CHAPTER 1. INTRODUCTION

A reliable tool for prediction of promoters has not been achieved [3 - 5]. Deriving a set of co-expressed genes might be critical for promoter characterizations as well as biomarker identification. However, obtaining such gene clusters has also not been straightforward or reliable. Some of the specific obstacles in these regards are listed below.

1. High diversity in the function and nature of promoters combined with lack of experimental characterization of promoters in most of the cases [3, 4, 6, 7].
2. High probability of random occurrence of the short cis-regulatory regions in genomic sequence [6].

3. Inaccurate determination of the transcription start site(s) leading to confusion in promoter sequence identification by several kilo base pairs [7].

4. The already existing databases used microarray data and did not permit an easy collection of co-expressed genes [Acharya et al, unpublished observations].

5. Exclusion of a large amount of microarray data from the already existing databases [8].

There have been ambitious approaches of trying to discriminate all promoters from all non-promoter regions. Developing specific algorithms with focus on one set of similarly functioning promoters at a time could help. Such promoter-clusters, i.e., promoters regulating expression in a similar set of location and conditions, are likely to have common features [9 - 11]

Clusters of promoters with potential functional similarities can be derived by first obtaining groups of co-expressed genes. However, the identification of such co-expressed genes was not straightforward. While mass scale gene expression studies, particularly those using microarray technology [12 - 14], have generated a huge amount of data, the reliability of such information has been questionable due to lack of reproducibility across studies [15, 16].

The current work is an attempt of making a better use of already existing gene expression data to solve the problems in predicting promoters, their functional features and gene expression patterns. The approach involved handling huge amounts of data as well as multiple software development and analysis. Hence, it was necessary to focus on one tissue at a time. Particular attention has been paid to the mammalian testis tissue and related diseases. This is because the testis is characterized by unique functions and drastic changes in the expression of genes to enable those functions that include a steady but controlled mitosis, meiosis, hormone production and responses, and spermiogenesis [17]. The goal was to use these testis specific gene-clusters to analyze the associated promoters, in comparison with other clusters, and then develop a better computational tool for promoter analysis and prediction. The intention was also to use the gene-clusters to shortlist candidate biomarkers for testicular conditions, especially for the non-obstructive azoospermia (NOA). NOA is a male infertility condition arising due to abnormalities in spermatogenesis in the testis tissue. Many molecular studies have been reported on this disease in the recent years [18]. However, the molecular mechanisms need to be explored further and new candidate marker molecules need to be identified for NOA.
A novel approach was developed with two main steps: a) developing databases with tissue-specific gene expression data from genome wide expression studies; b) using the newly developed meta-analysis method, which is incorporated into the databases, to derive a consensus for each gene’s expression pattern based on the extent of agreement and contradiction across studies. It was postulated that these methods lead to the identification of ‘genes with a higher reproducibility of their expression status’ - in each tissue and condition, and this information could eventually be used to ‘derive reliable sets of genes expressed or dormant’ - for specific tissues and conditions. The other related hypothesis was that the promoters associated with these refined and dependable clusters of co-expressed genes would possess characteristic sequence and other features. Exploring the above listed hypotheses would help to predict functional promoter features, expression patterns of genes as well as promoters in general. Such well-defined gene clusters were also expected to assist in a better identification of candidate biomarkers in various contexts.
CHAPTER 2. DEVELOPMENT OF TISSUE SPECIFIC MAMMALIAN GENE EXPRESSION DATABASE USING A NOVEL META-ANALYSIS METHOD

CONTENTS AND STRUCTURE

INTRODUCTION.

SECTION A. DEVELOPMENT OF A WEB-BASED TOOL TO ASSIST BIOCURATION.

SECTION B. DERIVATION AND APPLICATION OF NOVEL META-ANALYSIS METHOD.

SECTION C. DEVELOPMENT OF NEW MAMMALIAN GENE EXPRESSION DATABASES.

Observations of variations in microarray results across studies have caused uncertainty in researchers’ minds about the utility of the gene expression data obtained from array-based technologies. Novel and innovative approaches are needed to ensure that the huge amount of microarray gene expression data is not wasted or mislead researchers. Development of such a new computation method is presented in this chapter. The objective was to make maximum use of the available gene expression data for reliable prediction of gene expression patterns based on a consensus of the observations made for each gene, and each tissue and location of interest, from multiple reports. The work is presented here in three sections.

Section A. Development of the web-based system to assist biocuration.

Introduction: The awareness about the significance of biocuration has been on the rise in the recent years [19]. Since the initial observations indicated that existing gene expression databases miss a significant volume of data (also see 8, 20), it was necessary to screen published reports of such data and retrieve the needed information by biocuration.

It is known that biocuration is a painstaking and slow procedure. The process is also prone to errors and requires thorough evaluations at different stages. A computer interface can be developed to reduce the burden of biocurators, and that is what has been
done. An interface was developed along with necessary programs to assist in the various steps of the curation process.

**Method:** A web-based system was developed using PERL-CGI. Scripts were written to enable registering of the biocurators and for them to later interact with the back end programs to upload, review or edit the information. Software was created to facilitate 'creation of gene lists' using the gene expression data from public repositories, to improve the speed of biocuration. The microarray raw data could be downloaded from the GEO and ArrayExpress with partial manual interference. A program was created to process the data downloaded data using bioconductor packages such as **affy, codelink, lumi, limma and marray**. Scripts were written to facilitate the access and use of the processed data by biocurators, for creation of gene lists.

Two user-friendly forms were created as part of the system: one for uploading gene list and its associated information to the system, and the other for editing and/or validating the data uploaded in the system. Both upload and edit forms were designed to minimize typographical errors made during data uploading process, with the help of customized javascript functions. Options were made for the editing of a gene list for up to ten times. Codes were written to use the information uploaded in excel files and store in a computer readable format.

**Result:** The biocuration system was successfully created with 3 major components: 1) A tool to facilitate creation of gene lists from the gene expression data available in public repositories such as GEO, ArrayExpress. 2) A form to upload gene list and its associated information to the system. 3) A form to edit/validate the uploaded data in the system.

The interfaces, the storage and retrieval functions, data processing and display of information were verified and found to work satisfactorily. The system was then opened to biocurators. It has been used by 40 researchers over at least 2 years. It was used to upload 2077 gene lists related to various testicular conditions in human, mouse and rat, 2265 gene lists related to various uterine conditions in human, mouse and rat and 832 gene lists related to 27 healthy human tissues. The feedback of biocurators indicated the system reduced the errors and saved a good amount of time consumed for various steps.

**Section B. Development of a novel meta-analysis method and its application to the biocurated data.**

**Introduction:** Meta-analysis provided hope to the researchers around the world for not only addressing the common issues with microarray technology, but also promising more
reliable and generalized results. Many microarray meta-analysis methods had been developed [21 - 23], and they used various statistical approaches.

Most of the existing microarray meta-analysis methods can be applied only to a fraction of publicly available gene expression data and they usually do not attempt to derive a consensus expression status across the studies, particularly in terms of the binary absolute calls, viz., transcribed vs. not-detected - for each gene per tissue and condition of interest. These limitations prompted us to develop a new meta-analysis method. The new method is a vote-counting approach that derives a consensus expression pattern across most of the publicly available gene expression data and provides a quantitative measure (“reliability score”) for the reliability of consensus derived. The method was implemented on the gene lists that were uploaded via the newly developed biocuration system (described in the previous chapter).

**Method:** Scripts were written in the PERL language to implement the following steps of meta-analysis:

1) To parse the uploaded data in computer readable format to extract gene list(s) and the associated information.
2) To convert the homogenous or heterogeneous lists of gene identifiers into lists of NCBI Entrez gene identifiers using the GIDCon tool.
3) To assign reliability scores for each gene based on the gene-list-associated information. If a gene is reported to be expressed as per a gene list, then a score of 2 would be assigned, and -2 when a gene is reported to be 'not detected'. Eventually, these scores will be added up across gene lists into a final 'reliability score' for each gene per expression status for each location and condition (ESLC).
4) To store the result of the meta-analysis in specific MySQL table. A MySQL table was dedicated for every combination of species, condition and location.

The above mentioned four steps would be repeated for all gene lists. Each time when a new gene list is encountered, the reliability scores within the corresponding MySQL tables would be updated. This way, as and when new data sets are used, every MySQL table stays updated with consensus expression patterns and reliability scores.

**Result:** Each component of the meta analysis algorithm was tested manually and/or semi-automatically. Functions such as parsing the uploaded data, extracting gene list along with associated information, converting gene ids, and scoring as well as storing of the results were processed as expected on multiple types of data sets and platforms. The
programs in the system were successfully implemented on all uploaded gene lists corresponding to various testicular and uterine conditions and healthy human tissues.

The novel meta-analysis method could make maximum use of available gene expression data, due to its ability to handle a) data from multiple gene expression technologies such as microarray, RT-PCR and qRT-PCR, and b) data obtained from multiple microarray platforms (such as Affymetrix, Agilent, Codelink and Illumina) using different processing methods. Hence, it could be used upon independent data sets or as part of any gene expression database.

Section C. Development of new mammalian gene expression databases.

Introduction: Gene expression databases are immensely useful for research in various domains within life sciences. They provide easy access to existing gene expression data, and often assist in data analysis. Their use can lead to new hypothesis generation and discoveries in different aspects of molecular biology, and even aid in biomarker discovery. There were two major reasons that prompted development of the new mammalian gene expression databases: (1) limitations in existing gene expression databases such as the non-inclusion of published gene expression data and less user-friendly query features; (2) none of the published meta-analysis methods have been applied on large scale of publicly available gene expression data.

Limited coverage of gene expression data sets in databases is due to the lack of initiative from reporting authors to upload the data into databases. The only way to overcome this problem is to manually collect, verify and enter into the database. However, this cannot be done for all tissues and all species by a small team. Hence, there was a need to focus on one tissue at a time.

Mammalian testis was chosen as the tissue to focus on in the first round. Among adult tissues, testis shows unique cell differentiation and substantially more widespread transcription of the genome [24] and is an ideal tissue to study gene expression patterns and the mechanisms of such transcriptional regulation. Another reason to choose this tissue was the fact that testicular gene expression data was available following the biocuration, which in turn was performed using the newly developed computational system explained before (Chapter 2, section 1).

Method: The Mammalian Gene Expression Database for Testis tissue (MGEx-Tdb) was built on MySQL RDBMS, while its graphical user interface (GUI) was designed using
PERL-CGI scripts. The results of the meta-analysis of biocurated gene expression data formed the primary content. In addition to gene expression information derived by the new meta analysis method following information were downloaded provided as accessory information: gene related information from NCBI Entrez GENE, transcript related information from NCBI Refseq and protein related information from UniProt.

It was postulated that, even though it was time consuming to create such a database, it should be easier and faster to create new databases for other tissues - once all the scripts were in place. In fact, two other databases were quickly developed later: Mammalian Gene Expression Databases for the uterus (MGEx-Udb) & endometrial receptivity (HGEx-ERdb).

**Result:** MGEx-Tdb was developed and found to be fully operational with correct execution of all features. The database allows users to query in two ways,

1) Querying specific testicular conditions: this feature lists genes that are expressed and not detected in the condition of interest in the order of their reliability score. Each gene listed is hyperlinked to web pages providing additional information about the gene,

2) Querying specific gene: if the database is scanned with gene name, then the user can obtain the expression information of the gene in different testicular conditions and other gene related information.

The novel meta-analysis method along with MGEx-Tdb was published as a methodology article in BMC Genomic on August 2010 [1]. As on 15th march 2015, the database had 388 visits from researchers within India and 148 visits from international researchers. It has also been cited by 8 published research articles.

Databases were also successfully developed for a few other mammalian tissues later. MGEx-Udb, & a very specialized HGEx-ERdb, were published as research articles in PLOS ONE [2].

All the newly developed tissue-specific databases are freely available at the following URLs:

1. Testis (MGEx-Tdb): http://resource.ibab.ac.in/MGEx-Tdb/
2. Uterus (MGEx-Udb): http://resource.ibab.ac.in/MGEx-Udb/
3. Endometrial receptivity (HGEx-ERdb): http://resource.ibab.ac.in/MGEx-ERdb/

**Conclusion:**
1. A new web-based biocuration system was developed successfully. This system facilitated faster uploading of biocurated gene expression data and reduced errors significantly.

2. A new meta-analysis method was developed to assign a 'reliability score' for each gene for each expression status per location and condition (ESLC), based on the extent of consistency across existing gene expression data sets.

3. The meta-analysis approach was implemented on the biocurated gene expression data related to various testicular conditions. This was also used for a few other mammalian tissues later.

4. The output of the new meta-analysis method applied to the biocurated gene expression data were used to create gene expression databases for testis and other tissues.

5. Overall, the new approach of assisting biocurators via a computational system and creating databases with compiled gene expression data, such that a final consensus could be derived for each gene for different ESLCs - using the new meta-analysis method, is expected to help researchers to make a maximum use of publicly available gene expression data.
CHAPTER 3. VALIDATION OF THE NOVEL META-ANALYSIS METHOD.

The novel meta-analysis approach developed, and described in chapter 2, has a huge potential in terms of applications. It can be applied to gene expression data related to any species, tissue or condition - irrespective of the technology. It particularly has the potential to alleviate the problems arising from the variations in microarray experiments in terms of platforms, statistical methods of analysis, etc. However, a thorough validation of any new computational approach is indispensable. In this regard, both in silico and experimental validations of the new meta-analysis approach were performed. The work is presented in following five sections.
Section A. *In-silico* validation of the novel meta-analysis method: Comparison of the similarly expressed genes, identified using the novel meta-analysis method, with published sets of coexpressed genes.

**Introduction:** Last decade has witnessed several attempts from biomedical researchers to identify a reliable set of co-expressed genes using microarray data, especially housekeeping genes and tissue specific genes [25-33]. However, there seem to be significant differences in such lists. For example, even the most frequently generated set of ubiquitously transcribed genes has been varying across reports. The hypothesis behind the current work was that the variations in the reports stem from the use of a smaller number of data sets in each attempt made earlier. Efforts to make a better use of most of the microarray data sets, via the novel meta-analysis, should yield a more reliable clusters, and, if true, this type of newly derived co-expressed gene-clusters should have the highest agreement with other reported clusters of ubiquitously transcribed genes or tissue-specific gene-sets. These postulations have been tested in the current work, particularly by comparing the ubiquitous gene cluster and testis specific gene cluster obtained by the new method with other published lists.

**Method:** Derivation of ubiquitous and testis specific genes: In addition to gene lists related to testicular conditions, gene lists related to the normal condition of 27 healthy human tissues, reported from multiple studies each with several samples were used. These gene lists were prepared from data sets uploaded via the biocurator system and by applying the meta-analysis method. Using the cumulative reliability scores across the tissues - for the 'transcribed' and 'not detected' calls in healthy tissues for all protein coding genes, 2 types of clusters of ubiquitously transcribed genes were derived, a) genes which are expressed in all 27 healthy tissues [UBIQ2] and b) genes which are transcribed in at least 14 healthy tissues (50% cut-off) [UBIQ1]. In the same way, genes which are transcribed in testis tissue and not detected in all other tissues were selected as testis specific. All such gene clusters had a hierarchy of cumulative reliability scores.

Collecting published lists of ubiquitous and testis specific genes: A thorough literature search identified 9 research articles that reported lists of genes that were transcribed tissue specifically and/or those transcribed ubiquitously, based on DNA microarray studies.

Comparison of published gene lists with the gene lists derived in-house: Each of the nine published ubiquitous gene lists was compared with every other list, and with two newly
derived ubiquitous gene lists - using the match function within the Microsoft excel. Among the nine studies only three reported testis specific genes, and these published testis specific gene lists were also compared with each other as well as with the newly derived testis specific cluster.

Result: Ubiquitous and testis specific genes were derived using the new meta-analysis approach: Analysis of all the collected data identified 705 genes that were testis specific (i.e., transcribed in testis as well as not-detected in the other tissues). Similarly, 7567 genes were found to be transcribed with high scores in at least 50% of healthy tissues considered (UBIQ1) and 4056 genes in all 27 healthy tissues considered (UBIQ2).

Agreement of the in-house derived gene list with published gene lists: On comparison, it was found that UBIQ1 cluster included, on an average, about 80% of ubiquitous genes already reported by one or more of the earlier reported by nine studies. However, the highest agreement, among other published clusters of ubiquitously transcribed genes, was only 60%. Even the stringently selected UBIQ2 cluster included about 63% of reported genes from previously clusters, on an average.

The newly derived testis specific cluster of genes also showed the highest coverage (41%) among all comparisons. The results clearly indicate that the co-expressed gene clusters derived using the new meta-analysis method show a better reproducibility than the previously reported clusters. The newly derived clusters of co-expressed genes did contain new genes and these might be valuable new potential markers. Even the repeated findings of genes in the ubiquitous or tissue specific lists will be of higher value than before due to the additional information on the reliability scores per tissue for the transcribed vs. not detected calls.

Section B. In-silico validation of the novel meta-analysis method: Functional analysis of the derived co-expressed genes.

Introduction: Functional analysis of gene-set is routinely done to identify shared functions among co-expressed genes. This type of analysis helps to know about the key functions carried out by the set of genes. Such functions might be the reason for their shared expression patterns. Shared functions are suggested to indicate the generic nature of the genes or their specificity for a tissue/system [34]. Hence, a functional analysis was carried out on the new ubiquitous and tissue specific clusters of genes identified using the new meta analysis method, to challenge its utility.
**Method:** The top most scoring 500 genes were selected from both testis specific and ubiquitous clusters for the functional analysis. DAVID v6.7 (Database for Annotation, Visualization and Integrated Discovery) [35], an online tool for such analysis, was used to obtain enriched biological processes (GOTERM_BP_ALL) for testis specific and ubiquitous sets. Default cut-off criteria of p-value <0.05 and minimum gene count of 2 were chosen for the analysis.

**Result:** When the top 500 ubiquitous genes were uploaded to the tool, 220 biological processes found to be enriched with a p-value <0.05 and gene count >2. The top 20 enriched biological processes were related to ribosomal biosynthesis, protein metabolism, gene expression and cellular metabolism indicating that the cluster represented general cell functions expected to be carried out by genes transcribed ubiquitously.

Similarly, 49 biological processes were found to be enriched in top 500 testis specific genes with a p-value <0.05 and gene count >2. The top 20 enriched processes were related to germ cell development, gamete formation, fertilization, mitosis and sperm motility. These functions were also expected to be the ones where the testis specific genes would have a role. Thus, the results of analysis of both clusters supported that the gene clusters derived by the new meta analysis method are reliable.

**Section C. Experimental validation of the meta-analysis method:**
**Testing non obstructive azoospermia associated genes using reverse transcription PCR and real time PCR.**

**Introduction:** Although *in silico* validation approaches revealed the significance of co-expressed gene cluster obtained from the novel meta-analysis method, it is also important to test the predictions experimentally. Experimental testing of new samples is especially important from the perspective of the biomarker discovery potential of the new meta-analysis application. Experimental methods such as PCR techniques are still considered the gold standard for validating microarray results [36]. Hence, RT-PCR and qRT-PCR experiments were performed to test the reproducibility of genes predicted to be associated with the “Non obstructive Azoospermia” (NOA), due to differential expression in this disease condition when compared to normal condition.

**Method:**  Gene selection for experimental validation: Using MGEx-Tdb database, 12 genes were selected as potential markers. Eight genes were indicated to be transcribed in the testis samples from the donors with NOA disease and assigned a high reliability score, and they were similarly indicated to be 'not detected', again with high score, in
normal testis. Another four genes were assigned high reliability score by the database for being transcribed in normal testis and not-detected in testes of NOA patients.

Collection of clinical samples and RNA isolation: In compliance with the Institutional Biosafety Committee (IBSC) at the Institute of Bioinformatics & Applied Biotechnology (IBAB) and the ethical procedures of the Ankur hospital approved the procedure for collection of testicular biopsies from donors. Testis samples were collected from seven non obstructive azoospermic donors and were stored in RNA later solution. However, RNA for normal testis was purchased from Clontech. Electrophoretic and spectrophotometric analyses were carried to confirm the quality of the RNA isolated.

RT-PCR: cDNA synthesis was carried out using the total RNA as template. Primers to specific exon of each gene, were used for PCR. Products of RT-PCR were analyzed by image analysis.

Real time PCR experiment: SYBR green based RT-PCR kit was used for real time PCR experiments. The relative expression levels of genes in NOA and normal condition were estimated via the comparative Ct method [37].

Results: RT-PCR: Out of the eight genes tested for the transcribed status in NOA and not detected status in normal testis, none showed agreement for expression status in normal testis. Only one gene was found to be transcribed in all seven NOA samples. The other genes showed an agreement for expression status in 2 to 6 of NOA samples.

Motivation for further validation: This result of RT-PCR prompted us to continue the experimental validation of four genes (transcribed in normal testis and not detected in NOA) because RT-PCR results showed that in the majority of NOA samples the genes are transcribed. Thus, real time PCR experiments were later used to analyze the level of transcription of these four genes in NOA and normal conditions.

Real time PCR experiment: The 4 genes were found to be down-regulated in 2 to 4 of the 7 NOA samples compared with the normal testis.

Many genes did not show the expression trend predicted by the database, via the new meta-analysis method, in a significant number of cases. Given the lower sensitivity of the
microarray technique, it is not very surprising to note the contradictions in the cases of not detected status of genes. However, the other results were unexpected. Given the differences in the ethnicity of the sample donors (mostly Caucasian) of the studies used to predict the initial expression status vs. those used for validation (Indian), the number of genes considered for testing may be less. The predictions should perhaps be tested with a much higher number of genes.

**Section D. Experimental validation of the meta-analysis approach: Validation of MGEx-Tdb using RNA sequencing data of NOA and normal samples.**

**Introduction:** Results from validation of MGEx-Tdb using PCR experiments were inconclusive due to the low number of genes used. Hence, validation with another high throughput gene expression technology was needed. RNA sequencing, which is supposed to be more reliable compared to microarray technology, is used for the second round of experimental challenge of the new meta-analysis method. The clinical samples from NOA patients that we had access to and commercial testis RNA were subjected to deep sequencing. Comparison of RNA-seq gene expression data with results of meta-analysis can not only test the predictions by the new meta-analysis method, but also provide an account of the extent of correlations between the results from the two technologies.

**Method:** Good quality RNA from four Non-Obstructive Azoospermia (NOA) samples and one normal sample were used for RNA-seq. Paired end sequencing data of read length 72 base pairs corresponding to these samples were obtained from the Illumina Genome analyzer sequencing platform.

Data processing: Quality assessment was performed on the raw data using FASTQC tool. The sequencing reads were then aligned to the human reference transcriptome and genome Hg19. The aligned reads were assembled and quantified and FPKM values at gene level were obtained.

Comparison with MGEx-Tdb: Genes with FPKM value > 1 were considered to be transcribed. Expression status of each gene in NOA and normal testis conditions from MGEx-Tdb was compared with RNA-seq data. This provided an overall agreement between MGEx-Tdb and RNA-seq.
To evaluate the significance of reliability scores derived, the list of genes transcribed in NOA according to MGEx-Tdb was divided into ten blocks based on the reliability scores of genes. For example, the top 10% scorers formed the first block and the next 10% formed the second block. The agreement of expression status of all genes in each block was compared with that from RNA-seq data. Similar comparisons were made on genes transcribed in normal testis, genes not detected in NOA and genes not detected in normal testis.

**Result:** Overall agreement of RNA-seq data with MGEx-Tdb: When the transcribed genes from MGEx-Tdb were compared with those from RNA-seq, 84.39% agreement was found for normal testis condition, while 72.75% agreement was found for NOA condition. Similarly, when not detected genes from MGEx-Tdb were compared, 66% agreement was found in case of normal testis and 76% agreement was found in case of NOA condition. This indicated that most of the genes predicted to be transcribed using multiple microarray data sets and the new meta-analysis method, are likely to be reproduced by an independent experiment - even if the technology and ethnicity of samples used are different.

**Significance of reliability score:** Block wise comparison revealed that, genes with high reliability score for expression from MGEx-Tdb are reproduced more consistently in RNA-seq than those with low reliability score. The results support the utility of the reliability scores for the expression status of genes.

**Section E. Identification of most suitable reliability scoring system for meta-analysis approach.**

**Introduction:** Both the *in silico* and experimental validations indicate that gene expression predictions from the new meta-analysis method is reliable and, hence, will be useful to the biomedical researchers. However, the results of small scale experiments indicated that the scoring system is not 100% reliable. This suggested a possible scope for improving the reliability scoring system. Hence, multiple variants of the new scoring system adopted earlier were contemplated and their performance was assessed.

**Methods:** Five major types of reliability scoring methods were designed for testing. 1) The basic scoring method, the originally implemented method (Chapter 2, Section 2) in which reliability score for each gene was based on the agreement of expression status across all the gene lists related to a specific condition, irrespective of the number of studies. 2) A variant of the basic scoring method, in which the score was increased if the
agreement was noticed across studies. 3) Hybridization-based scoring method, in which the reliability score for each gene is based on the agreement of expression status across the gene lists, the number of hybridizations considered to derive the cumulative/representative gene lists as well as the studies. 4) Sample size based scoring method, in which the reliability score for each gene is based on the agreement of expression status across the gene lists, number of samples considered for the experiment and agreement across the studies related to a specific condition. 5) A combination of the sample size and hybridization based scoring methods. Under each type of reliability scoring method, different variants were introduced based on amount of score or 'weight' given for different parameters. A total of 450 scoring system variants were tested on 6 sets of gene lists of different types and condition. Each set contained 8 to 10 gene lists from 3 to 5 studies. The performance of each scoring system method was assessed by using a machine learning method, the “leave one out” method. PERL scripts were written to execute all the 450 scoring system on six set of gene lists and perform about a million calculations. Each scoring method was then ranked based its performance on each set of gene lists.

Result:
Hybridization and sample size based scoring topped when gene lists having expression information from more than three hybridizations were considered. While basic scoring performed best when gene lists having expression information from three or less hybridization. When cumulative ranks were considered for all scoring systems, basic scoring method remained best. Thus, basic scoring method is retained.

Conclusion:
1. The new meta-analysis method was applied on previously reported microarray gene expression data from healthy human tissues and two important new clusters were derived: genes transcribed ubiquitously and testis specifically.
2. Ubiquitous genes and testis specific clusters derived de novo showed higher agreement with comparable clusters reported earlier by individual experimentation or using meta-analysis of multiple data sets. This observation fortified the superior quality of the newly developed meta-analysis method.
3. The functional analysis of genes in the two de novo clusters also supported the reliability of the clusters.
4. Experimental validation of non obstructive azoospermia through RT-PCR, however, showed significant percentages of contradictions, particularly for the ‘not-detected’ status. Nevertheless, RNA-seq experiments showed that most of the genes do follow the expression pattern predicted by the new meta analysis method.
In fact, higher proportion of high-scorers revealed the expected transcription status compared to low-scorers. Thus, the experimental observations also supported the reliability of the predictions by the meta analysis method.

5. Based on RT-PCR experiments, FETUB and CYP2C9 were identified as potential biomarkers for NOA condition as they were consistently found to be expressed in normal condition and not detected in NOA condition, even in seven of the Indian donors.

6. Attempt to identify a better scoring method out of 450 new scoring method failed; the original scoring method was found to be perform the best, again validating the results obtained using the meta analysis method.
CHAPTER 4. DEVELOPMENT OF PROMOTER ANALYSIS TOOLS, ANALYSIS OF PROMOTER SETS AND DEVELOPMENT OF A MAMMALIAN PROMOTER ANALYSIS PLATFORM.

CONTENTS AND STRUCTURE

INTRODUCTION.

SECTION A. DEVELOPMENT OF MOTDET FOR POSITION-SPECIFIC DETECTION AND ANALYSIS OF SHORT DNA MOTIFS.

SECTION B. DEVELOPMENT OF GREAM TO SHORT LIST REPEAT ELEMENTS THAT MAY INFLUENCE GENE CO-EXPRESSION.

SECTION C. TRANSCRIPTION FACTOR BINDING SITE (TFBS) ANALYSIS OF PROMOTERS OF CO-EXPRESSED GENES.

SECTION D. ANALYZING DISTRIBUTION OF CIS-ELEMENTS IN THE PROMOTER REGION OF CO-EXPRESSED GENES.

SECTION E. ANALYSIS OF THE NEIGHBORHOOD OF CO-EXPRESSED GENES FROM NOVEL MICROARRAY META-ANALYSIS: CpG ISLAND AND REPEAT ELEMENT ANALYSIS OF CO-EXPRESSED GENE PROMOTERS.

SECTION F. DEVELOPMENT OF AN INTEGRATED PROMOTER PREDICTION/ANALYSIS PLATFORM.

Differential regulation of genes has attracted the attention of many scientists for obvious reasons. While many aspects of molecular mechanism of the gene expression control are known, a lot remains unknown. Identification of different reliable clusters of co-expressed genes presented an exciting opportunity to understand the mechanism behind their differential regulation. As part of the work presented in this chapter, testis specific and ubiquitous genes have been subjected to a thorough promoter analysis to identify the regulatory elements associated with them, using existing tools and the newly developed
tools. Promoter features such as transcription factor binding sites, CpG islands and repeat elements have been explored in the current study. A more useful computational platform was developed after integrated multiple promoter prediction/analysis programs. The work is described in the below mentioned sections.

**Section A. Development of MotDet, a tool for position-specific detection of short DNA motifs.**

**Introduction:** Identifying functional DNA motifs within the promoter sequences is often an essential first step in efforts to understand gene expression regulation. However, identification of potential functional DNA motifs is complicated because they are short (5-20 nucleotides long) [38] and could have high degeneracy [39]. Thus, most of the computation tools end up with high false positives. Motif features such as position specificity [40] and variant preference [41 - 43] were reported to be signatures of functional motifs. Thus by analyzing both of these features in tandem, functional motifs can be identified, this idea prompted us to develop new motif detection tool.

**Method:** PERL/CGI scripts were written and the program-modules were deployed on a Linux server located at IBAB. Assuming normal distribution for the over-representation of the DNA motifs in the string of sequences, Z-score statistic [44] is used to list significantly enriched motifs in the query sequence.

**Result:** MotDet is successfully developed and tested for the identification of position specific motif variant that may influence gene co-expression. It can accept set of promoter sequences and one or motifs in IUPAC form as input. It divides the promoter sequences into blocks and scans them for the existence of the motif(s) and their variants. It lists motif variants that are position specific and statistically over-represented within the promoter region. The tool can also list co-occurring motif variants across given set of promoter sequences as well.

**Section B. Development of GREAM, a tool to short list repeat elements that may influence gene co-expression.**

**Introduction:** Some of the repeat elements have been reported to play a role in gene expression regulation [45 - 49]. However, although more than 60 tools exist for repeat
element analysis [50], none of them facilitates analyzing repeat element distribution in the neighborhood region of a gene-set and identifying over-represented repeat elements. Hence, a new computational tool has been developed to enable such analysis: The 'Genomic Repeat Element Analyzer for Mammals' (GREAM).

**Method:** GREAM is a web-based tool developed using PERL-CGI. Information about genomic repeat elements (such as repeat annotation track for mammalian genomes, repeat element sequences and class and family information for repeat elements) was obtained from NCBI Entrez GENE, RepBase. The graphical user interface (GUI) was developed using user PERL-CGI scripts; Web technologies such as cascade style sheets (CSS) and JavaScript (JSS) were utilized for user-friendly features. PHP was used to create pictorial representation of repetitive element distribution in the neighborhood region of all query genes.

**Result:** GREAM is successfully developed and tested. It is currently hosted on a server at IBAB. GREAM accepts set of co-expressed/co-regulated genes as input. It scans the user defined neighborhood region of all submitted genes for existence of repeat elements. Repeat annotation track downloaded from NCBI is used for repeat element detection. For each repeat element identified, observed frequency is estimated, compared with expected frequency (pre-computed population statistics) and using binomial probability significance of over-representation is estimated.

Finally, the tool lists repeat elements that are over-represented within gene neighborhood region of input genes. GREAM facilitates repeat analysis on gene-sets from 17 mammalian species. It can also identify repeat elements enriched in orthologous gene-sets.

**Section C. Promoter analysis of co-expressed genes derived using the new meta-analysis method: Transcription factor binding site analysis of co-expressed genes.**

**Introduction:** Transcription initiation is the first and most important checkpoint of gene expression regulation. Binding of transcription factors at regions near to transcription start site can either promote/repress the transcription initiation [51]. Genes which are co-expressed or co-regulated are known to be controlled by same set of transcription factors. Thus, investigation of the promoters of co-expressed genes for shared transcription factor binding sites (TFBS) is an important step towards understanding mechanism of gene
expression regulation. TFBS analysis was performed on the promoters of most reliable testis specific and ubiquitous genes, derived using the newly developed meta-analysis method.

**Method:** TFBS analysis was performed using Match tool [52], which scans vertebrate positional weight matrices from TRANSFAC [53]. Promoter sequences of length 1600 (-1100 to +500) were obtained from DBTSS [54] for top 25 testis specific genes. To identify over-represented TFBS in the promoters of testis specific genes, TFBS distribution within testis promoters is compared with TFBS distribution within promoters of three random set of genes of same size (25 genes).

TFBSs that are significantly enriched (p-value <0.05 and matched promoter p-value<0.1) in the testis specific promoters by 1.3 fold compared to all three random sets were identified. To know the functional importance of these enriched TFBS, expression status of their respective TFs in normal testis condition was analyzed. Similar exercise was carried out to identify TFBS associated with ubiquitous gene-sets.

**Result:** Based on the frequency of occurrence and number of promoters having them, 58 TFBS were identified to be over-represented in testis specific promoters. Among them 39 would be bound by transcription factors which are expressed in normal testis according to MGEx-Tdb. Similarly, 32 TFBS were identified over-represented in ubiquitous promoters. Among the 29 would be bound by transcription factors, which are expressed in most of the healthy tissues.

Thus TFBSs, which could be responsible to drive the testis specific co-expression and constitutive expression, have been short-listed.

**Section D. Promoter analysis of co-expressed genes from novel microarray meta-analysis: Analyzing distribution of cis-elements in the promoter region of co-expressed genes.**

**Introduction:** One of the most important features of functional DNA motif is position specificity with respect to transcription start site [55]. As mentioned in previous section, TFBS analysis of promoters of co-expressed genes led to identification of that are over-represented. Analyzing position specific nature of these TFBS, could aid in recognizing
their functional importance. The newly developed tool MotDet facilitates this type of analysis.

Core promoter is also an important component of gene expression regulation [56]. DNA motifs found in this region are often referred to as core promoter elements (CPE). Testis specific and ubiquitous promoters were also scanned for the existence of well known CPE such as TATA box, Initiator (INR), TFIIB recognition element (BRE), downstream promoter element (DPE), Motif ten element (MTE) and XCPE1.

**Method:** Promoter sequences of testis specific and ubiquitous genes were obtained from DBTSS.

The analysis was carried out using position weight matrices (PWM) of over-represented TFBS from TRANSFAC and promoter sequences from DBTSS. The promoter sequences and PWMs were uploaded to MotDet. Minimum threshold frequency (percentage of sequences having specific motif) was set to 10, block size to be considered for analysis was set to 250.

The consensus sequence for TATA, INR, BRE, DPE, MTE and XCPE1 were obtained from published literature. The in-house tool, MotDet was used to identify CPEs within core promoter region (-100 to +100) of the promoter sequences.

**Result:** Out of 24 TFBS over-represented within the promoters of ubiquitous genes considered, three (V$AHRHIF_Q6, V$E2F1_Q3 and V$NMYC_01) showed position specificity. Similarly, out of 32 TFBS over-represented within the promoters of testis specific genes considered, seven (V$TBP_Q6, V$LMO2COM_02, V$GR_Q6, V$DBP_Q6, V$CIZ_01, V$CEBPA_01 and V$CEBP_Q2) were position specific.

INR and DPE were found to be equally associated with promoters of both testis specific and ubiquitous genes. While, BRE was found at higher frequency in promoters of ubiquitous genes (60%) compared to promoters of testis specific genes (32%).

Section E. Analysis of the neighborhood of co-expressed genes from novel microarray meta-analysis: CpG island and repeat element analysis of co-expressed gene promoters.

**Introduction:** Genomic region of about 1kb or so, having high GC content are known as CpG island. CpG islands are known to be one of the promoter features of housekeeping
and tissue specific genes [57]. Thus, promoter sequences of testis specific and ubiquitous genes were analyzed for CpG islands.

Genomic repeat elements are known to harbor TFBSs. These repeat elements are present at various locations, including the promoter and around it. It would be worth studying the repeat elements as well.

**Method:** CpG island analysis was carried out using CpGProD [58]. Promoter sequences of 25 most reliable testis specific and ubiquitous genes were obtained from DBTSS. Default parameters of CpGProD were retained for the analysis. GREAM was used to analyze repeat distribution in the neighborhood region of top-scoring testis specific and ubiquitous genes.

**Result:** The analysis revealed that all the 25 ubiquitous genes have CpG island within their promoters, while 18 out of 25 testis specific genes have CpG island in their promoters. This result is in line with the existing literature that promoters of housekeeping genes are enriched with CpG island.

Multiple repeat elements were found to be over-represented around the gene clusters. The type of repeat elements was different across the clusters. A DNA transposon, MER58B, was enriched in the neighborhood of ubiquitous genes, while LINEs such as L1M3d, L1PA5, L1M3 and L1ME3D were found to be over-represented in the neighborhood of testis specific genes.

**Section F. Development of integrated promoter prediction/analysis platform.**

**Introduction:** To understand the mechanism of gene expression regulation of a gene-set, one has to study not one but multiple promoter features. However, currently there is no tool available to permit multiple types of analysis. Hence, a platform has been created to facilitate researchers to analyze different promoter features associated with a set of genes and to suggest a probability that a specific input sequence is a promoter.

**Method:** Publicly available tools such as CLOVER for TFBS analysis using JASPAR position weight matrices, CpGProD for CpG island analysis, MEME for novel motif discovery, BLAST for sequence comparison with known promoter sequences and newly
developed MotDet and GREAM were integrated to develop the promoter analysis platform. The promoter sequence databases from DBTSS and UCSC were incorporated in the tool.

The platform was developed using PERL-CGI. Javascript, and CSS were used to implement some of the user-friendly features.

**Result:** The platform is successfully developed and tested for its performance. It can either accept gene names and extract promoter sequence from in house promoter database or receive promoter sequences as input. It allows user to analyze different promoter features through the web interface using above-mentioned tools.

**Conclusion:**

1. Two new promoter analysis tools, MotDet and GREAM were developed for identifying position specific motifs and repeat elements that may be associated with gene expression regulation.

2. By means of over-representation and position specificity, potentially important TFBSs were identified in the promoter sequences of testis specific and ubiquitous genes.

3. CpG island association with promoters of ubiquitous genes was reiterated. Nevertheless, CpG islands were also observed in a portion of testis-specific promoters.

4. No significant difference in the occurrence and distribution of core promoter elements was observed across the differentially transcribed gene clusters.

5. The relative distribution of few genomic repeat elements was also found to be differently associated across the two clusters analyzed. The possibility that they are the cause for the differential expression needs to be further explored.

6. The results indicate multiple differences in the promoter regions of differently expressed genes, particularly the TFBS distribution in the promoter region. However, more studies are warranted to confirm their suggested role in differential transcription.

7. A new integrated promoter analysis/prediction platform was developed. This now enables analysis of promoters of co-expressed genes from multiple perspectives.
The new platform also enables detection of promoters or promoter-like regions from sequences of interest from any mammalian species.