transcription initiation [Müller et al., 2007; Juven et al., 2010; Kadonaga et al., 2012;]. It ranges from -50 (50 base pairs upstream of transcription start site) to +50 (50 base pairs downstream of the transcription start site) [Thomas et al., 2006; Müller et al., 2007; Anish et al., 2009; Juven et al., 2010]. Core promoters that surround single transcription start site are known as focused core promoters, while those that surround cluster of transcription start sites separated by few nucleotides are known as dispersed core promoters. Genes that show constitutive expression are known to be more associated with focused core

![Figure I-2. Types of regulatory regions around genes coding for proteins (mRNAs), rRNAs and tRNAs](image-url)
promoters, while differentially regulated genes are known to be associated with dispersed core promoters [Juven et al., 2008]. Core promoter region harbors binding sites for RNA polymerase II and general transcription factors such as TFIIA, TFIIIB, TFIID and TAFs [Juven et al., 2008].

2. Proximal promoter: It is about 250 base pairs upstream of transcription start site. This region possesses binding sites for transcription factors, which facilitate stable formation of pre-initiation complex (PIC) during transcription.

3. Distal promoter: It is located further away from the transcription start site, composed of binding sites (such as enhancer or other TFBSs) for transcription factors that influence gene expression via long-range interactions.

Based on their chromatin architecture, promoters are classified into

1. Open promoter: It refers to a large nucleosome depleted region (~ 150 base pairs) in the upstream or around the transcription start site. Mostly promoters of constitutively expression genes show this type of chromatin architecture.

2. Covered promoters: It refers to the region covered by nucleosomes. The covered region includes transcription start site (TSS), region flanking TSS and binding sites of transcription activators. Highly regulated genes show this type of chromatin architecture.

Besides these classifications, few promoters are known as ‘bidirectional promoters’ as these promoters are located between genes that are transcribed in opposite strands and they influence expression of both genes [Yang et al., 2007; Jorge et al., 2010; Orekhova et al., 2013]. Many genes also use alternative transcription start site for their expression in different tissues/cell types/condition and thereby use alternative promoters [Davuluri et al., 2008; Jacox et al., 2010]

**Transcription Factor Binding Sites (TFBSs) form major functional elements within the promoters.** These DNA motifs of 5-20bp length [D'haeseleer, 2006] are distributed over the core promoter, promoter-proximal elements, and distance-independent elements and are recognized by sequence-specific DNA-binding proteins (transcription factors). A specific constellation of such trans-acting proteins is required for RNA polymerase II to initiate and achieve maximal rates of transcription of each target gene [Griffiths, 2000; Lodish et al., 2000a; Alberts et al., 2002a]. The binding of these general and specific transcription factors as well as certain cofactors, in various combinations, at the
promoters of different genes eventually leads to gene co-expression, cellular differentiation, and normal and abnormal physiology of the tissues [Greenblatt, 1991; Martin, 1991; Frankel et al., 1991; Sindhu et al., 2012; Slattery et al., 2014; Iwafuchi et al., 2014]. Various in silico, experimental as well as combinatorial approaches have been used to analyze the TFBSs across species [Pavesi et al., 2004; Elnitski et al., 2006; Ambesi et al., 2006; Merkulova et al., 2007; Hannenhalli, 2008; Liu et al., 2009; Ladunga, 2010; Soler et al., 2011].

**I-A.2. Mammalian promoters** are complex and many have been characterized mainly based on the TFBSs, CpG islands and the epigenetic modifications. Mammalian gene expression regulation is highly complex due to the diversity of tissues and cell-types. Knowledge on the mechanisms of transcriptional regulation has grown in various contexts in the past few decades [Mitchell et al., 1989; Müller, 1995; Miano, 2002; Savage et al., 2003; Ma, 2006; Spear et al., 2006; Wilson et al., 2009; de la Torre-Ubieta et al., 2011; Cano et al., 2014]. In fact, a significant portion of mammalian genomes is used to code for transcription factors. For example, in the human genome, ~6% of protein coding genes express transcription factors [Vaquerizas et al., 2009].

In addition to transcription factor binding sites, two other functional features are known to be important in mammalian promoters: a) the CpG islands (CGI) and b) the histone marks. A CpG island is a short stretch of DNA approximately 1000 base pairs long, showing an elevated C+G composition than the normally expected ratio [Cooper et al., 1983; Bird et al., 1985; Gardiner-Garden et al., 1987; Antequera, 2003; Deaton et al., 2009; Illingworth et al., 2009]. CGIs are known to be found in the promoter region of most (60-70%) of human genes [Larsen et al., 2006].

Post-translational modifications of histones of the nucleosomes can differentiate activated promoters from repressed promoters [Berger, 2007]. H3 (9, 27), H4 (20) characterize the gene inactive/transcriptionally repressed promoters, while lysine methylation of H3 (4, 36, 79) characterize active genes and thus active promoters and lysine ubiquitination at H2A (119) characterized repressed gene and thus inactive promoters. In fact, histone-codes seem to exist with multiple combinations of histone modifications in the promoters for different needs of transcriptional control of genes [Geiman et al., 2002; Zhou et al., 2011; Petty et al., 2013].

**Mammalian promoter prediction:** The process of identifying promoter region in the genome or identifying a DNA sequence as promoter based its feature is known as promoter prediction.
There are many tools available for promoter prediction, each of them use different sequence features to identify promoters. For example CpGcluster [Hackenberg et al., 2006] and CpGProD [Ponger et al., 2002] can be used to predict promoter based on CpG island, ReLA [González et al., 2012], Promoter 2.0 [Knudsen, 1999] and PromoterScan [Prestridge, 1995] can be used to predict promoter based on local distribution of TFBS, ProSOM [Abeel et al., 2008a] and EP3 [Abeel et al., 2008] can be used to predict promoter based on DNA structural properties and, ARTS [Sonnenburg et al., 2006] and DrangonPF [Bajic et al., 2002] can be used to predict transcription start site. Some tools like Eponine [Down et al., 2002], NNPP2.2 [Reese, 2001], Nscan [Gross et al., 2006] and Wu-method [Wu et al., 2007] use multiple sequence features to predict promoters.

However, most of these tool’s predictions were not accurate. A recent review [Abeel et al., 2009] reported that, the results of even better performing promoter prediction programs show less overlap among them.

Although recent advances in sequencing technologies have led to large scale mapping of transcription start sites and their usage in human and mouse primary cells [FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014], there is a need for better computational approaches for improving the efficiency of promoter prediction.

I-A.3. The DNA microarray technique has been the most widely used method for gene expression profiling, and has generated a huge amount of data.

Microarrays form a molecular biology tool that allows simultaneous detection of expression levels of thousands of genes [Watson et al., 1998]. A microarray can be either a glass slide or a wafer of glass or plastic spotted with oligonucleotides, referred as probes. Each spot can vary in diameter from 10 to 150 micrometers and contain millions of copies of such a probe, which may be 25 to 80 nucleotides long, depending on the type of manufacturer [Kawasaki, 2006]. A DNA microarray can be used to detect the abundance of each gene by measuring the amount of the corresponding cDNA molecules hybridized with the specific complementary probes on the chip [Stekel, 2003].

Microarray chips are built using two technologies [Stekel, 2003].
a) Robotic spotting: This technology uses spotting robots to print PCR products, cDNAs or pre-synthesized oligonucleotides on the glass slide. The first series of microarray chips were made using this technology.

b) In situ synthesis: In this technology, the probes are built on the array surface base by base. Most of the commercial microarray chips are designed in this way. Example, DNA chips from Affymetrix, Nimblegen, Rosetta, Agilent and so on.

Microarray experiment involves four major laboratory steps [Berrett et al., 2003],

1. Sample preparation and labeling: this step involves extraction of RNA from tissue of interest. The RNA isolated is reverse transcribed into cDNA, during which nucleotides are tagged with fluorescent molecules (e.g., Cy3, Cy5 or both) and/or biotin (in case of Affymetrix arrays).

2. Hybridization: This step involves spreading the hybridization solution on the microarray chip at optimum temperature for over a period of 12-24 hours. The hybridization solution comprises of labeled cDNA, Na\(^+\), and formamide.

3. Washing: once the hybridization is completed, the slides are washed to remove excess of hybridization solution, this reduces the cross hybridization of cDNA with probes. Washing is carried out by using buffers with low salt concentration.

4. Image acquisition: after washing, the microarray slide is placed in the scanner, where the hybridized dye is excited with the appropriate wavelength of light. The scanner records the signal intensity at each spot in the array in the form of an image. The image can be converted into a digital output known as microarray raw data.

**Microarray data analysis:** The analysis of microarray data relies on computational methods for normalization and summarization of the data. Although microarray raw data can be directly used to identify genes that are showing differential expression, the result is influenced by both biological variations across samples and also non-biological variations (systemic bias) that have crept into the data during the experiment. The potential sources of systemic bias include variability in sample preparation and hybridization conditions, spatial effects due to irregular robotic spotting of the probes, irregular scanner settings and bias due to experimenter [Berrett et al., 2003]. Thus, microarray raw data is subjected to normalization, a process that is designed to eliminate systemic bias from the data. Many normalization methods exist, but they can be divided into the following two types.
Within array normalization method: These normalization methods are used on raw data from double channel experiment (a microarray experiment that involves hybridization of two samples labeled with different fluorescent dyes onto same microarray chip). Loess (locally weighted scatter plot smoothening) normalization is most popular within array normalization method. It works on the assumption that majority of the genes on the microarray platform are similarly expressed. It uses nonlinear regression method to calculate normalized log ratio of signal intensity (Cy3/Cy5) for each feature on the microarray chip [Stekel, 2003].

Between array normalization method: These normalization methods are used on microarray raw data obtained from single channel experiments, where labeled cDNA from single sample is hybridized on single microarray chip. Scale normalization and quantile normalization methods are most popular between array normalization methods.

In case of scale normalization method, based on variation in expression level of subset of probes (either corresponding to housekeeping genes or spiked controls) a scaling factor is derived. Based on the assumption that systemic bias is same for all genes in the array, signal intensity of each gene in the array is normalized using the scaling factor [Dobbin et al., 2005].

In case of quantile normalization, the aim is to make distribution of signal intensity same across multiple array data. The method involves ordering of signal intensity values in each array data, obtaining the lowest value from each array data and computing their average. Now this average value is placed at lowest value position in all array data. The process is repeated with second lowest value and so on [Bolstad et al., 2003]. The main advantage of quantile normalization method is, signal intensity range becomes same across multiple arrays and they become comparable.

After the normalization of microarray raw data, it can be used to identify differentially expressed genes based on either fold change or detection calls and by using statistical approaches such as t-test, ANOVA, the significance of differential expression can be estimated.

Applications of microarray: Since its inception, microarray has been widely used as molecular biology tool to address wide variety of biological questions, which includes a) identifying genes in differential expression pattern both in normal and abnormal conditions, b) finding the
genes that indicates similar or differential behavior in different stages of disease conditions, c) classifying specific disease into subtypes based on differential expression profiles, d) analyzing the gene expression patterns in developmental stages of a tissue and e) identifying genes that are tissue-specific or constitutive expression.

**Limitations of microarray:** Even though the microarray technology is extensively used by research community, its findings are not reliable, due to the fact that 1) the results from microarray experiments are not often reproducible [Ntzani et al., 2003], and 2) the results are not robust enough to withstand the slightest data perturbations [Michiels et al., 2005; Ein-Dor et al., 2005]. While some researchers moved on to new technology for analyzing genome-wide gene expression profiles such as RNA sequencing, and others developed novel statistical approaches to obtain reliable gene expression information from existing data by means of meta-analysis.

**I-A.4. RNA sequencing** seems to be a more reliable method of gene expression profiling.

RNA sequencing (RNA-seq) or whole transcriptome sequencing is the technology that has been rapidly replacing microarray for genome-wide gene expression profiling. RNA-seq involves sequencing of total RNA found in a given sample using one of the Next Generation Sequencing (NGS) techniques [Wang et al., 2009; Costa et al., 2010; Marguerat et al., 2010; Oshlack et al., 2010; Ozsolak et al., 2011; Fang et al., 2011; McGettigan, 2013]. It has advantages over microarray [Korpelainen et al., 2015], as a) it doesn’t depend on the existing knowledge about genome sequences [Wang et al., 2009], b) it can detect the expression of mRNA at low level [Fu et al., 2009], c) unlike hybridization based methods it doesn’t face issues such as cross hybridization, limited dynamic range of detection and c) its expression results show high reproducibility [Marioni et al., 2008]

The major steps involved in RNA-seq are as follows [Korpelainen et al., 2015]

1. RNA isolation: RNA is extracted from the freshly obtained tissue sample using commercial RNA isolation kits. The quality of RNA isolated can be measured by nanodrop or other devices.
2. Library construction: All the RNA molecules or only a fraction of it (for example, only RNA with Poly A tail) is considered for library preparation. In this step, every RNA molecule is converted into cDNA, adaptor molecules are annealed at either end of cDNA and the all cDNA molecules are enriched by PCR, using sequences from the adaptors as primers. The protocol for library construction varies from NGS platform to platform.

3. Sequencing: sequencing of cDNA libraries generated can be carried out on different platforms. It can be performed only at one end of the cDNA (single read) or either end of cDNA (paired end). Most commonly used NGS platforms are
   a. Illumina: Single sequencing run in Illumina Hi-seq 2500 may take 3 to 12 days and can produce 600GB of pair end 100 nucleotide data. It uses sequencing by synthetic chemistry and detects fluorescent signals to record nucleotides. It is most popular among RNA-seq platforms.
   b. Roche 454: It uses sequencing by synthesis chemistry and detects fluorescent signals to record nucleotides. It can allow sequencing of longer reads (up to 1000 nucleotides) than other platforms. It can generate up to 700MB of sequence data per run in 1 day.
   c. Ion torrent: It uses sequencing by synthesis chemistry and detects change in pH to record nucleotides. It can generate up to 10GB of 200nt read data per single run. A sequencing run takes 2-4 hours. It is commercially affordable than other platforms.
   d. SOLiD: It uses sequencing by ligation chemistry and detects fluorescent signals to record nucleotides. Due to ligation chemistry, it can provide sequencing accuracy up to 99.9%, which could be pivotal for detecting SNPs. Single sequencing run can generate about 320GB of data.

4. Differential expression analysis: First step in RNA-seq data analysis is to perform quality control analysis of raw sequence reads. Tools like FASTQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] can aid in, analyzing sequence duplication levels, N content of the reads, sequence length distribution, quality scores of the sequence reads, GC content of the sequence reads, etc. Quality control analysis is followed by the preprocessing of reads, where low quality bases and artifacts such as adaptor sequences are removed. In the next step, cleansed sequence reads were mapped to reference genome, which leads to sequence alignment. The alignment file obtained will reveal the number of reads mapped to specific transcript and gene. Thus, the expression of each transcript or gene can be quantified using the read alignment information.
expression level of each gene or transcript can be either represented as read counts or as normalized units such as RPKM (reads per thousand nucleotides in transcript per million reads) or FPKM (fragments per thousand nucleotides per million mapped reads). Once the expression level of each gene or transcript is obtained in different samples/conditions, differences in expression level of a gene or transcript can be easily estimated.

**Applications of RNA-seq:**

1. Identification of novel transcript isoforms and novel protein coding genes
2. Quantification of transcript expression level in specific condition.
3. Analyzing correlation of gene expression variations with SNPs
4. Identification and quantification of fusion genes in specific conditions
5. Scanning single nucleotide polymorphisms within the protein coding region of the genome.
6. Identification and quantification of non-coding RNA.

**I-A.5. Mammalian testis** forms an ideal tissue for studying gene expression regulation.

Testis, being the male gonad, forms a major component of the male reproductive system [Knobil et al., 2000]. Each testis has an outer capsule comprising mass of seminiferous tubules, which are heavily coiled. In between tubules, there is interstitial tissue comprising blood vessels and Leydig cells [Knobil et al., 2000]. Spermatogenesis occurs within the seminiferous tubules. Each seminiferous tubule contains sertoli cells and various germ cells such as spermatogonia (diploid), primary spermatocyte (diploid), secondary spermatocyte (diploid), spermatid (haploid) and spermatozoa (haploid) [Figure I-3]. Sertoli cells provide nourishment for transforming spermatogonia, by secreting various proteins such as growth factors, hormones, enzymes and androgen binding protein [Griswold, 1998].

During spermatogenesis, each spermatogonium undergoes meiosis to give rise to spermatids forming intermediate cell types such as primary and secondary spermatocytes. Spermatids differentiate into sperm by a process known as spermiogenesis, which involves losing their cytoplasm, acquiring a flagellated tail, chromatin condensation and forming cap like structure
at the tip of the head called acrosome. The sperm uses mitochondria at the mid piece of its body to generate energy for its movement. The sperm is then released into the lumen of seminiferous tubules with a secreted fluid generated by accessory glands. The entire process of spermatogenesis requires 64 days in humans [Silverthorn, 2013].

Spermatogenesis is controlled by hormones such as follicle stimulating hormone (FSH) and leutinizing hormone (LH). FSH signals sertoli cells to produce paracrine molecules required for mitosis of spermatogonia, while LH targets Leydig cells to produce testosterone [Knobil et al., 2000].

Testis being the only male tissue where cells undergo meiosis, and having many other peculiar aspects such as the existence of blood-tissue barrier and continuous cycles of post-pubertal mitosis, testis must use a unique set of proteins across its cell types. It is known that testis has a more widespread transcription and complexity compared to many tissues, including brain, heart, liver and kidney in mammalian species [Soumillon et al., 2013]. In fact, transcriptional regulation in mammalian testis has drawn the attention of many researchers in the past [Reddi et al., 2002; Reddi et al., 2003; Grimes, 2004; Kimmins et al., 2005; Acharya et al., 2006; Soumillon et al., 2013].

I-A.6. Non obstructive azoospermia: There is a need to study more about NOA in terms of diagnostics, therapeutics and so on.

Non obstructive azoospermia is a type of male infertility condition. Infertility is the failure to conceive even after months of regular unprotected intercourse [Edmund et al., 2010]. The male infertility condition where the absence of sperm is observed in the ejaculate is known as azoospermia. Two major subtypes of azoospermia are

1. Obstructive azoospermia: In this case, spermatogenesis is normal, but infertility occurs either due to mechanical obstruction within the genital tract between the epididymis and the ejaculatory duct or due to absence of vasa deferentia [Esteves et al., 2013]. This forms 40% of the azoospermia cases [Wosnitzer et al., 2014]

2. Non obstructive azoospermia: is caused due to severely impaired spermatogenesis [Esteves et al., 2013]. This forms 60% of the azoospermia cases [Wosnitzer et al., 2014].
Prevalence of NOA in India: The prevalence of primary infertility in India has been estimated to be 10-20% and in 40% cases, male factor alone accounts for it [George et al., 2003].

Diagnosis and treatment methods available for NOA: For diagnosis of NOA in men with no history of OA or prior fertility or toxic exposure two common tests used [Wosnitzer et al., 2014] are:

1. Karyotype analysis: This analysis checks for chromosomal abnormalities associated NOA, such as Klinefelter syndrome (XXY), Robertsonian translocations.

2. Y chromosome microdeletion (YCMD) test: This tests integrity of three azoospermia factor (AZF) regions in the q arm of Y chromosome.

While for men with congenital hypogonadotropic hypogonadism, genetic testing of specific genes (KAL1, FGFR1) is carried out.

Currently NOA patients with Klinefelter syndrome or AZFc microdeletion are treated with Micro-dissection testicular sperm extraction (microTESE) in combination with IVF/ICSI. While men with hypogonadotropic hypogonadism are treated by providing gonadotropin supplements such as hCG, recombinant FSH, GnRH.

Efforts towards understanding the molecular mechanism of NOA and identification of diagnostic and therapeutic biomarkers: In the last decade, researchers in pursuit of understanding non obstructive azoospermia made following discoveries,

1. Chromosomal aberrations leading to NOA were identified [Gallego et al., 2014; Mirfakhraie, 2010; Zaimy et al., 2013]

2. Lian [Lian et al., 2009] and Wu [Wu et al., 2012; Wu et al., 2013], reported over-expressed microRNAs, which affect the expression of spermatogenesis related genes.

3. Ferfouri [Ferfouri et al., 2013], reported CpG islands that show differential methylation in normal and NOA conditions.

4. Mutations associated with increased or decreased risk of NOA were reported [Jin et al., 2013; Asadpor et al., 2013; Qin et al., 2014; He et al., 2012; Norambuena et al., 2012].
5. NOA susceptibility loci were identified via GWAS [Lu et al., 2014; Hu et al., 2014; Hu et al., 2011]

6. Potential biomarkers were reported using qRT-PCR, microarray and spectral counting analysis [Almeida et al., 2013; Dorosh et al., 2013; Malcher et al., 2013; Wang et al., 2013; Batruch et al., 2012; Bonaparte et al., 2010]

But there is a further need to study this condition, particularly to explore the potential new markers for prognostics, diagnostics and possible therapeutics [Kumar, 2013; Glina et al., 2013]. The reported potential biomarkers have not been extensively followed up and/or reported in a good number of patients, particularly among Indians.

**B. CURRENT WORK AND OBJECTIVES.**

**I-B.1. Gene expression control:** Establishing gene expression patterns and understanding the mechanism of expression control remain challenging tasks.

The invention of the high throughput genome-wide expression profiling technique microarray [Schena, 1996; van Hal et al., 2000; Schulze et al., 2001; Noordewier et al., 2001; Barrett et al., 2003; Kolbert et al., 2003] resulted in high ambition among researchers to explore gene expression profiles in general as well as for specific diseases [Weeraratna et al., 2004; Chittur, 2004; Choudhuri, 2004; Stoughton, 2005; Hofman, 2005]. In fact, microarray has been extensively used by molecular biologists to identify the genes associated with various physiological/experimental conditions in different species, and such studies have yielded promising results [Cooper et al., 2007; Gong et al., 2007; Sørensen et al., 2010].

However, with the increased use of microarray, researchers found out its limitations as well [Brazma et al., 2001; Forster et al., 2003; Russo et al., 2003; Draghici et al., 2006; Ramasamy et al., 2008]. These observations has led to two outcomes, a) many researchers desired and/or explored new technologies such as SAGE [Velculescu et al., 1995] and MPSS [Brenner et al., 2000]; and b) several researchers tried different statistical approaches to either reduce the variations in the microarray data from different studies [Tseng et al., 2012] or derive a consensus across data sets from comparable studies [Rhodes et al., 2004, 2004a]. However, most such meta-analysis methods could only address the issues partially.

But then, around these times, when limitations of microarray technology were evident, the RNA-seq technology surfaced and the hopes have resurged even stronger in recent years [Wang et al., 2009; Costa et al., 2010; Marguerat et al., 2010; Ozsolak et al., 2011;
Nevertheless, two problems remain: a) a huge amount of data has been produced using the microarray technology for various tissues and conditions across multiple research models; b) RNA-sequencing is very expensive, particularly if researchers would like to make the maximum potential use of this technique.

I-B.2. Promoter analysis and predictions also need to be improved.

Specific DNA motifs (TFBSs) and chromatin modifications along with coordinated protein-DNA interactions within the promoters make the transcriptional regulatory codes for different genes' expression and this code is not yet understood completely [Heintzman et al., 2007]. However, a comprehensive analysis of promoters of similarly expressed genes should aid in identifying functional regulatory signals and thereby facilitate promoter prediction. Many efforts have been made to develop computational aids for promoter prediction using various algorithms and parameters [Hannenhalli et al., 2001; Bajic et al., 2004; Kanhere et al., 2005; Burden et al., 2005; Gangal et al., 2005; Narang et al., 2005; Wu et al., 2007; Wang et al., 2007; Abeel et al., 2008, 2008a, 2009; Rani et al., 2009; Gan et al., 2009; Askary et al., 2009; Lin et al., 2011; Gan et al., 2012; Xiong et al., 2014]. Such efforts on in silico predictions and analysis of promoters have also been frequently reviewed [Werner, 1999; Pedersen et al., 1999; Werner, 2001; Zhang, 2003; Heintzman et al., 2007; Hernandez-Garcia et al., 2014]. However, a preliminary analysis (unpublished data, Appendix D) showed that the existing promoter analysis and prediction tools are less efficient - a conclusion also drawn independently by Bajic [Bajic et al., 2004] and Abeel [Abeel et al., 2009].

I-B.3. Better methods: The current work is based on a few postulations related to the methods of gene expression data analysis.

The thought of making a better use of the existing data by tissue-specific database creations along with the development of a novel meta-analysis method, to derive a consensus expression status of each gene across comparable data sets, has been one of the main motivations for the current study.

Such an approach was thought not only to enhance the possibility of better biomarker discovery, but also paves way to identify refined clusters of differentially expressed genes and hence promote a better analysis of corresponding genes.
Based on differential expression of genes, identified by the high throughput technologies it is possible to list a set of potential biomarkers for specific conditions of interest, such as human diseases. However, it is also possible to use the existing data sets to establish expression profiles for normal conditions across tissues in important research models such as the mouse, rat and human species. The usually scattered existing gene expression profiles were being compiled for selected mammalian tissues [Neelima Ch, Ongoing doctoral thesis work] and this provided an opportunity to cluster the mammalian genes based on their profiles and explore the corresponding promoters for the features that may be associated with the differential expressions. Since the main functions of mammalian promoters is to regulate the differential expression of genes across tissues under normal conditions, it may be possible to identify the functional features of promoters corresponding to different co-expressed genes such as those expressed exclusively in a tissue vs. those expressed in most tissues.

The previous approaches [Hannenhalli et al., 2001; Bajic et al., 2004; Kanhere et al., 2005; Burden et al., 2005; Gangal et al., 2005; Narang et al., 2005; Wu et al., 2007; Wang et al., 2007; Abeel et al., 2008; Abeel et al., 2008a; Abeel et al., 2009; Rani et al., 2009; Gan et al., 2009; Askary et al., 2009; Lin et al., 2011; Gan et al., 2012; Xiong et al., 2014] of promoter analysis and predictions have not followed such an approach. Many computational methods have been developed earlier for promoter prediction/analysis. The computational methods employed by earlier researchers seem to be robust and include support vector machine, pattern search and artificial neural network. They also found to analyze submitted DNA sequence for one or more of the variety of promoter features such as the TSS, CpG island, DNA structural properties, transcription factor binding sites and first exon. Still, the tools have not been very successful in general [Abeel et al., 2009]. They may have failed due to the ambitious goal of separating all promoter sequences from non promoters, while using limited data on promoters.

The degenerate nature of TFBSs and diverse sequence and structural properties among promoters themselves form a huge roadblock for reliable promoter prediction/analysis. Yet, position specific analysis of such motifs may be playing an important role in transcriptional regulation. Similarly, repeat elements in the neighborhood of TSS are also likely to have a functional role. Considering such novel features may help to analyze the functional features and/or predict mammalian promoters.

I-B.4. Objectives: The current work intends to improve the compilation of
available gene expression data and their meta-analysis, derive a reliable set of co-expressed gene clusters and efficiently analyze the corresponding promoter clusters.

It was hypothesized that making use of all or most of the expression data and quantifying the reproducibility in expression patterns would help to short list genes with consistent differential expression across tissues and conditions. It was also postulated that using such reliable sets of co-expressed genes can directly enhance efficient promoter prediction/analysis, as well as help in identification of potential biomarker genes. The promoters of each reliable set of differentially expressed genes are expected to have unique features in terms of the occurrence of motifs, repeat sequences and other elements, and/or their distribution. Thus, the comparative analysis of such promoter clusters can help in identification of the unique features that may be responsible for differential expression of genes.

The main objectives of the current work are:

1. To develop a novel meta-analysis method that can make maximum use of publicly available gene expression data, and quantify the reproducibility across experiments.
2. To obtain two mammalian gene-clusters, based on the transcription pattern across normal tissues: testis-specific genes and ubiquitously transcribed genes
3. To validate the reliability of the newly derived gene clusters.
4. To perform a thorough analysis of the regions around the TSS for the gene-clusters to identify DNA elements that may influence the transcription pattern of the genes.
5. To develop a new and efficient promoter prediction/analysis program.