CHAPTER 7: OVERVIEW AND CONCLUSION

7.1 Summary

The work illustrates that it is possible to make a better use of the existing gene expression data compared to the already reported methods/approaches. The initial estimate that a significant amount of published gene expression data may not be available in the already existing databases was found to be correct, and in fact, none of the databases had more than 50% of the gene expression data that is published. Painstaking biocuration enabled compilation of most of the existing gene expression data for the mammalian testis tissue. A database later developed using the compiled data was compared systematically with many other gene expression databases. The results established the superiority of the biocuration approach.

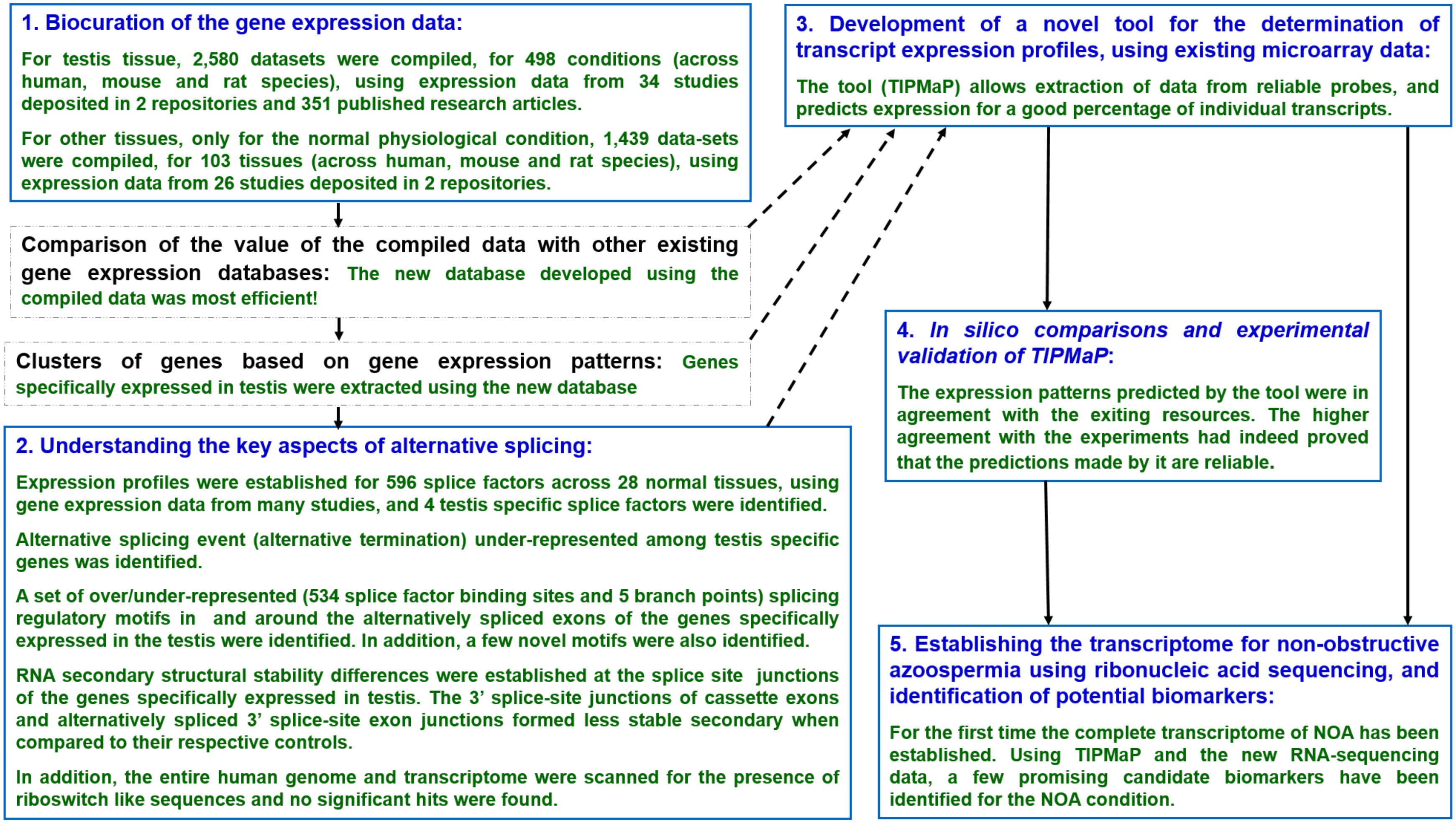
Establishing expression of the transcript isoforms in various physiological and disease conditions is important, as most of the human genes undergo alternative splicing and produce an enormous number of transcript isoforms, and these isoforms are also associated with diseases. Therefore, to determine the transcript expression profiles, a tool named 'Transcript Isoform Profiles from Microarray Probes' (TIPMaP) was developed. The tool reanalyzes the microarray data using ‘good’ quality microarray probes and determines the expression of specific transcripts. Reverse transcription polymerase chain reaction (RT-PCR) and *in silico* analyses were performed to test the entire approach of using the TIPMaP tool for transcript-expression-profiling. The results revealed a high reliability of the approach, particularly for the transcripts with higher consistency in report-status across microarray hybridizations. TIPMaP and additional sequencing of ribonucleic acid (RNA) by next generation sequencing were also used to identify potential biomarkers for a type of male infertility, non-obstructive azoospermia (NOA). Additionally, a systematic analysis of the co-expressed gene cluster (i.e., genes with similar expression pattern) provided new insights that can help in better understanding of alternative splicing regulation across differently expressed genes (figure 7.1, p. 220 illustrates the overall flow of the work and important conclusions).

7.2 Biocuration of the gene expression data

Laborious manual screening of research articles and gene expression databases helped in the compilation of a large volume of data. A total of 10,879 hits from PubMed and 940 from standard gene expression databases were carefully screened, and the information was processed to obtain 2,580 gene expression datasets. The gathered information corresponded to 498 testicular physiological conditions across human, mouse, and rat species.

This data was used by other researchers in the group (Institute of Bioinformatics and Applied Biotechnology (IBAB)) to create a mammalian gene expression database for testis tissue (MGEx-Tdb, http://resource.ibab.ac.in/MGEx-Tdb/). The new database and the existing gene expression databases were compared systematically. This comparison was first of its kind and was much needed to assist biologists in making objective decisions while using gene expression profiles. The new database made using gene expression data was found to be superior to others in more than one way. Thus, biocuration of the tissue-specific data formed an extremely useful first step in making the best use of the existing gene expression data. The analysis of the compiled data yielded many useful statistics and provided useful information about the distribution of the types of available gene expression data, particularly for the mammalian testis. These results have already been published in a peer-reviewed journal [1], which has also been cited seven times by other researchers. MGEx-Tdb has been accessed globally more than 1300 times.

Figure 7.1: Schematic representation of the overall flow of work and conclusions



1. Biocuration of the gene expression data:

For testis tissue, 2,580 datasets were compiled, for 498 conditions (across human, mouse and rat species), using expression data from 34 studies deposited in 2 repositories and 351 published research articles.

For other tissues, only for the normal physiological condition, 1,439 data-sets were compiled, for 103 tissues (across human, mouse and rat species), using expression data from 26 studies deposited in 2 repositories.

Comparison of the value of the compiled data with other existing gene expression databases: The new database developed using the compiled data was most efficient!

Clusters of genes based on gene expression patterns: Genes specifically expressed in testis were extracted using the new database

2. Understanding the key aspects of alternative splicing:

Expression profiles were established for 596 splice factors across 28 normal tissues, using gene expression data from many studies, and 4 testis specific splice factors were identified.

Alternative splicing event (alternative termination) under-represented among testis specific genes was identified.

A set of over/under-represented (534 splice factor binding sites and 5 branch points) splicing regulatory motifs in and around the alternatively spliced exons of the genes specifically expressed in the testis were identified. In addition, a few novel motifs were also identified.

RNA secondary structural stability differences were established at the splice site junctions of the genes specifically expressed in testis. The 3’ splice-site junctions of cassette exons and alternatively spliced 3’ splice-site exon junctions formed less stable secondary when compared to their respective controls.

In addition, the entire human genome and transcriptome were scanned for the presence of riboswitch like sequences and no significant hits were found.

3. Development of a novel tool for the determination of transcript expression profiles, using existing microarray data:

The tool (TIPMaP) allows extraction of data from reliable probes, and predicts expression for a good percentage of individual transcripts.

4. *In silico comparisons and experimental validation of TIPMaP*:

The expression patterns predicted by the tool were in agreement with the exiting resources. The higher agreement with the experiments had indeed proved that the predictions made by it are reliable.

5. Establishing the transcriptome for non-obstructive azoospermia using ribonucleic acid sequencing, and identification of potential biomarkers:

For the first time the complete transcriptome of NOA has been established. Using TIPMaP and the new RNA-sequencing data, a few promising candidate biomarkers have been identified for the NOA condition.

Blue and black fonts indicate the main phases and tasks, while the green fonts indicate the outcome of the work.

7.3 Understanding the key aspects of alternative splicing

Alternative splicing is influenced by various factors such as splice factors, splicing regulatory motifs, RNA structural stability at the splice-site junctions, and riboswitches. However, very little is known about such aspects in the context of regulation of alternative splicing in the mammalian testis tissue. In the current study, the expression profiles were established for 596 splice factors across 28 normal tissues, using expression data from many studies. The analysis has shown that most of the splice factors were ubiquitously expressed. Four testis-specific splice factors have been identified too. A set of genes specifically transcribed in testis (GSTT), were derived using MGEx-Tdb and analyzed from the perspective of aspects that may influence the type of alternative splicing. Following insights were obtained:

a. The relative frequency of most of the alternative splicing events in GSTT was similar to that observed among other genes (non-GSTT genes, i.e., genes other than GSTT). However, the alternative termination event, which is low in occurrence, seemed to be even lesser (p-value < 0.005 or 0.02) in GSTT.

b. Several splice factor binding sites and branch points were found to be over/under-represented in and around the alternatively spliced regions of GSTT when compared to constitutively spliced regions and alternatively spliced exons from non-GSTT genes. In addition, a few novel motifs have been identified by Multiple Expectation maximization (Em) for Motif Elicitation (MEME) analysis. These motifs might be the ones that regulate alternative splicing in this set of genes.

c. The average free energies of the alternatively spliced 3’ splice site exon junctions and 3’ splice site junctions of cassette exons of GSTT were significantly higher than the splice site junctions of the control sets. Thus they formed less stable secondary structures.

d. In addition, the entire human genome and transcriptome was screened for the presence of riboswitch-like sequences and no traces of these sequences were identified.

The results obtained here forms the first step to obtain better insights into the alternative splicing regulation in testis tissue.

7.4 Development of a novel tool for the determination of transcript expression profiles, using existing microarray data

A large number of mammalian genes undergo alternative splicing and produce more than one transcript isoform. However, most of the gene expression data do not address expression at the transcript level. It should be possible to derive transcript-specific expression profiles using the gene-level data from standard arrays. Hence, an online web server (Transcript Isoform Profiles from Microarray Probes, i.e., TIPMaP) has been developed with advanced query and display features. TIPMaP re-analyzes the expression data using good probes and determines the expression of individual transcripts. TIPMaP particularly relies upon the expression data from the standard 3’ Affymetrix gene arrays and determines expression as simple detection calls or as differential regulation calls. The tool can help in the identification of specific transcripts with reliable expression profiles for different conditions. The information provided in the results about the number of probes used to derive a call and consistency in the detection status across hybridizations can be extremely useful for data analysis. Thus, the tool can help in making a better use of available gene expression data from standard microarray platforms for a variety of conditions, including many diseases. This tool has been published in BMC Genomics [2] and is accessed 676 times so far globally, and also been cited by other researchers.

7.5 In silico comparisons and experimental validation of TIPMaP

In silico comparisons and experimental validations have been performed, to determine the reliability of the predictions made by TIPMaP. The high agreement with the existing resource, published data and with the experimental results indicate the predictions made by it are indeed reliable. Therefore, TIPMaP can be reliably used to predict transcripts that are differentially regulated or detected in a physiological condition. These findings have also been peer-reviewed [2].

7.6 Establishing the transcriptome for non-obstructive azoospermia using ribonucleic acid sequencing, and identification of potential biomarkers

NOA is a type of male infertility and not much is known about the etiology of the disorder. Though a lot of mass scale studies have been performed (using microarrays and mass spectrophotometry), gene expression profiles have not been completely established in the context of this disorder, particularly among Indian patients. The current study forms the first effort to sequence RNA and analyze the transcriptomic differences in NOA patients. Two things were done for the first time in this study: RNA-sequencing technology was applied to NOA and expression profiles were determined in the 'Indian' NOA samples. The expression profiles were established for a huge number of known and novel transcripts in the human testis in both NOA and normal conditions, and potential biomarkers have been identified for NOA condition. These genes have to be further studied to understand the molecular mechanisms of the disease. More experiments have to be particularly carried out across more NOA and normal testicular samples, to validate and further select important biomarkers. Thus, the study provides a good basic data for biomarker discovery by listing promising biomarkers, and it may also trigger studies on the molecular mechanisms at a completely new level, i.e., at the level of alternatively spliced transcript isoforms.

7.7 Future perspectives

The work presented in the thesis can be further extended in various directions:

a. Biocuration is a continuous process. As and when scientists generate gene expression data corresponding to testicular physiological conditions, the data have to be curated. Moreover, in recent times a lot of expression data have been generated by RNA sequencing and these data have to be included, for the development of a robust gene expression platform. In addition, features can be created in the databases to accept new datasets by other researchers.

b. More in *silico* analysis and experiments have to be performed to obtain better insights into the alternative splicing regulation of genes specifically transcribed in testis. The results provided in the thesis might be helpful to initiate a new analysis.

c. TIPMaP was developed to reanalyze the gene expression data from the standard 3’ Affymetrix arrays, and derive more useful information. The web server can be enhanced by incorporating additional features, using which the expression data from other microarray platforms can be analyzed.

d. RNA sequencing was done, and the transcriptome of NOA has been established, using only a few samples. To establish the complete NOA transcriptome in a reliable way, more number of samples have to be considered. Further, additional experiments have to be performed on a wider number of samples, to validate the expression pattern of the biomarkers, and for considering them as biomarkers in real time diagnosis or prognostics.

By and large, the resources developed and the results obtained in this thesis, provide a good motivation and platform for further studies and a better understanding of the regulation of genes in testis tissue in normal and other conditions, particularly in NOA.

7.8 Bibliography

1. Acharya KK, Chandrashekar DS, Chitturi N, Shah H, Malhotra V, Sreelakshmi KS, Deepti H, Bajpai A, Davuluri S, Bora P, Rao L. A novel tissue-specific meta-analysis approach for gene expression predictions, initiated with a mammalian gene expression testis database. BMC Genomics 2010, 11:467.

2. Chitturi N, Balagannavar G, Chandrashekar DS, Abinaya S, Srini VS, Acharya KK. TIPMaP: a web server to establish transcript isoform profiles from reliable microarray probes. BMC Genomics. 2013 Dec 27;14:922.