

# Contents

- 1 Introduction 1
  - 1.1 Next Generation Sequencing . . . . . 2
  - 1.2 General pipeline in NGS . . . . . 6
  - 1.3 Applications of NGS. . . . . 11
  
- 2 Motivation, aims and main contributions 17
  - 2.1 Motivation . . . . . 17
  - 2.2 Specific aims . . . . . 19
  - 2.3 Main contributions . . . . . 21
  
- 3 Quality Analysis of RNA-Seq technology 27
  - 3.1 Introduction . . . . . 28
  - 3.2 Objectives. . . . . 29

3.3 NOISEq. . . . .	30
3.4 Sequence Quality Control (SEQC) project . . . . .	38
3.5 Discussion. . . . .	54
<b>4 Data integration in NGS</b>	<b>57</b>
4.1 Introduction . . . . .	58
4.2 Methods . . . . .	60
4.3 Results and discussion . . . . .	69
4.4 Conclusions . . . . .	78
<b>5 Functional characterisation of long non-coding RNAs</b>	<b>81</b>
5.1 Introduction . . . . .	82
5.2 Objectives. . . . .	83
5.3 Functional characterisation of long non-coding RNAs . . . . .	84
5.4 spongeScan: A web for detecting microRNA binding elements in lncRNA sequences . . . . .	100
<b>6 General discussion and conclusions</b>	<b>115</b>
6.1 Overview . . . . .	116
6.2 Discussion and conclusions . . . . .	116
6.3 Reach and relevance . . . . .	121

# List of Figures

1.1	Evolution of whole human genome sequencing cost over the years. Courtesy: National Human Genome Research Institute. . . . .	2
1.2	General bioinformatics pipeline in NGS experiments. . . . .	7
1.3	Quality score across all the bases of a sample FastQ file before and after cleaning low quality reads. The first figure shows a sample containing reads of very low quality. The second figure corresponds to the same sample after filtering out those low-quality reads. . . . .	8
1.4	Alignment section example of the SAM format specification. . . . .	9
1.5	RNA-Seq analysis can benefit from the data integration of other omics such as ChIP-Seq, Methyl-Seq, etc. Special algorithms are needed to assign each regulatory region to the corresponding an- notated genes. For regions such as the one in red might be unclear which gene it should be associated to. . . . .	14

3.1	Outline of NOISeq package functionalities. . . . .	31
3.2	S4 classes used in NOISeq package. . . . .	32
3.3	Biodetection plot from NOISeq. . . . .	38
3.4	PCA analysis of FastQC output . . . . .	42
3.5	PCA of SEQC samples analysed by NOIseq read count quality parameters. Lanes and replicates are shown as different entities. Data are coloured by sample type. . . . .	43
3.6	PCA of SEQC samples analysed by NOIseq read count quality parameters. Lanes and replicates are shown as different entities. Data are coloured by sample type. Samples E & F were excluded. . . . .	44
3.7	PCA of SEQC samples analysed by NOIseq read count quality parameters. Lanes and replicates are shown as different entities. Data are coloured by laboratory. Samples E & F were excluded. . . . .	45
3.8	PCA of SEQC samples analysed by NOIseq read count quality parameters. Lanes and replicates are shown as different entities. Data are coloured by sequencing depth. Samples E & F were excluded. Yellow indicates higher sequencing depth than red colours. . . . .	46
3.9	Correlation between replicates of sample B in two different laboratories. Upper triangular matrix shows gene correlations and lower triangular matrix shows transcript correlations. . . . .	47

- 
- 3.10 Correlation of gene expression values for the same samples run at different laboratories. Mean expression values across 4 replicates are used to calculate correlations between laboratories. Upper triangular matrix shows gene correlations and lower triangular matrix shows transcript correlations. . . . . 48
- 3.11 The effect in the number of differentially expressed genes in samples A and B in function of the number of lanes being used. . . . . 49
- 3.12 The number of transcripts detected by an increasing number of replicates at different transcript expression intervals. Each bar represents the number of transcripts detected simultaneously by at least the indicated number of replicates, averaged through all possible replication sets of that replicates number. Transcripts were identified using Cufflinks and expression measured in FPKM. Data for the AGR site. . . . . 51
- 3.13 The number of junctions detected by an increasing number of replicates at different sequencing sites. Stacked bars indicate the relative frequency of the major junction in case of annotated alternative splicing events at the junction. . . . . 53
- 3.14 The number of junctions detected by Illumina sequencing of sample A across different sequencing sites at different levels of replication. Each bar represents the average number of junctions jointly detected by the indicated number of sites, considering all possible combinations of that site number. For each level of replication, one replication set was randomly selected per site and compared with the replication sets of all remaining sites. . . . . 54

4.1	Definition of the areas of a gene used by the RGmatch algorithm. . . .	61
4.2	Examples of two different situations that would result in a region being associated with more than one gene. <b>a</b> Two overlapped genes with different isoforms. <b>b</b> Two different genes with common areas overlapping the region (quasi-overlapping genes) . . . . .	62
4.3	Flowchart describing the rules used by RGmatch to decide the gene area to annotate the region-transcript association (default algorithm options) . . . . .	63
4.4	Venn diagram showing the number of region-gene associations obtained with the HOMER, RGmatch, and CisGenome methods . . . . .	75
5.1	Expression values of two random protein-coding and two long non-coding RNA genes to show that, in general, the expression values of protein-coding genes are almost two orders of magnitude higher than long non-coding RNAs. . . . .	90
5.2	PCA of coding and long-non coding RNAs across a wide range of tissues. Counts were corrected by sequencing depth. . . . .	91
5.3	PCA of coding and long-non coding RNAs across a wide range of tissues. Data were batch-corrected and normalised using the quantile normalisation approach. . . . .	92
5.4	Density plots applied over the expression values using quantile normalisation. Red line indicates the minimum threshold used for both biotypes to consider them as expressed. . . . .	93
5.5	Number of tissues the lncRNAs are specific in. . . . .	95

5.6	The number of lncRNAs specific per tissue. Tissues that were not specific of any lncRNAs were discarded from the representation. . . . .	96
5.7	Biological processes of tissue-specific lncRNAs. . . . .	97
5.8	Molecular functions of tissue-specific lncRNAs. . . . .	98
5.9	Biological processes of non-tissue-specific lncRNAs. . . . .	99
5.10	spongeScan architecture. . . . .	102
5.11	Flowchart showing the main strategy behind the spongeScan application. K-mers of 6, 7 and 8 nucleotides are searched for by using sliding windows of different sizes. Different k-mer frequencies are obtained for each pair k-mer – lncRNA. Highly enriched k-mers are reported and checked for correspondence with a miRNA canonical seed. Pairwise predictions are then represented in spongeScan. . . . .	103
5.12	Main view of the spongeScan web application. . . . .	109
5.13	Form to perform a new prediction analysis with the default example options loaded. . . . .	110

5.14 spongeScan output generated for the example data set. (A) Table showing pairwise enrichments of miRNA canonical seeds in lncRNA sequences. This view only shows a few of the total possible columns containing data and scores. (B) Expression data representation for the first pair CDR1-AS and miR-7-5p. The expression data are grouped by tissue and, when clicked, it will show the expression of all the samples in the tissue. (C) Expression levels of mRNA targets of miR-7 for different tissues as a function of the CDR1-AS expression. Red box-plots correspond to tissues where the lncRNA is not significantly expressed, whereas the green colour indicates expression of the lncRNA in the tissue. . . . . 113



# List of Tables

3.1	Sequencing depth of the samples per laboratory and replicate. . . . .	40
3.2	Differentially expressed genes in common between laboratories for samples A (upper quadrant) & B (lower quadrant). . . . .	48
4.1	Table showing the results at the exon level for the example shown in Figure 4.2 . . . . .	65
4.2	Table showing the results at the transcript level for the example shown in Figure 4.2 . . . . .	66
4.3	Table showing the results at the gene level for the example shown in Figure 4.2 . . . . .	66
4.4	Comparison of the functionalities of the different algorithms . . . . .	70
4.5	Equivalences between the gene areas defined by RGMATCH and HOMER . . . . .	76

4.6 Annotations for the region location within the gene returned by  
RGmatch (columns) and HOMER (rows) . . . . . 77