

## CONTENTS

Abstract .....	vi
List of tables .....	xii
List of figures.....	xv
List of boxes.....	xvii
List of abbreviations.....	xviii
1 OVERVIEW AND BACKGROUND	1
1.1 GENE EXPRESSION AND TECHNIQUES AVAILABLE FOR EXPRESSION PROFILING	3
1.1.1 Techniques available to determine gene expression profiles	3
1.1.1.1 Microarray	4
1.1.1.1.1 Affymetrix microarray chips	8
1.1.1.2 RNA sequencing	8
1.1.2 Prediction of gene expression profiles	11
1.2 THE MAMMALIAN TESTIS TISSUE	11
1.2.1 Notes on infertility	12
1.2.1.1 Azoospermia	13
1.2.1.1.1 Non-obstructive Azoospermia (NOA)	14
1.3 SPLICING AND ALTERNATIVE SPLICING	15
1.3.1 Alternative splicing (AS)	16
1.4 BIOMARKERS	19
1.5 BIBLIOGRAPHY	19
2 BIOCURATION OF THE GENE EXPRESSION DATA	30
2.1 INTRODUCTION	30
2.2 METHODS	33
2.2.1 Screening of research articles	33
2.2.1.1 Gene expression profiling studies related to testis tissue	33
2.2.1.2 Gene expression profiling studies related to other normal tissues	34
2.2.1.3 Curation of gene expression data	35
2.2.2 Consensus derivation	40

2.2.3	Comparison with other gene expression databases	42
2.2.3.1	Checking the reliability of the expression status across databases using manually curated data (MCD)	43
2.2.3.2	Comparison of coverage of gene-centric information, particularly their expression	43
2.2.3.3	Comparison of information availability about the expression of genes in various physiological conditions	44
2.2.3.4	Semi-quantitative comparison of information availability	45
2.3	RESULTS	46
2.3.1	Biocuration	46
2.3.2	Use of the curated data	51
2.3.3	Comparison of MGEx-Tdb with other gene expression databases	57
2.3.3.1	Checking the reliability of the expression status across databases using MCD	57
2.3.3.2	Comparison of coverage of gene-centric information, particularly their expression	58
2.3.3.3	Comparison of information availability about the expression of genes in various physiological conditions	58
2.3.3.4	Semi-quantitative comparison of information availability	59
2.4	DISCUSSION	60
2.5	BIBLIOGRAPHY	62
3	UNDERSTANDING THE KEY ASPECTS OF ALTERNATIVE SPLICING	68
3.1	INTRODUCTION	68
3.1.1	Splice factors and their expression	68
3.1.2	Splicing regulatory motifs	69
3.1.3	RNA secondary structure	69
3.1.4	Riboswitches	70
3.2	METHODS	71
3.2.1	Splice factors and their expression profiles	71
3.2.1.1	Compilation of splice factors	71
3.2.1.2	Extraction of splice factor expression profiles	71
3.2.2	Co-expressed gene cluster: genes specifically transcribed in testis (GSTT)	72
3.2.3	Mapping of AS events to genes	72

3.2.3.1	Is there a preference for an AS event in GSTT?	73
3.2.4	Identification of constitutively spliced exons and introns	73
3.2.5	Identification of over/under-represented motifs, which might be important for splicing	73
3.2.5.1	Splice factor binding site analysis	74
3.2.5.2	Branch point analysis	77
3.2.5.3	Identification of new motifs by Multiple Expectation maximization (Em) for Motif Elicitation (MEME) analysis	78
3.2.6	RNA secondary structure stability	80
3.2.7	Screening for riboswitches	81
3.3	RESULTS	82
3.3.1	Splice factors and their expression profiles	82
3.3.2	Genes specifically transcribed in testis (GSTT), a co-expressed gene cluster	84
3.3.3	Is there a preference for an AS event in GSTT?	84
3.3.4	Splicing regulatory motifs	87
3.3.4.1	Splice factor binding sites	87
3.3.4.2	Branch point sites	107
3.3.4.3	New motifs through MEME analysis	107
3.3.5	RNA secondary structural stability	111
3.3.6	Riboswitches signature sequences in human genome	113
3.4	DISCUSSION	116
3.5	BIBLIOGRAPHY	118
4	DEVELOPMENT OF A NOVEL TOOL FOR DETERMINATION OF THE TRANSCRIPT EXPRESSION PROFILES, USING EXISTING MICROARRAY DATA	124
4.1	INTRODUCTION	124
4.2	IMPLEMENTATION	125
4.2.1	Data downloading	125
4.2.2	Aligning probes to transcript sequences	126
4.2.3	Generating transcript-specific-probe-clusters	126
4.2.4	Creating new CDF files	126

4.2.5	Normalizing gene expression data	126
4.2.6	TIPMaP workflow	126
4.2.7	Comparison with existing possible alternatives	139
4.3	RESULTS AND DISCUSSION	139
4.3.1	Data compiled	140
4.3.2	BLAST results	140
4.3.3	Quality check	141
4.3.4	Resource usage – a case study	141
4.3.5	Comparison with existing computational resources	153
4.4	BIBLIOGRAPHY	157
5	IN SILICO COMPARISONS AND EXPERIMENTAL VALIDATIONS OF TIPMAP	161
5.1	INTRODUCTION	161
5.2	METHODS	161
5.2.1	Re-analyze the gene expression data using TIPMaP	161
5.2.2	In silico comparisons	162
5.2.2.1	Extracting expression information from MGEx-Tdb	162
5.2.2.2	Extracting expression information from literature	162
5.2.3	Experimental validation by RT-PCR	163
5.2.3.1	Selection of transcripts	163
5.2.3.2	Clinical sample collection and storage	163
5.2.3.3	RNA isolation	164
5.2.3.4	RT-PCR	164
5.3	RESULTS	164
5.3.1	Analyzing microarray gene expression data using TIPMaP	164
5.3.2	In silico comparisons	166
5.3.2.1	Comparison with MGEx-Tdb	166
5.3.2.2	Comparison with published literature	169
5.3.3	Experimental validation by RT-PCR	172
5.3.3.1	RNA volume, concentration and quality	172

5.3.3.2	Selected transcripts	172
5.3.3.3	Expression pattern validation by RT-PCR	172
5.4	DISCUSSION	173
5.5	BIBLIOGRAPHY	187
6	ESTABLISHING THE TRANSCRIPTOME FOR NON-OBSTRUCTIVE AZOOSPERMIA USING RIBONUCLEIC ACID SEQUENCING, AND IDENTIFICATION OF POTENTIAL BIOMARKERS	190
6.1	INTRODUCTION	190
6.2	METHODS	191
6.2.1	RNA sequencing	191
6.2.1.1	Sequencing	191
6.2.1.2	Verification of the expression profiles with the existing literature	192
6.2.2	Identification of potential biomarkers	193
6.3	RESULTS	196
6.3.1	RNA volume, concentration, and quality	196
6.3.2	RNA sequencing	197
6.3.2.1	Raw data and their quality	197
6.3.2.2	Alignment results	198
6.3.2.3	Clustering	198
6.3.2.4	Expression profiles of transcripts	198
6.3.2.5	Functional annotation	201
6.3.2.6	Verification of the expression profiles with existing literature	201
6.3.3	Identification of potential biomarkers	201
6.3.3.1	Comparison of the expression profiles from TIPMaP and RNA sequencing	201
6.4	DISCUSSION	211
6.5	BIBLIOGRAPHY	215
6	OVERVIEW AND CONCLUSION	219
6.1	SUMMARY	219
6.2	BIOCURATION OF THE GENE EXPRESSION DATA	219
6.3	UNDERSTANDING THE KEY ASPECTS OF ALTERNATIVE SPLICING	221

6.4	DEVELOPMENT OF A NOVEL TOOL FOR THE DETERMINATION OF TRANSCRIPT EXPRESSION PROFILES, USING EXISTING MICROARRAY DATA	221
6.5	IN SILICO COMPARISONS AND EXPERIMENTAL VALIDATION OF TIPMAP	222
6.6	ESTABLISHING THE TRANSCRIPTOME FOR NON-OBSTRUCTIVE AZOOSPERMIA USING RIBONUCLEIC ACID SEQUENCING, AND IDENTIFICATION OF POTENTIAL BIOMARKERS	222
6.7	FUTURE PERSPECTIVES	222
6.8	BIBLIOGRAPHY	223
	Publications and poster presentations .....	224
	Appendix.....	225