# Analysis of genomic Regions of IncreaseD Gene Expression (RIDGE)s in immune activation

*Lena Hansson*

Doctor of Philosophy

Institute for Adaptive and Neural Computation
School of Informatics
and
Division of Pathway Medicine
Medical School
University of Edinburgh

2009

# Abstract

A RIDGE (Region of IncreaseD Gene Expression), as defined by previous studies, is a consecutive set of active genes on a chromosome that span a region around 110 kbp long. This study investigated RIDGE formation by focusing on the well-defined, immunological important MHC locus. Macrophages were assayed for gene expression levels using the Affymetrix MG-U74Av2 chip are were either 1) uninfected, 2) primed with IFN-γ, 3) viral activated with mCMV, or 4) both primed and viral activated. Gene expression data from these conditions was studied using data structures and new software developed for the visualisation and handling of structured functional genomic data. Specifically, the data was used to study RIDGE structures and investigate whether physically linked genes were also functionally related, and exhibited co-expression and potentially co-regulation.

A greater number of RIDGEs with a greater number of members than expected by chance were found. Observed RIDGEs featured functional associations between RIDGE members (mainly explored via GO, UniProt, and Ingenuity), shared upstream control elements (via PROMO, TRANSFAC, and ClustalW), and similar gene expression profiles. Furthermore RIDGE formation cannot be explained by sequence duplication events alone.

When the analysis was extended to the entire mouse genome, it became apparent that known genomic loci (for example the protocadherin loci) were more likely to contain more and longer RIDGEs. RIDGEs outside such loci tended towards single-gene RIDGEs unaffected by the conditions of study. New RIDGEs were also uncovered in the cascading response to IFNγ priming and mCMV infection, as found by investigating an extensive time series during the first 12 hours after treatment. Existing RIDGEs were found to be elongated having more members the further the cascade progress.

# Acknowledgements

I would like to thank the entire team at the Division of Pathway Medicine. A special acknowledgement to those involved with the experiments analysed in this study; Sara Rodriguez Martin, Andrew Livingston, Kevin Robertson, Thorsten Forster, Paul Dickinson, and Garwin Kim Sing. A further thanks to Marilyn Horne for providing vital support.

Finally I would like to thank my two supervisors, Dr Douglas Armstrong and Professor Peter Ghazal, for giving me this oppertunity and for all the time and effort they spent on the project.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Lena Hansson*)

# Contents


**1 Introduction** **1**

1.1 Background . . . 3

1.1.1 Chromatin structure . . . 4

1.1.2 Possible explanations for non-random gene organisation . . . 9

1.1.3 Chromatin loops . . . 14

1.1.4 Regions of IncreaseD Gene Expression (RIDGE) . . . 15

1.2 Summary . . . 21

**2 Materials and Methods** **23**

2.1 Experimental methods and datasets . . . 23

2.1.1 Gene expression data . . . 23

2.1.2 Active genes . . . 24

2.1.3 Data sources . . . 24

2.1.4 Probe-to-gene-projection . . . 24

2.1.5 Biological experiments . . . 25

2.2 Bioinformatics methods . . . 26

2.2.1 RIDGE determination . . . 26

2.2.2 Gene function . . . 28

2.2.3 Sequence comparisons . . . 29

2.2.4 Architecture . . . 31

2.3 Statistics . . . 31

2.3.1 The RIDGE activity score . . . 32

**3 The Conceptual Framework** **33**

3.0.2 Requirements . . . 33

3.0.3 Existing software resources . . . 34

3.0.4 Architecture . . . 35

3.1 SORGE DB . . . 36

3.1.1 The genomic database . . . 37

3.1.2 The database of functional annotation . . . 41
3.2 The data processing layer . . . 45
3.2.1 Probe-to-gene projection . . . 45
3.2.2 Determination of active genes . . . 49
3.2.3 SORGE DATA . . . 50
3.3 SORGE Visualisation . . . 51
3.3.1 Method . . . 52
3.3.2 Results . . . 54
3.4 Functionality of SORGE . . . 57
3.5 Discussion . . . 58

**4 RIDGE definition and characterisation 61**
4.1 RIDGE definition . . . 61
4.1.1 RIDGE, loop, dimensions . . . 62
4.1.2 The RW/GL and the MLS models . . . 62
4.1.3 Additionally suggested RIDGE dimensions . . . 64
4.1.4 RIDGEs are $110 \pm 30$ kbp long genomic regions . . . 66
4.2 RIDGE characterisation . . . 66
4.2.1 RIDGE members . . . 67
4.2.2 RIDGE distributions . . . 69
4.2.3 Genomic organisation of RIDGEs . . . 72
4.2.4 ClustalW analysis of RIDGE member sequences . . . 75
4.3 Evaluation of RIDGE dimensions . . . 76
4.3.1 RIDGE dimension $80 \pm 20$ kbp . . . 76
4.3.2 RIDGE dimension $123 \pm 16$ kbp . . . 77
4.3.3 RIDGE dimension $150 \pm 50$ kbp . . . 77
4.3.4 RIDGE dimension $220 \pm 40$ kbp . . . 79
4.3.5 The chosen RIDGE dimension, $110 \pm 30$ kbp . . . 79

**5 RIDGE analysis of the MHC locus 81**
5.1 The MHC locus . . . 81
5.1.1 Rationale . . . 81
5.1.2 The biology . . . 82
5.1.3 Locus definition . . . 84
5.1.4 Experimental data . . . 86
5.2 Identification of RIDGEs in the MHC locus by SORGE . . . 87
5.2.1 Overlapping RIDGEs . . . 89
5.2.2 Discussion . . . 92

5.3 RIDGE analysis in the MHC locus . . . 92
5.3.1 Gene expression profiles for RIDGE members . . . 94
5.3.2 RIDGE gain in macrophages that were both primed and viral activated . 95
5.3.3 RIDGE loss in primed macrophages . . . 97
5.3.4 RIDGE gain in activated macrophages . . . 99
5.3.5 RIDGE loss in activated macrophages . . . 100
5.3.6 Static RIDGEs . . . 101
5.4 RIDGE characteristics for the observed RIDGEs . . . 103
5.4.1 RIDGE gain, RIDGE loss, and RIDGE members in a flux . . . 103
5.4.2 Static RIDGES . . . 104
5.4.3 Quantitative data for RIDGEs and RIDGE members . . . 104
5.4.4 Functional associations between RIDGE members . . . 105
5.4.5 Protein Interaction Network (PIN) analysis of RIDGE members . . . 106
5.4.6 Regulatory control of RIDGEs . . . 107
5.4.7 Number of silenced genes in a RIDGE . . . 108
5.5 Concluding remarks . . . 108

**6 Genome-wide RIDGE analysis** **111**
6.1 Genome-wide RIDGE analysis of the macrophage activation dataset . . . 111
6.1.1 Non-random chromosome organisation of RIDGEs . . . 111
6.1.2 Immune system genes . . . 114
6.1.3 RIDGEs on chromosome 17 . . . 116
6.1.4 RIDGEs on chromosome 11 . . . 117
6.2 Additional datasets . . . 118
6.2.1 The time series dataset . . . 118
6.2.2 The tissue dataset . . . 121
6.2.3 Immune system RIDGEs . . . 126
6.2.4 Discussion . . . 127
6.3 Additional loci . . . 128
6.3.1 The Protocadherin locus on chromosome 18 . . . 129
6.3.2 The Immunoglobin locus on chromosome 6 . . . 130
6.3.3 The Immunoglobin locus on chromosome 12 . . . 131
6.3.4 Random regions . . . 131

**7 Discussion** **135**
7.1 RIDGEs . . . 135
7.1.1 Evolutionary linked units . . . 135
7.1.2 RIDGE definition . . . 135

7.1.3 Consecutive RIDGE analysis . . . 136
7.1.4 Immune system RIDGEs . . . 137
7.1.5 Housekeeping RIDGEs . . . 137
7.1.6 Functional associations between RIDGE members (and RIDGEs) . . . 138
7.1.7 Time series analysis . . . 138
7.1.8 RIDGE gain, RIDGE loss, static RIDGEs, and RIDGEs in a flux . . . 139
7.2 Further Work . . . 140
7.2.1 Are RIDGEs conserved over evolution? . . . 140
7.2.2 Longer time series with less time in between time points . . . 140
7.2.3 Biological replicates . . . 140
7.2.4 Predictive biology . . . 140
7.3 Conclusion . . . 141

**A ER diagrams and a JAVA class diagram** **143**
A.1 ER diagram of the project part of the genomic database . . . 144
A.2 ER diagram of the bootstrap part of the genomic database . . . 145
A.3 ER diagram of the functional annotation database . . . 146

**B Immune system genes** **147**

**C Observed RIDGEs with zero, one, and two gaps** **149**
C.1 RIDGEs in the MHC locus identified by SORGE . . . 150
C.1.1 RIDGEs with no silenced genes . . . 150
C.1.2 RIDGEs with one silenced gene . . . 151
C.1.3 RIDGEs with two silenced genes . . . 151

**Bibliography** **153**

**Bibliography** **169**

# List of Figures


1.1 Chromatin organisation . . . . . 5
1.2 Nuclear architecture . . . . . 6
1.3 The lac operon . . . . . 12
1.4 Rosettes . . . . . 14
1.5 The loop-and-scaffold model . . . . . 15
1.6 The MLS model . . . . . 17
1.7 Four levels of chromatin organisation . . . . . 18

2.1 The macrophage activation dataset . . . . . 25
2.2 The time series dataset . . . . . 26
2.3 Physically linked genes . . . . . 28

3.1 Architecture for SORGE . . . . . 35
3.2 ER diagram of the genomic DB . . . . . 40
3.3 The interactions in the molecular interation database . . . . . 45
3.4 Example of a GFF input file . . . . . 53
3.5 Chromosome Views . . . . . 55
3.6 Expression plots . . . . . 56

4.1 The loop-and-rosette model . . . . . 61
4.2 Expected number of RIDGEs . . . . . 69
4.3 Expected number of RIDGEs with alternative RIDGE dimensions . . . . . 70
4.4 Expected number of consecutive RIDGEs . . . . . 71
4.5 The distribution of RIDGE activity scores . . . . . 72
4.6 Gene lengths . . . . . 74
4.7 The inter-gene distances . . . . . 74
4.8 The gene score and the UTR score . . . . . 75

5.1 RIDGEs found in the MHC class I locus . . . . . 88
5.2 RIDGEs in the Ier3:H2-L region . . . . . 89

5.3 Network for RIDGE A15 . . . . . 97
5.4 After merging of the two networks for A07 and A20. . . . . 106
5.5 After merging of the two networks for A01 and A03. . . . . 107

6.1 RIDGEs cluster on chromosomes . . . . . 112
6.2 The immune system genes on chromosome 17 . . . . . 115
6.3 The immune system genes on chromosome 11 . . . . . 115
6.4 RIDGEs on chromosome 17 . . . . . 116
6.5 RIDGEs on chromosome 11 . . . . . 117
6.6 Number of RIDGEs per time point. . . . . 119
6.7 Number of RIDGEs per tissue . . . . . 125
6.8 The Protocadherin locus . . . . . 129
6.9 The IG locus on chromosome 6 . . . . . 131

A.1 ER diagram of the project part of the genomic database . . . . . 144
A.2 ER diagram of the bootstrap part of the genomic database . . . . . 145
A.3 ER diagram of the functional annotation and molecular interaction database . . 146

# List of Tables


3.1 The genomic and functional annotation DB . . . . . 36
3.2 Data from potential data sources . . . . . 38
3.3 Molecular interactions . . . . . 44
3.4 Data sources for the probe-to-gene projections . . . . . 47
3.5 Hit ration and genome coverage per data source . . . . . 48
3.6 Genome wide coverage per microarray chip . . . . . 48
3.7 The difference in using the mean or the median expression value. . . . . 49

4.1 Number of active genes and RIDGEs . . . . . 67
4.2 Gaps, silenced genes, and RIDGE dimensions . . . . . 68
4.3 Genomic data for chromosomes . . . . . 73

5.1 The MHC locus . . . . . 85
5.2 Probe-to-gene projections . . . . . 86
5.3 RIDGE presence . . . . . 93
5.4 Gene expression for RIDGEs . . . . . 94
5.5 Gene regulation . . . . . 94

6.1 Chromosome characteristics . . . . . 113
6.2 Gene content for random regions . . . . . 132

C.1 Observed RIDGEs . . . . . 150
C.2 Observed RIDGEs with one silenced gene . . . . . 151
C.3 Observed RIDGEs with two silenced genes . . . . . 151

# List of symbols

| | | | |
|---|---|---|---|
| APC | Antigen Presenting Cells | MDS | Macrophage activation DataSet |
| ATP | Adenosine Triphosphate | Mb | mega base |
| bp | base pairs | Mbp | mega base pairs |
| CIITA | Class II Transactivator | MHC | Major Histocompatibility Complex |
| CMV | CytoMegalo Virus | MLS | Multi-Loop Subcompartment |
| CT | Chromosome Territory | MOE | medial olfactory epithelium |
| DB | Database | OO | Object-Oriented |
| DNA | Deoxyribo Nucleic Acid | PSMB | Proteasome subunit Beta |
| DPM | Division of Pathway Medicine | PPC | Protein-Protein Complex |
| ER | Entity-Relationship | $\mu$m | micro meter |
| ER | Endoplastic Reticulim | mRNA | messenger RNA |
| *gene score* | coding sequence similiary score | RNA | Ribo Nucleic Acid |
| GUI | Graphical User Interface | RW | Random Walk |
| H4 | Histone 4 | RW/GL | Random-Walk/Giant-Loop |
| HSDF | Hematological System ... | RIDGE | Region of IncreaseD Gene Expression |
| | ... Development and Function | SAR | Scaffold Attachment Region |
| IC | InterChromatin | TAP | Transporter associated with Antigen Processing |
| ICD | InterChromatin Domain | TCR | T Cell Receptor |
| IE | Immediate Early | TF | Transcription Factor |
| IFN | Interferon | TFBS | Transcription Factor Binding Sites |
| IG | Immunoglobulin | TGT | Target Intensity |
| ILSDF | Immune and Lymphatic ... | $T_H2$ | Helper Type 2 |
| | ... System Development and Function | | |
| INF-$\gamma$ | Interferon-gamma | TNF | Tumor Necrosis Factor |
| IG | Immunoglobulin | UAS | Upstream Activator Sequence |
| IRES | Internal Ribosomes Entry Sites | UTR | UnTranslated Region |
| kbp | kilo base pairs | *UTR score* | upstream sequence similarity score |
| LCR | Locus Control Region | VMO | vomeral nasal organ |
| MAR | matrix attachment region | | |

# Chapter 1

# Introduction

A fundamental question in biology concerns the extent of the relationship between the regulation of biological processes and spatial and temporal aspects of chromatin architecture. This structure/function relationship is well known in the prokaryotic operon. (Okuda et al., 2006) To what extent co-regulated and co-located genes are involved in the same biological processes in eukaryotes remains an under-explored area.

Evidence supporting an association between domain structure, genomic islands, and function has been published for several gene families. For example, the developmentally expressed homeobox (Hox) genes and globin locus are known to be co-regulated with conserved order along the chromosomes. (Krumlauf, 1994; Laat de Wouter, 2003) Further examples are found associated with the mammalian immune system; these include the Major Histocompatibility Complex (MHC) (The MHC sequencing consortium, 1999; Trowsdale, 2002), the Immunoglobulin (IG) $V_H$ locus (Cook et al., 1994), and the T-cell receptor (TCR) loci. (Hodges et al., 2003)

It is known that chromatin organisation can influence DNA replication, recombination, repair, transcription, centromere function, and chromosome segregation (Alsford and Horn, 2004). Furthermore the chromatin architecture is responsible for chromosome packaging via loops, scaffolds, and domains (McClean, Philip, 1997). Two models have been proposed to account for this structure of which the latter is more widely accepted (reviewed in (Albiez et al., 2006));

- the Random-Walk/Giant-Loop (RW/GL) (Sachs et al., 1995) (alternatively referred to as the chromosome territory (CT) model), and
- the Multi-Loop Subcompartment (MLS) (Munkel and Langowski, 1998), (also known as the Chromosome Territory (CT)-Interchromatin Compartment (IC) model).

There are a number of biochemical structures, such as the loop-and-scaffold model (Laemmli, 1979; Sumner, 2003), the Rosette model (Okada and Comings, 1979), and the looping, linking,

and tracking models that support the existence of co-expression of genomic regions. (Bulger and Groudine, 1999; Tolhuis et al., 2002; Spilianakis and Flavell, 2004; Masternak et al., 2003)

The hypothesis presented here is that gene order also matters, in that genes may be regulated as a block unit. Evidence includes the claim that all members of the same species, with rare exceptions, have the same order of genes along the chromosomes since this order is essential for pairing at meiosis (Trowsdale, 2002). In addition, gene order might be a defence against recombination and chromosomal mutations. (Purves et al., 2001) It is known that eukaryotic gene order is not random (Hurst et al., 2004) and that the genome forms complex structures. (Yamashita et al., 2004) In fact gene order can only be random if the positioning of genes is not important for transcriptional regulation; otherwise the high rate of genome re-arrangement would lead to the complete randomisation of gene order in a short period of evolutionary time. (Huynen et al., 2001)

The central hypothesis for this thesis is:

> *There are sub-genomic loci in the genome consisting of physically linked genes that are functionally related, and exhibit co-expression and co-regulation.*

The null hypothesis is that genes are essentially randomly scattered with respect to their functions and expression profiles.

The definition and determination of these sub-genomic loci, Regions of IncreaseD Gene Expression (RIDGEs), requires the use of real data which, in comparison to synthetic data, add multiple layers of complexity. One such layer is the projection between different identifiers, for example the Affymetrix probe identifier and the Ensembl gene identifiers. Another layer is the manual curation, and definition, of interactions between genes. A third is the determination of active genes in a specific biological condition. A framework for gene expression analysis and specifically for the definition, determination, characterisation, and visualisation of genomic regions has been implemented. This framework enables the correlation of functional relevance and structural chromatin data by integrating the DNA sequence, chromosomal location, gene function, gene expression, and molecular interaction data.

RIDGE formation could potentially be used to predict what genes will be active in a given situation. This would represent a step toward predictive biology. Another possible outcome is the usage of the RIDGE analysis as quality control for future gene expression studies. For instance one issue with microarray data is that not only genes which are direct targets are expressed, but there is off-target expression as well. (Marshall, 2005) RIDGE structures can therefore be used in the normalisation step. For instance if genes A, B, and C form a RIDGE, but only B and C are observed, then tweaking could detection of gene A as well, thereby reducing the number of false positives and negatives.

A RIDGE has been defined as a consecutive set of genes in 2D that cover about 123 kbp of DNA (Sachs et al., 1995; Knoch et al., 1998; Munkel and Langowski, 1998; Masternak et al.,

2003; Spilianakis and Flavell, 2004) and where the entire chromatin loop, Rosette (Okada and Comings, 1979), is co-expressed, co-transcribed, and co-translated. Previous works have shown that these RIDGEs (Caron et al., 2001) are present both in the Drosophila melanogaster and the human genome. (Caron et al., 2001; Lercher et al., 2002, 2003a; Oliver et al., 2002; Spellman and Rubin, 2002; Versteeg et al., 2003; Weitzman, 2002) This study focused on the major genetic loci of the mammalian immune system - the MHC locus. One reason is that it is conceivable that the immune system might benefit from a a looped organisation, since it may lead to a quicker immune response.

This study has focused on examining the above hypothesis and is structured as follows: the first chapter describes the biological background (and most of the literature review) and the second chapter the bioinformatics methods. Following these are four result chapters; chapter three presents the implemented framework, chapter four discusses RIDGE characteristics and RIDGE definitions, chapter five investigates RIDGEs in the MHC locus, and chapter six generalises the results to the entire genome, additional datasets, and additional loci. The final chapter of the thesis discusses these results, possible future work, and formulates a final conclusion.

The remainder of this chapter will discuss the background for this study; such as non-random chromatin and gene organisation, the latter specifically in relation to gene function and gene activity. Following; two chromatin loop models - the Rosette model and the loop-and-scaffold model - are presented, leading into previously proposed models such as the Random-Walk/Giant-Loop and the Multi-Loop Subcompartment model. Furthermore gene clustering based on gene expression is discussed. Finally the RIDGE model is presented.

## 1.1 Background

Genomic material is made up by nucleotide base sequences. In humans there are at least $4.6 * 10^7$ base pairs (bp) of DNA (stretching 14000 $\mu$m), that has to be packed into the 2 $\mu$m long nucleus during mitosis, this requires a packaging ratio of 7000. (McClean, Philip, 1997) This remarkable feat is accomplished by the chromatin. (Forsberg and Bresnick, 2001; University of Manitoba, 2005; McClean, Philip, 1997) Prokaryotic genomes tend to be small and do not need to be as tightly packed as eukaryotic chromosomes (Lee and Sonnhammer, 2003), yet these contains operons whereas no equivalent is to date found in eukaryotes.

Tissue-specific gene products evolve on average twice as fast as those that are ubiquitously expressed (Duret and Mouchiroud, 2000), and immune system genes evolve about twice as fast as non-immune genes. (Hurst and Smith, 1999) Although genes with interacting products should logically have a lower evolutionary rate due to the constraints imposed by the interaction. It has been shown that conserved gene pairs have a higher degree of sequence conservation (Versteeg et al., 2003), although adjacent gene pairs are less conserved in eukaryotes

than in prokaryotes. The overall genome rearrangement rate appears higher in eukaryotes than in prokaryotes. (Huynen et al., 2001)

### 1.1.1 Chromatin structure

Chromosomes have been assumed to have everything from no order to highly ordered arrangements. (Cremer and Cremer, 2001) Two interacting higher levels of chromosomal organisation were proposed: the state of chromatin and its position within the nucleus (Hurst et al., 2004), and both of these will be discussed here. The packaging of DNA into the chromatin controls all nuclear processes, furthermore chromatin is partly responsible for gene expression. In order to transcribe a gene, it has to be in a transcriptionally competent region. This in turn requires the DNA sequence to be positioned on the outside of the chromatin, making histone modification and the opening of the chromatin vital functions. Both the structure, and dynamics, of chromatin play an important role in establishing and maintaining a stable pattern of gene expression and differentiation in eukaryotic cells. (Munkel and Langowski, 1998; Munkel et al., 1999) Gene expression, on the other hand, determines cell fate and metabolic state and division. (Niehrs and Pollet, 1999) Furthermore, correct gene expression requires the presence of intact coding sequences and the appropriate regulatory elements; the promoter region, enhancers and a permissive local chromatin environment. (Kleinnjan and von Heyningen, 1998)

#### 1.1.1.1 Packaging of DNA into the chromatin

DNA packaging into chromatin controls all nuclear processes. This involves DNA metabolism (Forsberg and Bresnick, 2001), DNA replication, recombination, transcription, chromosome segregation, centromere function, as well as repair of DNA damage. Furthermore the process is relevant to the pathological progression of cancer and viral disease. (Sachs et al., 1995; Richmond and Davey, 2003; Alsford and Horn, 2004)

The chromatin consists of nucleosome core particles occurring every 200 bp, which in turn consists of a histone octamer. (McClean, Philip, 1997) This is made up by 146 bp of DNA wrapped around two subunits each of the four core histones (Forsberg and Bresnick, 2001) and almost two complete left-handed turns of double-stranded DNA. (McClean, Philip, 1997)

The linker histone binds the nucleosomes, facilitating chromatin condensation and regulatory functions, where the resulting structures assemble into increasingly condensed structures. (Forsberg and Bresnick, 2001) First the 10 nm filament, or fibre, that looks like beads-on-a-string in an electron microscopy (University of Manitoba, 2005; Cook, 1995) and has a packaging ratio of about 6 (McClean, Philip, 1997). Second, 6 nucleosomes are coiled into a left-handed 30 nm thick helix, the solenoid, (University of Manitoba, 2005) with a packaging ratio of 40. The final packaging is the organisation of the fibre into loops, scaffolds, and domains, with a packaging ratio of about 1000 in interphase chromosomes and 10000 in mitotic

chromosomes. (McClean, Philip, 1997)

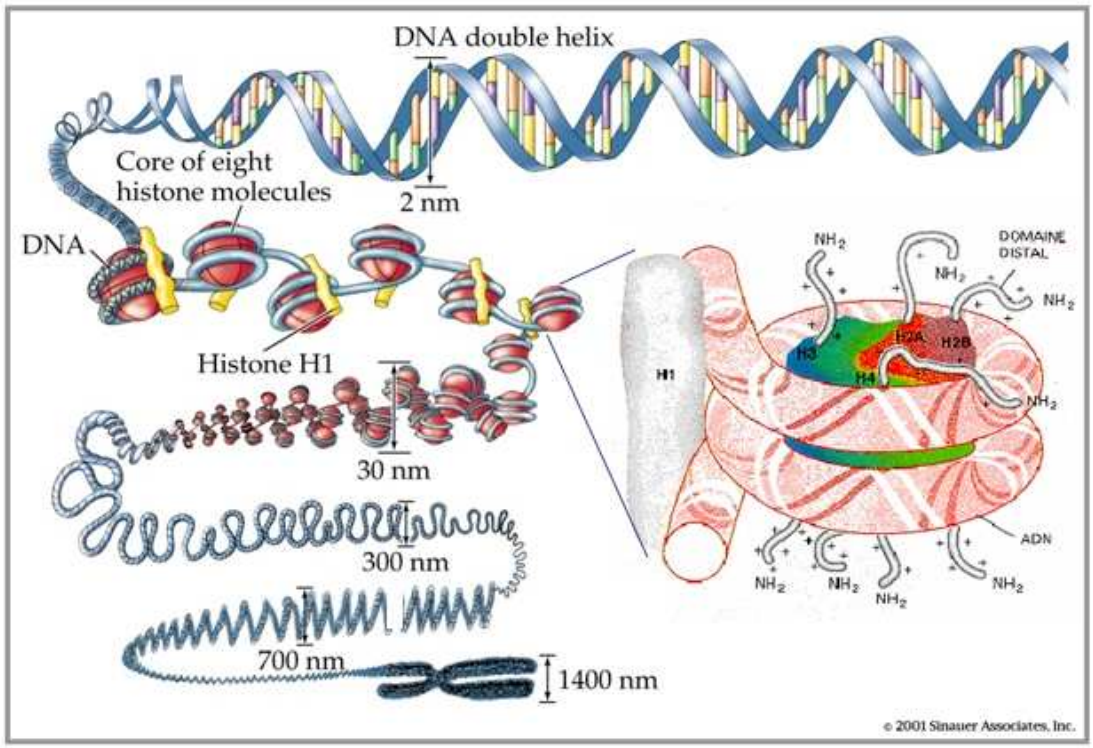



Figure 1.1: Chromatin organisation; from the DNA strand to the nucleus. Taken from (Israe Fortin, 2005).

#### 1.1.1.2 Chromatin organisation

Eukaryotic chromatin organisation consists of tightly wound heterochromatic structures and a more open and accessible euchromatic state, highly enriched in transcriptionally silent and active sequences respectively. (Kleinnjan and von Heyningen, 1998; Forsberg and Bresnick, 2001) The closed conformation correspond to the 30 nm supercoil with six to seven nucleosomes per turn, making 1.2-1.4 kbp of DNA at least partially exposed on the surface of the last superhelical turn. (Hebbes et al., 1994)

Mammalian chromosomes show a banded pattern of early-replicating and mid-to-late-replicating bands, Giemsa-light and G-dark respectively. The former have a high gene density and contains both housekeeping and tissue-specific genes, whereas G-dark bands are gene poor and contain only tissue-specific genes. (Cremer and Cremer, 2001) There is also an alternative base pair banding pattern. (Saitoh and Laemmli, 1994)

#### 1.1.1.3 Chromosome Territory (CT)

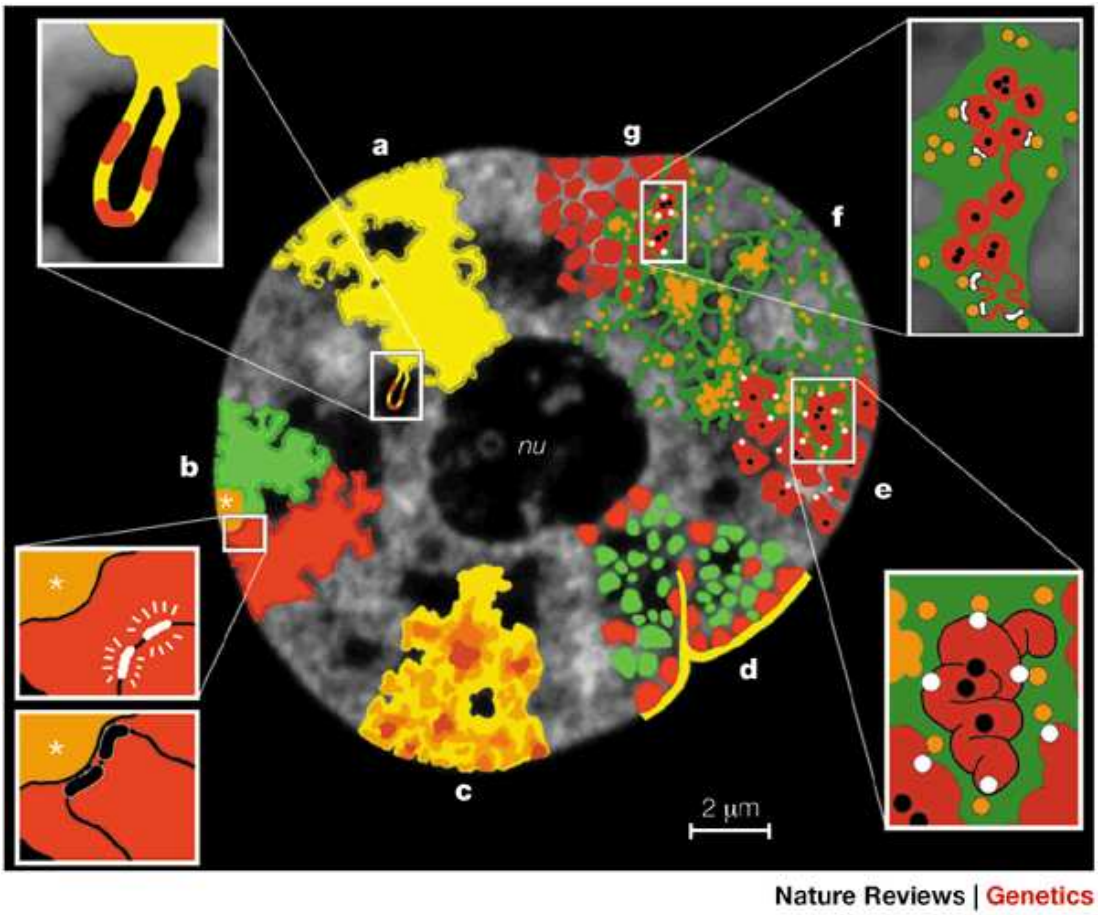



Figure 1.2: Functional nuclear architecture of the folded CT structure. Region a) A giant chromatin loop, with several active genes (red) was expanded from the CT into the IC. Region b) CTs contain separate centromeric and arm domains (asterisks). Top, actively transcribed genes (white) located on a remote chromatin loop. Bottom, recruitment of these genes (black) to the centromeric heterochromatin silences them. Region c) CTs have variable chromatin density; from high (dark brown) to low density (light yellow). Region d) CT showing early-replicating gene-rich chromatin domains (green) and mid-to-late-replicating gene-poor chromatin domains (red). Furthermore gene-poor chromatin is preferentially located at the periphery in contact with the nuclear lamina (yellow). Region e) Higher-order chromatin structures. Active genes (white) are at the surface of the fibre, whereas silenced genes (black) are located toward the interior. f) According to the CT-IC model, the IC (green) contains complexes (orange) and large non-chromatin domains (aggregation of orange dots) for transcription, splicing, DNA replication and repair. Region g) A CT with 1-Mbp chromatin domains (red) and IC (green) in between. At the bottom a closed 100-kbp domain was opened before transcriptional activation. Taken from (Cremer and Cremer, 2001).

Chromosomes occupy discrete territories in the cell nucleus (Cremer and Cremer, 2001) in association with the nuclear matrix (Ma et al., 1999), and are maintained as distinct individuals during interphase. (Dietzel et al., 1998) For example: mammalian and plant DNA is not distributed throughout the entire nucleus but limited to a territory - a subvolume of the nuclear space. (Munkel and Langowski, 1998) Moreover, transcription sites and processing components are both spread in discrete regions. (Jackson and Cook, 1993)

Each physically distinct expression domain contains a gene, or gene cluster, with its corresponding *cis*-regulatory elements. Specialised elements at the borders of these domains are proposed to prevent cross-talk between domains. (Laat de Wouter, 2003) Small proteins, like individual transcription factors are found within these territories but not larger structures. (Munkel et al., 1999)

CTs have complex folded surfaces where actively transcribed genes are located on a chromatin loop that is remote from centromeric heterochromatin and targeting of genes to the periphery, or to a centromeric region, induces silencing. Smaller, human, chromosomes are generally situated toward the interior and larger chromosomes toward the periphery of the nucleus. Gene content is a key determinant of CT positioning; CTs with similar DNA occupy distinct exterior and interior nucleus positions. (Cremer and Cremer, 2001) Factories localise preferentially either to the outside of the chromatid or the inside, but individual genes do no occupy fixed positions. (Cook, 1995)

#### 1.1.1.4 Histone acetylation and the opening of chromatin domains

Chromatin is partly responsible for the regulation of gene expression in association with histone modifications. (Litt et al., 2001) In addition, hyperacetylation of the core histones is required for making a domain transcriptionally competent. (Hebbes et al., 1994) Acetylation of lysines 5, 12, and 16 of histone 4 (H4) was shown to be involved in the initiation of chromatin opening, whereas acetylation of lysine 8 is important for its maintenance. (Litt et al., 2001) There is a close correspondence between the 33 kbp region of sensitive chromatin and the extent of acetylation. (Hebbes et al., 1994)

Transcriptional activation requires potentiation of chromatin which is linked to the activity of the ATP-dependent chromatin re-modeling complexes and to histone acetyl transferase. (Boutanaev et al., 2002) Long-range chromatin re-modeling correlates with the spreading of histone acetylation from the promoter to as far as 16 kbp upstream and is associated with bi-directionally acting transcripts. (Masternak et al., 2003)

Most chromatin is compacted into folded fiber, but the open chromatin has the ability to quickly unfold and, in the presence of the transcription machinery, maintain its steady-state - the 30-nm fiber. (Bystricky et al., 2004) In tissues where a gene is inactive, the chromatin is thought to be in a closed conformation of tightly packed nucleosomes, where transcription

factors (TF)s are unable to bind and the DNA is resistant to DNase I cleavage. The chromatin appears to be in a more open conformation allowing TF binding when a gene is active, as reflected by the presence of hypersensitive sites. The increased sensitivity extends far beyond the region of transcribed DNA, and thus the transcriptional unit could be interpreted as part of the chromosomal structural domain. (Williams et al., 1995)

There are two models for how boundaries function: 1) Boundaries act as roadblocks, obstructing proteins associated with enhancers or silencers from acting on genes, or regulatory elements, in adjacent domains. Thus, boundaries only have an indirect role in sub-dividing the chromosome and defines higher order domains by virtue of their ability to confine the progressive spread of active or silenced chromatin. 2) Boundaries define the physical end-points of looped higher order domains; either by interacting with each other along the main axis, or by interacting with another nuclear structure. (Blanton et al., 2003; Hebbes et al., 1994) RIDGEs could be an example of the latter by defining the genomic region that will loop out, to become either transcriptionally accessible or inaccessible.

#### 1.1.1.5 Gene regulation

Correct gene expression requires the presence of intact coding sequences and the correct regulatory elements, furthermore gene regulation only function correctly in a permissive local chromatin environment. These regulatory elements are; 1) the promoter region - where the basal transcription machinery loads onto the DNA and transcription is initiated; and 2) the enhancers and silencers - short DNA regions containing binding sites for transcription factors. (Kleinnjan and von Heyningen, 1998)

The transcription process is both slow and costly; it takes 50 milliseconds (Ucker and Yamamoto, 1984; Izban and Luse, 1992) and two ATP molecules to transcribe a nucleotide. This might provide selective pressure to make genes as short as functionally possible and the more copies of a gene that is required the stronger this pressure would be. Housekeeping genes are shorter than tissue-specific genes thus indicating a selection for compactness. Selection toward shorter genes should have eliminated introns in highly expressed genes unless they also have important roles such as splicing regulation. This therefore alludes to a balance between the advantageous contribution of the introns and the selective pressure for shortening them. (Eisenberg and Levanon, 2003)

Direct co-regulation is only one possible cause of co-expression. Additional causes include conserved expression patterns (for instance duplication of regulatory elements together with the coding regions) (Lercher et al., 2003b), and the nuclear topology (this might affect the transcriptional status of genes). (Cremer and Cremer, 2001) Finally the chromatin region is important for the expression of individual genes as shown when otherwise identical transgenes were inserted into different chromosomal sites and showed varying levels of expression.

(Spellman and Rubin, 2002)

### 1.1.2 Possible explanations for non-random gene organisation

There are large quantities of research into the relationship between gene function and gene organisation (for example reviewed in (Hurst et al., 2004)). Examples of non-random gene organisaiton include; 1) HOX genes, immunoglobulin genes, hemoglobin genes, and RNA binding genes; these examples are diverse both in terms of the organism in which they occur as well as the mechanism they use to obtain expression of downstream genes. (Blumenthal, 1998) 2) the histone and HOX genes are conserved in clusters (Huynen et al., 2001), 3) clusters of housekeeping as observed in human (Lercher et al., 2002), 4) most of the analysed eukaryotic pathways also clustered (Lee and Sonnhammer, 2003), 5) muscle genes are concentrated on chromosomes 17, 19, and X (Bortoluzzi et al., 1998), 6) non-random patterns of sperm gene distribution in Drosophila (Boutanaev et al., 2002) and mouse (Wang et al., 2001), 7) genes of known similar functions are clustered in budding yeast and human (Eisen et al., 1998), and 8) the seven linked genes involved in quinic acid utilization in fungi. (Hurst et al., 2004)

Expression of genes at the appropriate place and time in development and differentiation could be coordinated by linkage, as it is in the HOX gene cluster. (Zakany et al., 2001) Genes could also be linked to facilitate functional interaction of the products of polymorphic alleles. This could facilitate sequence exchange between similar nucleotide stretches from related, non-allelic genes. In fact, a consistent gene order is essential for the assembly of somatically re-arranged genes, such as immunoglobulins, T-cell receptors, and the protocadherins. (Wu and Maniatis, 1999) Genes that are imprinted may also be tightly clustered (for example the *Igf2* loci) to facilitate the establishment and maintenance of the epigenetic marks crucial for imprinting. (Trowsdale, 2002) Housekeeping genes are likely subject to the strongest selection of adjacent co-expressed genes; they are not only broadly but also highly expressed, a pattern that probably requires little regulation. (Singer et al., 2005) Another group of interesting genes are those encoding proteins of the immune system. These are constantly subject to intense selection for disease resistance as a result of interactions with pathogens. (Trowsdale, 2002)

Two linked genes, A and B, will, on average, stay together $\frac{1}{r}$ generations (where $r$ is the recombination frequency), for example if $r = 0.1\%$ then they would remain linked for 1000 generations before being separated by a crossing over event. For closely linked genes, where $r$ is small, the AB type will increase and become fixed; thus the closer the linkage, the greater the tendency to construct co-adapted complexes. (Motoo, 1994) For example; co-expressed gene pairs in *S.cerevisiae* are twice as likely to be preserved in *C.alibicans* as non co-expressed gene pairs (Huynen et al., 2001; Hurst et al., 2002); therefore gene pairing is an adaptation. (Singer et al., 2005)

The relative position of genes with respect to their Locus Control Region (LCR) contributes

to their respective levels of expression. In transgenes, the presence of a specific globin gene proximal to the LCR inhibits transcription of another, more distal, globin gene, but if the latter is inserted closer to the LCR it is expressed at higher levels but partially inhibits the expression of genes located downstream of it in both positions. (Bulger and Groudine, 1999) It is conceivable that genes have been evolved to be optimally expressed in their current environment, where they are connected to the necessary *cis*-acting regulatory elements, even in competition with closely located genes. (Kleinnjan and von Heyningen, 1998)

#### 1.1.2.1 Natural Selection

Gene expression clusters tend to contain fewer chromosomal breakpoints between human and mouse than expected by chance, indicating that they are being held together by natural selection. This conclusion applies to clusters defined on the basis of broad (housekeeping) expression, and on the basis of correlated transcription profiles across tissues, whereas genes with high expression are not clustered and therefore not conserved during evolution. (Singer et al., 2005)

According to the neutralist hypothesis gene expression clusters are functionally unrelated, but *cis*-acting regulatory elements cause the transcription of one gene to influence the transcription of its neighbor. A selectionist hypothesis on the other hand, proposes that co-regulation of the genes is required and that a chance re-arrangement in the past have brought them together. Once the co-expression proved advantageous the new gene order reached fixation. (Singer et al., 2005) Co-expressed genes in yeast tend to be linked and in close proximity; highly co-expressed gene pairs are twice as conserved as random pairs and physically close genes tend to be conserved more often, although it only accounts for a small proportion of the enhanced degree of conservation of co-expressed gene pairs. (Hurst et al., 2002)

Clusters defined by expression height are not conserved and probably therefore not maintained by natural selection. Housekeeping clusters in the human genome tend not to be broken up in the mouse genome (and vice verse), supporting the hypothesis that these clusters are advantageous and therefore preserved by purifying selection. There appears to be a selective benefit to the clustering of co-expressed and broadly-expressed genes in the human and mouse genomes. Housekeeping genes and co-expressed genes are independently clustered in the human and mouse genomes, and these clusters are maintained by negative selection, that is non-random gene arrangement is the product of natural selection. (Singer et al., 2005)

#### 1.1.2.2 Leaky gene regulation

Non-random gene order need not necessarily be the consequence of selection, another possible explanation would be that gene expression is a noisy process; specifically the opening of chromatin might allow leaky expression of linked genes. (Spellman and Rubin, 2002) During

the chromatin opening the histone modification is spread, and the whole chromosome region is made accessible, or inaccessible, for transcription. Whether these genes are actually expressed depends on other factors such as DNA methylation status, nuclear position, available transcription factors and *cis*-Upstream Activator Sequence effects. (Cohen et al., 2000)

Co-regulation within a neighborhood may be due to incidental interactions between promoters and transcriptional enhancers. (Oliver et al., 2002) A problem with this model is that increased chromatin accessibility is just as likely to facilitate the binding of repressors as activators, with the result that some genes would be up-regulated and some down-regulated (Oliver et al., 2002; Hurst et al., 2004) The neighborhood control is extended to inserted transgenes as well as co-evolved genes. Chromosome deletions and inversions could therefore alter a neighborhood. (Oliver et al., 2002)

Gene clusters correspond to regions of active chromatin (Hebbes et al., 1994), where the transcriptional machinery can access two co-expressed genes more efficiently if they are close together than if they are far apart. This beneficiary juxtaposition would then be selected for. (Roy et al., 2002) Or alternatively when the region is opened to express a single target, it might increase the accessibility of the promoters and enhancers for other genes, leading to a modest parallel increase in gene expression. In summary, regulation of transcription may be precise when it is needed and leaky when it is not. (Spellman and Rubin, 2002)

#### 1.1.2.3 The looping, linking, and tracking models

Eukaryotic transcription can be regulated over tens or even hundreds of kilo bases, involving spatial interactions between transcriptional elements, where intervening chromatin loop out. (Tolhuis et al., 2002)

The looping model postulates that parts of an LCR act together as an integral unit, a holocomplex, to interact directly with individual genes. This type of interaction is influenced by the distance between the LCR and its target genes and the availability of specific transcription factors. (Spilianakis and Flavell, 2004) Proteins bound to distal regulatory regions engage in direct protein-protein interactions with promoter-bound factors. The intervening DNA, regulatory elements, genes, enhancers, and promoters loops out (Tolhuis et al., 2002; Masternak and Reith, 2002), and the individual gene promoters are thought to compete for LCR activity. (Bulger and Groudine, 1999) The MHC class II transactivator association with both the distant Y'-S' motif and promoter-proximal S-Y region support the looping model. (Masternak et al., 2003)

The linking model proposes that the distal control regions serve as entry sites for chromatin re-modeling factors such as histone acetylase. These factors then catalyse the spreading of chromatin re-modeling by repeated cycles of nucleosome modification to adjacent regions. (Masternak et al., 2003) The lack of long-distance enhancer activity in yeast could be explained

by this model. (Bulger and Groudine, 1999)

The tracking model is a variant of the linking model. Chromatin re-modeling factors are associated with Pol II complexes that track along the DNA from the LCR down to the promoter, thus leading to the progressive spread of chromatin re-modeling. (Masternak et al., 2003)

#### 1.1.2.4 Prokaryotic operons

A distinct feature of the bacterial genome is the operon - a set of co-transcribed genes that typically provide a single metabolic function. (Lawrence, 1997) An operon consists of structural genes, a promoter, and an operator (see figure below). Prokaryotes shut down transcription by placing an obstacle between the promoter and the structural genes, the operator can bind to the repressor to create such an obstacle. (Purves et al., 2001) To protect against transcription termination polycistronic pre-messenger RNA is processed to monocistronic mRNAs, by 3' cap end formation and trans-splicing. (Blumenthal et al., 2002) With a few exceptions operons provide non-essential functions. (Lawrence, 1997)

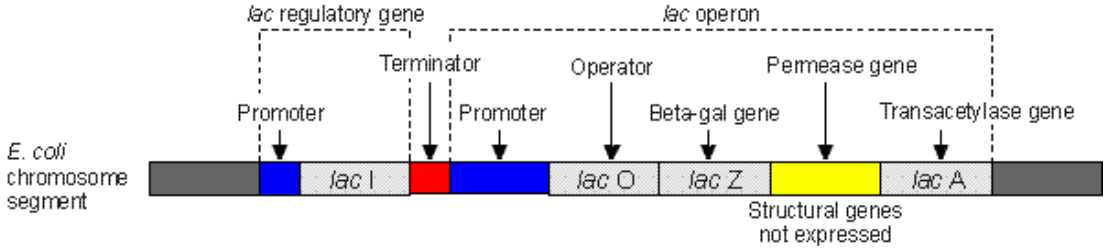



Figure 1.3: The *lac* operon. Taken from (Pearson Education, 2005).

In *E.coli* the *lac* operon is only required when lactose is present and glucose absent. (Nicklin et al., 2002) It consists of three adjacent structural genes, transcribed into a single mRNA molecule. (Purves et al., 2001)

There are five models for the origins of operons: the Natal model, the Fisher model, the Co-regulation model, the Selfish operon model (Lawrence, 1997), and the Trans-splicing model. (Blumenthal, 1998)

The Natal model explains operons by duplication and divergence. Operons encode proteins belonging to the same family (for example the mammalian globin gene clusters). The proximity of genes is an historical property and provides no direct individual benefit. This model can not explain bacterial operons as these evolved by assembly of previously un-linked ancestral genes.

In the Fisher model, operons result from co-adoption, and offer a selective benefit to the individuals in a genetically variable, freely recombining population.

In the Co-regulation model operons co-ordinate expression and regulation, providing a selective benefit to the individual. Co-regulation can only provide selection for the maintenance,

not the origin, of operons.

The Selfish operon model predicts that operons are made up by bacterial genes providing for weakly selected functions, thereby allowing efficient horisontal transfer among genes otherwise susceptible to loss by genetic drift. A selfish operon provides a new metabolic function; for instance, the host gains the ability to exploit a novel ecological niche more efficiently. Gene proximity does not provide a selective benefit to the individual, but does enhance the fitness of the cluster. (Lawrence, 1997)

A Trans-splicing event could be beneficial if the promoter of the upstream gene could transcribe both genes and the reduction in genome size could lead to selection for the deletion. If the two genes are functionally related, this operon might provide functional benefits as well. (Blumenthal, 1998)

Operons are poorly conserved but the eukaryotic progenitor probably had operons. For instance bacteria and archae share numerous operons with the same gene order. The polycistronic transcription units found in *C.elegans* are most likely distinct from the bacterial operons, and a relatively recent innovation (Blumenthal, 1998), so the mechanisms involved in processing of polycistronic mRNAs are quite different in *C.elegans* compared to bacterial genomes. (Blumenthal et al., 2002) The difference between *C.elegans* and *C.briggsae* in gene clustering supports a lineage-specific origin for the *C.elegans* genes. The high rate of chromosomal rearrangements in *Caenorhabditis* means that most ancestral loci would have been scrambled in the absence of selection for their maintenance. (Miller et al., 2004) A comparison of gene order between prokaryotes and *C.elegans* revealed only three cases of interacting, neighboring genes. (Huynen et al., 2001)

Genes that participate in the same functional pathway are often packaged into operons. (Lee and Sonnhammer, 2003) Although this works well in prokaryotes, operons appear to be very rare in eukaryotes and operon-like structures have only been discovered in a few organisms, most notably the nematode worm. (Spieth et al., 1993; Zorio et al., 2002; Blumenthal, 1998; Blumenthal et al., 2002) Polycistronic transcription units in *C.elegans* could be a method of coordinating expression of genes with related functions, as they are in bacteria. For example, the lin-15A and lin-15B proteins collaborate in a single process. (Huang et al., 1994; Clark et al., 1994)

Given the relative compact *C.elegans* genome, operon evolution may have been driven by constraints on chromosomal structure or organisation. The polycistronic transcription units in *C.elegans* co-regulate functionally related proteins. Examples include; 1) a gene encoding a Proteasome subunit found with a ubiquitin ligase complex subunit (which confer substrate specificity), 2) a gene encoding Transcription Factor II C (TFIIC) found with a RNA polymerase III subunit, 3) a gene encoding a regulator of ribosome synthesis found with a RNA polymerase I subunit, and 4) a gene encoding a vesicle docking and trafficking protein found

with a GRIP domain. (Blumenthal et al., 2002)

### 1.1.3 Chromatin loops

Interphase chromosomes consist of a string of nucleosomes running between or looping from more-adhesive factories. During mitosis, a compact stable structure is created through sticky-end aggregation where loops are tied through transcription factors or RNA polymerases to factories (Cook, 1995) and anchored to components of the nuclear matrix, or chromosome scaffold by S/MARs (scaffold/matrix attachments regions). These repeating loop domains are about 50-200 kbp long. (Ma et al., 1999) Super-coiling is another evidence for looping; it implies that DNA is tied down (looped) to prevent rotation. (Cook, 1995)

#### 1.1.3.1 The Rosette model

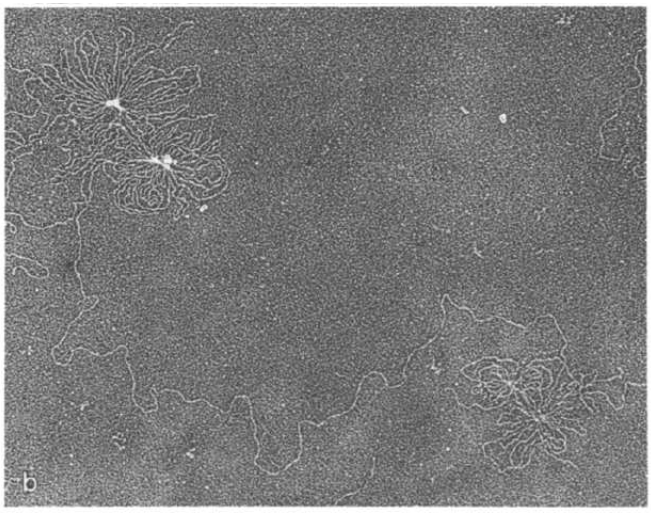



Figure 1.4: Rosettes of DNA loops clustered around an attachment point. Rosettes can have either short (two rosettes in top left corner) or long (from top left to bottom right) inter-rosette linkage. Taken from (Okada and Comings, 1979).

A regular series of Rosettes, connected by inter-rosette linking DNA, were found in the DNA of the Chinese hamster. These rosettes tended to be of similar size and the chromatin fibers were spread uniformly around a central structure. The mean length of inter-rosette linking DNA was $4.4 \pm 1.7$ $\mu$meter. The mean length of a rosette of DNA (measured for all loops) was $13.7 \pm 4.8$ $\mu$meter. The mean number of rosette loops was $20.7 \pm 5.3$. Numerous rosettes composed of loops of DNA cluster around an elevated center. (Okada and Comings, 1979)

Similar rosettes have been observed in the spreading of viral, bacterial, mitochondrial, and kinetoplast DNA. (Okada and Comings, 1979) 20 years later an electron tomography study demonstrated higher order chromatin structures but no Rosette structure. (Munkel et al., 1999)

#### 1.1.3.2 The loop-and-scaffold model

Laemmli et al (Laemmli, 1979) proposed the loop-and-scaffold model. Here histone and non-histone proteins (scaffolding proteins) yield different structural contributions to the higher-order folding of the nucleoprotein. The latter consists of highly folded DNA chains held together by scaffolding proteins. The scaffold is composed of scaffolding proteins that extends throughout most of the chromatid, forming a halo of DNA (see figure below). The chromatin fibre was proposed to be folded to form loops. These loops then wound so that the base form the central axis of the chromatid, where scaffolding proteins are responsible for stabilising and cross-tying these loops.

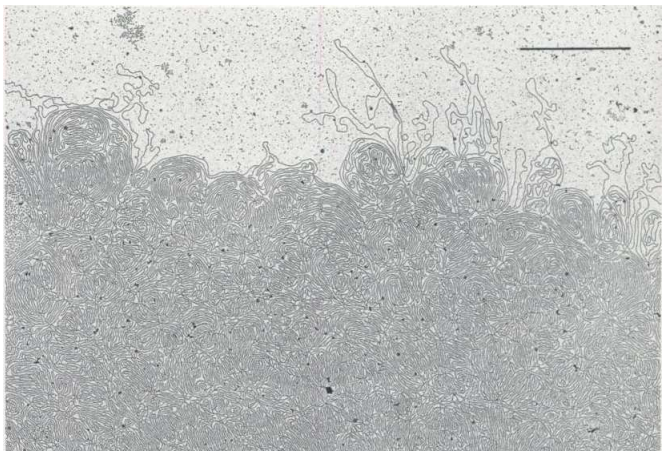


Figure 1.5: Chromosomes consist of central, fibrous, scaffolds to which the DNA is bound, forming a halo. (Individual DNA loops are seen at the top of the picture.) Taken from (Laemmli, 1979).

A great variability in loop size not related to gene-rich or gene-poor regions was seen. A possible correlation between loops and genes was suggested since the number of loops is of the same order as the number of genes in human cells. The final packaging of the loop-and-scaffold model is estimated to 12000, and since this exceeds the required 10000 for condensed chromosomes the model is adequate to produce the final level of condensation for metaphase chromosomes. (Sumner, 2003)

### 1.1.4 Regions of IncreaseD Gene Expression (RIDGE)

#### 1.1.4.1 The Random Walk/Giant Loop (RW/GL) model

In 1995 Sachs et al (Sachs et al., 1995) found evidence for large (around 3 million bp long) flexible chromatin loops along a random-walk backbone of human chromosome 4 during the G0/G1 stage of the cell cycle. The chromatin geometry corresponded to a simple random walk

on scales from 0.1 Mbp to 1.5 Mbp. Observed deviations at larger genomic scales could be explained by a polymer model, in which the DNA of a chromosome is confined to a spherical subvolume. (Sachs et al., 1995)

The chromatin fibre is organised into giant loops, each comprising several Mbp of DNA and connected to a backbone. A random walk is assumed for the folding of the backbone, the folding of the chromatin fibre (Dietzel et al., 1998; Sachs et al., 1995), and for the loop attachment points. The loops form by intrastrand protein connection linking sites that are several Mbp apart. The connections form at, or near, the same DNA sequences disregarding cell type. (Yokota et al., 1995; Sachs et al., 1995) The distance between points depends on their respective distance to the nearest loop attachment point, and on the distance along the random walk between the points. (Yokota et al., 1995) The distance distribution is in turn dependent on the loop size and the length of the backbone segments (protein linkers). (Dietzel et al., 1998) The backbone could be entirely composed of chromatin links around 0.2 Mbp if these links are very short. (Yokota et al., 1995)

The human interphase chromosomes are in agreement with a GL/flexible backbone model. Evidence for this includes; 1) the distance distributions indicate overall random folding; 2) a striking bi-phasic relationship between mean-square interphase distance and genomic separation, indicating two levels of random walk polymer behaviors; 3) the narrow spread of data points around the bi-phasic relationship is consistent with the systematic position of DNA within a periodic, higher order organisation; and 4) physical distance briefly decreases as genomic distance increase, indicative of a looped structure.

The data also support the hypothesis of sequence-specific loop attachment points, where the nucleus is partitioned into irregular shaped chromosomal domains with exactly two levels of higher-order chromatin structure. (Yokota et al., 1995)

#### 1.1.4.2 The Multi-Loop Subcompartment (MLS) model

Munkel et al (Munkel and Langowski, 1998) proposed the MLS model in 1998, which describes the chromosome territory in terms of structural flexibility with a high degree of compartmentalisation. Chromosomes consist of subcompartments connected by small fragments of chromatin. (Dietzel et al., 1998)

The MLS model is based on electron microscopial evidence for clusters of 120 kbp loops forming R- and G-band domains where protein linkers are thought to hold together each cluster. Thereby connecting the loop bases removing the need for a protein backbone. (Dietzel et al., 1998) The exact number of loops in a subcompartment (about 10) was derived from the observed band pattern (with a band size of about 1.5 Mbp). (Munkel and Langowski, 1998)

The human interphase chromosomes were described as flexible fibers with a higher order structure. Folding of 120 kbp loops were found, although similar 3D distances were obtained

for a doubled loop size of 240 kbp. These loops were arranged into rosette-like subcompartments. (Munkel et al., 1999)

Modifications of the MLS model have also been suggested. For example one model suggests a structural modification where several successive subcompartments were opened and the corresponding chromatin formed a single giant-loop domain in the Mbp range. (Munkel and Langowski, 1998) A modified version of the MLS model is the spherical 1-Mbp chromatin domain model, which assumes that CTs are built up from 1-Mbp domains, and that the relative fraction of the nuclear volume occupied by each CT is directly proportional to the number of domains. (Bolzer et al., 2005) This model does not make any assumptions about the internal structure of the 1 Mbp chromatin domains. (Cremer and Cremer, 2001)

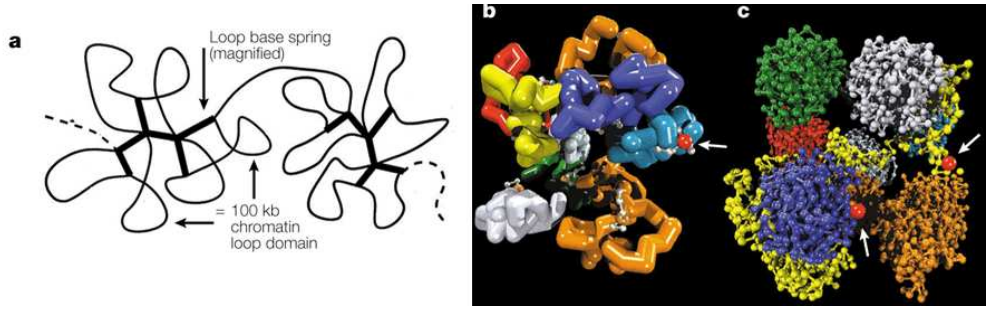



Figure 1.6: The Multi-Loop Subcompartment model. a) 1-Mbp chromatin subcompartments linked by the chromatin fibre. Each domain is built up as a rosette of 100 kbp looped chromatin fibres. b) The nucleosome chain compacted into a 30-nm chromatin fibre (cylinders) and folded into ten 100-kbp-sized loops. The arrow points to a transcription factor complex. c) Each of the ten domains modeled as per a RW nucleosome chain, where a dot represents an individual nucleosome. Taken from (Cremer and Cremer, 2001).

Both the original model (100 kbp) and the modified model (120 kbp) predict similar distances in the 200 Mbp range. Sub-compartments of 120 kbp sized loops are formed by stiff springs between the loop bases. Roughly ten loops form a subcompartment, which is connected by small chromatin fragments of about 120 kbp. (Munkel et al., 1999)

#### 1.1.4.3 Interchromatin models

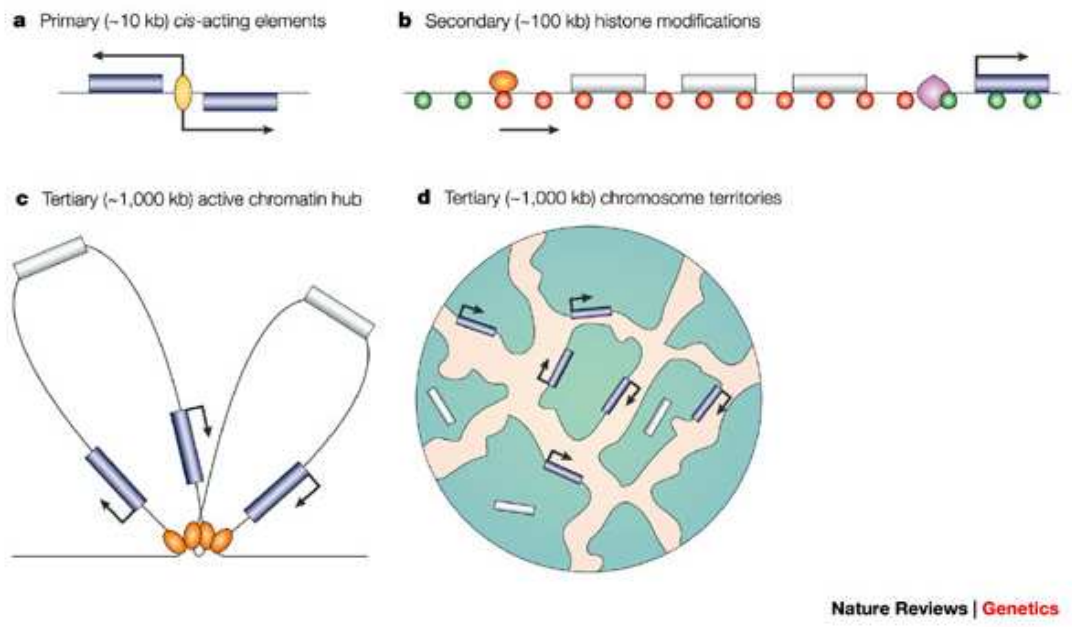



Figure 1.7: Four levels of chromatin organisation; a) *Cis*-acting elements directly affect the transcription of neighboring genes. b) Histone modifications spread from a LCR (orange) down the chromatin fibre until they are stopped by boundary elements (pink). Histone modification (red) suppress transcription of the intervening genes (grey), whereas un-modified histones (green) beyond the boundary element retain an open chromatin and allow transcription of neighboring genes (blue). c) *Cis*-acting elements (orange) form the node of the loop. Genes close to the hub (blue) are accessible to transcription, whereas genes further away (grey) are not. d) Transcription is restricted to territory surfaces (blue) in compact CTs, but suppressed within the interior (grey). Taken from (Hurst et al., 2004).

In 2001 Cremer et al (Cremer and Cremer, 2001) argued for the existence of a 3D interconnected IC with distinct structural and functional properties that co-evolved with a dynamic chromatin-domain architecture.

A higher-order chromatin re-modeling is required for long-term changes in gene expression patterns, reflecting the re-positioning of the genes in open, or closed, higher-order compartments. Open domains have enhanced accessibility to transcription complexes and most of the chromatin in the closed domains remains inaccessible to larger complexes. Regulatory regions and coding sequences of active genes can only interact with the transcription machinery when they are positioned at the surface of chromatin domains that line the IC; or alternatively on chromatin loops that extend into the IC. Long-term, or permanently, silenced genes are located in the interior and inaccessible to the transcription machinery. Whereas gene-rich and early-

replicating chromatin is readily accessible to the transcription and replication machinery. In addition, differentiating cells establish a cell-specific pattern of gene locations with respect to certain nuclear compartments, such as heterochromatin, the IC, and the nuclear lamina. (Cremer and Cremer, 2001)

Chromatin fibres expand throughout most of the nuclear space, resulting in a non-territorial interphase chromosome organisation. In the RW/GL model, chromatin loops of several mega bases are randomly folded to an underlying backbone, whereas it in the MLS model is assumed that chromatin domains are built up from a Rosette of random chromatin-loop domains around 100 kbp long. Both models assume a RW folding of chromatin loops and both are compatible with the assumption that specific positions of genes are not required for activation or silencing. Although Cremer et al (Cremer and Cremer, 2001) found more evidence for the latter model.

#### 1.1.4.4 Gene clusters based on expression levels

Lercher et al (Lercher et al., 2002) found that genes expressed in a given tissue do not map to random locations but instead resolve to clusters. The human genome is organised into subregions specialised in housekeeping genes. Therefore the apparent clustering of genes with high expression rates is a consequence of the clustering of housekeeping genes since housekeeping genes tend to have above-average expression strength. (Lercher et al., 2002; Eisenberg and Levanon, 2003; Singer et al., 2005)

Gene location could be viewed as an adaptive property related to regional base composition and chromosome structure. Here selective pressure would favour the concentration of housekeeping genes in genomic regions with particular structural properties, most likely in order to facilitate access to the transcription machinery. (Lercher et al., 2003a) Furthermore housekeeping gene clusters exhibit a lower proportion of chromosomal break points than elsewhere in the mouse genome. Consecutive mouse gene pairs have a higher expression breadth than randomised gene pairs, indicating a significant level of organisation of housekeeping gene pairs. (Singer et al., 2005) The organisation of housekeeping genes was also found in *Drosophila* by Oliver et al (Oliver et al., 2002) but not by Spellman et al. (Spellman and Rubin, 2002)

Co-expressed genes in *Drosophila* showed about 1700 testes-specific genes, and a third of these clustered into groups of three or more. (Boutanaev et al., 2002) Correlated pairs and triples, but not quadruplets have also been found.

Adjacent genes (disregarding orientation) tend to be co-expressed, as do divergent pairs. These co-expressed genes do not form operons since the two genes often occur on opposite strands, making polycistronic transcription impossible. (Cohen et al., 2000)

Spellman et al (Spellman and Rubin, 2002) found around 200 groups of adjacent and similarly expressed genes in *Drosophila* in more than 80 experimental conditions. Each cluster having between 10 and 30 members (with an average of 10 genes) accounting for over 20% of

the assayed genes. A cluster cover between 20 and 200 kbp of genomic sequence with a mean group size of about 100 kbp. These groups were not explained by polytene banding patterns or other known chromosomal structures, nor were the genes functionally related. A pair-wise Pearson correlation of gene expression in sliding windows resulted in an effect on gene expression that extended well beyond group of ten genes, where gene regulation at the level of the chromatin structure was believed to be the reason behind these clusters. (Spellman and Rubin, 2002)

#### 1.1.4.5 Other Work

There is a significant tendency for genes from the same metabolic pathway to cluster. This has been shown in the genomes of human, worm, fly, yeast, and *Arabidopsis thaliana*. However, not all pathways cluster, and the fraction that do is depending on the species and range from 30% for *D.melanogaster* to 98% for yeast. (Hurst et al., 2004) Tissue-specific gene pairs and triplets have been found to have common putative transcription factor binding sites and to share common biological functions. (Vogel et al., 2005) A study by Fukuoka et al (Fukuoka et al., 2004) found that in six eukaryotic species, there was a consistent tendency that neighboring genes were likely to be co-expressed. Additional examplex of clustering of co-expressed genes include; 1) Genes with similar functions tend to occur in adjacent positions along the chromosomes (Cohen et al., 2000); 2) Muscle-expressed genes cluster into groups of up to five genes in worm (Roy et al., 2002); 3) Co-expressed pairs of neighboring genes, also in worm, within a distance range of 20 kbp (Lercher et al., 2003b); 4) Gene clusters were partially regulated by the same transcription factors, shared biological functions and were characterised by non-housekeeping genes. (Vogel et al., 2005); 5) Human skeletal muscle and adipose tissue, are located on certain chromosomes (Bortoluzzi et al., 1998); 6) Testes-specific genes cluster in fruit flies ((Boutanaev et al., 2002)); 7) Highly expressed genes in liver is conserved among human, mice, and rat (Yamashita et al., 2004); 8) There are few clusters on the X chromosome in both mouse and human (could be an artificial result due to the density of available data); perhaps this is related to the recombination on sex chromosomes, resulting in a decrease of cluster formation. (Singer et al., 2005); 9) Not only are certain genes prevalent on the X chromosome (Wang et al., 2001), but in worm genes of similar expression profiles tend to be clustered (Blumenthal et al., 2002); and 10) Co-expressed genes in *C.elegans* ((Roy et al., 2002; Miller et al., 2004)).

#### 1.1.4.6 RIDGEs

In 2001 Caron et al (Caron et al., 2001) were the first to define a RIDGE, as a cluster of highly expressed genes. A quantitative definition of a RIDGE was not possible as they saw a continuum between small and very large clusters. 27 RIDGEs were identified in humans,

with an average expression level (per gene) up to seven times that of the genomic average. A RIDGE was defined as a region in which 10 consecutive moving medians are four times the genomic median. The probability of finding 27 RIDGEs under a random permutation was very low, and could not be explained by variations in the distribution of highly expressed genes.

Lercher et al (Lercher et al., 2002) instead clustered human genes into RIDGEs based on the hypothesis that tissue-specific genes resolve to clusters. Here housekeeping genes were found to cluster and they concluded that the apparent clusters of genes with high expression rates are a consequence of the clustering of housekeeping genes. (Lercher et al., 2002)

The human chromosomal gene expression profiles revealed a clustering of highly expressed genes in about 30 RIDGEs. Here RIDGEs were defined as gene-dense regions with high GC content, high SINE repeat density, low LINE repeat density, and significantly shorter introns. Clustering of weakly expressed genes in domains with fully opposite characteristics (ANTIRIDGEs) were also found. Both types of domains are open to tissue-specific regulation. RIDGEs were determined to be an integral part of a higher order structure in the genome related to transcriptional regulation and had an overall high expression in all analysed tissues. (Versteeg et al., 2003)

## 1.2 Summary

Eukaryotic chromatin consists of tightly wound heterochromatic structures and a more open, accessible euchromatic state, highly enriched in transcriptionally silent and active sequences respectively. (Forsberg and Bresnick, 2001; Kleinnjan and von Heyningen, 1998) Mammalian chromosomes also show a banded pattern of early-replicating and mid-to-late-replicating bands (Cremer and Cremer, 2001), and an alternative base pair banding pattern. (Saitoh and Laemmli, 1994) These examples of chromosome-level organisation make it unlikely that genes are randomly organised onto chromosomes. In fact it is no longer feasible to assume that gene order in eukaryotes is random; every analysed genome have similar or co-ordinated expression of gene clusters. (Hurst et al., 2004) In pioneering work in the field, Cohen et al (Cohen et al., 2000) and Kruglyak and Tang (Kruglyak and Tang, 2000) independently showed that in yeast (*S.cerevisiae*) adjacent pairs of genes show correlated expression. Co-expression of co-localised genes in higher eukaryotes have since then been found in *C.elegans*, *Drosophila*, *homo sapiens* and *mus musculus*, and *Arabidopsis thaliana*. (Caron et al., 2001; Lercher et al., 2002; Versteeg et al., 2003; Singer et al., 2005; Williams and Bowles, 2004; Roy et al., 2002; Boutanaev et al., 2002; Spellman and Rubin, 2002)

Clustering of genes into evolutionary linked units is supported by; 1) chromatin organisation most notably into the loops-and-scaffolds model (Laemmli, 1979) or the Rosette model (Okada and Comings, 1979); 2) non-random gene order as observed by chromosome synteny,

histone and HOX gene clustering (Huynen et al., 2001), and 3) the cost of protein synthesis; the transcription process is both slow and costly, taking 50 milliseconds. (Ucker and Yamamoto, 1984; Izban and Luse, 1992)

Co-expressed gene pairs in *S.cerevisiae* are twice as likely to be preserved in *C.alibicans* as neighbors that are not co-expressed and thus gene pairing is an adaptation and not a chance event. (Huynen et al., 2001; Hurst et al., 2002; Singer et al., 2005)

Evidence both for and against clustering of housekeeping genes has been presented. (Spellman and Rubin, 2002; Oliver et al., 2002) Whereas, in the mouse genome both housekeeping and immunogenic genes have been found to clusters. (Williams et al., 2002)

This thesis examines the following hypothesis:

> *There are sub-genomic loci, RIDGEs, in the genome consisting of physically linked genes that are functionally related, and exhibit co-expression and co-regulation.*

# Chapter 2

# Materials and Methods

This method chapter has been divided into three sections; Experimental methods, Bioinformatics methods, and Statistics. Experimental methods cover gene expression data (options, benefits, and drawbacks), determination of active genes, required data (what data is necessary for the analysis, and which data sources are used), and finally the experiments (biological conditions) used. Bioinformatics methods cover the tools, methods, and algorithms used; such as sequence comparisons of genes within RIDGEs and their regulatory regions, or the algorithm implemented to detect RIDGEs. Statistics cover the scoring of active RIDGEs, and the implemented permutation analysis.

## 2.1 Experimental methods and datasets

### 2.1.1 Gene expression data

This study has used microarrays to measure gene expression levels. Microarrays quantify gene expression levels by measuring the hybridisation of DNA to mRNA. A single array can capture thousands of spots, and thereby simultaneously quantify thousands of genes. But even a time series only provides snapshots of a specific developmental stage in a single specimen.

#### 2.1.1.1 Normalisation

Scaling and normalisation are used to correct for variations between arrays before arrays can be compared. Biological variations may arise from many different sources, such as genetic background, growth condition, dissection, time, weight, sex, and age. Technical variations may be due to experimental variables, such as the quality, or quantity, of the hybridised target, reagents, stain, or alternatively handling errors.

The microarray probe sets are scaled to a specific target, followed by normalisation of the baseline versus the scaled probe sets. (Affymetrix Inc, 2001) Here the genes were scaled using

the Target Intensity (TGT) method, with the target set to 100, as suggested for example by Scopes et al. (Scopes, 2002) First the 2% lowest, and highest, values on the array are removed and the remaining probes averaged; then the scale factor is calculated using the target intensity (in this case 100). (Affymetrix Inc, 2002) The TGT method is good at dealing with higher intensity signals, thus reducing the number of false positives. Values below 20 are considered as background. Another benefit from using the TGT method is that it is possible to add in additional arrays during analysis, without having to recalculate the scaled values. (Dickinson, 2007)

### 2.1.2 Active genes

Each probe on a microarray chip is associated with an intensity signal and a detection call (Present (P), Marginal (M), or Absent (A)). For this study, multiple thresholds of gene activity were tested for both methods (see section 3.2.2 for details).

The treatment of biological replicates also influences the determination of active genes. Here the median of the replicate intensity signals has been used (as suggested by for example Anderson et al (Anderson et al., 2006)). The median reduce the effect of outliers, otherwise the results might reflect sampling errors rather than actual expression levels. See section 3.2.2 and table 3.7 for a discussion about the difference between the median and mean values as observed for this study.

### 2.1.3 Data sources

The main data source used for this study is the Ensembl Gene Database. (Hubbard et al., 2005; Birney et al., 2006) In addition, other data sources are also included such as the UniGene database, the UniProt database, the GeneCard database, and the MGI database (see section 3.1 for specifics). Ensembl is chosen as the main data source because it incorporated most of the critical data (such as exact physical gene location and the nucleotide base sequence for the chromosomes) and because it had easy access to this data (via the BioMart interface). The data was retrieved and stored in an in-house PostgreSQL database along with associated meta-data and information for the current project/conditions (see section 3.1.1.2 for specifics).

### 2.1.4 Probe-to-gene-projection

Since the generation of the probe sequences, it is possible that the genome annotation have been refined, therefore a projection between an Affymetrix probe identifier and an Ensembl gene identifier is implemented. Ambiguous and duplicate mappings between sets of identifiers are usually deleted (Kirov et al., 2005), however this data can beneficiary be used to score projections. In our method, every possible identifier-to-identifier projection is considered. Fur-

thermore, a BLAST sequence comparison was performed between the probe and the gene sequences. Each possible projection is then associated with a reliability score, based on our confidence in each (see sections 3.2.1 for more details).

### 2.1.5 Biological experiments

Three different experiments is used in this study, one small specifc dataset on a well-defined, important, biological question, a time series, and a wide-ranging tissue dataset.

#### 2.1.5.1 The macrophage activation dataset

To investigate how the immune system in mice (*Mus musculus*) responds to mCMV infection, macrophages where used in vitro. Specificially, to explore whether or not mCMV infection is associated with RIDGEs we have chosen to key the MHC immune locus using gene expression data from mouse macrophages assayed using the Affymetrix MG-U74Av2 chip. Eight treatments (with three biological replicates), were investigated including control, mCMV infection and IFN-γ priming (which are the ones presented in detail for this thesis).

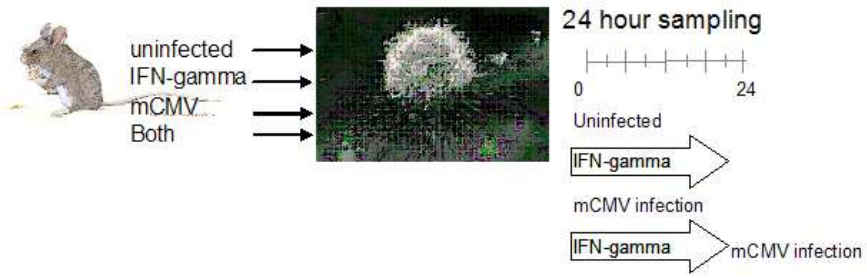



Figure 2.1: Macrophages were either 1) uninfected, 2) primed with interferon-γ (IFN-γ) for 24 hours, 3) viral activated with mCMV, or 4) first primed for 24 hours, then infected, and harvested. (The healthy macrophage was from (5M Enterprises Ltd, 2007)).

This dataset contains enough data quantities to find significance, however not so much that the manual analysis becomes intractable.

#### 2.1.5.2 A time series on macrophage activation

Macrophages were investigated for the first 12 hours (the first 720 minutes) after activation. Macrophages were either primed with IFN-γ, viral activated with mCMV, or both primed and viral activated. Samples were prepared every 30 minutes which resulted in 25 distinct time points for each of the three biological conditions. The data was assayed onto the Agilent chip and normalised via the within-chip normalisation LOESS; followed by a between-chip

normalisation (Median-Absolute-Deviation). A gene is considered active if it has a value above zero.


Figure 2.2: The time series dataset. Macrophages were either 1) primed, mCMV infected, or 3) both primed and viral activated. (The healthy macrophage was taken from (5M Enterprises Ltd, 2007)).

#### 2.1.5.3 An extensive tissue dataset

The Genomics Institute of the Novartis Research Foundation (GNF) has performed an extensive gene expression analysis of 61 tissues in mouse each with 2 replicates (see chapter 6.2.2). For this dataset a gene is considered active if the majority of detection calls was set to present. (GNFb; GNFa)

## 2.2 Bioinformatics methods

The hypothesis that *there are sub-genomic loci, RIDGEs, in the genome consisting of physically linked genes that are functionally related, and exhibit co-expression and co-regulation* requires the determination of:

1. physically linked genes - determined through the implementation of a sliding window algorithm in JAVA.
2. functionally related genes - determined through a number of sources, such as GO, lists of housekeeping genes, and lists of genes known to interact in the same pathway.
3. co-expression between genes - determined by gene activity over a set of conditions.
4. co-regulation of genes - determined by a sequence similarity score of the upstream regions in combination with known transcription factor binding sites.

### 2.2.1 RIDGE determination

To determine if there are RIDGEs in the mouse genome, a sliding window algorithm is implemented in JAVA. A RIDGE is considered as a set of active neighboring genes located on the same chromosome with a loop dimension around 110 kbp ($110 \pm 30$ kbp). The algorithm requires a number of parameters; such as should the analysis be restricted to genes with reliable probe-to-gene projections, should all members in a RIDGE be present on the same strand,

should silenced genes be allowed inside a RIDGE, how is physical linked genes determined, and very important how are active genes determined. Below a short discussion about each parameters is presented along with the choosen default value.

#### 2.2.1.1 Gene orientation is not considered

A looped chromosome organisation implies that gene orientation is less important since all genes in the loop are made accessible for transcription. This assumption is supported by the literature. Gene pairs and gene triplets have been found on both the same and divergent strands, (Kruglyak and Tang, 2000; Huynen et al., 2001; Vogel et al., 2005) as has co-expressed genes. (Cohen et al., 2000) Divergent genes are in fact more likely to belong to the same regulatory unit. (Kruglyak and Tang, 2000) Convergently transcribed gene pairs in eukaryotes are more often conserved than divergently transcribed pairs and at least as often as genes transcribed in the same direction. (Huynen et al., 2001)

Thus RIDGE members are not restricted to the same transcriptional orientation.

#### 2.2.1.2 Restrict analysis to genes with expression data

The analysis of the mouse genome is restricted to genes that have associated expression data, in other words genes that are present on the current microarray. The MG-U74Av2 array only cover about one third of the mouse genome. This means that gene expression profiling of the mouse genome is incomplete, and the resulting RIDGE structures too noisy. This problem is also discussed for example by Singer et al (Singer et al., 2005) who also choose to restrict their analysis to genes with expression data.

#### 2.2.1.3 A RIDGE should not contain silenced genes

A gene inside a loop could be silenced because of; 1) gene competition - the genes compete for the same regulatory elements, 2) the gene might lack a probe, or probe projection, for the current array (a potential false gap), or 3) the silenced gene(s) are interwening looped out genes (a true gap). For example, almost 50% of the muscle-expressed gene clusters have a gene that is not detectably expressed in them. (Roy et al., 2002) Since genes without reliable probe-to-gene projections are allowed inside a RIDGE, a RIDGE may not contain silenced genes.

#### 2.2.1.4 Physically linked genes

Genes can be defined as physically linked according to a number of criteria including chromosome, strand, and physical location. The physical location is most commonly sorted according to start position. Here genes are sorted according to their midpoint (see figure below) in order to try to correct for potentially multiple regulatory regions both in the 3' and 5' end.

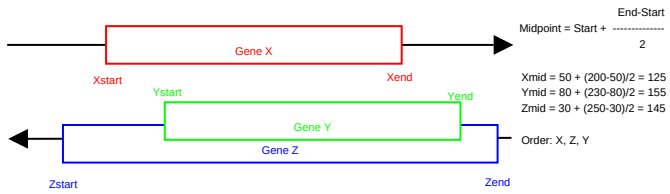



Figure 2.3: Physically linked genes are in this study defined according to midpoint. Thus making Z and Y physically linked, but not X and Y.

### 2.2.2 Gene function

In order to link structure and function a method for determining if genes are functionally related or not was required, although no such objective method exist. Ambiguity arise from multiple data types (such as pathway associations, gene-gene interactions, protein-protein interactions, gene-protein interactions, alleles affecting the same trait) and furthermore these are not mutually exclusive nor necessarily overlapping. (Hurst et al., 2004)

Here the focus is on the immune response and genes are classified according to: 1) their UTR regions (see 2.2.3), 2) housekeeping properties (see below), 3) GO-descriptions (molecular function, biological process, and cellular component) (The Gene Ontology Consortium, 2000), 4) UniProt descriptions (such as subcellular location, development stage, tissue specificity, similarity) (Bairoch et al., 2005), 5) KEGG pathways (Kanehisa, 1997; Kanehisa and Goto, 2000; Kanehisa et al., 2006), 6) GeneCard entries (such as aliases and family) (Rebhan et al., 1997), MGD data (such as symbols and names) (Eppig et al., 2005), 7) manually curated interaction partners, and 8) network associations from the literature, although this latter is even more subjective than the previous suggestions.

GO-terms are used as the only source of functional data in a number of scientific research, for example in (Cohen et al., 2000; Spellman and Rubin, 2002; Eisenberg and Levanon, 2003; McCarroll et al., 2004; Yamashita et al., 2004; Fukuoka et al., 2004; Hurst et al., 2004; Vogel et al., 2005; Coppe et al., 2006).

#### 2.2.2.1 Housekeeping genes

Housekeeping genes are responsible for energy production, chromatin structure, cytoskeletal function, RNA processing, protein expression, genes encoding oxidoreductases, transcription factors and proteins involved in lipid, fatty acid and sterol metabolism. (Roy et al., 2002)

Generally a gene is tissue-specific if expressed in fewer than six tissues (Williams and Hurst, 2002), and a housekeeping gene if expressed in more than 9 tissues. (Lercher et al., 2002)

A table of housekeeping genes was attempted; but even since six different online sources

were merged ((Ge HealthCare), (SuperArray Bioscience Corporation), (Functional Glycomics Gateway), (Eisenberg and Levanon, 2003), (Z-lab), (Dorus et al., 2004)) only 978 entries were found.

For the purpose of this study a housekeeping gene (or RIDGE) is expressed in at least ten different tissues or present in the above table.

#### 2.2.2.2 Pathway analysis

RIDGE formation could benefit genes participating in the same biological process. A number of external pathway databases exist; such as KEGG (Kanehisa, 1997; Kanehisa and Goto, 2000; Kanehisa et al., 2006), DAVID (Dennis et al., 2003), and Ingenuity (Ingenuity Systems). The latter was used to suggest networks, and pathways, for RIDGE members, whereas KEGG did not result in any pathways for the RIDGEs in the MHC locus.

In addition, three immune related pathways are stored in the database as based on the manual annotation of the MHC class II antigen presentation pathway (see discussion in 3.1.2.3).

- MHC Class II genes - the genes and gene products found in any of the following articles (Rudensky et al., 1991; Morris et al., 1994a; Muhlethaler-Mottet et al., 1997; Panjwani et al., 1999; Wagle et al., 1999; Siemasko and Clark, 2001; Ting and Trowsdale, 2002; Boss and Jensen, 2003).

- Immune associated genes - the genes and gene products found in any of the following articles (see 3.1.2.3 and appendix B).

- Genes in the Jak-Stat pathway - the 90 genes, or gene products, found in either Meraz et al or Aaronson et al. (Meraz et al., 1996; Aaronson and Horvath, 2002)

The Jak-Stat pathway is of interest because IFN-$\gamma$ mainly signals through the Jak-Stat pathway (Schroder et al., 2004) and because STAT1 deficient mice are unable to resolve infections by microbial pathogens and viruses. (Meraz et al., 1996)

### 2.2.3 Sequence comparisons

#### 2.2.3.1 The *coding sequence similarity score*

ClustalW (Chenna et al., 2003) is a heuristic algorithm that aims at finding a good-enough solution reasonably fast, and is normally used to align nucleotide and protein sequences in order to identify conserved regions. Here it is used to calculate two different scores for a RIDGE, the first is the *coding sequence similarity score* (gene score) and the second an *upstream sequence similarity score* (UTR score).

An assumption for the purpose of this project is that the absence of conserved regions indicates that the sequences are independent of each other, being neither sequence duplications nor share regulatory regions. ClustalW ignore local alignments and only focus on global alignments, which corresponds well with the gene score but less well with the UTR score. A distribution of scores for random genes were created. The actual score is considered significant if $p < 0.05$, where the analysis is restricted to gene groups with the same number of genes.

To determine if genes in a RIDGE are in fact, recent sequence duplications, a gene score is calculated based on a sequence alignment for the gene sequences. Differences in sequence lengths result in worse (higher) scores and in addition the more RIDGE members the worse the score. The UTR scores are lower than the gene scores which is probably due to the inclusion of fewer gaps between aligned sequences because the sequences are of the same lengths.

#### 2.2.3.2 Co-regulation of RIDGE members

For the purpose of this study the combination of PROMO and ClustalW is used to determine if the RIDGE members are regulated by the same element(s). For both methods the analysis is restricted to 5 kbp of the 5' untranslated region (UTR). 5 kbp were chosen as the longest region from the following papers: (Elefant et al., 2000; Aerts et al., 2003; Kreiman, 2004). A high similarity score reported by ClustalW (Chenna et al., 2003) for this regulatory region imply a shared regulation through shared regulatory elements.

PROMO (Messeguer et al., 2002; Farré et al., 2003) identifies potential transcription factor binding sites (TFBSs) shared by all the RIDGE members. For this study the dissimilarity score was lowered from the default 15% to 5%, for a more stringent matching. CORG, as used by Vogel et al (Vogel et al., 2005), is not suitable for this project since it is only based on TFBS in non-coding regions conserved between human and mouse. (Vogel et al., 2005)

A single-gene RIDGE is considered to have many TFBS if it has more than 31 TFBS; a two-gene RIDGE has to share 26 TFBS, a three-gene RIDGE 22, a four-gene RIDGE 19, a five-gene RIDGE 15 TFBS.

#### 2.2.3.3 Sequence comparisons via BLAST

For sequence comparisons via BLAST the most important parameter is the threshold for a true match. This is normally defined by the e-score, where literature suggestions range from 0.2 down to $10^{-60}$ where both the size and percentage of matches are considered (in decreasing order: (Fukuoka et al., 2004; Hurst et al., 2002; GeneSpeed; Kutchma et al., 2006; Cheung et al., 2003; Bortoluzzi et al., 1998; Lercher et al., 2003b))

Here, each hit with an e-score below $10^{-16}$ is considered, which is a more stringent match than used by for example by GeneSpeed. (GeneSpeed; Kutchma et al., 2006)

### 2.2.4 Architecture

The implemented framework (chapter 3) is platform independent. It was implemented under Linux (Red Hat 9 through Scientific Linux 3-6), but also tested under Windows XP and Mac OS 10.4. Most of the coding has been done in JAVA 1.4.2, but has subsequently been moved to 1.6. Backward compatibility requires recoding of certain Vectors, since class conversions are better dealt with in the later version.

The data is stored in a PostgreSQL database (version 8.1.8), although originally created for an earlier version (7). The database upgrade lead to the renaming of the table *Array* to *Arrayes,* since Array is a restricted word in the newer version. The database currently (2007-09-13) requires 70 GB of storage, whereas only 2 GB is required to store the files that are downloaded from the internet. The larger part of this is taken up by the results from the permutation tests (see 2.3).

In addition to JAVA; Perl, C, and MatLab has been used where appropriate. For example Perl is used for most of the string parsing (including the creation of the database), C for speed crucial implementation (such as the initial randomisation of the genome), and MatLab is mostly used to create graphs and determine p-values of real scores in comparison to a distribution of scores. During the first year Perl was used as the main programming language, but as the complexity of the project grew it soon became apparent that an object-oriented language was necessary. In addition a visualisation module was added so a language that could easily handle GUI programming was necessary. Thus the code was re-implemented in JAVA.

There are a number of required libraries used in this project all of which are publicly available; 1) The Math library from the Jakarta project (The Apache Software Foundation, 2008) is required to perform certain statistical methods. 2) A database connection library is required to connect to the database (for example pg74.215.jdbc1.jar). 3) The simple API for XML (SAX) is required to parse the configuration files stored in XML. (The SAX project, 2004) 4) The JUnit library (JUnit.org, 2007) is used to evaluate the test cases. and 5) The ANT libraries have been used during the development. (The apache ant project, 2007)

## 2.3 Statistics

All statistics reported for this study are based on random genomes created by randomising the physical start position of the genes.

For a permutation, or bootstrap test, a null hypothesis has to be defined in order to later on test for deviation. Often the gene location is randomised, and then the test function is recalculated for the randomised genome. This is repeated many times (N times), to generate the null distribution, to which the observed value is then compared. If there are $n$ random values, and $r$ have a test score that is equal to, or greater than, the observed value, then $p = \frac{r+1}{n+1}$.

Each randomised genome is then subjected to the sliding window RIDGE detection algorithm and the corresponding RIDGEs saved to the database. These random RIDGEs are in turn quantified and the distribution of an observed RIDGE characteristic is compared to the actual value for a real RIDGE. For instance the number of RIDGE members, the ClustalW gene score, or the RIDGE activity score.

In the literature N varies from 100 to 100000. [100 ((Williams and Hurst, 2002),(Lee and Sonnhammer, 2003)); 200 ((Lee and Sonnhammer, 2003)); 1000 ((Singer et al., 2005)); 10000 ((Williams and Hurst, 2002), (Roy et al., 2002), (Vogel et al., 2005), (Caron et al., 2001)); 100000 ((Vogel et al., 2005), (Singer et al., 2005))].

#### 2.3.0.1 Randomise the start position of genes

For this project a gene has its physical start position randomised; although all other gene features such as chromosome, strand, length, gene identifier, and gene expression levels remain the same. Also two genes are excluded from having the same start position although they are allowed to overlap. N was set to 30000.

### 2.3.1 The RIDGE activity score

Ideker et al (Ideker et al., 2002) was trying to determine if a specific sub-network was more active than other sub-networks. Here we want to determine if a specific sub-structure is more active than random gene groups.

A Pearson pair-wise correlation $p$-value is calculated for each gene $i$ in a group (be it a network or a RIDGE). Each $p_i$ is then converted to a z-score $z_i = F-1(1-p_i)$, where $(F-1)$ is the inverse normal CDF. This yields a uniform distribution (0,1) for random data, and the $z$-scores follow the standard normal; smaller $p$-values correspond to larger $z$-scores.

To produce an aggregate z-score, $z_A$ for an entire group of genes, the individual $z_i$ are summed for all genes in the group and background corrected. For multiple conditions the scores are first sorted and significance calculated using a binomial order statistics. Finally, a standard z-score is obtained via the inverse CDF ($\phi^{-1}$), which is rank adjusted. The maximum of these values in then the final group score, and the higher the score the more active the group.

The distribution of RIDGE activity scores for the random genomes is calculated on the macrophage activation dataset. Since this deals with invasive treatments (such as viral infection and priming) the scores are probably higher than if another cell line and another set of biological conditions had been used. This is probably why low-scoring non-immune related RIDGEs are observed although high-scoring immune related RIDGEs are not.

# Chapter 3

# The Conceptual Framework

The Structural ORGanisation of Expression (SORGE) software package correlates region and function relationships at the chromatin level in eukaryotes, as per the MLS/CT-IC model (Munkel and Langowski, 1998; Albiez et al., 2006).

SORGE is implemented in order to elucidate an answer to the main hypothesis for this study: *There are sub-genomic loci, RIDGEs, in the genome consisting of physically linked genes that are functionally related, and exhibit co-expression and co-regulation.* For this purpose SORGE integrates a number of disparate data sources, and has a built-in exploration tool facilitating data query, validation, evaluation, and export.

### 3.0.2 Requirements

In order to define and categorise genome sub-regions a software package, SORGE, was implemented. The three main user requirements for SORGE, were *speed,* a complete analysis *pipeline,* and a *visualisation software.* Whole-genome analysis - the retrieval, analysis, and visualisation of tens of thousands of genes, and/or hundreds of thousands of proteins simultaneously) is a time-consuming task, therefore code and database optimisation became central design issues. A whole-genome analysis usually consists of multiple experimental conditions, for example a time series. A pipeline is required since it has to be possible to analyse oeach of these conditions with multiple parameter settings. A year into the project it became apparent that existing visualisation softwares, such as GeneSpring (GeneSpring Analysis Platform), were not suitable. For example, every time a genomic region was visualised it took a week to retrieve and upload the genomic sequence from the database, and manually annotate the genes with name and function. Requirements not met in GeneSpring include the possibility to color code genes according to different features, such as expressed/suppressed, and to zoom in on a specific genomic loci, such as the MHC class II locus.

#### 3.0.2.1 Data requirements

To find out if a gene falls in a genomic sub-region required the determination that gene G1 comes before gene G2 on chromosome C. To determine why genes form genomic sub-regions we had to be able to specify whether, or not, the genes are regulated by the same transcriptional elements or if they participate in the same biological processes. The following data are collected;

a) all chromosome sequences,

b) all gene sequences,

c) the upstream, regulatory regions of the genes,

d) physical, not chromosomal, start and end positions for all genes,

e) gene length,

f) all known protein targets, transcripts, for a gene,

g) physical start and end positions for each intron and exon,

h) strand orientation (sense or anti-sense),

i) gene and protein function

j) molecular interactions of genes and proteins

### 3.0.3 Existing software resources

Existing software resources, for example those provided by EBI (Brooksbank et al., 2003, 2005) and NCBI (Wheeler et al., 2007), usually focus on a single object; a gene (Entrez Gene (Maglott et al., 2005)) or a even a complete pathway as in KEGG (Kanehisa, 1997; Kanehisa and Goto, 2000; Kanehisa et al., 2006). These object-centric solutions focus on user-oriented, well-developed interfaces, but offer limited capabilities to scale beyond the level they were designed to operate at.

More integrated resources exist, such as GenomeMap (Sato and Ehira, 2003), GenomeViz (Ghai R, 2004), Genome2D (Baerends et al., 2004), the Microbial Genome Viewer (Kerkhoven et al., 2004), ChromoViz (Kim et al., 2004), and GeneViTo (Vernikos et al., 2003). We reviewed each of these resources in turn with respect to our requirements. Of these, GenomeMap (Sato and Ehira, 2003), GenomeViz (Ghai R, 2004), and Genome2D (Baerends et al., 2004) only include bacterial genomes; the Microbial Genome Viewer (Kerkhoven et al., 2004) and ChromoViz (Kim et al., 2004) are only available as web-based tools; GeneViTo (Vernikos et al.,

2003) and GenomeAtlas (Pedersen et al., 2000) do not support gene expression data. Commercial resources, such as Spotfire (Spotfire) and GeneSpring (GeneSpring Analysis Platform) require expensive licenses that were not appropriate for a PhD project.

### 3.0.4 Architecture

SORGE is implemented as a three-tier architecture (see figure 3.0.4); a database layer, a data processing layer, and a GUI layer, each layer is further divided into encapsulated modules. SORGE can run as a stand-alone application or via the command line.

The database layer consists of SORGE DB and associated SQL-queries (defined in an XML-file). The data processing layer has three parts; 1) the projection of probe identifiers onto gene identifiers, 2) the determination of expressed genes, and 3) SORGE DATA with region and function specific models. The GUI layer is both responsible for dealing with user interactions, but mainly for the implementation of SORGE Visualisation, which graphically super-impose gene annotations on chromatin data.

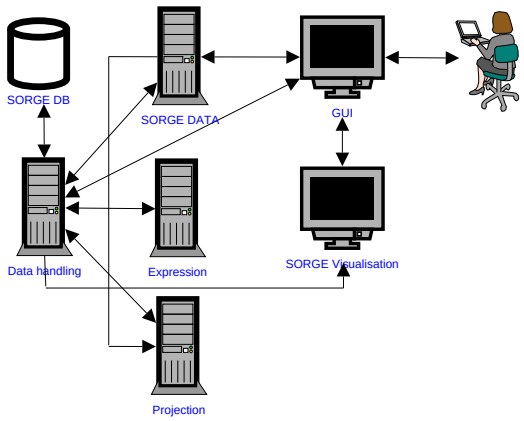



Figure 3.1: The three-tier architecture of SORGE with a database layer (SORGE DB and data handling), a data processing layer (SORGE DATA, Expression, and Projection) and a GUI layer (user interactions and SORGE Visualisation).

SORGE is a standalone package that requires a number of JAVA-libraries (Jakarta Maths libraries, PostgreSQL connection library, SAX, and JUnit) and a number of XML-files (config, db, SQL, and createdb), all of which come packaged with the program. In config.xml all paths, settings, microarray chips, and loci are specified. In db.xml, the databases, user-names, and passwords are specified. In sql.xml, all SQL-queries are specified as preprocessed statements. In createdb.xml, the db region is specified as *create table*-statements.

#### 3.0.4.1 Technology

At first, we assumed that SORGE would only consist of SORGE DB and SORGE DATA, and therefore SORGE DATA was implemented in Perl. The benefit of Perl is that it is quick to implement small programs, but as we later realised that three additional parts were needed; 1) projecting probes to genes, 2) retrieving gene annotations, and 3) visualising chromosomes, we choose to re-implement SORGE in JAVA version 1.4. The choice of the object-oriented language JAVA has been invaluable due to the *research* component, new data and requirements are constantly added, leading to an iterative analysis, design, and implementation phase. Although JAVA is not the obvious choice in a speed critical application it is acceptable, especially in combination with C, Perl, and MatLab where necessary. SORGE Visualisation has greatly benefited from an OO language; it started out as a very simple application; now SORGE displays transcript orientation, gene names, colors genes according to a number of features, and provide additional data, such as interaction partners. Furthermore it had to be possible to make SORGE publicly available, which meant that SORGE had to be platform independent which made JAVA a good choice.

The optimisation of SORGE DB was crucial in order to fulfill the first requirement; speed. SORGE DB is an in-house modular PostgreSQL 8.1.8 warehouse. A warehouse does not depend on external networks or connection speed, the drawback to a warehouse is the time and effort spent to keep the warehouse up to date.

## 3.1 SORGE DB

| Genomic data | Functional annotation |
|---|---|
| Probes | GO description data |
| Transcripts | UniProt description data |
| Exons | Synonyms |
| Genes | Pathway associations |
| Chromosomes | Molecular interactions |

Table 3.1: Division of the data into a database for genomic data and one for functional annotations.

The database layer, SORGE DB, is a modular warehouse that consists of two separate databases; 1) a species specific database (mouse), and 2) a database focused on biological function, including molecular interactions and pathway associations. The database division was made to make it easy to add additional genomes, for example human or worm, later. This division also

speeds up data retrieval and makes it easier to overview the tables.

### 3.1.1 The genomic database

#### 3.1.1.1 Requirements

This database was designed to answer the following questions:

1. *What genes are present in a specific genome?*
   (a) Distinguish between chromosomes and/or strands
   (b) Gene locations
2. *What genes are active?*
   (a) When is a gene active? (see section 3.2.2)
   (b) Connect gene expression data to project descriptions and analysis method

3 *What genes fall into a RIDGE?*

4 *Is the length of a RIDGE equal to a chromatin loop?*

   (a) What is the physical length of a genomic sub-region? (see section 3.2.3)

5 *Are the potential sub-regions similar to operons in their regulation?*

   (a) Do a group of genes have a high sequence similarity in their gene sequences?
   (b) Do a group of genes have a high sequence similarity in their upstream regions?
   (c) ...and if so is it because they share transcriptional factor binding sites?
   (d) How complex is the transcriptional control of a sub-region ?

In order to answer the above questions, the database stores data about *Arrays* (2a and 2b), *Chromosome* (1, 3, and 4), *Gene* (1, 2, 3, and 5), and *Project* (2b). Derived data include Gene Expression Profiles (2), External identifiers (2 and 5), and Common names (5).

ClustalW has been used to determine if genes are in fact sequence duplications of each other (5) - do the genes in a sub-region have a high similarly score. PROMO and TRANSFAC (see section 2.2) are used to determine if the genes share, known, transcription factor binding sites (TFBS). The number, and features, of transcripts, introns and exons are used to determine if the transcriptional complexity is considered complex, or not. One of the drawbacks with Ensembl is the extent of alternative transcripts, in fact the genome average is $1.21 \pm 0.75$, and 3297 genes actually lack a known transcript.

#### 3.1.1.2 Ensembl as the main data source

Publicly available whole genome databases were investigated with respect to the requirements; Entrez Gene (Maglott et al., 2005) at NCBI, Ensembl (Birney et al., 2006) at EBI, USCS Genome Browser (Kent et al., 2002) at USCS Genome Bioinformatics, MGI (Eppig et al., 2005) at Mouse Genome Informatics, GeneLynx (http://www.genelynx.org) and GeneCard (Rebhan et al., 1997). There are also databases requiring licenses, such as ERGO (Overbeek et al., 2003) at Integrated genomics.

| Source | Name | Type | Syn. | Location |
|---|---|---|---|---|
| Entrez Gene 12/09/2007 | Proteasome (promose, macropain) subunit, beta type 9 (large multi-functional peptidase 2) | Protein coding (PC) | Lmp2 Lmp-2 | 17 B1; 18.59 cM |
| Ensembl release 46 | (same as Entrez) | PC | | 33.792.336- 33.797.608 |
| USCS jul2007 | Proteasome (prosome, macropain) subunit beta | | | 34.319.044-34.324.275 |
| GeneLynx v 1.99 | (same as Entrez) | | | 17 18.59 cM |
| MGI 12/09/2007 | (same as Entrez) | | Lmp2 Lmp-2 | 18.59 cM; 33.792.336- 33.797.608 |

Table 3.2: Example data from the different potential data sources; here data for the protein coding region of Psmb9 on chromosome 17 in mouse is shown.

Table 3.2 shows that neither GeneLynx nor Entrez are appropriate since they only provide the map position, and not the actual physical location. Furthermore, MGI and GeneCard are species specific databases *(Mus musculus* and *Homo sapiens* respectively) and were therefore not selected.

That left Ensembl and USCS, and of these I choose Ensembl for a number of reasons, including accessibility to the chromosomal location, sequence data for chromosomes, genes, exons, and upstream regions, and the number of external identifiers (including Affymetrix probe identifiers). Additional benefits include the fact that Ensembl is part of both the CCDS and the VEGA projects. The CCDS project ((NCBI)) is a collaborative effort between EBI and WTSI (Birney et al., 2006). The VEGA project (Loveland, 2005) is a manual effort to curate the whole-genome with genes and splice variants. One drawback of using Ensembl is the lack of gene names via the BioMart interface (Durinck et al., 2005).

For a warehouse, one of the most important design decisions is how to keep the database up-to-date. This project uses downloaded text-files, via the BioMart interface. Ensembl has

further access to a) database dumps, b) tab-delimited reports, c) APIs for common programming languages such as Java and Perl, d) DAS, e) CVS, and f) FTP. (Birney et al., 2006; Dowell et al., 2001) Thus it would be possible to implement proper extraction tools if desired. UCSC only comes with a web-interface and a database dump thereby limiting the extraction choices. (Kent et al., 2002)

#### 3.1.1.3 Database structure and population

*Chromosome*-centric data include DNA sequences {*Chromosome_seq*}, length, centromere positions, and telomere positions. *Gene*-centric data include common name {*GeneName*}, physical position (start and end), strand orientation, transcripts, and exons. Further, *Gene* has two additional derived data columns, length and midpoint. *Exon*-centric data holds data about the transcript, and the positions of the exons. Array-centric data {*Affymetrix, Arrayes*} include provider, species, and individual probes. *Blast* stores the result of a specific blast run (probe vs. transcript or transcript vs. probe). A lot of mapping data, for example {*GO*, *UniGene*} and external identifiers, are stored in a number of mapping tables.

The ER diagram for the project module of the genomic database can be seen in appendix (**??**). *Project*-centric data is primary investigator and project description. *Treatment*-centric data stores the meta-data about the specific biological condition and analysis parameter settings. *ExpressedGene* stores the genes that were considered active for the current biological condition and analysis, whereas *Ridge* stores the genes that fall inside a RIDGE, and whether or not the current gene is considered silenced for the current biological condition.

The data for SORGE DB are obtained both from primary experimental studies, manually curated data, and publicly available data. A unified data-model is created and the database is populated via the BioMart approach which consists of three steps; 1) manually choose the data and format, 2) download the resulting text-files, and 3) run the software that extracts and stores the data. Stein et al (Stein, 2003) claimed that the most ambitious attempt at a warehouse database structure survived a year before collapsing because the import software had to be rewritten on average once every two weeks. This project combined human sequence data with genetic and physical maps.

Most entities {Color Red in Figure 3.2} are created by two JAVA files (CreateDB and Insert), some are derived {*GeneName*, *Translate;* Color Green} two created by a Perl call from Insert {*Chromosome_seq*, *IMGT;* Color Blue} (heavy string parsing required), {*Translate*} also gets data from the BLAST program (Ye et al., 2006) {*Blast;* Color Purple}, and one table is populated manually {*Arrayes;* Color Yellow}. All entities in the project part (except *Gene*) {Figure A.1} are populated through the analysis phase (section 3.2.3). The bootstrap section (Figure A.2) is populated via the bootstrap methods described earlier (section 2.3).

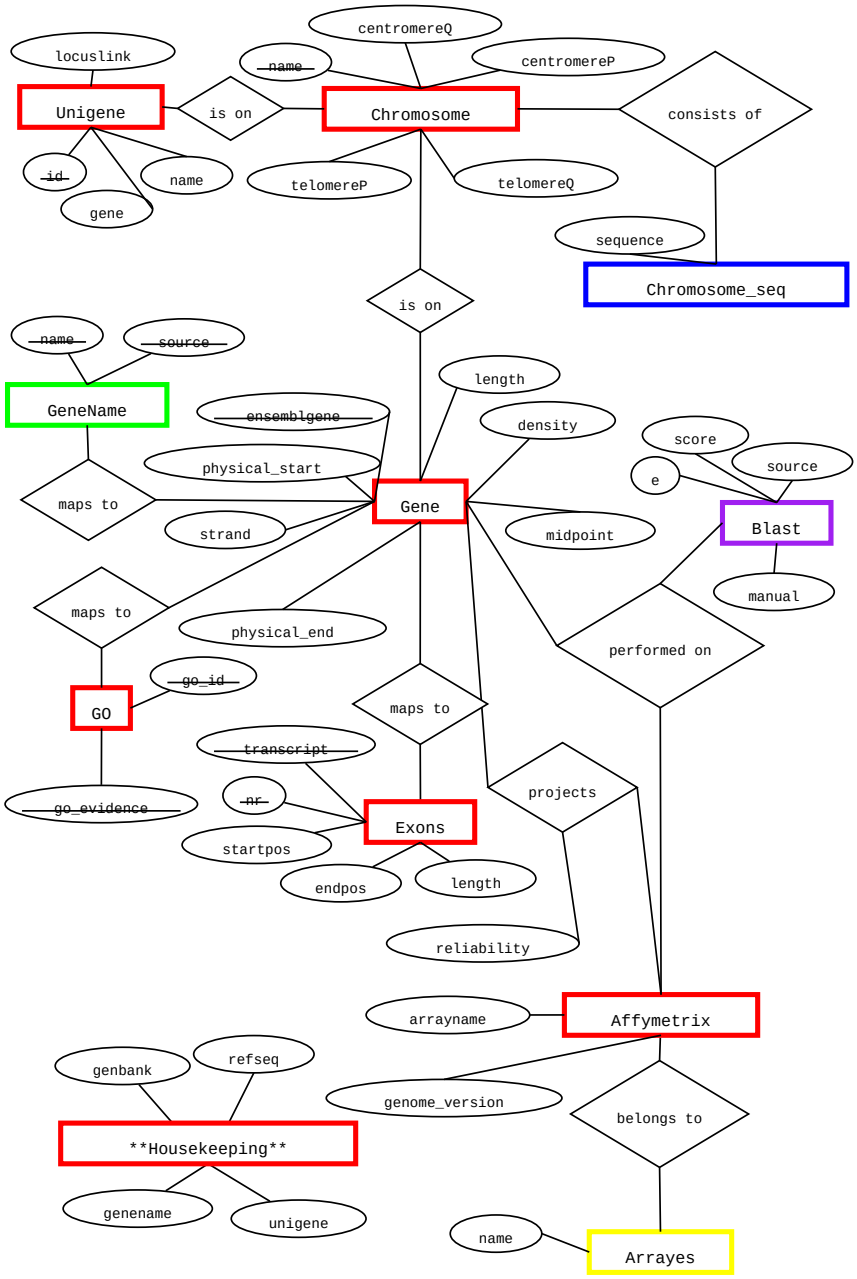



Figure 3.2: ER diagram of the genomic database. Red entities are automatically created, and populated, via JAVA. The green entity contains derived data (for optimisation purposes). The blue entity is created via Perl (heavy string parsing). The purple entity is created via the BLAST sequence comparisons. The yellow entity contains array specific information from Affymetrix.

#### 3.1.1.4 Optimisations and bottlenecks

There are four main optimisations attempts: 1) indexes on columns, 2) columns and tables storing derived data, for example *Gene*:Length, *Translate,* and *Gene Name*, 3) prepared statements, and 4) preprocessing of datasets. The preprocessing of a dataset is simply the projection of the probe identifier onto an Ensembl Gene Identifier and the storage of the latter rather than the first. In spite of these optimisations the retrieval of all known biological annotation for an entire chromosome is still a time consuming task. Each gene takes about 2 seconds, multiply this with around 1000 genes on a single chromosome (2000 seconds = 33 minutes) and it is a bottleneck, although because of heavy database, and code, optimisation it only takes between 6 and 7 minutes to visualise all known biological annotation on top of all the genes in a specific chromosome.

#### 3.1.1.5 The bootstrap database

Section 2.3 describes the performed permutations; in effect gene positions within the genome were randomised and potential sub-regions stored in the database. All tests are then performed on this permutation data. This makes it possible to compare results across tests and to add additional tests later. These results are also stored in the bootstrap database (see figure A.2). Basically the analysis parameters are stored in a table similar to *Treatment*, followed by the genes that fall inside a *Ridge*; these tables are re-created inside a new bootstrap schema in order to easily separate real RIDGE data from randomised data.

### 3.1.2 The database of functional annotation

The second database within SORGE contains functional annotation used to classify genes, primarily focused on immune related biology. This database is further divided into two logical modules, one with external data, and one with manually curated data.

The motivation for this separate database was speed although it is still slow. A molecular interaction database was needed in order to store more specific molecule types (normally few types, for example gene, protein, complex) and, more importantly, to be able to store more specific references (in order to quicker access the reason a specific interaction was added to the database).

#### 3.1.2.1 Requirements

1) *Are RIDGE members functionally related?*

   (a) Do the genes participate in the same biological process/pathway?

   (b) Do the genes exhibit similar function(s)?

(c) Do the genes act in the same cellular location?

Ingenuity (Ingenuity Systems), DAVID (Dennis et al., 2003; Huang et al., 2007), and KEGG (Kanehisa, 1997; Kanehisa and Goto, 2000; Kanehisa et al., 2006) are all used to determine if the genes in a genomic sub-region participate in the same pathway. (see section 2.2.2.2) GO, UniProt, GeneCard, and MGI are used to determine if genes share a biological function, in addition to the manually curated molecular interaction data. The sub-cellular location of a gene is currently taken from the UniProt tissue specificity field. Each sources can suggest biological function(s), it is up to an expert in the field to determine if the genes have a functional correlation.

#### 3.1.2.2 Database structure

The database structure for both derived and manually annotated data can be found in the appendix (see figure A.3).

The biological function of a gene is mainly derived from GO (which is a common way to determine gene function (McCarroll et al., 2004)) and UniProt.

**GO** (The Gene Ontology Consortium, 2000) provides biological process, molecular function, and cellular component.

**UniProt** (Bairoch et al., 2005) for example provides allergen, catalytic activity, cofactors, developmental stage, disease, domain, enzyme regulation, function, pathway, polymorphic, similarity, sub-cellular location, subunit, tissue specificity, and toxic dose.

In addition, the gene name also contains relevant data, for example Psmb9 [Proteasome (prosome marcopain) subunit, beta type 9 (large multifunctional peptidase 2)]. In addition data from GeneCard (Rebhan et al., 1997) include name, symbol, alias, family, previous names and symbols, data from MGD (Eppig et al., 2005) include names and symbols. The expert then decides what function(s) to assign to a gene.

Biological function can also be derived from interaction partners, the manually curated interaction data comes from these articles (appendix B). An interaction, with type and location, consists of two molecules and an associated PubMed reference. These interactions are then classified into types according to their biological level (such as genetic interactions or protein interactions). The interaction type makes it possible to extend simple interactions into more complex pathways across biological levels. Interactions are further defined by associations between the molecules such as acetylates, attracts, cleaves, enhances, inhibits, recruits, requires, or transports.

*Laboratory*, *journal*, *date*, and *reason* (for reading this paper) are used to determine the validity of the paper. *Cell line* and *cell type* (when available) are used to prioritise interactions,

for example, this PhD project focuses on macrophages, therefore results from the cell line macrophage are more valid than, for instance, fibroblast result. The *page* number, the *column* (left—middle—right) and the *paragraph* are used to quickly find the relevant section in a paper. The more meta data available to the expert the more valid the decision to rely on, or to ignore, a certain interaction is. In order to quickly come up with a set of interactions of interest for the MHC class II antigen presentation pathway, the database also hold cited data (*original_cited*).

#### 3.1.2.3 The molecular interaction database

A molecule is defined as one of the following 16 kinds {APC, cell, complex, DNA, enzyme, family, gene, interaction, location, molecule, peptide, protein, receptor, regulatory element, result, other}, this gives more specificity than the normal 3-4 kinds. For example KEGG (Kanehisa, 1997; Kanehisa and Goto, 2000; Kanehisa et al., 2006) has gene, protein, enzyme, compound, and map, whereas Reactome (http://www.reactome.org) has PhysicalEntitiy, CatalystActitivy and Event (although each of these then comes with textual sub-specifications). Of these 16 molecule types; *result* is a free text "final state", for example "Development of the thymus TCR repertoire"; *location* specifies a cellular location, for example the "ER", the "nucleus", or the "cell surface"; *molecule* are those molecules that I have not been able to classify; and *other* are those that do not have a category on its own, for example "IFN-gamma treatment", "Amino Terminal", "Jak-Stat pathway". Finally every interaction is considered a molecule (of type *interaction*) and this is how we can build networks of interactions.

An interaction is restricted to 44 different types, in alphabetical order {acetylates, antagonists, and, assembles, associates with, attracts, binds, cleaves, contains, detects, determines, digests, does not control, encodes, enhances, exists, has, homologs, in, inhibits, initiation of, interactions, leads to, orthologs, paralogs, partners, part of, phosphorylates, presents, produces, proteolyses, recognises, recruits, regulates, relates to, releases, removes, required for, requires, sensitive to, signals via, to, transports, other}.

Together these two tables allow the user the flexibility to add any kind of molecular interaction, yet the specificity to exactly determine what, why, and where an interaction happen (Table 3.1.2.3 and 3.1.2.3).

n_Interaction

| ID | Mol. | Interaction | Location | Text | Partner |
|---|---|---|---|---|---|
| 768 | 1475 | produces | Binding groove | | 335 |
| 766 | 476 | inhibits | | | 1375 |

n_Molecule

| ID | Name | Type | Free_Text |
|---|---|---|---|
| 335 | CLIP | Peptide | |
| 1475 | 766 | Interaction | |
| 476 | Catin L | Molecule | Lysosomal proteases |
| 1375 | CD74 | Protein | MHC class II-dedicated |

Interaction example 1: Inhibition of the protein CD74 by Catin L leads to the production of the peptide CLIP in the binding grove.

n_Interaction

| ID | Mol. | Interaction | Loc. | Text | Partner |
|---|---|---|---|---|---|
| 560 | 98 | And | | | 2 |
| 561 | 1066 | Enhances | | Transcription of | 1 |

n_Molecule

| ID | Name | Type | Free_Text |
|---|---|---|---|
| 98 | CBP | Protein | |
| 2 | CITTA | Molecule | |
| 1066 | 560 | Interaction | |
| 1 | DRA | Gene | |

Interaction example 2: The protein CBP in combination with the molecule CIITA enhances the transcription of the gene DRA

Table 3.3: Two interactions from the database

The molecular interaction database contains 1.035 interactions with 852 molecules (out of 989 molecules) from 53 articles (appendix B). These interactions have been used to determine the gene function, but also helped determine the MHC loci borders where the subjective "sum of the functional annotation" have been used to determine, if a gene should be considered a part of the loci. For example the gene Atp5g2 is probably not part of the MHC since it has no known immune associations. In contrast the gene Tnf (GO:humoral immune response, positive regulation of I-κB kinase/NF-κB cascade, programmed cell death and interacts by signaling via c-Rel (Ting and Trowsdale, 2002)) is.

The molecular interactions from the database were uploaded into BioLayout. It became apparent that two of the largest interactions hubs were CITTA and CBP (yellow circles in figure 3.3), as is expected since CIITA is the MHC class II transactivator (Spilianakis et al., 2003; Masternak et al., 2003), and the CBP genes are co-activators, nuclear factors, and involved in histone acetylase (Spilianakis et al., 2003). Another feature is the fact that the molecule type *result* fall on the outside of the interaction diagram.

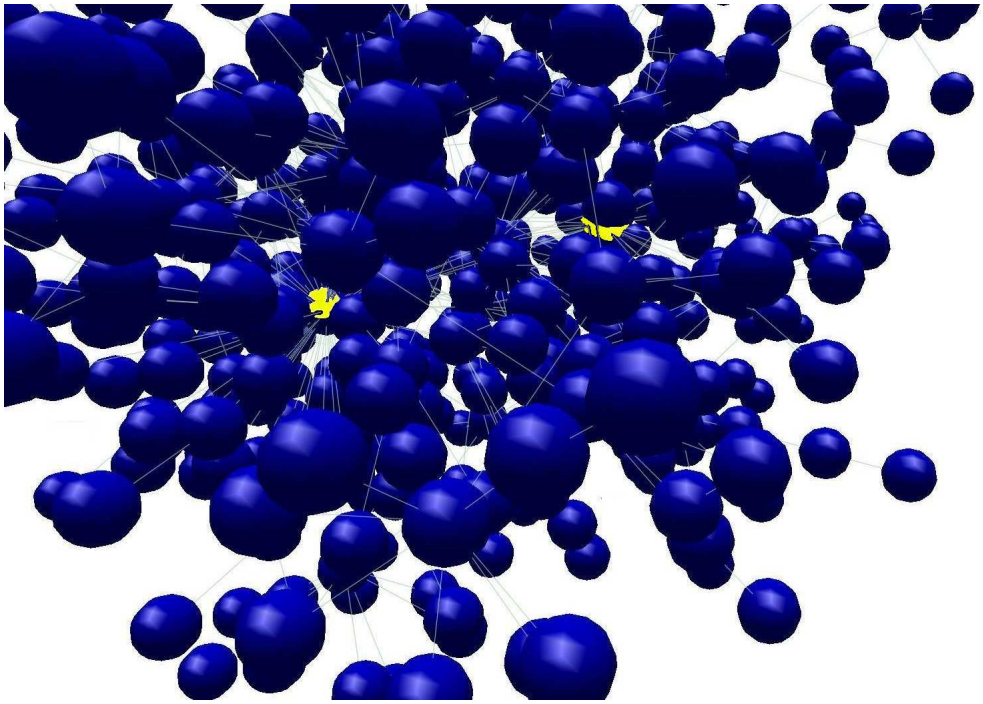


Figure 3.3: The interaction map from BioLayout representing the molecular interaction database. From this figure, the two hubs (CIITA and CPB (Spilianakis et al., 2003; Masternak et al., 2003)) of the MHC class II antigen presentation pathway are easily spotted (here highlighted in yellow). Spheres on the outside of the interaction diagram are usually of the molecular type *result*.

The manually curated interaction data has also been used as a 'requirement specification' for a large pathway database that is under development at DPM.

## 3.2 The data processing layer

The implementation of the data processing layer for SORGE has been relatively straightforward, mainly because it was first, partially, implemented in Perl.

### 3.2.1 Probe-to-gene projection

The genome annotation have been refined since Affymetrix generated the probe sequences, therefore many genes now have no probes and many probes lack specificity. To map between the gene identifiers in Ensembl and microarray probe identifiers, it is possible to cleanse duplicate or ambiguous entries, as Kirov et al did. (Kirov et al., 2005) However, it is advantageous to retain this data because it can be used to score probe-to-gene projections.

#### 3.2.1.1 Other methods

There are large quantities of suggestions for how this projection should be implemented;

- Verdugo et al (Verdugo and Medrano, 2006) mapped the vendor-provided gene symbols onto an unique Entrez gene identifier for mouse. The probe sequences were not used since not all platforms supply the sequence of a probe, but they did suggest that this would be an better approach, which is what we used here.
- Genes in the same UniGene cluster, and with the same chromosomal coordinates were joined, as were clusters of genes in the same orientation with at least one common exon, and gene clusters within 1 Mbp. A reliability scale was used when the cluster mapped to multiple chromosomal locations (RefSeq, mRNA, spliced EST). A gene with only a single EST was removed from the cluster. (Fujii et al., 2002)
- RefSeq and Ensembl was used for mouse. (Cheung et al., 2003)
- GeneALaCart associated Affymetrix HG133A probes with a unique human gene (in the GeneAnnot database) and to Entrez gene IDs. (Coppe et al., 2006)
- Ensembl and CrossMatch sequence comparisons resulted in a 67% coverage of the Ensembl human dataset. (Vogel et al., 2005)
- The SAGEmap tag-to-gene-mapping-table assigned RefSeq to LocusLink and SAGE to UniGene. The UniGene clusters were then mapped down to human chromosomal positions. (Yamashita et al., 2004)
- UniGene clusters mapped to LocusLink and Ensembl genes (via EnsMart), the LocusLink IDs that mapped to multiple UniGene clusters were discarded, (both mouse and human). (Singer et al., 2005)
- Adjacent gene pairs, within 10 Mbp, were collapsed into TRIBE families, linking mouse Ensembl genes to Affymetrix tags. (Singer et al., 2005)
- GenBank, FlyBase, Resourcered was linked at TIGR (Fukuoka et al., 2004),
- WormBase (Blumenthal et al., 2002),
- Ensembl, WormBase, FlyBase, Arabidopsis, KEGG, BLAST, and MSPcrunch could be integrated (Lee and Sonnhammer, 2003),
- HOVERGENE, LocusLink, Mouse Genome Informatics (MGI), UniGene, and EST data was connected (Williams and Hurst, 2002),
- GenBank and PILEUP (Williams and Hurst, 2002)

- LENS (Linking ESTs and their associated Name Space), Human Gene Index at TIGR, BodyMap, UniGene, and LDB are additional suggestions. (Bortoluzzi et al., 1998)

#### 3.2.1.2 Implemented method

All suggested identifier-to-identifier projections are stored to the database in addition to a BLAST of all probe sequences against the current genome release, and vice verse (for a discussion about thresholds see 2.2.3.3). These projections are stored in the *Translate* table (3.2), and assigned a reliability score based on our confidence in each. For instance, alignments between the probe sequence and the genome are assigned a higher weight, whereas the UniGene transcriptional sequence-to-locus assignments (Pontius et al., 2003) are assigned a lower weight. Furthermore, the more ambiguity in a projection the lower the reliability score.

| Source | URL and Identifiers |
|---|---|
| Ensembl | http://www.ensembl.org/Multi/martview<br>Ensembl, GO, Entrez, External, RefSeq (RS), UniProt (UP), Affymetrix, Agilent |
| UniProt | http://www.ebi.uniprot.org/database/download.shtml<br>UP |
| UniGene | ftp://ftp.ncbi.nih.gov/repository/UniGene/Mus_musculus/Mm.data.gz<br>UniGene (UG), GenBank (GB), LocusLink/Entrez |
| GeneCard | http://www.gene.ucl.ac.ul/nomenclature/data/gdlw_index.html<br>HGNC, Entrez, GB, RS, Ensembl, MGD |
| MGD | ftp://ftp.informatics.jax.org/pub/reports<br>MGD, Entrez, GB, RS, UG |
| NetAffy | http://www.affymetrix.com/support/technical/byproducts.affx?cat=array&Mouse<br>Affymetrix, Ensembl, Entrez, GB, RS |
| GNF | http://symatlas.gnf.org/SymAtlas<br>GNF, Ensembl |

Table 3.4: Data sources for the probe-to-gene projections

#### 3.2.1.3 Evaluation of data sources

The comparison of data sources was based on “hit” ratios and coverage. Hit ratio compares the number of reliable projections to the number of total projections per data source. Coverage compares the number of reliable projections to the number of genes in the genome.

| | MG-U74Av2 | | Mouse430_2 | |
|---|---|---|---|---|
| Source | Hit Ratio | Coverage | Hit Ratio | Coverage |
| Affymetrix | 93.3% | 25.49% | 96.11% | 52.16% |
| BLAST | 94.36% | 26.48% | 95.92% | 57.32% |
| Ensembl | 90.65% | 27.34% | 93.23% | 59.60% |
| Entrez | 94.42% | 28.95% | 96.73% | 57.62% |
| GenBank | 92.91% | 16.71% | 96.87% | 19.23% |
| UniGene | 90.44% | 26.32% | 94.54% | 58.59% |

Table 3.5: Hit ratio and genome coverage for the MG-U74Av2 and Mouse 430 2.0 chips per data source.

Entrez has the best hit ratio (although closely followed by BLAST) for the MG-U74Av2 chip, whereas GenBank has the best for the Mouse 430 2.0 chip and UniGene the best coverage. The worst hit ratio is seen for GenBank and Ensembl respectively, and the worst coverage for GenBank for both.

#### 3.2.1.4 Results

Only 8817 genes (31.3%), out of the 28157 in Ensembl version 44, are unambiguous represented on the MG-U74Av2 chip, the rest are either control genes, project to multiple Ensembl genes, or lack a probe altogether. This means that less than one third of the genes have associated gene expression levels for this chip. The best coverage is provided by the Agilent chip, and the least by the MG-U74Av2 chip.

| Chip | Probes | Mapped | % | Coverage |
|---|---|---|---|---|
| MG-U74Av2 | 12488 | 8445 | 67.62% | 29.99% |
| 430_2 | 21179 | 17569 | 82.96% | 62.40% |
| Agilent | 19659 | 19580 | 99.60% | 69.54% |
| GNF | 14975 | 14941 | 99.77% | 53.06% |

Table 3.6: Genome wide coverage per microarray chip

The model implemented by Verdugo et al (Verdugo and Medrano, 2006) resulted in a genome coverage of 75.6% with the Affymetrix 430 2.0 chip, whereas here a coverage of 63.3% is found. The difference is that this method associates each projection with a reliability score. So although the total coverage is less than for the Verdugo approach, this method made it apparent that a lot of ambiguity exist for the excluded projections.

Both under, and overrepresented regions are found. One example of the former is the protocadherin locus on chromosome 18 (where only 10% of the genes are reliable projected, as seen in figure 6.8). One example of the latter is the MHC locus on chromosome 17 (where 40% of the genes are reliable projected).

### 3.2.2 Determination of active genes

#### 3.2.2.1 Method

There are a number of possible methods for determining active genes;

1. The detection flags set by Affymetrix. The user has to choose if *all* or the *majority* of detection calls should be set to present. If no majority of detection call could be reached, then the gene is assumed to be marginal and M is returned. For example; {P, P, P, M} will return P if on the *majority* setting, but M if *all* should be used, whereas {P, P, M, M} would return M in both as would {P, P, A, A}.

2. The expression values - Genes with an *intensity value* above or equal to the specified threshold will be considered active. SORGE provides the option of using all replicates or the mean expression level.

3. Change from control to treatment - Compares two treatments to each other (for example control vs. treatment). This requires that the user set a p-value and a fold-change for when a gene should be considered differentially expressed.

(see 4.2.1.1 for a discussion on how the method influence the results)

#### 3.2.2.2 Replicates

| Gene | Chrom | Rep1 | Rep2 | Rep3 | Mean | Median |
|---|---|---|---|---|---|---|
| ENSMUSG00000051853 | 15 | 175.1 | 172.57 | **188.6** | 178.76 | 175.1 |
| ENSMUSG00000068615 | 2 | 3.8 | **10.1** | 2.2 | 5.37 | 3.8 |
| ENSMUSG00000017721 | 2 | 188.3 | 166.2 | **208** | 187.5 | 188.3 |
| ENSMUSG00000030245 | 6 | **68.1** | 40.4 | 43.5 | 60.67 | 43.5 |
| ENSMUSG00000024124 | 17 | **66.2** | 21.1 | 25.3 | 37.53 | 25.3 |
| ENSMUSG00000040610 | 11 | 5.4 | **8.2** | 2.9 | 5.5 | 5.4 |
| ENSMUSG00000039236 | 7 | 68.3 | **75** | 67.8 | 70.34 | 68.3 |
| ENSMUSG00000050347 | 7 | **2340.2** | 2202.6 | 2133.1 | 2225.3 | 2202.6 |

Table 3.7: The difference in using the mean or the median expression value.

The above table show 8 random genes, as taken from the uninfected macrophages. Usually a single replicate (in bold) exhibit a substantially higher signal than the other two replicates, and these are not due no sampling errors, since they are not found for the same replicate. This means that the median is more appropriate than the mean, since it is less influenced by outliers.

Gene ENSMUSG00000030245 (in magenta) would be considered active if a threshold value of 50 and the mean expression was used but would be considered silenced if the median was used. Gene ENSMUSG00000040610 (in blue) is on the other hand an example of very good correlation between the median and mean. For this study the median value has been used.

### 3.2.3 SORGE DATA

A JAVA interface *Analyser* forces all child classes to *BasicAnalyser* to implement an *analyse* method that returns an array of genomic sub-regions, making it easy to add additional analysis methods. The GUI class *Run* deals with the user-defined parameters and calls the appropriate sub-class.

#### 3.2.3.1 Loop sized based RIDGE analysis

The distance based method requires the user to specify the window size, which here is represented as a midpoint and a radius.

**Algorithm 1** Pseudo code for the RIDGE detection algorithm

```
Get the expressed genes for the current treatment
for each chromosome in the genome
    sort the genes according to midpoint [start + (length/2)]
    for each gene on the current chromosome
        if the current gene is expressed
            if (potential + current gene) < (center + radius)
                add to potential
        if not
            if (potential + current gene) > (center - radius)
                save
    if not
        if current gaps < allowed gaps
            add gene to potential as GAP
            increase current gaps
        else
            if potential RIDGE within length requirement
                save the potential RIDGE
    if genes in potential
        move the "pointer" to the second gene in potential
```

#### 3.2.3.2 Conclusion

Initially the goal of this project was to use these simple methods as a basis for more complicated algorithms. This soon became unfeasible as we realised the extent of the missing data and the uncertainty in the analysis stage. Even on a whole-genome microarray chip a lot of the genes lack expression data, the determination of expressed genes influence the results, there are not enough transcripts available in Ensembl, determining transcriptional binding sites is a PhD project on its own, the manual curation of molecular interactions took time as well. As we did not foresee that these would be so difficult and time-consuming we soon changed the focus to the analysis of biology instead. SORGE is written in such a fashion that adding additional methods is straight-forward.

## 3.3 SORGE Visualisation

The SORGE framework is divided into three parts, SORGE DB, a data processing layer, and a GUI layer. The latter is further divided into two modules; SORGE Visualisation and user interaction, I/O, classes.

SORGE Visualisation provides a schematic 2D representation of a chromosome, the "chromosome view", overlaid with both textual and graphical annotations. Features include gene data such as names, location, transcription orientation, length, function, and molecular interaction partners; and gene expression level(s) from a set of user-defined treatment(s). (figure 3.5 and 3.6 respectively)

SORGE Visualisation supports the possibility to:

- Find interaction partners of a gene.
- Display results from multiple experimental conditions - for example a time-series data set using the bar plots. If extra annotation (such as Affymetrix present, absent, or marginal calls) is available then the data are color coded accordingly. These plots can be shown independently or overlaid onto the chromosome view.
- Color code genes in the chromosome view according to a number of features; for example gene *expressed* or *down regulated* (which is dependent on the current parameter setting and the experimental condition) or interaction partners. Importantly, genes that do not have a good *gene-to-probe-projection* are highlighted with an alternative colour.
- Zoom in/out on a genomic region.
- Specify genomic regions, loci, or special interest, for example the MHC class II locus.
- Save graphs and plots as PNG files.

- Store and print functional annotation for a gene/set of genes.

But most importantly SORGE provides a convenient method of performing on-the-fly analysis, visualisation, and categorisation of genomic sub-regions.

### 3.3.1 Method

#### 3.3.1.1 The GUI layer

The GUI layer is implemented in a separate folder and contains a main class (*MainGUI*) with its children, where each child is responsible for a specific action. For instance the visualisation of chromosomes and expression levels is contained in *GUI_Display*; sequence comparisons via ClustalW in *GUI_SeqSim*, and the report functions in *GUI_LocusReport.*

#### 3.3.1.2 SORGE Visualisation

SORGE Visualisation is responsible for the chromosome representation and associated expression plots. There are two "main" classes; *SORGE* and *UserView*. *UserView* accepts a number of user parameters and calls the appropriate classes/methods, the main method mimic the behavior of *GUI_Display*. *SORGE* is called by *UserView* and is responsible for determining the layout of the chromosome representation. *UserView* can also call *ReadFromGFFFile* (before it calls *SORGE*), that reads in data from a GFF-formatted file instead of taking data from the database.

The three main classes of SORGE Visualisation are *ChromosomePanel, BasicRenderer*, and *Coloring*. *ChromosomePanel* show a single chromosome panel to the user. *BasicRenderer* renders the strands; if centromere data is available then show the chromosome as an X, otherwise as two horisontal lines. *Coloring* connects a gene with a color depending on the current parameter settings and user input.

#### 3.3.1.3 Gene Feature Format (GFF)

GFF-import is implemented in order to open up the usability of SORGE to additional fields.

| name of chromosome/gene | application | call | start | end | strand | | | optional attributes |
|---|---|---|---|---|---|---|---|---|
| chromosome:1 | SORGE | . | 0 | 10000 | . | . | . | . |
| gene:ENSMUSG00000000001 | SORGE | P | 10 | 150 | + | . | . | density “25” |
| gene:ENSMUSG00000000003 | SORGE | A | 50 | 78 | - | . | . | density “3” |
| gene:ENSMUSG00000000028 | SORGE | P | 63 | 453 | + | . | | color “green” ; density “36” |
| chromosome:2 | SORGE | . | 0 | 131654 | . | . | . | . |
| gene:ENSMUSG00000000031 | SORGE | P | 365 | 987 | + | . | . | density “56” |
| gene:ENSMUSG00000000037 | SORGE | M | 13 | 136 | - | . | . | density “25” |

Figure 3.4: Example of a GFF input file

A GFF file should have an identifier (chromosome name or gene identifier) a Present/Absent/Marginal call, the physical start and end position, the transcription orientation (strand). Then there are a number of optional attributes, the most useful of these being the color, which allows the user the ability to color genes according to a pre-defined criteria.

The ’color-by-filtered-gene-list’ attribute can be used to analyse and visualise additional features, for example the result of a cluster analysis. The optional attributes field could also be used to further annotate a primary experiment with user-defined comments

#### 3.3.1.4 Parameters

There are two groups of parameters, the choice of dataset and display parameters.

To select a dataset, choose the array (for example the MG-U74Av2 chip), the project, the control treatment, and finally *optional choose the treatment(s) to compare to the control.*

To select the output, choose

- which chromosome(s) to display, (for example chromosome 1 and 3)
- what genes should be displayed, (those in a genomic sub-region or all)
- *optionally focus the visualisation to one, or more, regions of interest*
- whether to show the gene common names in the panel, (default is to show)
- which parts the user wants to see:
  - a genomic view (show everything in a single window),
  - expression plots (only the expression plots are shown),
  - genes in RIDGEs (for comparisons between treatments),
  - a combination of the first three options (preferably only 2 treatments, otherwise the screen overflows),

- show a locus (will potentially divide the display over multiple windows),
- show the differentially expressed genes, or
- do not color genes.

#### 3.3.1.5 Platform changes

About a year into the project, the focus was shifted from the implementation of RIDGE models, to the validation of data, methods, ands results. This meant that the first choice of programming language, Perl, no longer was valid, also at this point the unstructured Perl code was unreadable. We therefore choose to re-implement SORGE DATA in JAVA, which is an object-oriented programming language with a well-defined GUI-component for user interaction. At this time, we encapsulated each RIDGE model, and each validation step, into a separate class, making it easy to add additional models and validations later; for example the probe-to-gene projection. Another highly used benefit of encapsulation is the ability to make large changes in the GUI with small changes in the code. We also added the model SORGE Visualisation at this time. SORGE has survived an upgrade of both PostgreSQL and JAVA, with only minor changes necessary for both forward and backward compatibility (see section 2.2.4).

### 3.3.2 Results

#### 3.3.2.1 Chromosome view

The chromosome view shows a schematic, linear, representation of a chromosome, or a genomic region/locus.

A configuration setting defines how much data to display in each chromosome panel, the default is 1.6 Mbp. Genes are graphically annotated with colors according to a number of criteria, for example gene *expressed* or *down regulated,* depending on the experimental condition(s). Importantly, genes that do not have a good *probe-to-gene-projection* are highlighted with an alternative colour, default grey, it is up to the user to define the threshold for a good projection. Additional features, such as interaction partners for a specific gene, can also be highlighted.

All data for a gene/protein in SORGE DB is available to the user via a simple left-click on the gene of interest.

#### 3.3.2.2 Expression Bar Plot view

The expression bar plot can either be displayed simultaneously with the chromosome view, or separately. Results can be displayed from multiple experimental conditions, for example along a time-course using bar plots in the bar plot view. If extra annotation is available (for example Affymetrix present, absent, or marginal calls), then the bars are color coded accordingly.

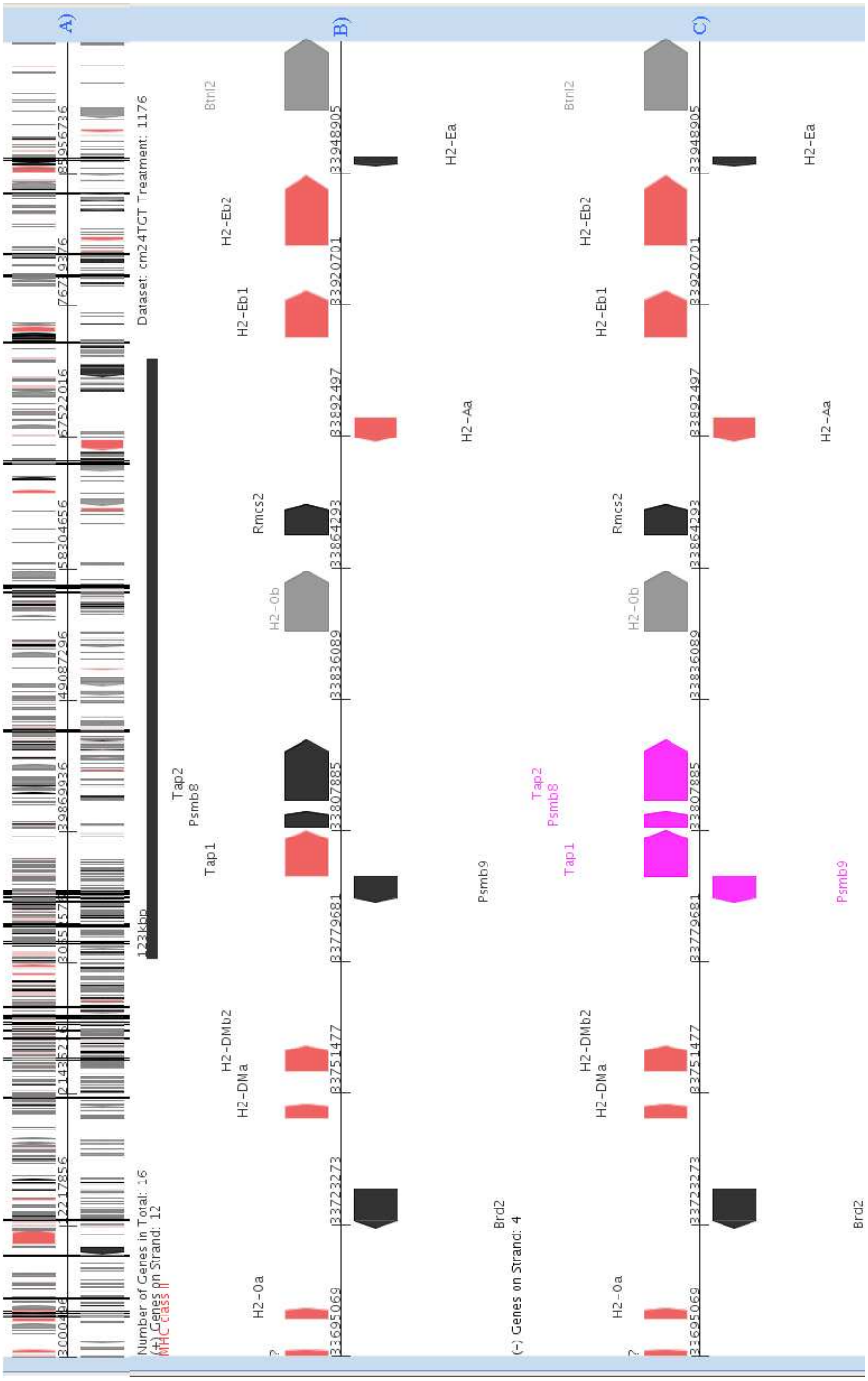



Figure 3.5: Panel a) shows the entire chromosome 17 in mouse. Expressed genes are shown in black and silenced genes in red. The grey genes lack a reliable probe-to-gene-projection. Panel b) zooms in on the class II MHC Loci (see text in the top left corner) on chromosome 17 and use the same color scheme as panel a. Panel c) demonstrate the visualisation of additional features through highlighting. For example; if you are interested in interaction partners for a specific genes (in this case Tap2) then simply right-click that gene and the interaction partners will be shown. For Tap2 in the MHC class II Loci the interaction partners are Tap1, Psmb8, and Psmb9 (Wagle et al., 1999; Chen et al., 1997)

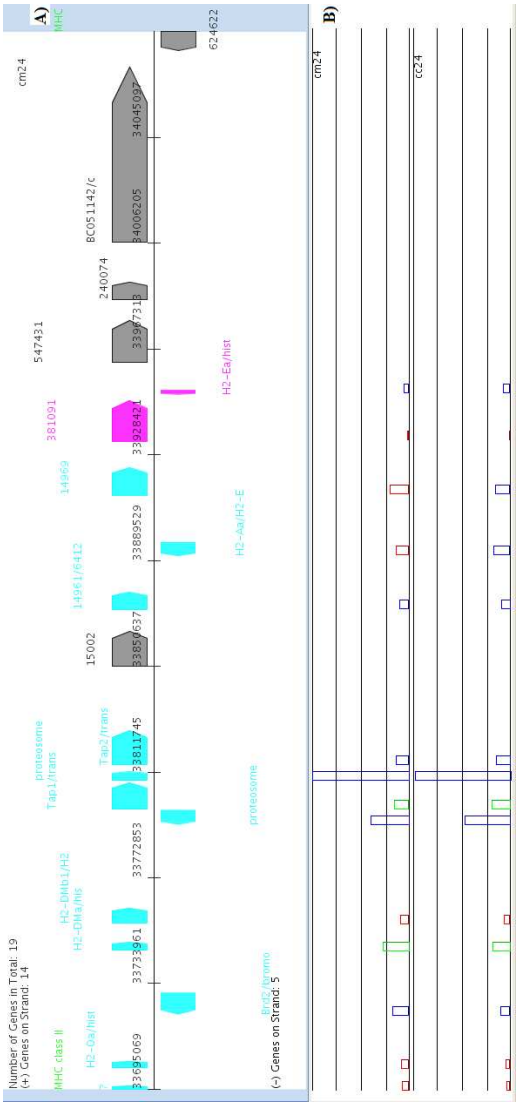



Figure 3.6: In panel b the expression plot for the uninfected macrophages (upper row) is shown in correlation to the expression levels for the viral activated macrophages. (here genes are considered active is the median intensity is above 50)

#### 3.3.2.3 Evaluation

Due to the time constraints imposed by a PhD project, we choose not to perform a full user evaluation of SORGE, since the implementation of the graphical part was not a central part of this project, just a necessary tool.

## 3.4 Functionality of SORGE

SORGE supports the following actions:

- Identify and visualise the relationship between genomic region(s), gene expression levels, and functional annotation.

  The visualisation of RIDGEs is the main use of SORGE.

- Map between a large set of identifiers.

  For example; unigene to ensembl, or genbank accession to uniprot.

- Different methods to determine active genes

  - Signal intensities, detection calls, and intensity values

- Identify all genes, expressed in a specific tissue, that share common function.

  Tissue specificity is suggested in UniProt, but common function of a set of genes is a subjective user action.

- Select all genes that are associated with a specific function, for example the immune system or the Jak-Stat pathway, grouped by chromosomal location.

  Genes are organised depending on the chromosome, strand, and physical position.

- Return the sequence and UTR region for a set of genes in FASTA format and either send to the sequence comparisons Blast (Ye et al., 2006) or ClustalW (Chenna et al., 2003) *or* retrieve potential transcription factor binding sites.

  These sequence alignments could be used to determine if a set of genes are in fact sequence duplications. PROMO (Messeguer et al., 2002) could be used to get shared transcription factor binding sites.

- Produce reports for a genomic region, containing any combination of features such as transcripts, proteins, gene function, molecular interactions, statistical evaluation

All ”relevant” data for a gene, or a set of genes, can be retrieved with a single method call, either from the GUI or from the commandline.

- Zoom in on a genomic region or loci.

  During the course of the analysis it became apparent that the genomic context, such as the MHC locus or the Hox locus, might influence gene expression, therefore SORGE incorporates the possibility to restrict an analysis, or a view, to a specific locus. Loci are defined in the configuration file, making it easy to add, remove, or update.

## 3.5 Discussion

Datasets obtained from high-throughput sequencing projects or post-genomic technologies, such as microarrays, have enabled an exploratory analysis at the genome level. There is an increasing need for biological researchers to contextualise, manipulate, and visualise ever more complex, interrelated data sets.

Existing tools developed for whole-genome analysis are usually either specific for bacterial genomes, web-based, or of limited availability due to license costs. Those specific for bacteria [GenomeMap (Sato and Ehira, 2003), GenomomeViz (Ghai R, 2004), and Genome2D (Baerends et al., 2004)] typically have scaling issues, since the bacterial genomes are smaller and simpler than eukaryotic. Additionally, the transcriptional complexity (introns/exons) of a gene might influence its regulation. Web-based genome resources [the Microbial Genome Viewer (Kerkhoven et al., 2004) and ChromoViz (Kim et al., 2004)] are not suited for exploratory analysis under multiple experimental conditions and different settings of parameters, or for inclusion in an analysis pipeline. It is notable that some resources (such as GeneViTo (Vernikos et al., 2003) and GenomeAtlas (Pedersen et al., 2000)) do not support gene expression data and therefore not the investigation of co-expression. Applications like Spotfire (Spotfire) and GeneSpring (GeneSpring Analysis Platform) fulfill most of these requirements but require an expensive commercial license.

In this study, testing the hypothesis that gene activity and chromatin architecture are correlated in eukaryotes (*Mus musculus*), as in prokaryotes (operons), required visualisation of co-regulated gene expression clusters on chromosome architecture. To determine if these gene clusters were functionally associated; a database of biological annotation, such as common names, synonyms, known functions, and interactions was created. Additionally, gene expression profiles, for multiple experimental conditions, were plotted to visualise trends in the data.

SORGE is a novel open-source genome visualisation tool optimised for use at the chromosome level. SORGE can be used for investigation of both eukaryotic and prokaryotic genomes.

SORGE allows the user to analyse and visualise complex, interconnected data sets spanning entire genomes. SORGE enables the integration of sequence and expression data in addition to providing a built-in exploration tool facilitating data query, evaluation, and export. SORGE can be used either as a standalone graphical desktop program, or accessed through the command line as part of an analysis pipeline. SORGE overlay the chromosome representation with a number of features including: loci annotation, genes, biological function, interaction partners, network associations and expression profile(s), from primary experimental results. All features can be exported via the report facility, including the results from the in-house sequence alignments. Although SORGE is specifically aimed at integrating genomic architecture with expression data, the flexible architecture and the use of a common data exchange format (GFF) provides a framework for extending the use beyond the initial scope. For example, SORGE could be used to visualise the result of a hierarchical clustering method.

SORGE is divided into three sections; SORGE DATA that defines the different RIDGE models, SORGE Visualisation visualises, and categories, these sub-regions, whereas SORGE DB hold the required data. The Ensembl database (Birney et al., 2006) stored the biological core data (physical location, chromosome sequence, external identifiers) and had an easy retrieval system, BioMart (Durinck et al., 2005) , additionally it includes most model organisms and is therefore used as the main source.

SORGE enabled us to locate potential Rosettes (Okada and Comings, 1979) in the mouse genome and provided evidence supporting the existence of chromatin regulation of gene order, as seen in the MLS/CT-IC (Munkel and Langowski, 1998; Albiez et al., 2006) model. A further benefit provided by SORGE is the assignation a reliability score to a projection between an Ensembl gene identifier and a specific Affymetrix probe.

# Chapter 4

# RIDGE definition and characterisation

In this chapter the RIDGE definition will be discussed in detail. Furthermore the statistical scores used to evaluate RIDGEs will be presented, for example the expected number of RIDGEs. Finally the impact of different RIDGE dimensions will be analysed. In the following two chapters the RIDGEs found with the chosen RIDGE dimension, $110 \pm 30$ kbp, are described and the motivation behind focusing on the MHC locus explained.

## 4.1 RIDGE definition

RIDGEs were initially investigated according to three different definitions; 1) a set of consecutive active genes covering a certain genomic distance (as based on the extensive literature review presented in the introduction and in this section), 2) a simplified model of, at least 2, consecutive active genes, disregarding length of loop, and 3) at set of consecutive genes with correlated gene expression levels. The resulting RIDGEs in the MHC locus was analysed for all three definitions for the macrophage activation dataset, and the first one made the most biological sense, and is therefore the only one presented in this thesis. According to all three definitions, the entire chromatin loop is expressed, transcribed, and translated simultaneously.


Figure 4.1: Chromatin organisation according to the loop-and-rosette model. A) Single elements (cyclosomes-globosomes) with single histone-depleted DNA loops and accumulations (duplications) of these in linear form. B) Arrangement of grouped scaffold elements into compact ring-like chromomeres (cyclomeres with loops). C) The small DNA loops (1.2-2.2 kbp) are further condensed around the scaffold elements, giving rise to superbeads. Taken from (Peter Engelhardt, 1998).

### 4.1.1 RIDGE, loop, dimensions

A RIDGE is defined as active neighboring genes that cover around 123 kbp of DNA. This definition would obviously benefit from 3D data, but this is currently not available (Bolzer et al., 2005), therefore we restricted the analysis to linear 2D data. Benefits of this model include strict definitions of RIDGEs, as opposed to simply neighboring genes, the drawback is that the determination of loop sizes requires an additional two parameters; a midpoint and the allowed variation (the radius). We propose that these Regions of IncreaseD Gene Expression (Caron et al., 2001) are spatially regulated and co-expressed genomic regions that cover about 110 kbp of DNA (Sachs et al., 1995; Knoch et al., 1998; Munkel and Langowski, 1998; Masternak et al., 2003; Spilianakis and Flavell, 2004) (validated by looking at the β-globin locus, the helper type 2 ($T_H2$), and by a wider literature review) where an entire chromatin loop, rosette (Okada and Comings, 1979) is co-expressed, transcribed, and translated.

### 4.1.2 The RW/GL and the MLS models

#### 4.1.2.1 The Random-Walk/Giant-Loop (RW/GL) model

Sachs et al (Sachs et al., 1995) used fluorescence in situ hybridisation data of distances between defined genomic sequences to construct a quantitative model for the geometric structure of human chromosome. They suggested that the large-scale geometry during the cell cycle, G0/G1, may consist of flexible chromatin loops, averaging 3 million bp (Mbp) with a random-walk backbone; the RW/GL. The DNA strand of a human chromosome has in the order of 100 Mbp arrayed along its contour. At the level of 0.001-0.01 Mbp the DNA is associated with proteins to form the chromatin fiber. The level of 0.1-1.5 Mbp could correspond to simple RW and observed deviation at larger levels could be explained by a polymer model in which the DNA is confined to a spherical subvolume. An alternatively explanation would be flexible giant loops, several Mbp long, with their base points along a RW. The authors argue against other models, based on the assumption that most of them are too complicated (involving too many parameters), which are too detailed for the current data set. So variations in loop size was not allowed, even though they did not expect all loops to be equally long. (Sachs et al., 1995)

#### 4.1.2.2 The Multi-Loop Subcompartment (MLS) model

Münkel et al (Munkel and Langowski, 1998) simulated human chromosomes by a polymer model. The chain was arranged into loops and several consecutive loops formed subcompartments, where each chromosome in the cell nucleus consisted of a single DNA fibre with its DNA limited to a specific territory - a subvolume. Previous results showed a crossover from the initial RW to a more compact structure and were interpreted as an additional backbone polymer, folding the fibre into giant loops, where the MLS model predicts structural flexibility

together with a high degree of CT compartmentalisation. The chromatin fibre is represented as a linear chain of segments (with individual stretching energy), and the rigidity of the fibre was modeled by Kuhn sized segments. Indications of loop sizes about 100 kbp were found, but the author focused the analysis on four Kuhn segments, 120 kbp, as supported by their literature review (although the results were insensitive to a doubling of the loop size). The exact number of loops that formed a subcompartment, about 10, was derived from the observed band pattern (1.5 Mbp). The notion of a highly compartmentalised CT predicted, and verified, by the MLS model is valid even on the level of subcompartments. A structural modification of the MLS model allow several successive subcompartments to be opened and their chromatin to form a single GL in the Mbp range. (Munkel and Langowski, 1998)

The MLS described human interphase chromosomes as flexible fiber, where higher-order structure consisted of 120 kbp loops arranged into rosette-like subcompartments. The number, and size, of these subcompartments were found to correspond with chromosome bands in early prophase. An agreement between human chromosome 15 and subchromosomal foci composed of either early or late replicating chromatin (which were observed at all stages of the cell cycle and that possibly provide a functionally relevant unit of CT compartmentalisation) was found. Computed distances of chromosome specific markers on both smaller, 1 Mbp, and larger, 10-100 Mbp, scales agree with fluorescent in situ hybridization measurements. In the 70th higher-order, rosette like structures were found. (Munkel et al., 1999)

#### 4.1.2.3 The RW/GL versus the MLS model

Gene regulation is closely connected to genome organisation in the nucleus. For example protein synthesis, structure maintenance, and cell duplication all depend on the precise structural arrangement of cellular components, in other words chromosomes occupy distinct CTs and exhibit non-random distribution of the chromatin fibre. Furthermore, since CTs are very compact active genes are transcribed mainly at the periphery and macromolecules are transported between adjacent territories. In the RW/GL model large chromatin loops, 3-5 Mbp, are bound to a nuclear matrix; whereas in the MLS-model 120 kbp loops form rosettes corresponding to the size of chromosomal interphase band domains of 1-2 Mbp, connected by chromatin where the structural support is not a protein matrix. This model allows easy decondensation, or condensation, of a chromosome from metaphase, since an entire loop is opened, formed, at its base. Best agreement between simulation and experiment was found for the MLS model with both a loop size and linker length of 126 kbp. The simulation of a whole human cell nucleus resulted in the formation of distinct CTs, but in contrast to the RW/GL model the MLS-model show low overlap between CTs and chromosome arms, in agreement with confocal image series analysis. (Knoch et al., 1998)

#### 4.1.2.4 Summary

Sachs et al (Sachs et al., 1995) proposed that DNA folds into a three-parametric polymer model for large, 3 Mbp long, flexible loops attached to a flexible backbone. Three years later, Münkel et al, (Munkel and Langowski, 1998), showed that the polymer chain of each chromosome was arranged into loops and that several consecutive loops formed a sub-compartment. The loop size was set to 120 kbp, which is equal to four segments of the polymer model, but their result were insensitive to a loop size increase by a factor of 2. (Munkel and Langowski, 1998) Later they adjusted the loop size to 126 kbp. (Knoch et al., 1998)

### 4.1.3 Additionally suggested RIDGE dimensions

#### 4.1.3.1 Loci length

The β-globin locus (200 kbp) could be expected to be a RIDGE long. (Tolhuis et al., 2002) The locus shows gene activity in the range of 70 kbp, and can be explained by a loop in the 100 kbp range where inhibition of RNA polymerase II confer dramatic changes in the chromosome structure. (Munkel et al., 1999) Both the mouse and human loci contain an upstream LCR and multiple β-like genes arranged from 5' to 3' in order of their developmental expression. In addition there are several distal hypersensitive sites; including a downstream 3' HS1 (20 kbp apart) and two upstream hypersensitive sites (60 kbp away). (Palstra et al., 2003) In contrast, the $T_H2$ locus spans 120 kbp, with an additional related gene 7 kbp upstream of this. (Spilianakis and Flavell, 2004)

Other genomic loci related examples includes isochores (a homogeneous up to 200 kbp stretch of DNA classified into five base composition classes with 40%-60% GC) (Saitoh and Laemmli, 1994) and tandem duplications (two related intrachromosomal duplicons located within 200 kb of one another). (Cheung et al., 2003) Furthermore, the transcription unit, or replicons, generally range in size from 50 to 200 kbp whereas longer separations approach the length of the nuclear radius. (Bystricky et al., 2004) Early S replication sites were assumed to contain several typically sized replicons, DNA loops, of 50-200 kbp. (Ma et al., 1999)

#### 4.1.3.2 A RIDGE is around 123 kbp long

130 kbp is the median of a lot of different propositions: 1) loops of 86 kbp of DNA are stable to variations in conditions (Jackson et al., 1990), 2) long chromatin fragments (>100 kbp) can electrolyte from encapsulated and permebilized cells (Jackson and Cook, 1993), 3) the mean DNA content of groups of co-expressed genes is 125 kbp (Spellman and Rubin, 2002), 4) hypersensitive sites spread over 130 kbp in the nuclear space (Tolhuis et al., 2002), 5) chromatin containing 150 kbp of DNA can electrolyte through agarose, 6) *Hind* III restriction fragments spread over 170 kbp of DNA (Palstra et al., 2003), 7) in *Drosophila melanogaster*, 20% of the

genes are organised into clusters with similar expression patterns up to 200 kbp in length (Spellman and Rubin, 2002), and 8) the majority of duplicated sequences occur intrachromosomally within 200 kbp of each other (Cheung et al., 2003).

More specifically suggested chromatin loop sizes include: 1) 75 kbp (Cook, 1995), 2) 86 kbp (Cook, 1995), 3) 100 kbp (Saitoh and Laemmli, 1994), 4) above 100 kbp (Fukuoka et al., 2004), 5) the Rosette chromatin fibre about 100 kbp long (Cremer and Cremer, 2001), 6) 146 kbp of core particle DNA (Horowitz et al., 1994), and 7) 3 mega bp-sized loops with the loop bases distributed in a RW (or as a chain of chromosomal sub-compartments) each comprising 10-20 loops of 120 kb. (Bystricky et al., 2004). Larger variations were found in the form of 1) 10 to 220 kb, or between 15 and 125 kbp ((Jackson et al., 1990) and references therein), 2) 50-200 kbp long repeating loop domains (Ma et al., 1999), 3) 100 and 500 kbp windows (Coppe et al., 2006), and 4) nucleoids centre around 12.5 kbp or are distributed between 50 and 250 kb (Jackson et al., 1990). Therefore a loop is likely to be around 100 kbp long, although with a great variability in length.

#### 4.1.3.3 Longer RIDGE suggestions

In the nucleus more than one active gene has to be transcribed in each factory, because there are fewer transcription factories than there are active genes and other transcription units. In fact, actively transcribed genes, separated by up to 40 Mbp, frequently co-localise in the same transcription factory. (Chakalova et al., 2005)

Previous suggested, longer, loop sizes are: 1) 100 and 500 kbp (Coppe et al., 2006), 2) co-expression of neighboring genes in human (500 kbp) and flies (200 kbp) (Lercher et al., 2003b), 3) 430 kbp which is the length of the unfolded DNA chain (Bystricky et al., 2004), 4) deletions can lead to changes in the spreading of heterochromatin over a $>$ 900 kbp (Kleinnjan and von Heyningen, 1998), 5) larger loops (up to several Mbp) extend outward from the surface of the CT and possibly into the IC (Cremer and Cremer, 2001), 6) the original model assumption was 3 Mbp (Sachs et al., 1995), and 7) loop sizes between 4.0 and 5.5 Mbp. (Dietzel et al., 1998)

#### 4.1.3.4 A 32 kbp variation in RIDGE length is allowed

Biological features that fall within the variation of 32 kbp include: 1) fibers had a modal DNA length of excess of 10 kbp, (Bednar et al., 1998), 2) muscle-expressed genes within 10 kbp of each other (Roy et al., 2002), 3) nucleoids contains loops centered around 12.5 kbp (Jackson et al., 1990), 4) the chromosomal fiber is relatively stiff over intervals of 10-20 kbp (Bystricky et al., 2004), 5) histone acetylation can spread at least 16 kbp upstream, and the binding of RFX and CITTA correlate with this spreading from the promoter (Masternak et al., 2003), 6) a 16 kbp condensed chromatin region (Litt et al., 2001), 7) pure DNA fragments of $<$ 20 kbp (Jackson and Cook, 1985), 8) above 20 kbp the relationship ceases to be linear (Jackson

and Cook, 1985), 9) *C.elegans* have many co-expressed neighboring gene pairs within 20 kbp, (Lercher et al., 2003b), 10) the mean DNA content of each polytene band is 25 kbp (Spellman and Rubin, 2002), 10) all inversions in *C.elegans* are $<$ 25 kbp (Lercher et al., 2003b), 12) correlated adjacent genes in the cell-cycle (26 kbp and 20 kbp respectively) (Cohen et al., 2000), 13) an excess of duplicates that are located close to the original gene ($<$ 30 kbp) (Lercher et al., 2003b), 14) most transposed DNA segments are $<$ 30 kbp long (Lercher et al., 2003b), 15) a broad 30 kbp hypoacetylated subdomain (Forsberg and Bresnick, 2001), and 16) gene activation is paralleled by acetylation of a 32 kb chromatin domain. (Elefant et al., 2000)

Furthermore, evidence for even large variations include the sensitive chromatin domain of 33 kbp which correlate with the extent of acetylation, where DNase I sensitivity extends from 10 kbp upstream, to 9 kb downstream. In fact the extent of nuclease-accessible chromatin corresponds to the MARs in chicken lysosome (20 kbp), and ovalbumin (100 kb), and for human apolioprotein B (47 kbp). (Hebbes et al., 1994) Even further distances are shown in the β-globin genes (β-major and β-minor) which spatially interact with the LCR, located 40-60 kbp away and where the inactive genes loop out. (Laat de Wouter, 2003)

### 4.1.4 RIDGEs are $110 \pm 30$ kbp long genomic regions

The initial suggestion was to define a RIDGE as $123 \pm 16$ kbp, but based on the above it became apparent that: a) longer RIDGEs should exist, and b) more variation in RIDGE size should be allowed. Furthermore a visual overview of gene expression in the MHC locus supported RIDGEs in the range of 80 kbp as well, and so the RIDGE dimension was refined to $110 \pm 30$ kbp.

## 4.2 RIDGE characterisation

Here the distributions, both from real and random genomes, are described. These are used to evaluate potential RIDGEs so the thresholds for significant values are also described.

As described in "Methods and Materials" section 2.3.0.1, the physical start position of a gene was randomised, although all other attributes (such as chromosome, length, strand, and gene expression) remained the same. Initially C was used to count the number of expected RIDGEs, but this was later re-implemented in JAVA to fit into the SORGE framework in order to store the resulting RIDGEs to the SORGE DB. All scores presented in this section, was then calculated on these random RIDGEs. The re-implementation in JAVA lead to a significant decrease in speed; before 10000 permutations took seconds, now a few hours. 30000 gene shuffling permutations have been run for the macrophage activation dataset (see 2.1.5.1), with around 15 million RIDGEs (and 42 million genes).

This section is divided into four parts; determination of active genes, RIDGE distribution

scores, genomic organisation of RIDGEs, and sequence similarity scores. RIDGE distribution scores include expected number of RIDGEs, the RIDGE activity scores, and the number of genes in a RIDGE. Genomic organisation of RIDGEs deal with gene lengths, inter-gene distances, number of transcripts, and the number of exons. Finally the two ClustalW scores, sequence similarity of RIDGE members and their 5' regions, are discussed.

### 4.2.1 RIDGE members

RIDGEs are dependent on the method for determining active genes (section 4.2.1.1), the number of silenced genes allowed within a RIDGE (4.2.1.2), and the RIDGE dimension (4.3). These will be presented for the macrophage activation dataset with the four biological conditions; rested macrophages, macrophages primed with IFN-γ, macrophages viral activated with mCMV, and macrophages that were both primed and viral activated.

#### 4.2.1.1 Method for determining active genes

| | majority of calls | | all calls | | mean>50 | | abs>50 | | mean>75 | | abs>75 | | mean>100 | | abs>100 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Genes | RIDGEs | | | | | | | | | | | | | | |
| uninfected | 3749 | 664 | 3381 | 590 | 3754 | 737 | 3533 | 654 | 3026 | 486 | 2837 | 434 | 2502 | 349 | 2340 | 305 |
| IFN | 3542 | 597 | 3085 | 470 | 3504 | 635 | 3035 | 502 | 2829 | 427 | 2366 | 317 | 2345 | 305 | 1921 | 208 |
| mCMV | 3903 | 707 | 3601 | 595 | 3805 | 754 | 3560 | 669 | 3078 | 490 | 2859 | 437 | 2549 | 349 | 2364 | 302 |
| both | 3890 | 694 | 3548 | 591 | 3675 | 701 | 3405 | 614 | 2929 | 448 | 2713 | 400 | 2471 | 336 | 2248 | 283 |

Table 4.1: Active genes, and RIDGEs, per biological condition and method for determining active genes. The left most column specifies the four biological conditions, either 1) uninfected (rested) macrophages, 2) macrophages primed with IFN-γ, 3) macrophages infected with mCMV, or 4) both primed and infected macrophages. The top row specifies the method for determining active genes; either the majority of detection calls are set to present, all detection calls are set to present, or alternatively the mean, or absolute, intensity signal above 50, 75, or 100 respectively. For each of these combinations the number of active genes (Genes) and the number of RIDGEs (RIDGEs) are shown. Note that these headers are only shown once, and that although the same number of RIDGEs were found in uninfected and infected macrophages (for the mean intensity signal above 100), at least 101 of these RIDGEs differ between the two conditions.

The most active genes was seen for infected macrophages, although a decrease in the number of active genes was expected for this condition (Capellini et al., 2006; Versteeg et al., 2006), and following, the most RIDGEs. The infected macrophages has the most active genes, and RIDGE, across all methods for determining active genes, except for when all replicates

had to have an intensity signal above 100. Here the most active RIDGEs are seen for uninfected macrophages. The least active genes, and RIDGEs, are consistently found for primed macrophages. This consistency implies that the choice of method for determining active genes do not influence the relative order of the conditions, although the increase, or decrease, in relative number of active genes was influenced. For example; for the majority of present detection calls, the number of active genes is decreased by 5.5% in primed macrophages (from uninfected macrophages), decreased by 6.7% for mean intensity signal above 50, and decreased 6.3% for mean intensity signal above 100, and for viral activated macrophages increased by 4.1%, 1.4%, or 1.9% respectively, and finally for macrophages that were both primed and viral activated increased by 3.8%, decreased by 2.1%, and 1.2% respectively.

Capellini et al (Capellini et al., 2006) speculated that the reduction in gene expression (as seen for *M. tuberculosis* associated genes) could be due to nucleic acid biosynthesis and that higher levels of active genes would in fact slow down the cell cycle. Versteeg et al (Versteeg et al., 2006) also found a reduction of active genes, from 46 to 40%, in murine hepatitis virus (MHV) infection. Here an increase in number of active genes was seen from uninfected to viral activated macrophages, and a decrease from uninfected to primed macrophages. One possible explanation for the latter is that the first wave of IFN-γ induced transcription occurs within 15-30 minutes; and many of the directly IFN-γ regulated genes are TFs responsible for driving the second wave of transcription. (Schroder et al., 2004) The response to IFN-γ priming cascade for at least 12 hours (as seen in 6.2.1) with both resting periods and periods of high expression. It could be the case that after 24 hours, the resulting cascade needs to be terminated, or at least halted, especially since no invading virus, or bacteria, was found, thus leading to the most active genes in infected macrophages.

#### 4.2.1.2 Silenced genes in a RIDGE

| $110 \pm 30$ | 0 gaps | 1 gap | 2 gaps | 2 gaps* | $80 \pm 20$ | $123 \pm 16$ | $150 \pm 50$ | $220 \pm 40$ |
|---|---|---|---|---|---|---|---|---|
| uninfected | 664 | 875 | 951 | 657 | 667 | 326 | 719 | 339 |
| IFN | 597 | 773 | 844 | 596 | 619 | 275 | 657 | 326 |
| mCMV | 707 | 913 | 1017 | 703 | 715 | 347 | 808 | 392 |
| both | 694 | 918 | 1024 | 709 | 713 | 348 | 781 | 392 |

Table 4.2: Number of RIDGEs found when changing the number of allowed silenced genes (gaps) in a RIDGE, or the RIDGE dimension. First no silenced genes were allowed in a RIDGE, then up to a single silenced gene was allowed, and the following two columns both allow 2 silenced genes, although the latter (*) also consider genes with no reliable probe-to-gene projection as silenced. The last four columns varies the RIDGE dimension.

As expected, the least number of RIDGEs is found in the narrow definition of a RIDGE, and the most when up to 2 silenced genes is allowed. For the analysis of active genes per condition, it was expected that the most RIDGEs should be found in the mCMV infected macrophages, but it is found either in infected or both primed and infected macrophages (although the difference between these conditions is less than 7 RIDGEs).

Furthermore that the increase from zero to one silenced genes includes twice as many new RIDGEs as the increase from one to two silenced genes were unexpected. Zero silenced genes and 2 (*) result in similar number of RIDGEs, but for the uninfected macrophages at least 241 of these RIDGEs differ.

### 4.2.2 RIDGE distributions

RIDGE distribution scores include; the number of RIDGEs expected in uninfected macrophages for all five RIDGE dimensions; the distribution of RIDGE activity scores per number of genes within a RIDGE, the expected number of RIDGE members, and finally the exon density of a RIDGE. The latter is a measurement of how much of the RIDGE was taken up by exons.

#### 4.2.2.1 The number of observed RIDGEs is significant

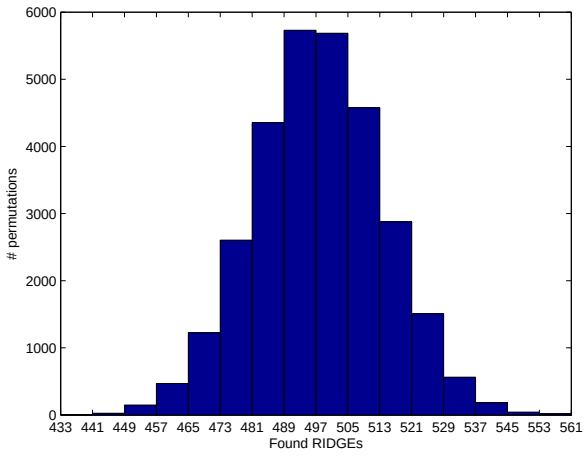
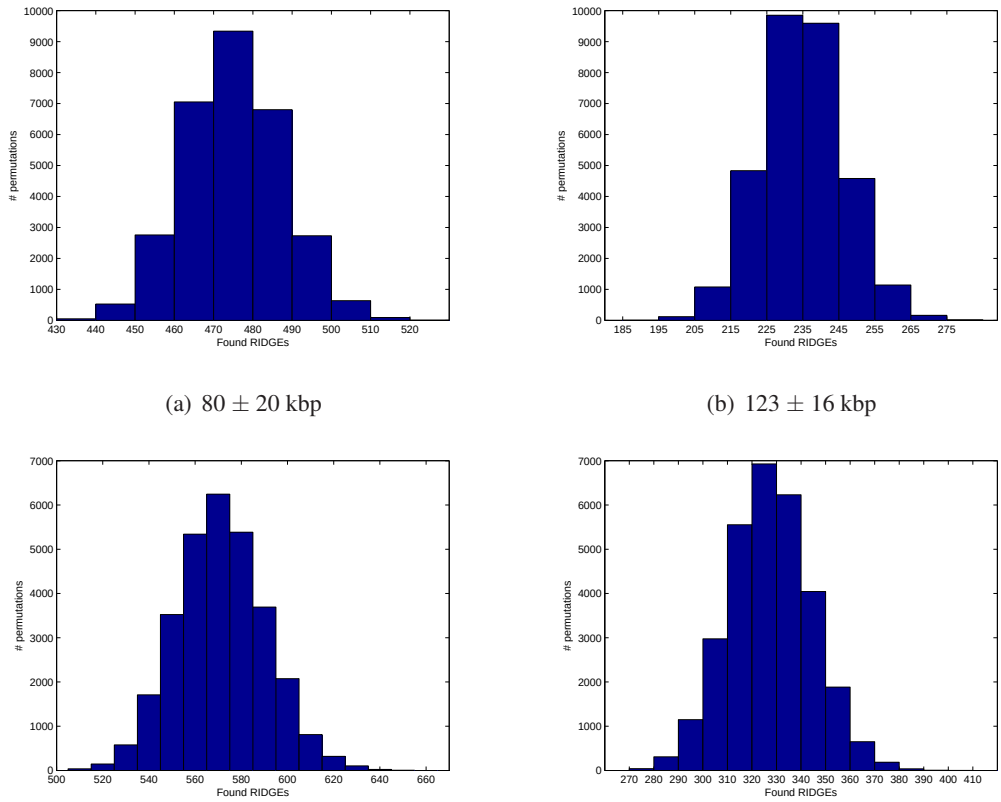



Figure 4.2: The distribution of the number of RIDGEs found with a RIDGE dimension of $110 \pm 30$ kbp (N=30000 and bin size=8).

The mean number of RIDGEs found is 497 and the standard deviation 16 ($497 \pm 16$). If the number of found RIDGEs was larger than 97.5% of the distribution, i.e. $>529$, then it was considered statistical significant and the genome as having non-random gene organisation. Alternatively if the actual number of RIDGEs had been less than 2.5% of the distribution, i.e. $<466$, then the mouse genome would not have shown RIDGE organisation.

According to the bootstrap analysis (see 2.3.0.1) $497 \pm 16$ RIDGEs (N=30000) were expected in uninfected macrophages. The highest recorded number of RIDGEs for these permutation is 560 RIDGEs. Since 664 RIDGEs were found, the mouse genome is likely to have RIDGE organisation, although it is not possible to assign a p-value to this. Another 30000 permutations were performed, and now the highest recorded number of RIDGEs was only 544. The same was true for all investigated RIDGE dimensions.




(a) $80 \pm 20$ kbp

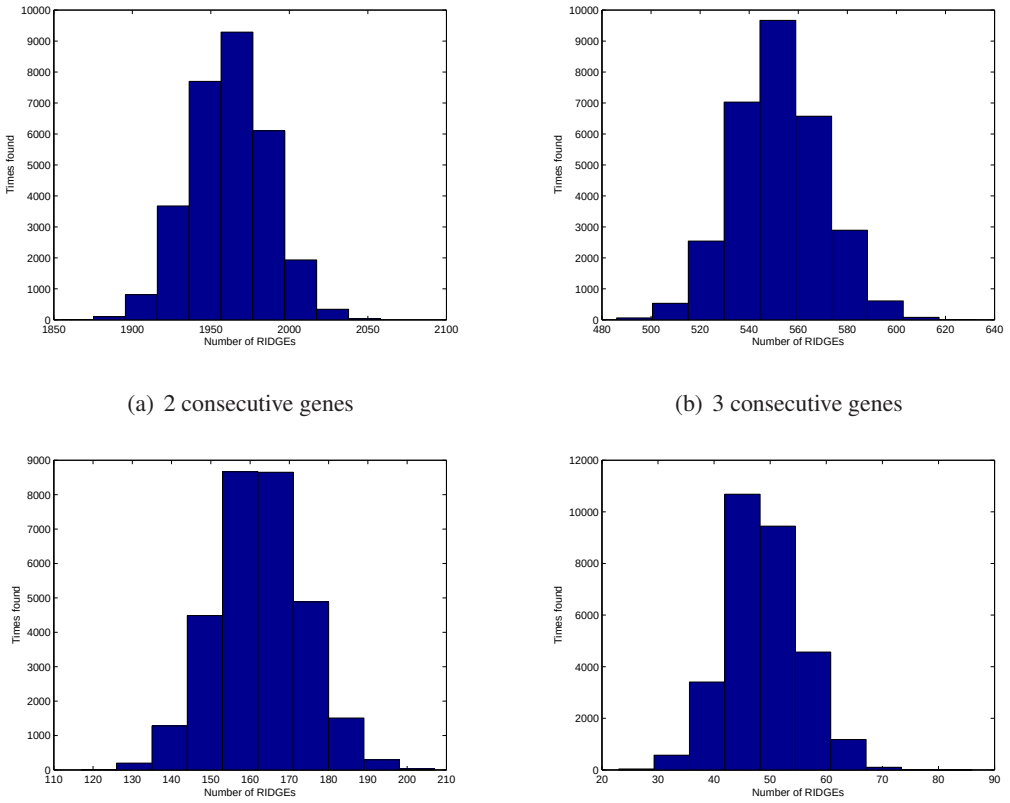
(b) $123 \pm 16$ kbp

(c) $150 \pm 50$ kbp

(d) $220 \pm 40$ kbp

Figure 4.3: Expected number of RIDGEs for alternative RIDGE dimensions with a bin size of 10. The expected number of RIDGEs for the smallest RIDGE dimension, $80 \pm 20$ ($mean \pm standarddeviation$) kbp (graph a), is $473 \pm 12$ RIDGEs; for the $123 \pm 16$ kbp definition $230 \pm 11$ RIDGEs (graph b); for the $150 \pm 50$ kbp definition $589 \pm 19$ RIDGEs (graph c); and finally for the largest definition, $220 \pm 40$, $338 \pm 18$ RIDGEs is expected (graph d).

For uninfected macrophages, 477 RIDGEs are expected for the shortest dimension, and 644 is found; for the narrow dimension 230 RIDGEs are expected and 326 found; for the second largest dimension 580 are expected and 710 found; and finally for the longest RIDGE dimension 220 RIDGEs are expected as compared to the 338 that is found.




(a) 2 consecutive genes

(b) 3 consecutive genes

(c) 4 consecutive genes

(d) 5 consecutive genes

Figure 4.4: Expected number of consecutive RIDGEs. For RIDGEs of 2 consecutive active genes (graph a) $1962 \pm 25$ RIDGEs ($mean \pm standarddeviation$) are expected; for 3 genes (graph b) $552 \pm 18$ RIDGEs are expected; for 4 genes (graph c) $163 \pm 11$ RIDGEs are expected; and for 5 genes (graph d) $49 \pm 7$ RIDGEs are expected.

For uninfected macrophages, 1962 RIDGEs are expected for 2 consecutive genes, and 1745 are found (which is statistical significant, although fewer than in random genomes); for 3 genes 552 are expected and 833 found; for 4 genes 163 are expected and 391 found; and finally for 5 consecutive genes 49 are expected and 181 found.

When 2, or more, consecutive active genes were investigated, 10% more are found for random genomes than actually seen in the genome, although 3, 4, and 5 consecutive genes are more abundant than random genomes, implying that this model is not stringent enough, that finding two gene pairs is not surprising (as seen below as well).

#### 4.2.2.2 Expected number of genes in a RIDGE

The number of RIDGE members was also investigated, where the fewest number of RIDGE members were consistently found for primed macrophages, and the most for macrophages that were both primed and viral activated. In other words, primed macrophages has both the fewest

and the shortest RIDGEs, whereas macrophages that are both primed and viral activated has the most and the longest RIDGEs. A RIDGE consist of between 1 and 4 genes; spanning from $1.94 \pm 0.98$ members per RIDGE in infected macrophages, to $2.92 \pm 1.76$ genes for 2 silenced genes (where missing genes are considered silent) in macrophages that are both primed and viral activated.

For the 30000 permutations for the chosen RIDGE dimension, $110 \pm 30$ kbp, 29% of the found RIDGE are single-gene RIDGEs, 59% two-gene RIDGEs, 11% three-gene RIDGEs, and the remaining 1% has 4 or more RIDGE members. Therefore a RIDGE has to have more than 3 genes to be considered statistically significantly long.

#### 4.2.2.3 The RIDGE activity score

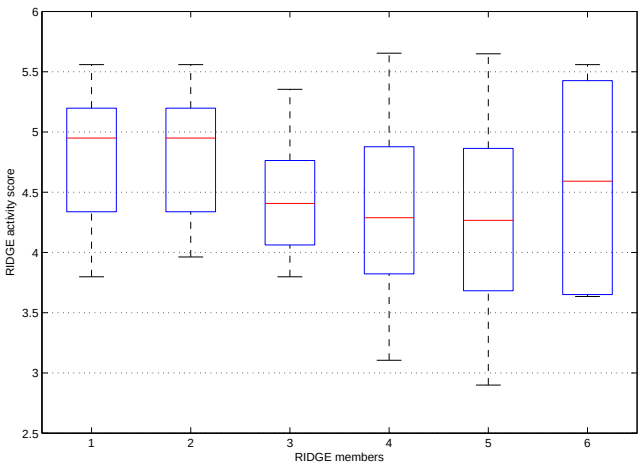



Figure 4.5: The distribution of RIDGE activity scores for the chosen RIDGE dimension. Here the z-scores for the randomised RIDGEs based on the number of genes in a RIDGE are shown.

### 4.2.3 Genomic organisation of RIDGEs

#### 4.2.3.1 Gene distance and density per chromosome in mouse

The inter-gene distance between RIDGE members should ideally be lower than the inter-gene distance between two random genes. Additionally the gene density should be higher inside than outside a RIDGE.

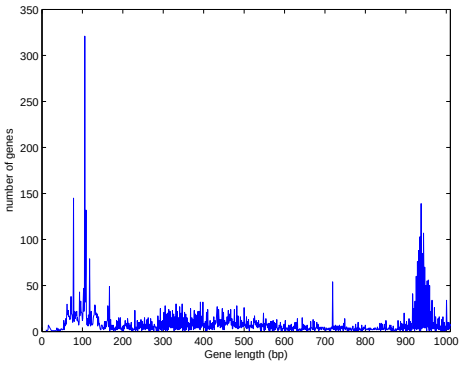
| | Length | Genes | Density | Distance | StdDev | Length | StdDev |
|---|---|---|---|---|---|---|---|
| 1 | 197 069 962 | 1592 | 2.28% | 84 368 | ± 212 206 | 37 439 | ± 80 125 |
| 7 | 145 134 094 | 2495 | 3.64% | 37 758 | ± 99 509 | 19 180 | ± 51 966 |
| 11 | 121 798 632 | 1990 | 5.32% | 34 485 | ± 114 588 | 25 170 | ± 58 243 |
| 17 | 95 177 420 | 1264 | 3.3% | 46 738 | ± 139 163 | 26 222 | ± 67 081 |
| 18 | 90 736 837 | 700 | 2.06% | 86 073 | ± 194 474 | 38 992 | ± 80 214 |
| 19 | 61 321 190 | 877 | 3.61% | 38 539 | ± 108 202 | 27 808 | ± 71 295 |
| Y | 16 029 404 | 29 | 0.35% | 60 821 | ± 146 079 | 33 990 | ± 83 600 |

Table 4.3: Gene distance, gene length, and gene density for a selection of chromosomes; where the chromosome with the highest (in red), and lowest (in green), value for each characteristics (column) are included. Chromosomes are numbered from the longest (1) to the shortest (19). Chromosome length is not found to correlate with gene density, inter-gene distance, or gene length. For example the longest chromosome do not have the longest inter-gene distance, or the lowest gene density. In addition to the 5 chromosomes that contained the extreme values, chromosome 17 (with the MHC locus) and chromosome 19 (the shortest non gametes) are also included.

Chromosome 7 had the most genes, the lowest standard deviation for the inter-gene distances, the shortest gene length, and the lowest standard deviation for gene length, whereas chromosome 1 is the longest chromosome and has the highest standard deviation for inter-gene distance (all of which are obviously related). Interestingly enough chromosome 18 has both the longest inter-gene distances and the longest genes. Chromosome Y is the shortest chromosome and has the lowest density and the highest standard deviation for gene length. But since it only contains 29 genes; most of the time, no RIDGEs are found.

#### 4.2.3.2 Gene length

A gene is expected to be 7272 bp long. Here the median is reported instead of the mean because almost one third of the genes are shorter than 1 kbp and 3% are longer than 200 kbp, skewing the mean, and standard deviation, toward longer genes ($29826 \pm 75627$).




| Length | Genes | % |
|---|---|---|
| <80 | 591 | 2 |
| <1010 | 8193 | 29 |
| <10 kbp | 15571 | 55 |
| >20 kbp | 4174 | 15 |
| >50 kbp | 4174 | 15 |
| > 100 kbp | 1921 | 7 |
| > 200 kbp | 717 | 3 |

Figure 4.6: 29% of the genes in the genome were shorter than 1010 bp (shown in graph to the left) and 3% are longer than 200 kbp. Because of these extreme values the mean and standard deviation are not appropriate measurements, instead the median was used.

A gene has to be shorter than 80 bp to be considered statistically significant short (2.1%), and longer than 225 kbp to be considered long (2.1%); therefore all RIDGE members, presented in this theis, are considered normal long genes.

#### 4.2.3.3 Inter-gene distances

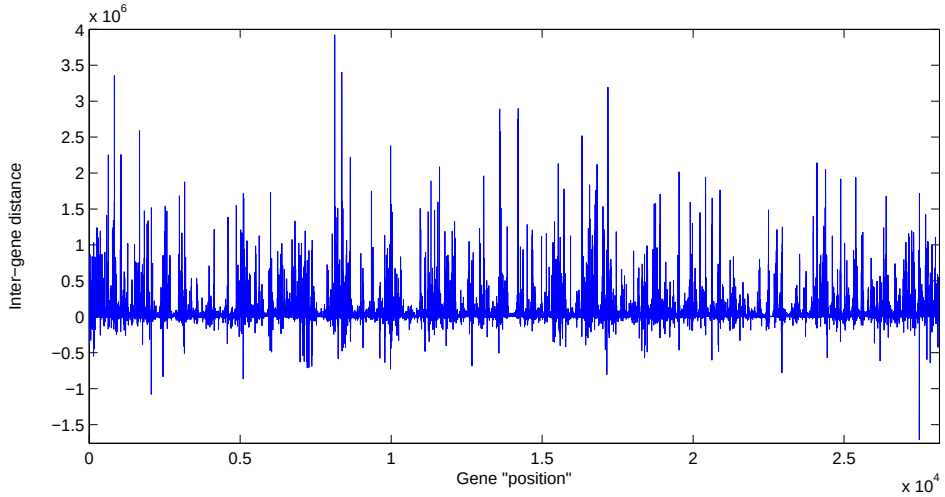



Figure 4.7: The inter-gene distances between genes (calculated as $gene2.startposition - gene1.endposition$). The X-axis shows the gene position (where the left most gene was the first gene on chromosome 1 and the right most gene was the last gene on chromosome Y)

There are 28136 inter-gene distances, of which 8% are below 0 (where the genes are overlapping). This means that for two genes to be considered close they actually have to overlap, therefore none of the inter-gene distances found for the RIDGEs are statistical significant. Fur-

thermore to be considered significantly long an inter-gene distance have to be longer than 500 kbp (which obviously none of the RIDGE members fulfill), and 30% of the distances are longer than 40 kbp.

#### 4.2.3.4 Number of transcripts

A gene has either 1 transcript (80%) or 2 transcripts (14%); 4% of the genes have 3 transcripts and only 2% have 4, or more, transcripts. Therefore a gene is said to have many transcripts (and therefore more complicated regulation of transcription) if it has 4, or more, transcripts. This is only found for the long, single-gene, RIDGEs.

#### 4.2.3.5 Number of exons

For the 28157 genes in the mouse genome 36007 transcripts were found. Of these 10% have 1 exon, and 6% have more than 40 exons. The maximum number of exons found, 604, were found for Neb, followed by 392 exons for Herc1. None of the RIDGE members therefore have significantly many, or few, exons.

### 4.2.4 ClustalW analysis of RIDGE member sequences

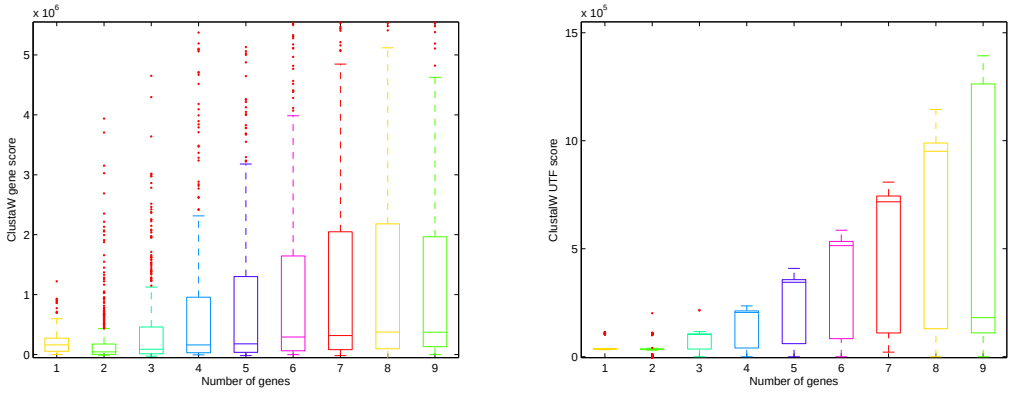



Figure 4.8: The two ClustalW scores, the gene score and the UTR score

The *coding sequence similarity score* (gene score) and the *upstream sequence similarity score* (UTR score) were both defined in chapter 2 (2.2.3).

The UTR score continues to increase exponentially the more and more genes are added to the sequence similarity comparison, whereas the gene score varies (as for example seen with the many outliers), although in general the more genes in a RIDGE, the higher the score. A RIDGE is considered to have a statistical significant score if the associated p-value (for the number of genes) is less than 0.05.

## 4.3 Evaluation of RIDGE dimensions

Here a short discussion about how the RIDGE dimension influence the results for the MHC locus. For example when the RIDGE dimension was narrowed to $123 \pm 16$ kbp, then 10 fewer RIDGEs are found. RIDGE with identifiers, for example A03, are presented in detail in the next chapter because these are also observed for the chosen RIDGE dimension, $110 \pm 30$ kbp. Once again the macrophage activation dataset, MDS, is used, with the four biological conditions; uninfected macrophages, primed macrophages with IFN-γ, viral activated macrophages with mCMV, and macrophages that were both primed and viral activated.

### 4.3.1 RIDGE dimension $80 \pm 20$ kbp

When the RIDGE dimension was restricted to shorter RIDGEs, no RIDGEs were found that are not also found for the chosen dimension, $110 \pm 30$, although in most cases in longer (elongated or combined) versions. 19 RIDGEs were found, including 2 overlapping regions with 14 RIDGEs; 1) (Pknox1, U2af1), A02, is only present in macrophages that were both primed and viral activated; 2) The static (unaffected by all four conditions) single-gene RIDGE (Brd4), A03; 3) (Akap8l, Wiz, and A430107D22), A04, that is absent in primed macrophages; 4) (Brd2, Psmb9, H2-DMB2, H2-DMa) that is only present in macrophages that were both primed and viral activated; 5) the static single-gene RIDGE H2-Q2, A09;

1. (Psmb8:H2-Ab1) a 69 kbp long region, part of A07, and present in all but uninfected macrophages, although all genes were significantly upregulated for all conditions, except for Tap2 in primed macrophages
   (a) (Tap2:H2-Ab1), 63 kbp long
   (b) (Psmb8:H2-Aa), part of A07
   (c) (Tap2:H2-Aa)
2. (Dhx16:Abcf1), A16, is static although with fluctuating gene regulation, for example; Abcf1 was downregulated after priming (as was Mrps18b) and upregulated after infection, in addition Mrps8b was upregulated after macrophages that were both primed and viral activated.
   (a) (Ier3:Mrps18b), A11, that is absent in viral activated macrophages
   (b) (Flot1:Mrps18b), A12, expressed in the same conditions as the previous RIDGE
   (c) (TUBB:Mrps18b), part of A13, expressed when A11 and A12 were expressed
   (d) (2310061I04:Abcf1), part of A16, 77 kbp, lacking after both.
   (e) (2310061I04:GNL1), part of A17

(f) (H2-T23:H2-L), A21, static although both genes upregulated for all conditions

(g) (H2-L), A22, expressed when A21 were expressed

(h) (Mrps18b:GNL1), part of A17

(i) (Abcf1:H2-T23), part of A19

### 4.3.2 RIDGE dimension $123 \pm 16$ kbp

For dimensions of $110 \pm 30$ kbp 22 RIDGEs are observed in the MHC locus. Of these 10 RIDGEs are not observed in the more strict $123 \pm 16$ dimension. Specifically the follow disappeared; A02 - (Pknox1, U2af1), A03 - (Brd4), A04 - (Akap8l, Wiz, A430107D22), A09 - (H2-Q2), A11 - (Ier3:Mrps18b), A12 - (Flot1:Mrps18b), A14 - (KIAA1949:Abcf1), A16 - (Dhx16:Abcf1), A21 - (H2-T23, H2-L), and A22 - (H2-L), meaning that the 3 single-gene RIDGEs are removed.

Of the 12 RIDGEs that are found; 1 have good, low, RIDGE activity score, 8 have functional associations between the RIDGE members, 3 are considered to have many TFs (A05, A06 and A13), 2 are considered to have significant UTF scores. For the 10 RIDGEs in the less strict definition; 2 more are considered to have good, low, RIDGE activity scores, 4 more are considered to have functional associated RIDGE members (4 RIDGEs in the KIAA1949:H2-T23 region, A15, A17, A18, and A19), 4 more are considered to have many TFs, and 1 more have a significant UTF score. Therefore the less strict RIDGE dimension is considered more appropriate.

### 4.3.3 RIDGE dimension $150 \pm 50$ kbp

When the RIDGE dimension was increased, the number of observed RIDGEs decreased from 22 to 21. Obviously the 10 RIDGEs that were lost in the stricter RIDGE dimension are lost here as well, where the remaining RIDGEs are either elongated or combined. For instance 14 of the 21 RIDGEs are found in the gene dense Ier:H2-L region, 300 kbp, (as discussed in 5.2.1.1);

A01 - (1500032D16, Pknox1, and U2af1); is only present in macrophages that were both primed and viral activated because Pknox1 is absent in uninfected, downregulated in primed, and absent in viral activated macrophages.

A combination of A05 and A06 (Myo1f, March2, and Rab11b) is static, i.e. present in all, although Myo1f is not static but significantly upregulated after both viral activated and both primed and viral activated macrophages.

A06 - (March2 and Rab11b) is static as were the genes.

A07 - (Psmb8, Tap2, H2-Ab1, H2-Aa, and H2-Eb1) is absent in uninfected macrophages, but otherwise present, because of H2-Aa and H2-Eb1. All genes were upregulated in all con-

ditions, but for Tap2 in primed macrophages, even extremely upregulated (foldchange=38) for H2-Ab1, H2-Aa, and

H2-Eb1 in both primed and viral activated macrophages.

1. A08 - (Tap2, H2-Ab1, H2-Aa, and H2-Eb1),
2. the elongation of A10 (Tcf19, Cdsn, Gft2h4) is only present in uninfected macrophages, although the only significant change in gene expression level is detected for Tcf19 that was downregulated (foldchange=6) fold in primed macrophages.
   (a) A10 - (Cdsn, Gtf2h4).
3. the elongation of A11 (Ier3:Abcf1) is static, although the genes are not. TUBB is constantly downregulated, Ier3 is downregulated after infection and both, Dhx16 is downregulated after priming and infection, Mrps19b is downregulated after priming but upregulated after both. 2310014H01 is upregulated after priming, Abcf1 is downregulated after priming and upregulated after infection. GNL1 is static although marginal in the three first conditions and present in the fourth.
   (a) elongation of A11 - (Ier3:H2-T24)
   (b) elongation of A12 - (Flot1:Abcf1)
   (c) elongation of A12 - (Flot1:H2-T24);
   (d) A13 - (TUBB:Abcf1),
   (e) elongation of A13 - (TUBB:H2-T23),
   (f) A14 - (KIAA1949:Abcf1),
   (g) elongation of A15 - (KIAA1949:H2-T23),
   (h) elongation of A17 - (Dhx16:H2-T23),
   (i) elongation of A18 - (2310061I04:H2-T23),
   (j) A19 - (Mrps18b:H2-T23),
   (k) A20 - (H2-T24:H2-L) is static although the RIDGE members are upregulated in all conditions, but for H2-T24 in primed macrophages.
   (l) part of A19 and A22 - (Abcf1:H2-L), and
   (m) part of A19 and A22 - (GNL1:H2-L).

By forcing a RIDGE to be longer, no new RIDGEs are added and the complexity of overlapping RIDGEs increased (see discussion in 5.2.1.1 for the 14 overlapping RIDGEs, and 5.2.1.3 for the collapse of RIDGE A05 and A06). Also the new member in RIDGE 4, Tcf19, is not

associated with skin (The Gene Ontology Consortium, 2000; Bairoch et al., 2005), as the other RIDGE members are, although it is located within the psoriasis susceptibility 1 (PSORS1) locus (Bowcock and Krueger, 2005) and therefore might potentially be a valid RIDGE member.

### 4.3.4 RIDGE dimension $220 \pm 40$ kbp

When the RIDGE dimension was doubled, around 240 kbp as per (Munkel et al., 1999), only 10 RIDGEs are found, of which 7 are overlapping RIDGEs in the Ier3:H2-L region, making this dimension too long;

1. combination of A05 and A06 - (Rab11b, Myo1f, March2) is static although Myo1f is significantly upregulated in both viral activated macrophages and both
2. (H2-Ea, Gpsm3, Pbx2), 255 kbp, is absent in viral activated macrophages
    (a) (H2-Ea, Gpsm3), 250 kbp, is static
3. combination of A11 and A19 - (Ier3:H2-T23),
    (a) combination of A12 and A19 - (Flot1:H2-T23),
    (b) combination of A13 and A19 - (TUBB:H2-T23),
    (c) combination of A15 and A20 - (KIAA1949:H2-L),
    (d) combination of A17 and A20 - (Dhx16:H2-L),
    (e) combination of A18 and A21 - (2310061I04:H2-L), and
    (f) combination of A19 and A22 - (Mrps18b:H2-L).

### 4.3.5 The chosen RIDGE dimension, $110 \pm 30$ kbp

To summarise this RIDGE dimension was chosen based on the literature review (4.1.1), and because;

- The shorter RIDGE dimension only contained shorter versions of RIDGEs present for this dimension,
- The stricter RIDGE dimension excluded RIDGEs with significant RIDGE activity scores, functional associations between RIDGE members (although only in the overlapping Ier3:H2-L region), had many TFs, and significant UTR scores,
- The somewhat larger RIDGE dimension mostly contained overlapping RIDGEs in the Ier3:H2-L region, and remaining RIDGEs were mostly elongations, or combinations, of RIDGEs that were found for the chosen dimension

- Almost no RIDGEs are found for the long dimension (except for 7 overlapping RIDGEs in the Ier3:H2-L region).

# Chapter 5

# RIDGE analysis of the MHC locus

## 5.1 The MHC locus

The aim of this PhD project has been to investigate whether *there are sub-genomic loci, RIDGEs, in the genome consisting of physically linked genes that are functionally related, and exhibit co-expression and co-regulation.* This chapter focuses on RIDGEs in a key immune locus, the MHC, in the common house mouse, *Mus musculus*. A specific case of RIDGE association with immune activation, e.g. the macrophage response, is described.

Immune system genes fall into a number of loci, such as the T cell receptor (TCR), the immunoglobulin (IG), and the MHC locus. These loci are, usually, further divided into a number of smaller loci, such as the MHC class Ia, the MHC class Ib, the extended MHC class II, the classical class II, and the MHC class III locus (Yuhki et al., 2003), and these five are all found next to one another on chromosome 17.

### 5.1.1 Rationale

The investigation of RIDGEs were restricted to the MHC locus because it is:

1. A known key immune locus where about half of the genes are known to be immune associated. (Yuhki et al., 2003)

2. Extensively studied and therefore the gene annotation (such as gene, gene products, and gene function) should be more extensive, and readily, available; furthermore more microarray probes on the microarray chips are expected.

3. A relatively gene dense region (The MHC sequencing consortium, 1999); about four times the genomic average.

Furthermore, genes in the immune system are subject to intense selection for disease resistance (as a result of interactions with pathogens) (Liu and Shaw, 2001; Trowsdale, 2002), this in

addition to the plasticity of the MHC molecules (Trowsdale, 2002) (they bind large numbers of peptides with high affinity and low specificy) imply that they are more likely to be organised into co-regulated units of functionally related genes. Finally, a RIDGE formation could reduce the energy (ATP) requirement of mounting an immune response. Mouse was chosen because it is the most closely related to human. (Twyman, Richard, 2002) Together these motivations imply that if there are RIDGEs in the genome, the MHC locus would be a good candidate for the identification and characterisation of RIDGEs before whole-genome analysis is attempted.

Many genes are already known to be functionally related in the immune response thereby, hopefully, making it easier to determine if RIDGE members are functionally related or not. Some genes are potentially regulated by the same TFs because they are known to have sequence similarities. Furthermore, a small, manageable, region is necessary for manual annotation of RIDGEs with gene function, gene characteristics, sequence similarities, and interaction partners. Finally the mouse genome is not fully sequenced, and because of the high importance of this region in disease resistance, and for the pharmaceutical industry, it is expected that the MHC locus would be better annotated than most regions. This would mean that more of the mouse genes are represented on the microarray chip, which is necessary since not even the "whole genome" array from Affymetrix (Mouse430) covers the entire mouse genome.

### 5.1.2 The biology

#### 5.1.2.1 Immunology

The immune system is the body's response to foreign materials, without it an organism would soon die from a virus or an infection. The skin acts as a barrier and is the first line of defence, stopping invaders from entering the body. Once the invader has successfully circumvented these walls, it needs to escape both the specific and the innate immune systems, and macrophages are one of the key components of the latter.

#### 5.1.2.2 mCMV (murine cytomegalovirus)

CMV is a member of the herpesvirus family that can establish both acute and chronic infections. (Lucin et al., 1992) For humans the symptoms are normally milder than the common cold, but in those with disrupted immune response (as in cancer) more serious symptoms are presented. (Martin and Hine, 2000) In fact human CMV is a major cause of morbidity and mortality in immunocompromised individuals. Even in immunocompetent hosts, mCMV cause disseminated acute infection, and persistently produces infectious virus in the salivary gland for up to months after induction of the specific immunity. This life-long latency suggest that the virus is able to evade, or modify, the immune systems response (Heise et al., 1998); and yet mCMV provokes strong responses by both the innate and the specific immune response (Heise

et al., 1998), for example by decreasing MHC expression. (Goldsby et al., 2003) Macrophages are a key cell type as it can act as a cellular site for latency and dissemination for mCMV in mouse. (Pollock et al., 1997; Stoddart et al., 1994; Popkin and Virgin, 2003)

#### 5.1.2.3 Macrophages

The microarray experiments used in this project investigate how the immune system (e.g. the macrophages) react to the introduction of foreign materials. Macrophages determines if a cell it comes into contact with is healthy, sick or dead; and whether or not the cell died from natural causes (such as aging) or, for example, from an infection. The main benefit of using macrophages is that they are in a rested state until they become primed (for example by IFN-γ) or excited (for example after a mCMV infection); this means that a comparison between control and a specific treatment is more likely to reflect changes related to the treatment, rather than off-target expression.

Differentiation from a monocyte to a macrophage involves a number of changes: 1) the cell is significantly enlarged, 2) intracellular organelles increase in both number and complexity, 3) phogocytotic ability increases, 4) more hydrolytic enzymes are produced, and 5) soluble factors are secreted. The macrophages are then dispersed throughout the body; some become fixed, i.e. they take up residence in a particular tissues, whereas others remain motile. (Goldsby et al., 2003)

Macrophages, normally found in their rested state, can be activated by a variety of stimuli, for example by 1) cytokines secreted by activated $T_H$ cells, for example IFN-γ, 2) mediators of the inflammatory response, or by 3) components of bacterial cell walls. Activated macrophages are more effective than rested in eliminating pathogens, because they exhibit increased: 1) phagocytotic activity, 2) ability to kill ingested microbes, 3) secretion of inflammatory mediators, and 4) ability to activate T cells. In addition, activated macrophages secrete various proteins that help eliminate a broad range of pathogens, including virus-infected cells, tumor cells, and intracellular bacteria. They also express higher levels of class II MHC molecules, allowing them to function more efficiently as antigen-presenting cells (APC). (Goldsby et al., 2003)

#### 5.1.2.4 Interferon

When cells are infected by a virus, they secrete proteins called interferons that increase the resistance of neighboring cells to infection by the same, or other, viruses. (Purves et al., 2001) These pro-inflammatory molecules are particularly important in limiting infection during the period when specific humoral and cellular immunity is developing (Lydyard et al., 2004), i.e. interferons are an integral part of the non-specific immune response. Interferons inhibit viral replication by binding to receptors in the plasma membranes. (Purves et al., 2001)

Interferons are divided into two groups, type I (IFN-α and IFN-β) and type II (IFN-γ), the latter is also called the immune IFN. In contrast to the broad and rather non-specific antiviral activity of type I, IFN-γ is primarily a cytokine of the adaptive immune system. It is produced by $T_H1$ cells, and natural killer (NK) cells. (Lydyard et al., 2004)

IFN-γ enhance the phagocytotic function of a macrophage, and enhance the function of professional APCs (Lydyard et al., 2004), such as macrophages and dendritic cells. Furthermore IFN-γ is crucial for macrophage function since it increase their ability to kill both intracellular bacteria and parasites. (Lydyard et al., 2004) IFN-γ have also been shown to induce expression of the class II transactivator (CIITA), thereby indirectly increasing expression of class II MHC molecules on a variety of cells, including non-APCs (e.g. skin keratinocytes, intestinal epithelials cells, vascular endothelium, placental cells, and pancreatic β cells) (Goldsby et al., 2003), and initiating the adaptive immune response.

### 5.1.3 Locus definition

Most vertebrates have an MHC structure and composition fairly similar to that of humans, although the gene composition and genomic arrangement varies. For example, chickens have one of the smallest known MHC locus with only 19 genes spanning 92 kbp (Yuhki et al., 2003), in comparison to the human locus that spans almost 4 mega bases and contains more than 200 genes, of which about half have known immunological function. (Trowsdale, 2002)

The five MHC loci have somewhat separate functions; both the MHC class I and class II loci encode heterodimeric peptide binding proteins, whereas the first also encodes antigen processing molecules such as TAP and Tapasin, and the latter proteins that modulate peptide loading onto MHC class II proteins; and finally the MHC class III locus encodes other critical molecules such as complement components (e.g. C2, C4, and factor B) and cytokines (e.g. Tumor Necrosis Factor (TNF)-α). (Goldsby et al., 2003) The MHC class I molecules are found on almost every nucleated cell, whereas the class II molecules only are found on specialised cell types (including macrophages, dendritic cells, activated T cells, and B cells; all of which are professional APCs). (Purves et al., 2001)

#### 5.1.3.1 Locus determination

The determination of the MHC loci borders is based on eight sources, in addition to all the data in the functional and molecular interaction database (section 3.1.2).

The first three sources (two articles and a book) focus on the entire MHC region: 1) the MHC Sequencing Consortium's update on the human HLA locus (The MHC sequencing consortium, 1999), 2) the complete gene map of the entire human HLA locus by Shiina et al (Shiina et al., 2004), and 3) the book Immunology with both the human and mouse HLA/MHC. (Goldsby et al., 2003)

In addition, more specialised sources are utilised, for example one article focused specifically on the MHC class II region in human, murine, and feline genomes. (Yuhki et al., 2003) To determine the MHC class Ia and class Ib borders, the following gene/gene group specific papers were used; 1) to determine the start of MHC class Ia three articles are used; Bartoloni et al (Bartoloni and Antonarakis, 2004), Huai et al (Huai et al., 2004), and Hui et al (Hui et al., 2006) and 2) to determine the end of MHC class I b only one article by Takade et al (Takada et al., 2003) was used.

#### 5.1.3.2 Locus description

The MHC locus in mouse stretch 5.8 Mbp, compared to 3.6 Mbp in humans, 3.8 in rat, 2.0 in pig, and 92 kbp in chicken. (Yuhki et al., 2003) The human HLA locus has 260 genes (Kelley et al., 2005), and the mouse MHC was, in this study, found to contain 250 genes. An earlier study showed that the human HLA locus contained 238 genes; 122 class I, 20 extended class II, 34 classical class II, and 62 class III genes. (Shiina et al., 2004)

| **Locus** | **Ia** | **Extended II** | **Classical II** | **III** | **Ib** | **MHC** |
|---|---|---|---|---|---|---|
| Physical start | 30 852 327 | 33 519 969 | 33 695 069 | 34 172 386 | 34 871 167 | 30 852 327 |
| Physical end | 33 455 642 | 33 676 696 | 33 977 115 | 34 861 733 | 36 681 079 | 36 681 079 |
| Genomic size | 2 603 315 | 156 727 | 282 046 | 689 347 | 1 809 912 | 5 828 752 |
| # Genes (+) | 64 (28) | 18 (9) | 16 (12) | 69 (36) | 77 (34) | 250 (121) |
| First Gene | Tff3 | Daxx | ? | Notch4 | H2-Q2 | |
| Last Gene | Cd320 | Col11a2 | Btnl2 | Ddx39 | Gabbr1 | |

Table 5.1: The MHC locus. For each of the 5 MHC loci, the genomic start, and end, positions, the length, the number of genes found, the number of genes found on the sense (+) strand, and the first, and last, gene is displayed. The MHC class Ia locus is the longest, but class Ib has the most genes, whereas the extended class II locus is the shortest, but the classical class II locus has the least number of genes.

When annotating these genes for immunological function the following was noticed:

- The interaction data in SORGEDB focus on the genes in the MHC class II pathway; where only 22, of the 250, genes have known interactions.
- Immune related genes are those additional 10 genes that were found in SORGEDB, although they had no associated interactions.
- 9 genes are class II associated (but as expected these mostly overlap with the other definitions).

- 5 genes have GO categories that involve the immune system directly, for example the GO category antigen presentation.

In summary, less than 25% of the genes found within the MHC are immune related according to the above programmatic definitions, whereas it is normally easier for the expert to manually determine this.

### 5.1.4 Experimental data

To explore whether or not mCMV infection is associated with RIDGEs the immune loci, MHC class I, II, and III were investigated using gene expression data from mouse macrophages assayed using the Affymetrix MG-UG74Av2 chip. Eight treatments (each with three biological replicates) were investigated (see 2.1.5.1 for details) and four are presented in detail; uninfected macrophages, macrophages primed with IFN-$\gamma$ (where the macrophage is primed for 24 hours and then harvested), macrophages viral activated with mCMV (where the macrophage is infected and harvested), and macrophages both primed and infected (where the macrophage is first primed for 24 hours and then infected and harvested).

#### 5.1.4.1 Microarray probes in the MHC locus

| Array | U74Av2 | | 430 | |
|---|---|---|---|---|
| Good projections | 102 | 41% | 184 | 74% |
| No known identifiers | 20 | 8% | 22 | 9% |
| Bad projections | 74 | 30% | 19 | 8% |
| Another gene for the probe | 25 | | 30 | |
| Badly designed probes | 96 | | 68 | |

Table 5.2: Probe-to-gene-projections in the MHC locus. Of the 250 genes in the MHC locus 192 genes have a good probe-to-gene projection (77%) when the two arrays are combined. There are 55 genes with a bad projection and 21 of these do not have any external identifiers (probably new additions to the Ensembl DB). The last two errors deal with badly designed probes, e.g. a probe with more than one Ensembl gene suggestion; for example the probe 102267_at has 15 different suggested Ensembl gene identifiers. Bad projections are those that do not fit in any other category, for example projections with a reliability score below the cutoff (0.5) meaning that the external sources disagree on the projection.

The ratio of validated Affymetrix probe identifier to Ensembl gene identifier projections (probe-to-gene projections as specified in 3.2.1) were expected to be higher inside the MHC locus

than outside, and it is. The percentage of good translations in the MHC locus are 41% for the Affymetrix MG-U74Av2 chip compared to 30% for the entire genome, and 74% for the 430 chip compared to 70%. In fact in the classical MHC class II region, 14 of the 16 genes can be translated (on the U74Av2 array), missing only H2-Ob and Btnl2.

That only 41% of the genes are found on the Affymetrix MG-U74Av2 chip means that the analysis is restricted to these 41%, e.g. an inherent caveat in the analysis.

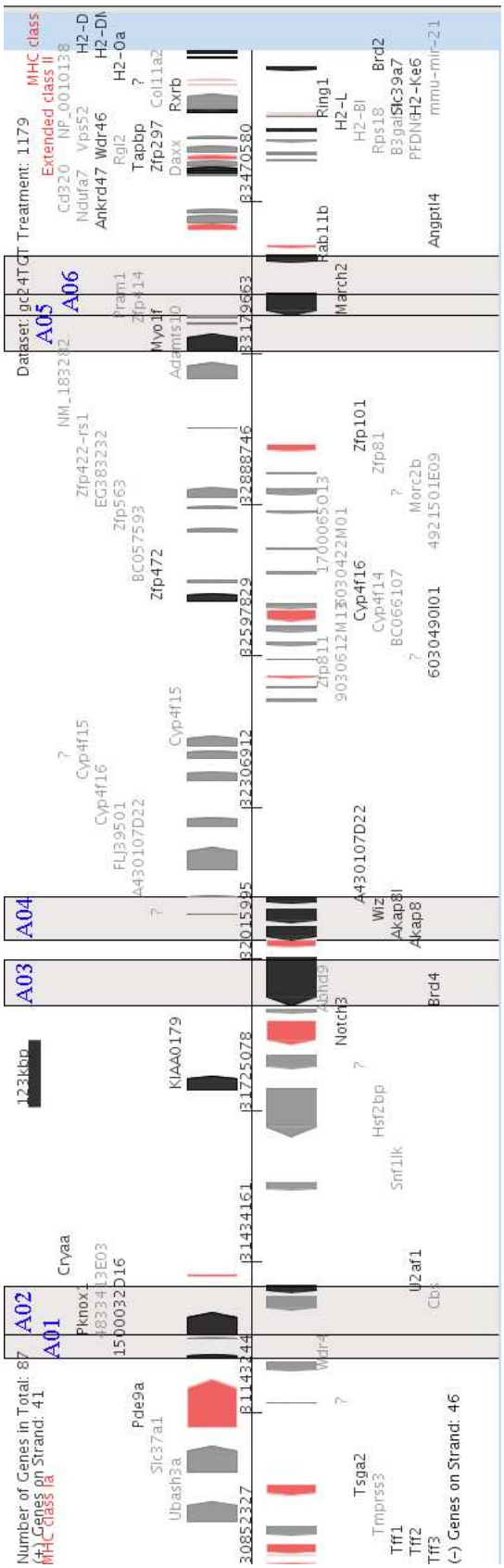
## 5.2 Identification of RIDGEs in the MHC locus by SORGE

For the remainder of this PhD thesis, unless otherwise specified, a RIDGE is defined in accordance with the Rosette model; a group of consecutive active gene(s) that cover around $110 \pm 30$ kbp of genomic data, where an active gene is defined as a gene where the majority of detection calls are considered present, and consecutive genes is according to 2D linear organisation of genes according to their physical positions.

The RIDGEs presented in this chapter are those found in either uninfected or activated macrophages (viral activation with mCMV, priming with IFN-γ, or the combination of both) restricted to the MHC locus in *Mus musculus,* whereas the next chapter describe the generalisation of RIDGEs to the entire genome and to additional datasets and loci.

RIDGEs can be divided into two classes; overlapping and non-overlapping. The regions of overlapping RIDGEs are analysed in this section and one suggested RIDGE is chosen for each region. These chosen RIDGEs, in combination with the non-overlapping RIDGEs, are presented in detail in the next section.

RIDGEs could be represented visually in the context of the genome (for example RIDGEs in the MHC class I locus as seen in figure 5.1 below), as simple present/not present textual representations (for example for the entire MHC locus in table 5.3 or alternatively as a more complex table where the RIDGE members individual gene expression profiles are shown (for example for the entire MHC locus in table C.1.




This visual representation of a genomic region, RIDGEs, and RIDGE members is created by SORGE Visualisation (section 3.3), although the RIDGE "names" (shown at the top of the figure, for example A01) were manually added.

The name of the current dataset (gc24TGT - macrophages that are both primed and viral activated) and the current treatment number (1179) is written in the top right corner. The desired RIDGE dimension, 123 kbp, is displayed as a box at the top.

RIDGEs are highlighted with a gray background (4 regions) and with a black border (6 RIDGEs). Active genes are shown in black (for example Pknox1), silenced in red (for example Notch3), and genes without a good probe-to-gene projection in grey (for example Cyp4f15).

The chromosome (17) is visualised as a horisontal line and annotated with evenly spaced genomic positions (for example the first RIDGE is located between position 311433244 and 31434161).

If the start of a loci is found within the current window, the loci name will be shown in red along the top of the panel (for example MHC class Ia in the top left corner).

Figure 5.1: RIDGEs found in the MHC class I locus

### 5.2.1 Overlapping RIDGEs

There are 4 overlapping RIDGE regions in the MHC locus for rested or activated macrophages (as seen in the table, C.1, in appendix C); 1) A01 and A02 overlap, 2) A05 and A06 overlap, 3) A07 and A08 overlap, and finally 4) the 12 RIDGEs in the Ier3:H2-L region, A11-A22, are all overlapping to some degree.

#### 5.2.1.1 RIDGEs in the Ier3:H2-L region - A11-A22

The Ier3:H2-L region is a 300 kbp long gene dense region with 22 genes, of which only 12 have a reliable probe-to-gene projection. If silenced genes are not allowed in a RIDGE then 12 RIDGEs, for the four conditions, are found in this region (see figure 5.2 below and table C.1). These 12 RIDGEs are all overlapping to some degree; for example the only difference between the two suggestions, A14 and A15, is that the latter also contains the gene GNL1 (and therefore only present in macrophages that are both primed and viral activated). A 300 kbp long region could potentially be divided into 3 RIDGEs around 100 kbp each, but this did not fit the data, instead two RIDGEs are created; A15 and A20 including 9 of the 12 projected genes.

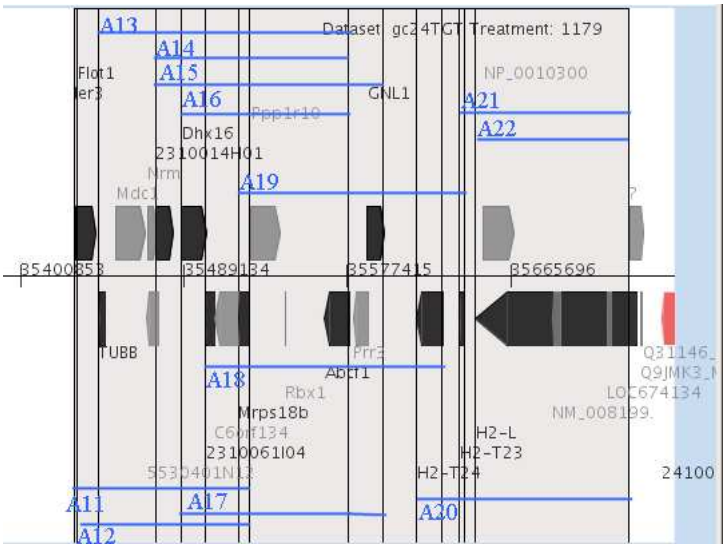



Figure 5.2: RIDGEs in the Ier3:H2-L region. The 22 genes in this region, from the 5' to 3', are (where genes without a reliable probe-to-gene projection (shown in grey in the figure) are shown in italic); Ier3, Flot1, TUBB, *Mdc1*, *553041N12Riken, Nrm*, 2310014H01 (KIAA1949), Dhx16, 2310061I04Riken, *C6orf134*, Mrps18b, *Ppp1r10*, *Rbx1,* Abcf1, *Prr3*, GNL1, *NP_0010300*, H2-T24, H2-T23, H2-L, *NM_008199,* and *LOC674134*.

The 12 projected genes have the following functions;

- Ier3 is associated with the cell cycle, cell death (apoptosis), DNA replication, recom-

bination, and repair (Ingenuity Systems), cellular development, tissue development, the immune response, (Ingenuity Systems) and is integral to the membrane. (The Gene Ontology Consortium, 2000)

- Flot1 is associated with the cytoskeleton, integrin signaling (Jonathan Kerr, 2006), protein binding, and is integral to the membrane. It is found in the caveola, the membrane, and in the flotillin complex. (The Gene Ontology Consortium, 2000)

- TUBB is associated with the cell cycle, cell death (apoptosis), DNA replication, recombination, and repair (Ingenuity Systems), GTPase activity, structural molecule activity, cellular function, immunological disease, hematological disease, (Ingenuity Systems) and binding (of MHC class I proteins, GTP, and nucleotide). (The Gene Ontology Consortium, 2000)

- The function of KIAA1949 is not known, but it is found in Ingenuity in a network with PPP1CA and there associated with cancer, cellular compromise, and nervous system development and function. (Ingenuity Systems)

- Dhx16 is involved with RNA splicing, mRNA splicing, ATP activity, and RNA post-transcriptional modification. (The Gene Ontology Consortium, 2000)

- The function of 2310061I04 is not known.

- Mrps18b participates in protein biosynthesis. (The Gene Ontology Consortium, 2000)

- Abcf1 is involved with ATPase activity, ATP binding, (The Gene Ontology Consortium, 2000) TF activity, translation, inflammatory response, and protein synthesis, for example biosynthesis. (Ingenuity Systems)

- GNL1 binds nucleotides and GTP. (The Gene Ontology Consortium, 2000)

- H2-T24 participates in the immune response. (The Gene Ontology Consortium, 2000)

- H2-T23 is involved with the immune response (The Gene Ontology Consortium, 2000) and with cellular growth, inhibition, assembly, and organisation. (Ingenuity Systems)

- H2-L is involved in antigen presentation, and is integral to the membrane. It is found in the MHC class I protein complex and in the membrane. (The Gene Ontology Consortium, 2000)

There are therefore two broad functional groups here; the first is involved in the cell cycle and the last three genes in antigen processing. Ideally, the last three genes (H2-T24, H2-T23, and H2-L) therefore constitutes one RIDGE, both based on functional associations and gene

expression profiles, A20 that is 115 kbp long. This would leave 185 kbp, e.g. potentially two other RIDGEs. Of the 12 suggestions (A11-A22) the last five (A18-A22) all contain one, or more, of the genes within the A20 RIDGE and are therefore not considered.

- On the basis of functional associations between Ier3 and TUBB, and Flot and TUBB, and the large (27 kbp) inter-gene distance between TUBB and KIAA1949, a break between TUBB and KIAA1949 is suggested. The Ier3:TUBB part is not long enough (17 kbp) to make up a RIDGE on its own, and 5' of Ier3 there is a 122 kbp inter-gene distance to a U6 gene (which does not have a reliable probe-to-gene projection) and even further 5' is Ddr1, a 157 kbp long region. This division leaves RIDGE suggestions A14-A17.

- Evidence in favour of RIDGE A11 is the large (40 kbp) inter-gene distance between Mrsp18b and Abcf1 containing two genes without reliable probe-to-gene projections.

- In favour of A15 and A17 is that they both contain all the genes associated with protein biosynthesis (Dhx16, Mrps18b, Abcf1, and GNL1).

- A15, A17, A18, and A19 all share the same 15 Transcription Factor Binding Sites (TFBS - see 2.2.3.2) in their upstream regulatory regions which is evidence in favour of collapsing these borders into one large RIDGE. But this region would then be 167 kbp long, and furthermore both A18 and A19 contains genes within RIDGE A20, arguing in favour of A15 and A17. In addition both A15 and A16 have significant ClustalW *upstream sequence similarity score* (UTR score), whereas neither have significant ClustalW *coding sequence similarity score* (gene score).

- In favour of A17 is the figure 5.3; where KIAA1949 (the first gene) participate in a network on its own, whereas the other genes all participate in the same network.

From the above data, both A15 and A17 could have been chosen as the representation of the region, but here A15 is chosen because it contains all the genes within A17, but also because GNL1 share the GO-term nucleotide binding with Abcf1 and intracellular location with Mrps18b.

Other examples of possible divisions include; 1) Ier3:GNL1 which is too long with 167 kbp, 2) Ier:Abcf1 which is also too long with 148 kbp, 3) Ier3:2310061I04 which is too short with 75 kpb, and 4) Flot:2310061I04 which is also too short with 74 kbp.

#### 5.2.1.2 A01 (1500032D16, Pknox1, and U2af1) or A02 (Pknox1, and U2af1)

RIDGE A02 is chosen to represent this region since A02 have more functional associations between the RIDGE members, although this could be due to the limited functional data available for 1500032D16. Furthermore RIDGE A01 has two missing genes, whereas RIDGE A02 only

has one, and is therefore more likely to be real RIDGE. Finally 1500032D16 have three long transcripts with many exons, probably making this RIDGE more complex from an transcriptional regulation point of view.

#### 5.2.1.3 A05 (Myo1f and March2) or A06 (March2 and Rab11b)

RIDGE A05 contains two missing genes, whereas A06 does not contain any, which is why it was chosen. Both RIDGEs have similar functions, but if they are collapsed the region becomes too long (185 kbp). Another piece of evidence in favour of collapsing these RIDGEs into one is that they have, more or less, the same TFBS in their regulatory regions.

#### 5.2.1.4 A07 (Psmb8, Tap2, H2-Ab1, H2-Aa, and H2-Eb1) or A08 (Tap2, H2-Ab1, H2-Aa, and H2-Eb1)

Two RIDGEs fall inside the classical MHC class II locus. The longest version, RIDGE A07, is chosen to represent the region as it contains all the genes in A08, but also because PSMB8 is known to interact with the other members and therefore more likely to be included than excluded. Both RIDGEs share the same TFBS in their regulatory regions. Both versions contains the same missing gene, H2-Ob.

### 5.2.2 Discussion

This manual analysis of overlapping RIDGEs in the MHC locus for rested or activated macrophages would not have been possible to perform for the whole genome. From these four examples it is clear that there are no rules governing the RIDGE representation, since this is mainly based on subjective measurement of functional relations between RIDGE members. For two examples (A07 and A15) the longest version is chosen, but not in the others. The first gene is not always included in the chosen RIDGE (A02, A05, and A15).

The initial RIDGE analysis was focused on a well-defined key immune loci in order to be make it possible to map the biology and determine if the results are biologically meaningful, and this is the reason that overlapping RIDGEs were pruned from the dataset in order to reduce the noise in the dataset. In the next section only the non-overlapping RIDGEs and those RIDGEs chosen to represent an overlapping region will be presented.

## 5.3 RIDGE analysis in the MHC locus

Chromosome 17 in *Mus musculus* is 95.2 Mbp long and contains 1264 genes. The five MHC immune loci; MHC class Ia, the extended MHC class II, the classical MHC class II, the MHC class III, and the MHC class Ib locus are all located on this chromosome, spanning 5.7 Mbp

(about 6% of the chromosome length) and contains 250 genes (about 20% of the genes on the chromosome). Most of the genes in the MHC are known to be immune system genes (Yuhki et al., 2003), although here only about 25% of the genes in the locus are classified as immune systems genes, indicating that functional data is missing even for this important locus. Of the 250 genes in the MHC locus only 41% are represented, with a reliable probe-to-gene projection, on the Affymetrix MG-U72Av2 chip, so only these 102 genes have associated gene expression data meaning that only these are considered in the RIDGE analysis. 22 RIDGEs are found in the MHC locus including 2 in the classical class II locus and 12 in the Ier3:H2-L region. When one silenced gene is allowed within a RIDGE 10 new RIDGEs are found, where 7 are inside the classical MHC class II locus.

| RIDGE identifier: Genes in the RIDGE | rested | IFN | mCMV | both |
|---|---|---|---|---|
| **A02**: Pknox1, U2af1 | | | | P |
| **A03**: Brd4 | P | P | P | P |
| **A04**: Akap8l, Wiz, A430107D22 | P | | P | P |
| **A06**: March2, Rab11b | P | P | P | P |
| **A07**: Psmb8, Tap2, H2-Ab1, H2-Aa, H2-Eb1 | | P | P | P |
| **A09**: H2-Q2 | P | P | P | P |
| **A10**: Cdsn, Gtf2h4 | P | | | |
| **A15**: KIAA1949, Dhx16, 2310061I04, Mrps18b, Abcf1, GNL1 | | | | P |
| **A20**: H2-T24, H2-T23, H2-L | P | P | P | P |

Table 5.3: The left most column shows the RIDGE identifier in bold (for example A02) and in genomic order the genes, members, defining the RIDGE. For each of the four conditions (uninfected macrophages, primed with IFN-γ, viral activated with mCMV, and both primed and viral activated macrophages) a RIDGE is marked with a P if it is present for that condition. The gene responsible for silencing the RIDGE is highlighted in red.

The RIDGE activity score (as described in 2.3.1) identifies groups of genes, RIDGEs, that are significantly more, or less, active than random genes. Most genes in the MHC locus, especially genes associated with the innate immune response, need to respond quickly to macrophage activation, and should therefore not display significant ($p < 0.05$) low RIDGE activity scores. Yet four RIDGEs are found, indicating that their members are less active after immune system activation; 1) A01 (1500032D16, Pknox1, and U1af1), 2) A02 (Pknox1 and U2af1), 3) A04 (Akap8l, Wiz, and A430107D22), and 4) A10 (Cdsn and Gtf2h4). Neither of these RIDGEs have members that are associated with the immune response. All four have functional associations between their members in the sense that they encode genes with general functions, neither has a significant *coding sequence similarity score* nor a significant *upstream sequence similarity score*, and neither is considered to have many shared TFBS in the upstream

regions.

### 5.3.1 Gene expression profiles for RIDGE members

| RIDGE identifier: Genes in the RIDGE | rested | IFN | mCMV | both |
|---|---|---|---|---|
| **A02**: Pknox1, U2af1 | AP | ↓↓ | AP | P↑ |
| **A03**: Brd4 | P | P | P | P |
| **A04**: Akap8l, Wiz, A430107D22 | PPP | PP↓ | ↑P↑ | ↑↓P |
| **A06**: March2, Rab11b | PP | PP | PP | PP |
| **A07**: Psmb8, Tap2, H2-Ab1, H2-Aa, H2-Eb1 | PPPAA | ↑P↑↑↑ | ↑↑↑↑↑ | ↑↑↑↑↑ |
| **A09**: H2-Q2 | P | P | P | P |
| **A10**: Cdsn, Gtf2h4 | PP | AP | AP | AP |
| **A15**: KIAA1949, Dhx16, 2310061I04, Mrps18b, Abcf1, GNL1 | PPPPPM | ↑↓P↓↓M | P↓PP↑M | PPP↑PP |
| **A20**: H2-T24, H2-T23, H2-L | PPP | P↑↑ | ↑↑↑ | ↑↑↑ |

Table 5.4: Significant changes in gene expression profiles levels for the RIDGE members where the arrow shows the direction of regulation, up or down.

| | rested | IFN | mCMV | both |
|---|---|---|---|---|
| Pknox1 | A=86 | **-2.6** | -1.1 | -1.6 |
| U2af1 | P=386 | **-3.0** | 1.0 | **1.4** |
| Brd4 | P=111 | 1.3 | 1.2 | 1.4 |
| Akap8l | P=57 | 1.3 | **1.6** | **1.9** |
| Wiz | P=82 | -1.1 | -1.4 | **-1.1** |
| A430107D22 | P=63 | -2.5 | **1.9** | 1.6 |
| March2 | P=412 | -1.4 | -1.2 | -1.3 |
| Rab11b | P=297 | -1.2 | -1.1 | -1.1 |
| Psmb8 | P=889 | **2.3** | **1.6** | **2.9** |
| Tap2 | P=89 | 1.1 | **1.9** | **4.6** |
| H2-Ab1 | P=30 | **7.7** | **3.4** | **40.7** |
| H2-Aa | A=64 | **9.8** | **2.3** | **38.4** |
| H2-Eb1 | A=39 | **8.3** | **2.7** | **46.9** |

| | rested | IFN | mCMV | both |
|---|---|---|---|---|
| H2-Q2 | P=248 | 1.4 | 1.4 | 1.8 |
| Cdsn | P= 54 | 1.0 | -1.2 | -1.0 |
| Gtf2h4 | P=137 | -1.5 | 1.1 | 1.2 |
| KIAA1949 | P=912 | **1.2** | -1.1 | -1.1 |
| Dhx16 | P= 94 | **-1.5** | **-1.5** | -1.6 |
| 2310061I04 | P=113 | -1.1 | 1.3 | 1.2 |
| Mrps18b | P=190 | **-1.7** | -1.0 | **1.4** |
| Abcf1 | P=351 | **-2.3** | **1.1** | 1.1 |
| GNL1 | M=146 | -1.3 | -1.2 | 1.1 |
| H2-T24 | P=134 | 1.1 | **2.3** | **2.2** |
| H2-T23 | P=126 | **3.1** | **2.3** | **3.9** |
| H2-L | P=270 | **3.3** | **3.5** | **4.5** |

Table 5.5: The level of regulation, foldchange, where bold letters denote significant changes according to $foldchange > 2$ or $p < 0.05$, and red an absent, or marginal, call. For the uninfected macrophages the detection call (Present, Absent, or Marginal) is shown along with the mean value signal intensity for the gene.

Of the above nine RIDGEs, four (e.g. A03, A06, A09, and A20) are considered static under these four condition, although as can be seen for RIDGE A20, the genes are not necessarily static. This RIDGE consists of genes involved in antigen presentation and would therefore be expected to change under these biological conditions, and the three RIDGE members are all upregulated for these conditions, except for H2-T24 in primed macrophages. Similar to RIDGE A20, all members are upregulated for all conditions for RIDGE A07, except for Tap2 in primed macrophages. The last three genes in this RIDGE (H2-Ab1, H2-Aa, and H2-Eb1) are around

40 times upregulated in macrophages that were both primed and viral activated, but note that although Psmb8 show a lower increase (2.9) in comparison, the actual gene expression level is still the twice as high for Psmb8 because of the difference in gene expression levels for the rested macrophages.

### 5.3.2 RIDGE gain in macrophages that were both primed and viral activated

There are two genes; Pknox1 and GNL1 that are only active after both priming with IFN-γ and viral infection with mCMV. These genes affect the corresponding RIDGEs, A01-2, A15, and A17-19. RIDGE A01 and A02 were overlapping, and therefore A02 is the only RIDGE presented here.

Evidence for the formation of the two Pknox1 RIDGEs are functional associations between the RIDGE members (both genes are associated with transcription and localise to the nucleus), sequence similarity in the upstream regions, and low RIDGE activity scores. Evidence for the four GNL1 RIDGEs are functional associations between the RIDGE members and low RIDGE activity scores.

#### 5.3.2.1 A02 - (Pknox1 and U2af1)

Pknox1 is involved in TF activity, RNA polymerase II TF activity and regulation of transcription. It is found in the TF complex (The Gene Ontology Consortium, 2000), and is a homeobox protein. (Bairoch et al., 2005) Genes involved in transcription, especially in a regulatory capacity, are potential targets for enhancing, or silencing, an immune response; here the gene is downregulated (albeit from an absent state) in primed macrophages and enhanced in macrophages that were both primed and viral activated. The Ingenuity Pathway Analysis software (Ingenuity Systems) further suggests that Pknox1 is involved in organ morphology, visual system development and function, genetic disorder, gene expression, cardiovascular system development and function, organismal development, survival, and binding of DNA, enhancer, and proteins. The second gene in the RIDGE, U2af1, is involved in RNA processing (The Gene Ontology Consortium, 2000) and RNA post-transcriptional modifications (Ingenuity Systems) in addition to metal ion binding. (The Gene Ontology Consortium, 2000) As this gene is also involved with transcription, it is also expected to change; both genes are downregulated after priming, but U2af1 is also upregulated in macrophages that were both primed and viral activated. The third gene in RIDGE A01, 1500032D16, is involved in cell signaling, energy production and metabolism of vitamins, minerals, and carbohydrate. (Ingenuity Systems) Energy production and metabolism are both functions that would be expected to change after immune activation, since resources are limited and should be redirected toward immunological resistance (Colditz, 2002); but the gene 1500032D16 is static for these four conditions.

RIDGE A02 consists of Pknox1 (41 kbp long) and U2af1 (719 bytes long) and has a significant low RIDGE activity score (1.38) (see 2.3.1 for theory and 4.2.2.3 for distribution of scores). The RIDGE is 94 kbp long, from the genomic start of Pknox1 to the genomic end of U2af1, with an exon density of 13%. This measures the length of all exons for all RIDGE members and divides it by the RIDGE length. The remaining 87% is a 41 kbp (44%) inter-gene distance between the two, in addition to introns. This inter-gene distance is occupied by one missing gene (Cbs), e.g. a gene that does not have a reliable probe-to-gene projection for the MG-U74Av2 chip . The genes have 2 and 1 transcript respectively of varying lengths (866-4083) and number of exons (8-21) (compare to distributions in 4.2.3).

Neither the ClustalW *coding sequence similarity score* (gene score) or the *upstream sequence similarity score* (UTR score) are significant, as seen by only 24 shared potential transcription factor binding sites (TFBS) as identified by PROMO (Messeguer et al., 2002); AhR, AP-1, C/EBPα, C/EBPβ, c-Fos, c-Jun, CRE-BP2, f(α)-f(ε), GAPA-1, GR, HES-1, HOXA5, JunD, MyoD, myogenin, NF-1, NF-κB, Pax-6, Sp1, TCF-2, TFE3-S, USF-1, YY1, and YY1. U2af1 and Pknox1 are both hubs in their respective networks as reported by the Ingenuity Pathway Analysis software (Ingenuity Systems) but these networks can not be merged.

#### 5.3.2.2 A15

GNL1 are found in 4 RIDGEs with 7 other genes in the Ier3:H2-L region, but here A15 is presented. (See 5.2.1.1 for a brief overview of functional associations for these genes.)

GNL1 is a GTP-binding protein (The Gene Ontology Consortium, 2000), i.e. a regulatory protein. GTP-binding proteins act as molecular switches and control a wide range of biological processes including receptor signaling, intracellular signal transduction pathways, and protein synthesis. (Gavel, Y, 2008) Heterotrimeric G proteins acts as molecular switches (Oldham and Hamm, 2008), and the G-domain in many regulatory GTPase acts as a molecular switch in many different cellular processes, such as control of cell growth and differentiation, protein trafficking, and signal transduction. This ability to act as a molecular switch (similar to GNL3L (Rao et al., 2006)) mean that it could be a viable target for immune activation. Yet, here no significant change in gene expression level was detected.

Functional associations between the RIDGEs were found, such as a number of overlap in GO-terms, although neither KIAA1949 nor 2310061I04 have any associated GO-terms. Additional function associations between Dhx16, Mrps18b, Abcf1, and GNL1 is that they are all associated with protein synthesis, and furthermore H2-T24 and H2-T23 are both involved in the immune response. The shared GO-terms are; Dhx16 and Abcf1 share both ATP binding, and ATPase activity; Abcf1 and GNL1 share nucleotide binding; and GNL1 and Mrps18b share intracellular location. (The Gene Ontology Consortium, 2000)

A15, has a RIDGE activity score of 2.88, which is not significant. The RIDGE is 123

kbp long which is the *de facto* RIDGE dimension. It has a high exon density (30%) even though most genes are short (<13 kbp) because the intra-gene distances for this RIDGE are also short (<14 kbp) expect for in between Mrps18b and Abcf1 (40 kbp). The shortest inter-gene distance, and the only significant one, is between Dhx16 and 2310061I04 (17 bp). The seven genes in these four RIDGEs have one transcript each, expect for Abcf1 and GNL1 that both have 2, long, transcripts. Dhx16 has one, long, transcript with many exons (20), whereas 2310061I04, and Mrps18b have few exons (4 and 7 respectively). 4 missing genes (C6orf134, Ppp1r10, Rbx1, and Prr3) are present in this RIDGE (see figure 5.2 for their location within the RIDGE).

This RIDGE is co-regulated, and all four RIDGE suggestions (A15, A17, A18, and A19) share the same 15 TFBS in their regulatory regions, among these C/EBPα and β, f(α)-f(ε), HOXA5, and YY1, suggesting that the RIDGE borders should be collapsed, i.e. this should be considered one long regulatory region stretching 167 kbp (see discussion in 5.2.1.1). Neither the gene score nor the UTR score are significant. The networks for the different genes can be merged in the Ingenuity Pathway Analysis software (Ingenuity Systems) (as seen below), supporting the decision to choose this RIDGE to represent the entire A11-A19 region (see 5.2.1.1).

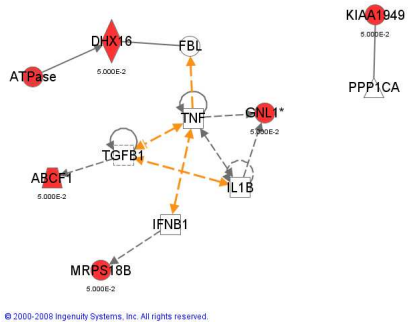



Figure 5.3: Network for the RIDGE members in A15 - Red nodes are RIDGE members and, orange arrows merge the five separate networks. Ingenuity Pathway Analysis software (Ingenuity Systems).

The first unknown gene, KIAA1949, is observed in the Ingenuity Pathway Analysis software (Ingenuity Systems) whereas the second, 2310061I04, is not.

### 5.3.3 RIDGE loss in primed macrophages

One gene, A430107D22, is active in all but primed macrophages, and so is the RIDGE A04 (containing Akap8l and Wiz as well). This RIDGE has a significantly low RIDGE activity score.

Akap8l is involved in binding (of nucleic acid, DNA, zinc ion, DEAD/H-box RNA helicase, and metal ion), cancer, reproductive system disease, cell death, kinase activity, and DNA replication, recombination, and repair. The gene is located intracellular in the nucleus, the chromosome, the nucleoplasm, the nuclear matrix, and in the cytoplasm. (The Gene Ontology Consortium, 2000) Akap8l belongs to the AKAP95 family, and could play a role in constitutive transport element mediated gene expression. (Bairoch et al., 2005) Wiz is involved in binding of nucleic acid, zinc ion, and metal ion, and located intracellular. (The Gene Ontology Consortium, 2000) According to iHOP (Hoffmann and Valencia, 2004, 2005) Wiz is a G9a/GLP-associating zinc finger molecule that interacts with G9a and GLP independently but is more stable in the G9a/GLP heteromeric complex, and it has two potential CtBP binding sites. GLP deficiency decreases G9a levels, suggesting that the Wiz/G9a/GLP tri-complex may protect G9a from degradation, links the G9a/GLP heteromeric complex to the CtBP co-repressor machinery, and finally the Wiz small inhibitory RNA also knocks down G9a. A430107D22 is involved in GTPase activator activity and regulation of small GTPase mediated signal transduction, and also located intracellular. (The Gene Ontology Consortium, 2000)

Akap8l and Wiz share the ability to bind large groups of molecules and the intracellular location, the latter also shared by A430107D22; all 4 GO-terms for Wiz are shared by Akap8l; nucleic acid binding, zinc ion binding, metal ion binding, and the intracellular location.

Small GTPases serve as molecular switches to regulate growth, morphogenesis, cell mobility, axonal guidance, cytokinesis and trafficking (Pierce), and GTPase is furthermore responsible for controlling protein synthesis. (Gavel, Y, 2008) GTPase are known to be immune related, for example Mx proteins are $\alpha/\beta$ IFN-inducible large GTPases with potent antiviral activities against specific groups of viruses, for example the murine Mxl protein is essential for protection against influenza viruses. (Johannes et al., 1993) Making it interesting that this gene is downregulated to become absent in primed macrophages and upregulated in viral activated macrophages.

RIDGE A04, has a significantly low RIDGE activity score. The RIDGE is one of the shorter RIDGEs, only 82 kbp long, with a high exon density of 32%, which follows from long ($>$13 kbp) genes and normal ($<$13 kbp) inter-gene distances. Yet one missing, unknown, gene is present in this RIDGE. Wiz has 4, long transcripts, and A430107D22 has 2, long, transcripts, whereas Akap8l only has one. The number of exons varies from normal to many (8-34). Neither the gene score nor the UTR score are significant, and only 21 TFBS were shared by the 3 genes; AP-1, C/EBP$\alpha$ and C/EBP$\beta$, c-Fos, c-Jun, COE1, CREP-BP2, f($\alpha$)-f($\varepsilon$), GR, HES-1, HOXA5, JunD, MyoD, myogenin, NF-1, NF-AT1, NF$\kappa$B, PU.1, Sp1, USF-1, and YY1. Akap8l and Wiz are found in two separate networks (and the third gene is not found at all in the Ingenuity Pathway Analysis software (Ingenuity Systems)), and Akap8l is a network hub. (Ingenuity Systems)

### 5.3.4 RIDGE gain in activated macrophages

One RIDGE, A07, is active in activated macrophage. For this RIDGE the method to determine active genes has a strong influence on gene presence; if the detection calls are used then Psmb8, Tap2, and H2-Ab1 are present, whereas H2-Aa and H2-Eb1 are absent; if the signal intensity has to be above 50, then Psmb8 and Tap2 are still present, whereas H2-Ab1 is absent and H2-Aa and H2-Eb1 are present.

Rmcs2 is the name found in SORGE DB for this gene, although the Ingenuity Pathway Analysis software (Ingenuity Systems) mapped it to H2-Ab1 and HLA-DRA. H2-Ab1 is a common knock-out target in MHC class II deprived mice (Mazmanian et al., 2005; Sakoda et al., 2007), and subsequently the gene is found to be significantly upregulated in activated macrophages. For both of the MHC class II RIDGEs (A07 and A08) functional associations and many regulatory regions, TFBS, are found.

There are a lot of functional overlap between these genes, for example; 1) all five genes are part of the antigen presentation pathway (The Gene Ontology Consortium, 2000), 2) all five genes are part of the signaling pathway, 3) all genes, but Tap2, participate in immune and lymphatic system development and function, 4) all genes, but Tap2, are involved in cell-to-cell signaling, 5) all genes, but Psmb8, participate in immunological disease (Ingenuity Systems), 6) all genes, but Psmb8, are integral to membrane (The Gene Ontology Consortium, 2000), 7) H2-Ab1, H2-Aa, and H2-Eb1 are involved in hematological system development and function, 8) H2-Aa and H2-Eb1 participate in IL-4 signaling, 9) both TAP2 and PSMB8 participate in the protein ubiquitination pathway, 10) Tap2 and H2-Ab1 are involved in diabetes, and 11) H2-Aa and H2-Eb1 participate in dermatological diseases and conditions. (Ingenuity Systems) All genes, but H2-Ab1, are interacting in the molecular interaction database (SORGEDB).

RIDGE A07 has a RIDGE activity score of 3.41 which is not significant, which is intriging. The RIDGE consists of immune system genes and would therefore be expected to have a high activity score under these conditions. The RIDGE is 115 kbp long and has an exon density of 8%, where the five genes are between 3 and 12 kbp long with varying inter-gene distances (3-45 kbp), and contains the missing gene H2-Ob. All genes, except for H2-Ab1 that has two, short transcripts, have a single transcript with few to normal number of exons (5-12). Neither the gene score nor the UTR score are significant, and still 18 shared TFBS were found for the 5 genes; AP1, C/EBP$\alpha$ and $\beta$, c-Jun, c-Fos, COE1, CRE-BP2, GR, HES-1, HOXA5, JunD, MyoD, myogenin, NF-1, NF-AT1, NF-$\kappa$B, PU.1, USF-1, and YY1. These RIDGE members are found in a network associated with immune response, immunological disease, and cell-to-cell signaling and interaction. (Ingenuity Systems)

### 5.3.5 RIDGE loss in activated macrophages

Both the projected genes (Cdsn and Gtf2h4) and the missing genes in RIDGE A10 encode hair and skin functions. Normal regulation of skin and hair functions could be considered less prioritised upon immune activation, which could explain why the Cdsn is suppressed after activation of the immune system. Another evidence for this RIDGE formation is the high RIDGE activity score.

Cdsn is involved with cell adhesion, found in the desmosomes (The Gene Ontology Consortium, 2000), and is expressed in the upper dermis and inner root sheat. It is an intracellular junction, cell-cell adhesion glycoprotein that anchors intermediate filaments to cell membranes, bridge adjacent keratinocytes, is involved in hypotrichosis simplex of the scalp (HSS) and associated with Netherton syndrome and psoriasis. (Davalos et al., 2005; McGrath, 2005; McGrath and Wessagowit, 2005; Schmitt-Egenolf et al., 2001) Furthermore, Cdsn is associated with hair and skin development and function, organ development, dermatological diseases and conditions, genetic disorder, and tissue development according to the Ingenuity Pathway Analysis software (Ingenuity Systems). The second gene, Gtf2h4 or TFIIH (Bairoch et al., 2005), is involved in transcription, TF activity, RNA polymerase II TF activity, DNA repair, and response to DNA damage stimulus. (The Gene Ontology Consortium, 2000) It is also a nucleotide excision and repair factor implicated in human DNA repair disorder, xeroderma pigmentosum, Cockayne syndrome, and brittle hair disease. (Lanning and Lafuse, 1999; May et al., 2004; Snoek et al., 2000) Point mutations in TFIIH, a DNA repair and basal transcription factor, can cause brittle hair, developmental abnormalities, reduced life span, UV sensitivity, and skin abnormalities. (de Boer et al., 1998) Gtf2h4 is furthermore associated with DNA replication, recombination, and repair, cell signaling, gene expression; part of the nucleotide excision and repair pathway and the estrogen receptor signaling pathway. (Ingenuity Systems) The gene is found in the nucleus and the TF complex, (The Gene Ontology Consortium, 2000) and is associated with MHC class II and IFN-$\gamma$ response.

RIDGE A10 has a high RIDGE activity score although no significant changes in gene expression levels are seen for these four conditions. When all nine conditions are considered the gene Gtf2h4 has a small upregulation in primed, viral activated and then rested for 24 macrophages, a small downregulation when it has rested for 72, hours and a downregulation in serum free, i.e. starved, macrophages.

The RIDGE is 122 kbp long which is close to the ultimate RIDGE dimension. Both genes are short ($<$6 kbp) and the inter-gene distance long (111 kbp) so would expect a lower exon density than the found 10%. Both genes have a single transcript; with 2 exons for Cdsn and 14 for Gtf2h4.

The large inter-gene distance between the two RIDGE members are occupied by 3 missing genes; 1) 230000M23, 2) Diffuse Panbronchiolitis Critical Region 1 (Dpcr1), and 3) Valyl-

tRNA synthetase 2-like (Vars2l). 230000M23 is most likely the mouse version of simian taste bud-specific gene (STG), as inferred from comparing the mouse and human PSORS1 locus. The gene is also expressed in other tissues, such as tonsils and skin. (Sánchez et al., 2004) Dpcr1, also known as CD26, is an antigen expressed on both B-cells, T-cells, and macrophages, and keratinocytes (and thereby found in skin). (COPE, 2008) Vars2l (also known as Bat6 or G7a) is involved in regulation of transcription and translation, specifically in deacetylation of misformed Thr-tRNA, it is further involved in synthesis of dinucleotide polyphosphate signaling molecules, tRNA processing, and can act as a cytokine. The protein product Vars2l is found in skin. (EMBL, a,d,c,b) The latter two genes are therefore both expressed in skin and act as cytokines, which is a specialised form of growth factors (e.g. regulatory peptide factors) (COPE, 2008), and Vars2l and Gtf2h4 are both involved in transcription.

Neither the gene score nor the UTR score are significant, and only 25 TFBS are shared by the 2 genes; AhR, AP1, c-Fos, c-Jun, C/EBPα, C/EBPβ, COE1, f(α)-f(ε), GATA-2, GR, HES-1, HNF-3β, HOXA5, JunD, MyoD, myogenin, Mitf, NF-1, NF-AT1, NF-κB, Pax-5, PU.1, Sp1, USF-1, and YY1. The genes are found in two separate networks that cannot be merged according to the Ingenuity Pathway Analysis software (Ingenuity Systems).

### 5.3.6 Static RIDGEs

The four static RIDGEs are one in the Ier3:H2-L region (A20, although there are 5 overlapping RIDGEs in this region that also are static), one in the Myo1f:Rab11b region (A06, as well as A05), and two single-gene RIDGEs (A03 and A09).

The first RIDGE, A03, is the single-gene RIDGE Brd4, with a RIDGE activity score of 4.54, which is not significant. The RIDGE is 89 kbp long and has a low exon density (7%). For a single-gene RIDGE it is not possible to determine the sequence similarity scores. Neither is a PROMO investigation of shared TFBS interesting (the gene have 29 TFBS in its upstream region). Brd4 is involved in protein amino acid phosphorylation, kinase activity, transferase activity, binding of DNA and protein (The Gene Ontology Consortium, 2000), it help govern chromosomal dynamics during mitosis, and is part of BET. Brd4 is ubiquitously expressed, but most abundant in midgestation embryo, testis and brain, and found in the nucleus. (Bairoch et al., 2005) Brd4 is a hub in a network involved with cell morphology, hematological disease, and DNA replication, recombination, and repair with 8 other molecules. (Ingenuity Systems)

The second RIDGE, A06, consists of March2 (Hnrpm) and Rab11b. The RIDGE activity score 4.40 is not significant, with no significant changes in gene expression levels. The RIDGE is 114 kbp long with a density of 8%. Both March2 (41 kbp) and Rab11b (14 kbp) are general binding genes, for example of nucleotides and both are found in the membrane. (The Gene Ontology Consortium, 2000) There are no missing genes in this RIDGE, even though there is a large (60 kbp) inter-gene distance. Myo1f has one transcript and March2 has two with many

exons. Neither the gene score nor the UTR score are significant, although the 2 genes share 26 TFBS; AP-1, c-Fos, c-Jun, C/EBPα, C/EBPβ, COE1, CRE-BP2, f(α)-f(ε), GATA-1, GATA-2, GR, HES-1, HOXA5, JunD, MyoD, myogenin, NF-1, NF-AT1, NFκB, Pax-5, PU.1, RelA, Tal-1, TCF-1(P), USF-1, and YY1. These two genes are found in two separate networks and both acts as hubs for these. (Ingenuity Systems)

The third RIDGE, A09, is another single-gene RIDGE, H2-Q2. The RIDGE activity score for this RIDGE is 4.54 which is not significant , and the RIDGE is very short (83 kbp long). The RIDGE has a low exon density of 5% meaning that 95% of the gene consists of introns. Again it is not possible to determine gene score, UTR score, or whether or not the members have many shared TFBS (H2-Q2 has 36; which is probably related to the four transcripts this gene has). H2-Q2 is involved in antigen presentation and processing, defense response, and protein binding. The gene is found in the MHC class I protein complex, the membrane, the plasma membrane, and on the cell surface. (The Gene Ontology Consortium, 2000) H2-Q2 is a single-pass type I membrane protein and a H-2D cell surface glycoprotein that belongs to the MHC class I family and is integral to membrane. (Bairoch et al., 2005) RIDGE A09 results in "no genes eligible for analysis" in the Ingenuity Pathway Analysis software (Ingenuity Systems).

RIDGE, A20, has a RIDGE activity score of 4.15 which is not significant. The RIDGE is 115 kbp long with an exon density of 14%. Neither the gene score nor the UTR score are significant, but still 21 TFBS are found in all three genes; AP-1, C/EBPα, C/EBPβ, c-Fos, c-Jun, COE1, f(α)-f(ε), GATA-2, GR, HES-1, HOXA5, JunD, MyoD, myogenin, NF-1, NF-AT4, NF-κB, PU.1, RelA, USF-1, and YY1. RIDGEs, A20-A22, consist of the genes H2-T24, H2-T23 (HLA-E), and H2-L. The three genes share a number of functions, such as antigen presentation, integral to membrane, located in the membrane, and MHC class I protein complex, furthermore H2-T23 and H2-T24 share immune response. (The Gene Ontology Consortium, 2000) H2-T23 is the only one found in the Ingenuity Pathway Analysis software (Ingenuity Systems) where it has been associated with cancer, gastrointestinal disease, hepatic system disease, cell signaling, hematological system development and function, and finally with cellular function, maintenance, growth, proliferation, assembly, and organisation. (Ingenuity Systems) This implies that the members of this RIDGE are involved in both antigen presentation and cell cycle regulation. This RIDGE is a core immune RIDGE with correlated gene expression profile that is considered static on the RIDGE level, but with changes in gene expression levels. H2-L is a long gene (83 kbp) with 5 transcripts and many exons (10, 19, 29, 37 and 47), whereas H2-T24 (14 kbp) and H2-T23 (3 kbp) are normal genes with 1 and 2 transcripts respectively. There is a significant upregulation of all genes under all conditions, except for H2-T24 in the primed condition.

## 5.4 RIDGE characteristics for the observed RIDGEs

664 RIDGEs were found in rested macrophages, and 14 of these (2%) fall inside the MHC locus. When all RIDGES found in any of the four conditions are counted, 22 RIDGEs are found in the MHC locus. In total 34 RIDGEs in the MHC locus have been investigated; the 22 found when a RIDGE may not contain any silenced genes, 10 additional RIDGEs found when a RIDGE may contain one silenced gene, and 2 additional RIDGEs found when a RIDGE may contain up to two silenced genes. When RIDGEs are found to overlap on one, or more, RIDGE members one RIDGE was chosen to represent the region (see 5.2.1), which has narrowed the analysis to the 9 RIDGEs presented here.

### 5.4.1 RIDGE gain, RIDGE loss, and RIDGE members in a flux

An important example of RIDGE gain is seen in RIDGE A07; a classical MHC class II locus RIDGE containing the genes Psmb8, Tap2, H2-Ab1, H2-Aa, and H2-Eb1. This RIDGE was absent in rested macrophages, but present in activated macrophage, and the RIDGE members were all upregulated under all three conditions except for Tap2 in the primed macrophages. This fits well with the expectation that the MHC class II genes are activated when the macrophage is activated. An example of RIDGE loss is RIDGE A10 (Cdsn and Gtf2h4) whose members are associated with skin and hair; functions that are probably less prioritised once an immune response is required and this correlates with the RIDGE presence; present in rested macrophages, but absent in activated. Both of these RIDGEs (A07 and A10) could be considered split RIDGEs where there is a clear distinction in behavior between each half of the RIDGE.

Another group of RIDGEs are the overlapping RIDGES A11-A19 in the Ier3:H2-L region where the RIDGE members are in a state of flux, e.g. where the members are both up and down regulated within the same condition, i.e. fine regulation within a RIDGE (one reason that the RIDGE analysis were also performed for RIDGEs with one, or two, silenced genes; see tables C.1.2 and C.1.3 in appendix C). The closer to the 3' end of this region, e.g. the higher the RIDGE number, the more and more correlated the expression profile is. For example RIDGE A11; in the primed macrophage three genes are downregulated (TUBB, Dhx16, and Mrps18b) and one upregulated (KIAA1949), after viral activation three genes (Ier3, TUBB, and Dhx16) are downregulated, and after both treatments two genes (Ier3 and TUBB) are downregulated and one (Mrps18b) upregulated. As compared to RIDGE A19; in the primed macrophages the first two genes (Mrps18b and Abcf1) are downregulated and the last two genes (H2-T24 and H2-T23) are upregulated, after viral activation three genes (the second gene and the last two genes) are upregulated, and after both treatments the first gene and the two last genes are upregulated. This latter example could be considered a split RIDGE and is therefore used to

support the division into two different RIDGEs. For A15; KIAA1949 is upregulated in primed macrophages, whereas three genes (Dhx16, Mrps18b, Abcf1) are all downregulated; in viral activated macrophages then Dhx16 is downregulated and Abcf1 upregulated; and finally in both primed and viral activated macrophages Mrps18b is upregulated.

### 5.4.2 Static RIDGES

There are in total 10 static RIDGEs, three single-gene RIDGES (A03, A09, and A22), and seven others (A05, A06, A13, A14, A16, A20, and A21). The first seven (in genomic order) have general functions, whereas the last three RIDGEs have very specialised gene functions involved in both antigen presentation and cell signaling. One of the RIDGEs significantly high *upstream sequence similarity score* and none have a significant *coding sequence similarly score* implying that these static RIDGEs are not operon like models of co-regulated genes nor false RIDGEs due to sequence duplications.

### 5.4.3 Quantitative data for RIDGEs and RIDGE members

The RIDGE profile, i.e. RIDGE presence/absence for the four conditions, was compared to the gene expression profile, i.e. up or down regulation of the RIDGE members, and three groups of behaviour was seen;

- three static RIDGEs with static gene expression (A03, A06, and A09) of which the first and last are single-gene RIDGEs,
- two RIDGEs with only positive regulation of members (A07 and A20) where the largest foldchanges are seen, and
- two RIDGEs in a state of flux (A04 and A15) where the largest foldchanges are seen in the primed macrophages.

In addition RIDGE A02 has the same, if any, regulation of gene expression within a condition, and RIDGE A10 is static on the gene level, although not on the RIDGE level.

Since IFN-$\gamma$ is known to activate the immune system we would expect to see a group of immune related RIDGEs react to immune activation, here RIDGE A07 is activated (and significantly upregulated), whereas A20 was already present but is upregulated, and A09 (H2-Q2 is a MHC class I gene) remains static. The RIDGE members of the RIDGEs, A07 and A20, are upregulated for all conditions expect for Tap2 and H2-T24 in primed macrophages. RIDGE A07 members are the most upregulated in primed and viral activated macrophages, and the least upregulated in viral activated macrophages (except for Tap2). This profile is repeated for H2-T23 in A20, whereas H2-L is the most upregulated in both, and the least upregulated in primed

macrophages, and finally H2-T24 is the most upregulated in viral activated macrophages, the least in both, and not at all in primed.

IFN is also known to activate the cell cycle and this is seen in RIDGE A15, although it required the combination of viral activation and priming. Viral activation with mCMV would also be expected to result in significant changes, and this is found for two of the overlapping RIDGEs, A11 and A12 (section C.1.3.1) involved in the cell cycle. The gene Flot1 is marginal for viral activated macrophages, but present in all other conditions. The lipid raft protein Flotillin-1 is associated with the cytoskeleton and integrin signaling, (Jonathan Kerr, 2006) and is part of the caveolae (that mediate the transcytosis of macromolecules, are the site of potocytosis, i.e. help transverse the plasma membrane, and may participate in the relay of extracellular signals to the cells interior by organising signal transduction molecules). (Bickel et al., 1997) The virus might disable either transport or signaling function via Flot1, although no significant change in gene expression level is seen.

Because the conditions investigated are concerned with the immune response in general, specifically macrophage activation, the RIDGEs that are not immune related are more likely to be static, and the largest foldchange is seen for RIDGE A02 in primed macrophages, for both U2af1 and Pknox1. Pknox1 is involved in transcription and is a Hox cofactor (Moens and Selleri, 2006), marking a wide variety of genes for activation by various transcription factors. It has been theorised that marking can be reduced, or even prevented, by restricting access to the nucleus for Pbx proteins. (Sagerström, 2004) Another RIDGE, A04, becomes absent in primed macrophages because A430107D22 is downregulated in primed macrophages, whereas upregulated in viral activated macrophages, as is Akap8l which is also upregulated in both. This leads to the question of whether or not this RIDGE encode genes that are more or less important for the priming of the immune system, although still important for an actual immune response. Similar, the activation of the genes Pknox1 and GNL1 (in RIDGEs A01 and A15) in macrophages that were both primed and viral activated might imply that the immune system has activated yet another set of genes to deal with the incursion, although neither gene have known immunological functions.

### 5.4.4 Functional associations between RIDGE members

Of the nine RIDGEs presented in detail here, two are single-gene RIDGEs and so functional associations between RIDGE members can no be qualified. Both genes in RIDGE A02 are involved in transcription and localise to the nucleus, all three genes in A04 are binding genes and localise intracellular, RIDGE A06 consists of general binding genes located in the membrane, RIDGE A07 consists of MHC class II genes, RIDGE A10 consists of genes expressed in skin, the genes in RIDGE A15 are involved in protein biosynthesis, and the genes in RIDGE A20 encode MHC class I genes. This implies that there is indeed functional associations between

RIDGE members, as all seven have functional associations to some degree.

When the analysis is expanded to include all 34 RIDGEs in the MHC locus (i.e. when overlapping RIDGEs (A01-22), RIDGES with one (B01-B10), or two (C01-C02), silenced genes are also considered) 21 RIDGEs (62%) are considered to have functional associations between their members, the above seven in addition to: A01 (longer A02), A04, A05 (similar to A06), A08 (shorter A07), A11-A21 (Ier3:H2-L region), B03-9 (MHC class II RIDGEs), and C01 (a MHC class II RIDGE).

### 5.4.5 Protein Interaction Network (PIN) analysis of RIDGE members

The RIDGE members in A07 (HLA-DQA1, HLA-DQB2, MHC2α, MHC II β, HLA-DRB1, MHC class II, Tap2, and PSMB8) are found in a network involved with cell signaling, immune response, and protein degradation with 24 other molecules. IFN-γ acts as both a hub and a scaffold (e.g. if this molecule is silenced the network is decomposed) and could, together with HLA-F, be used to merge this network with the one for A20, for which only H2-T23 (HLA-E) were found in the Ingenuity Pathway Analysis software (Ingenuity Systems). H2-T23 is a hub for a network involved in cell signaling, immune response, and cell death with 20 other molecules. In the merged network there are a total of 33 molecules, and 9 of them are found in these two RIDGEs (shown in red).

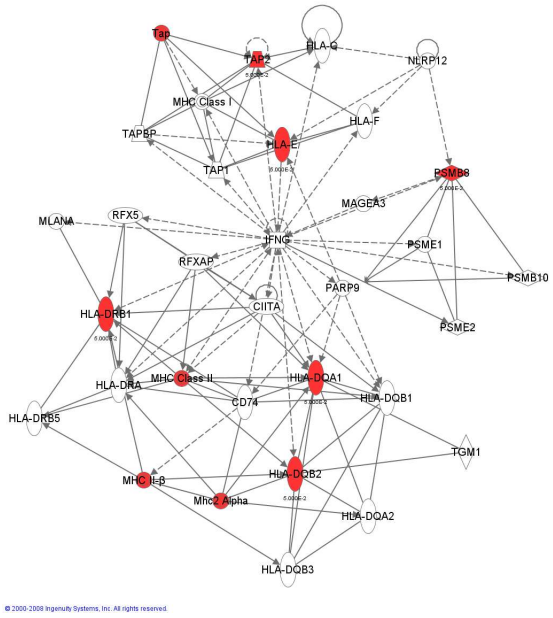
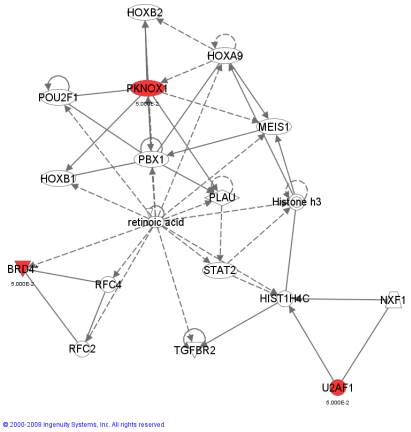



Figure 5.4: After merging of the two networks for A07 and A20.

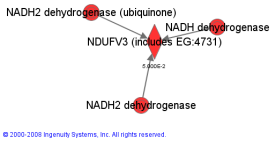
The RIDGE members for RIDGE A15 decompose into one network per member with one interaction each, except for Dhx16 and GNL1; both of these networks contain two additional

molecules. The hub for the merged network is TNF which is also located on chromosome 17 but 700 kbp away from Dhx16, so this could be an example of long-range influence of physical location.

The members of A02 and A03 all acts as hubs in a separate network, furthermore the A01 gene 1500032D16 constitutes all three molecules in that network. Once these two (A01 and A03) where merged, three scaffolds appear: retinoic acid, STAT2, and HIST1H4c. In addition fewer interactions for each RIDGE member is included, for example Pknox1 interacted with 18 molecules in its own network. This could be an example of an even higher order regulation between RIDGEs.




(a) Network 1




(b) Network 2

Figure 5.5: After merging of the two networks for A01 and A03.

The RIDGE members of A06, March2 and Rab11, both acts as the hub for their respective network. These two networks can be merged if an interaction between two of the edges (GTP in the Rab11b network and TNPO1 in the March2 network) is added.

The RIDGE members in RIDGE A04 form two separate networks which can not be merged, i.e. they do not share any molecules or known interactions between edges, furthermore none of the molecules in these two networks are found in any of the other RIDGE member networks.

### 5.4.6 Regulatory control of RIDGEs

RIDGE members might form a RIDGE because they share the same regulatory mechanism(s), and two methods were used to determine this (see 2.2.3 for a discussion of these). The first, PROMO, was used to retrieve the transcription factor binding sites (TFBS) that were located in the 5 kbp upstream region of all RIDGE members, and the second, ClustalW, was used to determine if members share control regions, e.g. was the *upstream sequence similarity score* significant. Ideally the results from these two methods should return the same RIDGEs,

although the first is subjective and the second an objective score.

Of the 34 RIDGEs 9 are considered to have significantly high *upstream sequence similarity scores* ($p < 0.05$) (A14, A18, A19, B01, B02, B04, B06, B08, and C02), and 8 of these were also considered to have many TFBS (with the exception of RIDGE B02 (Tapbp, Wdr46, and H2-K1) which have a significant UTR score but only 20 TFBS). In total 21 RIDGEs (A05-7, A09, A12-13, A15, A17-19, B03-B6, and C01-2) were considered to have many TFBS.

Note that because of re-sampling issues it could be argued that the p-value for ClustalW should be significantly lowered (for example 0.01), which would affect all of the results but for A13 (TUBB:Abcf1).

Both these methods indicate that there is indeed upstream regulatory mechanism(s) behind RIDGE formation (co-ordinated regulation of functionally linked genes).

### 5.4.7 Number of silenced genes in a RIDGE

The RIDGE analysis presented here is restricted to RIDGEs that may not contain any silenced genes, as this is the most stringent definition. This because the 10 additional RIDGEs found when one silenced gene is allowed are seven versions of the MHC class II RIDGE (A07), one RIDGE (B10) that consists of two inter-gene distances (in total the distances cover 114 kbp) and four short (<6 kbp) genes. When up to 2 silenced genes are allowed in a RIDGE, there are 2 additional RIDGEs, these contain too many missing genes in combination with the 2 silenced genes to, thereby lowering the likelihood that these are real RIDGES. In short, when the number of silenced genes (gaps) were increased, the number of RIDGEs were increased (from 664 to 875 and 951 respectively), but so was the noise (tables C.1.2 and C.1.3). All RIDGEs with significant RIDGE activity scores are found when no silenced genes are allowed. Many shared TFBS are found for all (expect for B01, B02, and B10) RIDGEs when one silenced gene is allowed, and most of these are considered to have significant UTR scores as well, and most of them have functional associations between the RIDGE members. This would imply that the correct definition is to allow up to one silenced gene, but most of the additions are in fact overlapping RIDGEs and as such would be removed from the analysis in favor of RIDGE A07, although RIDGES B01 (Zfp297, Tapbp, Wdr46, H2-K1) and B10 (?, Ddah2, Csnk2nb, Bat4) should potentially have been added.

## 5.5 Concluding remarks

The presence of RIDGEs as seen in the MHC locus, supports the hypothesis that there is indeed a connection between genomic structure and function.

This narrow focus on a well-defined region in the mouse genome has made it possible to map the biology and determine if the results are biologically meaningful; setting the stage for

the whole-genome analysis presented in the next chapter.

# Chapter 6

# Genome-wide RIDGE analysis

The first section of this chapter presents the genome-wide RIDGE analysis for the macrophage activation dataset with the four conditions; uninfected, primed, viral activated, and both primed and viral activated macrophages. In the following two sections two additional datasets, three genomic loci, and 21 random regions are also investigated.

No two genomic regions are exactly the same and for this study regions are categoried according to region length and gene content. Gene content is defined as the number of genes, the gene scarcity (calculated as $\frac{sizeOfLocus}{numberOfGenes}$), and the gene coverage (calculated as $\frac{\sum geneLengths}{sizeOfLocus}$). In this chapter it will be investigated if RIDGE formation is related to the gene content of a region, rather than higher-order chromosomal organisation, if RIDGEs are immune response specific units, housekeeping, or tissue-specfic.

## 6.1 Genome-wide RIDGE analysis of the macrophage activation dataset

### 6.1.1 Non-random chromosome organisation of RIDGEs

RIDGE formation could be related to the gene content of a region, rather than higher-order chromosomal organisation.

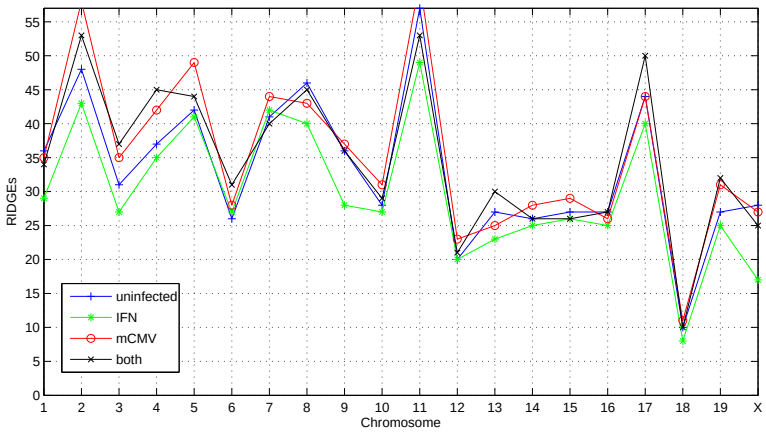



Figure 6.1: The 664 observed RIDGEs for uninfected macrophages exhibit non-random chromosome organisation in uninfected macrophages. The macrophage response is the highest to mCMV viral activated macrophages (the red line).

RIDGEs are observed on all mouse chromosomes except for chromosome Y. This chromosome has the fewest genes and, related, the highest gene scarcity and gene coverage.

| Chrom | Length | Genes | Projected | scarcity | coverage | RIDGEs | #Members | #Unique |
|---|---|---|---|---|---|---|---|---|
| 11 | 121798632 | 1990 | **719** (**36**%) | **61205** | **41** | **57** | **117** | **101** (86%) |
| 2 | 181976762 | 2290 | 665 (29%) | 79466 | 35 | 48 | 95 | 81 (85%) |
| 8 | 132085098 | 1399 | 423 (30%) | 94414 | 31 | 46 | 89 | 70 (79%) |
| 17 | 95177420 | 1264 | 409 (32%) | 75298 | 35 | 44 | 97 | 63 (65%) |
| 5 | 152003063 | 1583 | 519 (33%) | 96022 | 35 | 42 | 86 | 72 (84%) |
| 7 | 145134094 | **2495** | 665 (27%) | 58169 | 33 | 41 | 92 | 82 (89%) |
| 4 | 155029701 | 1639 | 506 (31%) | 94588 | 31 | 37 | 64 | 59 (92%) |
| 1 | **197069962** | 1592 | 500 (31%) | 123788 | 30 | 36 | 65 | 60 (92%) |
| 9 | 124000669 | 1509 | 447 (30%) | 82174 | 36 | 36 | 68 | 52 (76%) |
| 3 | 159872112 | 1335 | 444 (31%) | 119754 | 27 | 31 | 66 | 52 (79%) |
| 10 | 129959148 | 1263 | 380 (30%) | 102897 | 32 | 28 | 57 | 45 (79%) |
| X | 165556469 | 1273 | 271 (**21**%) | 130052 | 21 | 28 | 73 | 45 (62%) |
| 13 | 120614378 | 1096 | 295 (27) | 110050 | 29 | 27 | 58 | 47 (81%) |
| 15 | 103492577 | 1024 | 357 (35%) | 101067 | 31 | 27 | 56 | 41 (73%) |
| 16 | 98252459 | 882 | 257 (29%) | 111397 | 31 | 27 | 48 | 41 (85%) |
| 19 | 61321190 | 877 | 300 (34%) | 69921 | 40 | 27 | 59 | 51 (86%) |
| 14 | 123978870 | 1225 | 305 (25%) | 101207 | 33 | 26 | 45 | 38 (84%) |
| 6 | 149525685 | 1574 | 494 (31%) | 94997 | 34 | 26 | 51 | 45 (88%) |
| 12 | 120463159 | 1118 | 279 (25%) | 107748 | 19 | 20 | 29 | 27 (93%) |
| 18 | 90736837 | 700 | 191 (27%) | 129624 | 30 | 10 | 17 | 17 (**100%**) |
| Y | **16029404** | **29** | **8** (28%) | **552738** | **6** | **0** | **0** | **0** |

Table 6.1: Chromosome characteristics for uninfected macrophages [chromosome length, number of genes (and reliable projected genes), gene scarcity, and gene coverage] are not in linear relation to the number of RIDGEs (or RIDGE members be they overlapping or non-overlapping (#Unique)). The table is sorted according to the number of RIDGEs per chromosome and the extreme values are shown in bold.

The number of RIDGEs observed for a chromosome is not linearly related to either chromosome size or gene content. The chromosome with the most RIDGEs (chromosome 11) has the highest gene coverage, and the most reliable projected genes. Furthermore the chromosome with no RIDGEs (chromosome Y) has the highest gene scarcity, the lowest gene coverage, and the fewest reliable projected genes. But, the remaining 19 chromosome are not organised according to these criteria. This non-linear relationship provides evidence for a non-random higher-order gene organisation into RIDGEs.

Not only does chromosome 18 contain the fewest RIDGEs, it also has no overlapping RIDGEs (the number of RIDGE members equals the number of unique RIDGE members). The 10 RIDGEs for this chromosome are spread out on the chromosome, with the exception of two consecutive RIDGEs, where one of these RIDGEs map to the protocadherin locus. On the other hand more than half the RIDGE members on chromosome X are overlapping, followed by one third for chromosome 17. The latter is clearly demonstrated in the MHC locus (as for example seen in section 5.2.1).

### 6.1.2 Immune system genes

Immune system genes are defined as genes present in any of the manually curated articles focused on the MHC class II antigen presentation pathway (see 2.2.2.2 for a discussion). Jak-Stat genes are defined as genes present in either Meraz et al or Aaronson et al. (Meraz et al., 1996; Aaronson and Horvath, 2002) IFN-$\gamma$ is known to primarily signal via the Jak-Stat pathway (Schroder et al., 2004), and so a large overlap between immune system genes and Jak-Stat pathway genes were expected, but this was not found. Of the 6 Jak-Stat genes on chromosome 11 (Irf1, Tnip1, Tnfrsf13b, Tnfsf13, Traf4, Tnfaip1) only one gene, Irf1, were also considered an interferon response gene, and of the 8 Jak-Stat genes on chromosome 17 (Tnfrsf12a, Traf7, Tnf, Tnfrsf21, Tnfaip8l1, Tnfsf9, Cd70, and Tnfsf14) again only one gene, Tnf, overlap. One third of the genes in this pathway (31 of 90) are found on chromosome 4 and 20 are found on chromosome 5, thereby exhibiting preferential chromosomal organisation. 21 of the 31 Jak-Stat pathway genes on chromosome 4 are found in a 357953 long region around 88 Mbp with 21 interferon associated genes.

The mapping of immune system genes onto their chromosomal location would be expected to reveal a number of known immune loci, such as the immunoglobulin loci. Furthermore this mapping was used to conclude that immune system RIDGEs are observed but that not all observed RIDGEs are associated with the immune response.

Chromosome 17 does not have the most immune system genes, as might have been tempting to assume since the MHC locus contains 250 genes. Neither has chromosome 6 with an immunoglobulin locus containing 350 genes. Instead chromosome 11 is found to contain the most immune system genes, although chromosome 17 does have the highest percentage of immune system genes in relation to total gene number.

It was found that immune system genes do cluster onto chromosome. For instance chromosome 11 contains twice as many as expected (if a simple $\frac{numberOfImmuneSystemGenes}{numberOfChromosomes}$ is employed), and chromosome 16 not even half as many as expected.

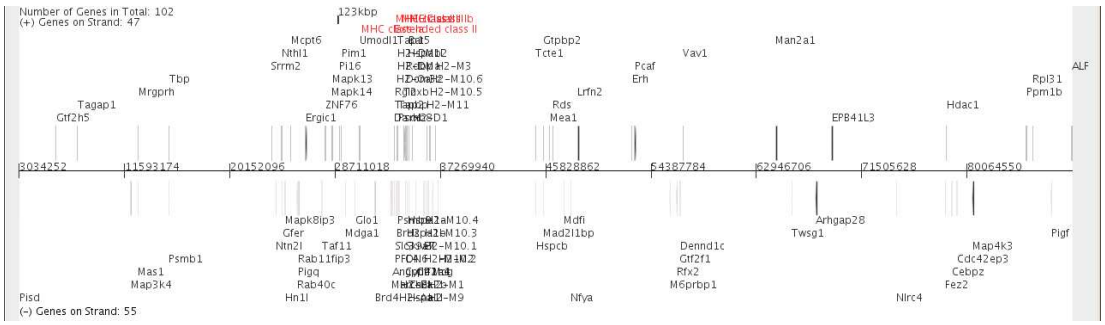



Figure 6.2: The 102 immune system genes that map on chromosome 17. The MHC locus clearly stands out as a very gene dense region.

102 immune system genes map to chromosome 17, although not the same 102 genes that can be projected for the locus. Given the above definition of immune system genes as either having one, or more, GO-terms associated with the immune response, or based on the molecular interaction data from the database genes only 42 genes in the MHC locus are considered immune system genes. Of the 60 immune system genes that do not map to the MHC locus some cluster into small groups, for instance 11 genes around 24 Mbp.

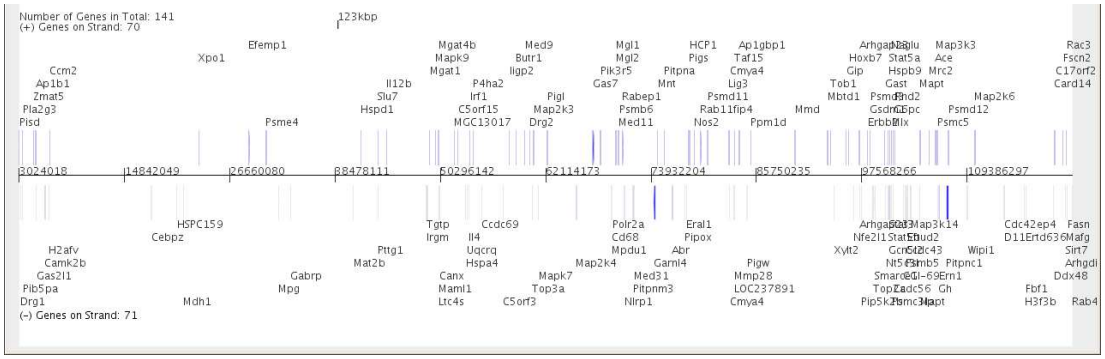


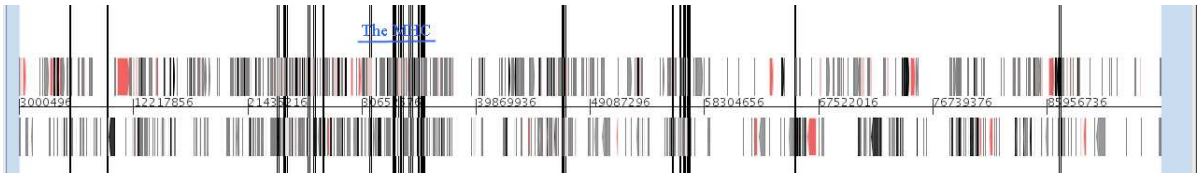

Figure 6.3: The 141 immune systems genes on chromosome 11, where one gene dense region around 100 Mbp are observed.

For the cluster of genes around 100 Mbp, there is no known annotation about this being a genetic locus containing Stat5a or Hoxb7. Similarly there is no known locus containing the genes Sirt2, Gmfg, or IL4R for chromosome 7, whereas the immune system genes on chromosome 2 do form clusters.

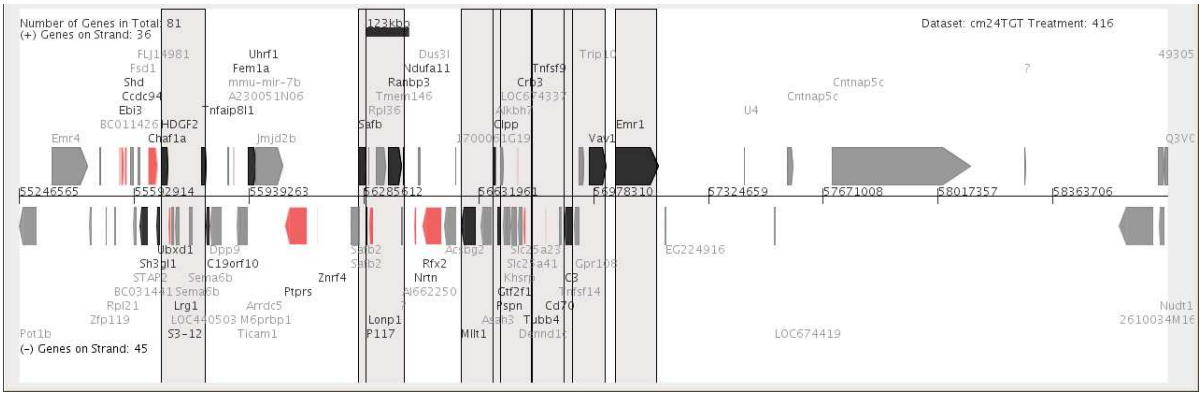
#### 6.1.2.1 RIDGEs do not exclusively consist of immune system genes

71 different RIDGEs are observed on chromosome 11 for the four macrophage activation state dataset, and of these only 7 consists of immune system genes; 23 RIDGEs are observed on chromosome 12 and of these only 4 consists of immune system genes.

### 6.1.3 RIDGEs on chromosome 17




(a) The entire chromosome 17



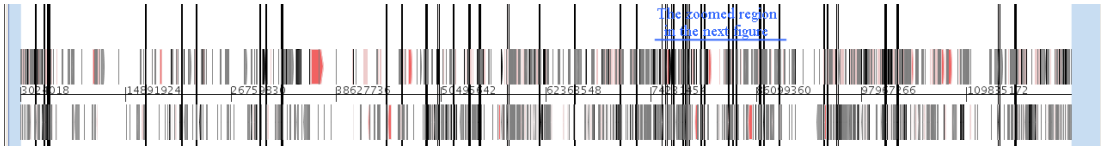

(b) RIDGEs on chromosome 17 - zoomed in around 56 Mbp

Figure 6.4: RIDGEs on chromosome 17 for uninfected macrophages. RIDGEs are shown as grey boxes (which at the zoomed out level looks like think black lines) and active genes are highlighted in black, genes without a reliable probe-to-gene projection in grey, and silenced genes in red.

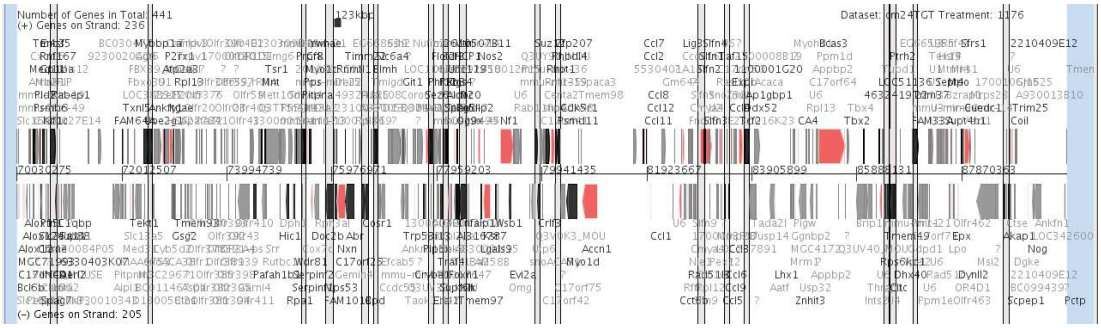
RIDGEs on chromosome 17 are not evenly spread along the chromosome, but clustered into three loci; the MHC locus, a region around 56 Mbp (shown in subfigure 6.4.b) and a region around 25 Mbp.

The RIDGE dense region around 56 Mbp was not a known genomic locus, and neither are the genes known to be of the same class.

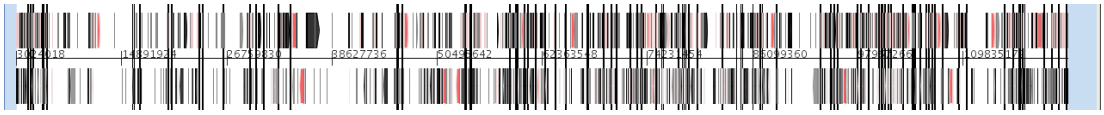
### 6.1.4 RIDGEs on chromosome 11




(a) RIDGEs on chromosome 11 for the macrophage activation dataset



(b) RIDGEs on chromosome 11 around 80 Mbp



(c) RIDGEs on chromosome 11 for the time series dataset

Figure 6.5: RIDGEs on chromosome 11 for uninfected macrophages. Subfigure a) shows the RIDGEs observed for the macrophage activation dataset. Subfigure b) shows the zoomed in region around 80 Mbp with an observed high RIDGE density for the same dataset. Subfigure c) shows the observed RIDGEs for the time series dataset. Note that the Agilent chip has a higher degree of probe-to-gene projections.

The observed RIDGEs on chromosome 11 are evenly spread across the chromosome, with a somewhat higher RIDGE density around 80 Mbp (as shown above in figure 6.5.b), and this spread is reinforced by the time series dataset.

Many genes without a reliable probe-to-gene projection are observed for this chromosome, for example the Ccl genes around 81.3 Mbp. These are interferon inducible chemokines of the intercrine beta family (Bairoch et al., 2005); intercrines are a general term for a family of small, activation-inducible, mainly chemotactic and pro-inflammatory proteins, now renamed as chemokines. (COPE, 2008) The CC chemokines are known to cluster on human chromosome 17 (q11-q32).

The even pattern of RIDGE dense and RIDGE sparse regions provide evidence supporting RIDGE formation. It could be argued that RIDGEs are too dependent on genome probe coverage, but by allowing RIDGEs to contain genes without reliable projections, this is somewhat

negated. Also RIDGE dense and RIDGE sparse regions are seen for the Agilent chip (the time series dataset) with around 70% coverage.

## 6.2 Additional datasets

Here RIDGEs observed in both the MHC locus (specific case) and in the whole genome (general case) are presented for:

1. an extensive time series spanning the first 12 hours after macrophage activation using the Agilent chip, and

2. an investigation of gene expression across 61 tissues using the GNF chip.

The time series measures the cascading effects of IFN-γ priming, or mCMV viral activation on macrophages during the first 12 hours after activation. By using time series data an assumption is made that RIDGEs consists of functionally related members, preferably associated with the cascade response to activation of the immune response. On the other hand the tissue specificity study focuses on gene expression levels across multiple tissues in order to identify potential housekeeping RIDGEs. In this study, examples of both immune specific RIDGEs and housekeeping RIDGEs are observed.

The time series use the Agilent chip and the tissue dataset the GNF chip (see 3.2.1.4 for coverage of these - around 70 and 50% respectively) making direct comparisons across the three datasets difficult.

Genes without a reliable probe-to-gene projection for the macrophage activation dataset are shown in parenthesis (), genes without a reliable probe-to-gene projection for the time series dataset are shown in brackets [], and genes without a reliable probe-to-gene projection for the tissue dataset are shown in curly brackets {}. For example [{H2-Eb2}] should be interpreted as the gene H2-Eb2 has no reliable probe-to-gene projection for the time series and tissue datasets, but is projected for the macrophage activation dataset, i.e. on the Affymetrix MG-U74Av2 chip.

### 6.2.1 The time series dataset

Macrophages were either 1) primed with IFN-γ, 2) viral activated with mCMV, or 3) both primed and viral activated and investigated for the first 12 hours after activation. Samples were prepared every 30 minutes which resulted in 25 distinct time points for each of the three biological conditions, although most RIDGEs observed in one of the three conditions are observed for all three. This section will therefore focus on the priming of macrophages.

A gene is, for this dataset, considered active if the mean intensity signal is above 0. (see 2.1.5.2)

49 RIDGEs are observed for this dataset that map to the MHC locus. Most RIDGEs observed in one of the three biological conditions, are observed in all, and mostly at similar time points.

#### 6.2.1.1 RIDGEs in the MHC locus for primed macrophages

During the first 3 hours alternatively 5 or 1 new RIDGE appear for every other time point. Even after 450 minutes 15 new RIDGEs appear, although these are all elongated versios of RIDGEs observed for previous timepoints.

The most RIDGEs (22) are observed at time point 630; and the least (4) at time point 60. There is a trend toward few (less than 8) RIDGEs during the first 3 hours and then the immune response is even more activated (by activating for example the MHC class II RIDGEs) and atleast 11 RIDGEs are observed per time point, although in most cases still fewer than the 22 RIDGEs that are observed in the macrophage activation dataset.

#### 6.2.1.2 Genome-wide RIDGE analysis

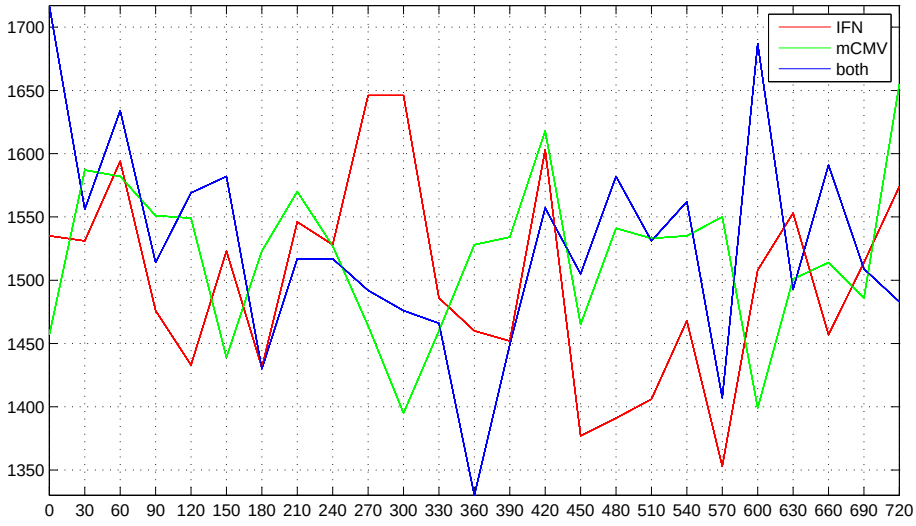



Figure 6.6: Number of RIDGEs per time point.

Although it is tempting to assume that the same RIDGEs are present at time point 270 and 300, this is not the case, for example there are 10 RIDGEs more on chromosome X in the latter. Most RIDGEs for primed macrophages were found at time points 270 and 300, and although the same number of RIDGEs was found, the same RIDGEs were not. Most RIDGEs for both primed and viral activated macrophages was seen at time 0, which would imply that the immune system responded directly when a macrophage was both primed and viral activated,

although a delay was seen when a macrophage was either. Viral activated macrophages had the most RIDGEs at time 720, in comparion to very few at time 0.

Only one RIDGE (Jarid1d and Eif2s3y) is ever observed for chromosome Y. It is observed at time 0, 30, 90-390, 480, and 540-600.

#### 6.2.1.3 Resting periods

For primed macrophages a resting period was initiated after 420 minutes that continued up to 600 minutes after priming.

The curves for primed macrophages and both primed and viral activated macrophages behave similarly, although the latter had more RIDGEs, except at 240-420. This is a highly active time in the interferon response pathway, although a resting period for the immune response to both priming and viral activation. For the MHC locus, the most RIDGEs were found after 630 minutes, and a lot more was found between 9 and 12 hours, than within the first 3 hours.

A resting period is observed for primed macrophages between 480 and 600 minutes after treatment where no new RIDGEs appear (as also seen at time point 300). This rested state hypothesis is supported by investigating the total number of RIDGEs at a specific time point (see figure 6.6).

#### 6.2.1.4 Cyclic response

For viral activated macrophages, the curve behaves varying (and no resting period was found) but corresponds to the other two conditions at time point 210 and 420 (a cyclic phenomenon) (figure 6.6).

Viral activated macrophages on the other hand exhibit two decreased time points: 300 and again at 600 minutes, potentially indicating a cyclic respone. (figure 6.6)

#### 6.2.1.5 Static RIDGEs

When investigating static RIDGEs, only the 5 RIDGEs present for time 0 can truly be considered static and RIDGE (Abcg1, Tff3, and Tff2) and RIDGE (Tff2 and Ubash3a) are both observed for all time points.

The second most static RIDGEs are, suprisinlgy enough, to immune system RIDGEs. RIDGE (H2-T24, H2-T23, and H2-L) becomes activated after 3 hours and remain active spanning the entire time series, and is therefore present at the last 18 time points. RIDGE (H2-Q8, H2-Q9, H2-Q10, and Pou5f1) has the same expression profile, except it is silenced after 12 hours (720 minutes).

#### 6.2.1.6 Elongated the further the cascade progessed

The same base RIDGEs are observed for the time series, with elongations and subtractions to the RIDGE members, as expected considering the underlying biology.

One intriguing example of overlapping RIDGEs was RIDGE 240.3 (Ddx39 and H2-Q2) that is also observed as 240.4 (H2-Q2), 270.1 (Nfkbil1, Atp6v1g2, Ddx39, H2-Q2), 270.2 (Atp6v1g2:Ddx39), 300.1 (Aif1, Lst1, Ltb, Tnf, Lta, Nfkbil1, Atp6v1g2, and Ddx39), 330.2 (extended 300.1 with Bat2), and four even longer RIDGEs found for time point 690 (690.1-4).

Here more and more genes are grouped into the RIDGE the longer the cascade was running. This could be an example of fine tuning within a RIDGE; where more and more genes were needed, to regulate, either up or down, the protein production.

#### 6.2.1.7 Summary

Both IFN-γ primed macrophages and macrophages that were both primed and viral activated exhibit a decrease in the number of observed RIDGEs at time point 570. Followed by a sharp increase in observed RIDGE numbers at the following time points. Viral activated macrophages on the other hand exhibit two decreased time points: 300 and again at 600 minutes, a cyclic response.

Both primed and primed and viral activated macrophages had a dip of number of RIDGEs at time point 570, followed by a sharp increase in RIDGE number at the next time point, whereas viral activated macrophages had two dips, one at 300 and one at 600 minutes after activation, and .

### 6.2.2 The tissue dataset

The Genomics Institute of the Novartis Research Foundation (GNF) has performed an extensive gene expression analysis of different tissues in mouse. (GNFb) (see 2.1.5.3) For these 61 tissues, only 23 RIDGEs are observed in the MHC locus, and 2988 RIDGEs across the whole genome, with around 644 RIDGEs ($644 \pm 189$) per tissue, corresponding with the 664 RIDGEs observed for uninfected macrophages.

The 61 tissues can broadly be divided into 7 categories (where the 26 tissues with observed RIDGEs that map to the MHC locus are shown in bold)

- 6 immune tissues: **adipose tissue**, **bcells**, brown fat (brown adipose tissue), **cd4$^+$ tcells**, **cd8$^+$ tcells**, and **thymus**.
- 10 brain tissues: amygdala, **cerebellum**, cerebral cortex, dorsal root ganglia, **dorsal striatum**, frontal cortex, hippocampus, hypothalamus, prepotic area, and substantia nigra.

- 10 developmental tissues: blastocysts, fertilised egg, **embryo day 6.5**, embryo day 7.5, **embryo day 8.5**, embryo day 9.5, embryo day 10.5, oocyte, placenta, and the umbilical cord.

- 9 endocrine and reproductive system tissues: **adrenal gland**, mammary gland, **ovary**, pancreas, pituitary gland, **prostate**, **testis**, thyroid, and uterus.

- 6 sensory organs: **digits**, **medial olfactory epithelium** (MOE), olfactory bulb, **retina**, trigeminal nerve, and the **vomeral nasal organ** (VMO).

- 9 bone, skeleton, and skin tissues: **bone**, bonemarrow, epidermis, skeletal muscle, snout epidermis, lower spinal cord, upper spinal cord, tongue epidermis, and **trachea**.

- 11 remaining organs: **bladder**, heart, **small intestine**, **large intestine**, **kidney**, **liver**, **lung**, **lymph node**, salivary gland, **spleen**, and stomach.

These should be enough to determine housekeeping properties of RIDGEs. Lercher et al (Lercher et al., 2002) defined housekeeping genes as present in more than 9 tissues, and this is the stratagem used here; a housekeeping RIDGE is present in at least 10 tissues and a tissue-specific RIDGE in at most 5. (Williams and Hurst, 2002)

#### 6.2.2.1 Housekeeping RIDGEs

1096 housekeeping RIDGEs are observed (which constitues 37% of the 2988 RIDGEs). Thus showing that RIDGEs are not restricted to housekeeping genes although both Lercher et al and Singer et al (Lercher et al., 2002; Singer et al., 2005) suggested they were.

For the MHC locus only one housekeeping RIDGE was found (Notch3 and Rrp1B). This RIDGE is observed in 14 tissues (adipose, adrenal gland, bcells, bladder, bone, digits, dorsal striatum, embryo at day 8.5, kidney, lung, ovary, retina, thymus and trachea). The RIDGE is also observed in the time series for 14 time points; Rrp1B is only sparsely expressed (in 15 time points) whereas Notch3 is expressed in 69 of the 75 time points.

The RIDGE is 130 kbp long and has a RIDGE activity score of 4.55, which is not significant (although a low score might have been expected for a housekeeping RIDGE). The RIDGE has one long inter-gene distance, 60 kbp, between the genes in which one missing, unknown, gene locates. The gene score is significant, but the UTF score is not, implying that the RIDGE members formed through sequence duplication events. Both genes are involved in superoxide dismutase activity, rRNA processing, and are small subunit precursor nucleolar preribosomes. (The Gene Ontology Consortium, 2000) Notch3 was further involved in protein binding, superoxide metabolism, and metal ion binding. (The Gene Ontology Consortium, 2000)

#### 6.2.2.2 Tissue specific RIDGES

A RIDGE is tissue specific if present in five, or less, tissues. (Williams and Hurst, 2002) 1716 tissue-specific RIDGEs are observed which constitues 57% of the RIDGEs. Examples of tissue-specific RIDGE include; (Adamts10 and NM_183282) and (H2-Q2).

#### 6.2.2.3 RIDGEs in the MHC locus for the tissue dataset

When the RIDGEs are grouped by tissue, the following is observed; the RIDGE (Notch3 and Rrp1B) is observed to be the only RIDGE in 8 (of the 61) tissues (adrenal gland, digits, dorsal striatum, embryo at day 8.5, kidney, lung, ovary, and retina), and the single-gene RIDGE (Pde9a) is observed as the only RIDGE in 2 of the tissues (large intestine and prostate). In addition;

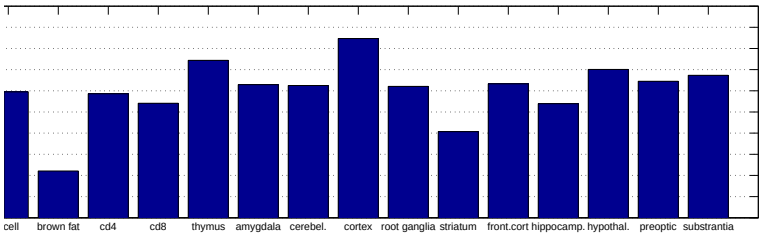
- The adipose tissue has 9 RIDGEs; 7 are consisting of MHC class II genes, one RIDGE with 3 H2-Q genes, and finally (Notch3 and Rrp1B).
- The bcells has 6 RIDGEs; 1) (Notch3 and Rrp1B), 2-3) (Akap8, Akap8l, Wiz, and A430107D22), and the shorter version excluding Akap8, 4-5) two MHC class II RIDGEs, and 6) (H2-L).
- The bladder has 2 RIDGEs; (Notch3 and Rrp1B) and (9030612M13 and Cyp4f15).
- There are 3 RIDGEs in bone; the Notch3 RIDGE, 1 MHC class II RIDGE and (Adamts10, Myo1f, and Pram1).
- The cd4 and cd8 tcells have 8 RIDGEs; 7 MHC class II RIDGEs and H2-L.
- Embryo at day 6.5 has 3 RIDGEs; (Ddx39 and H2-Q2), (H2-Q2), and (H2-L).
- Small intestine has the (Pde9a) and (Brd2, Psmb9, Psmb8, H2-DMB2, and H2-DMa) RIDGEs.
- The lymph node has 8 RIDGEs, all in the MHC class II region.
- The liver has one RIDGE; (9030612M13 and Cyp4f15).
- The MOE and VMO tissues have the same 4 RIDGES in the Tsga2 region (Tff2, Tff1, Tsga2, Tmprss3), (Tff1, Tsga2, Tmprss3), (Tsga2, Tmprss3), (Tsga2 and Slc37a1), and (9030612M13 and Cyp4f15) but VMO also has the Notch3 RIDGE.
- The spleen has 2 MHC class II RIDGEs.
- The testis has 3 RIDGES; (Tff1, Tsga2, and Tmprss3), (Tsga2 and Tmprss3), and (Adamts10 and NM_183282).

- The thymus has 5 RIDGEs; 3 MHC class II RIDGEs, (H2-L), and the Notch3 RIDGE.

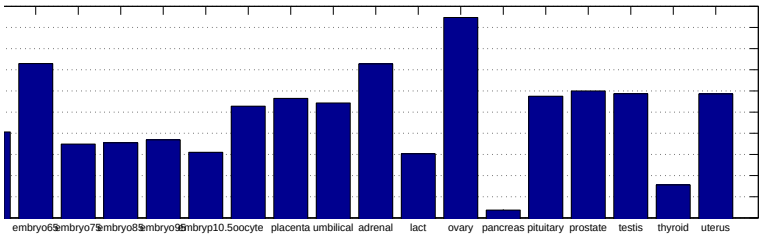
- The trachea has 4 RIDGEs; 2 MHC class II RIDGEs, the Notch3 RIDGE, and one in the Tsga2 region.

Only one housekeeping RIDGE is observed in the MHC locus (the Notch3 RIDGE), and at most 9 overlapping RIDGEs are seen in a single tissue (for adipose) and at most 4 non-overlapping RIDGEs (for bcells).

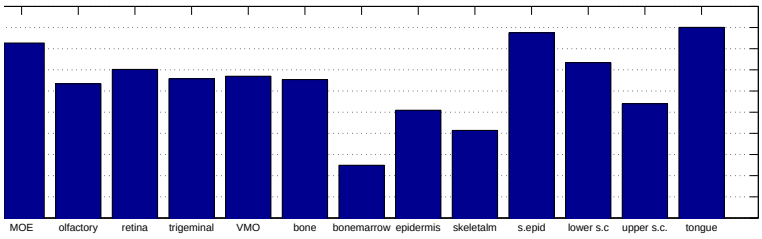
#### 6.2.2.4 Genome-wide RIDGE analysis



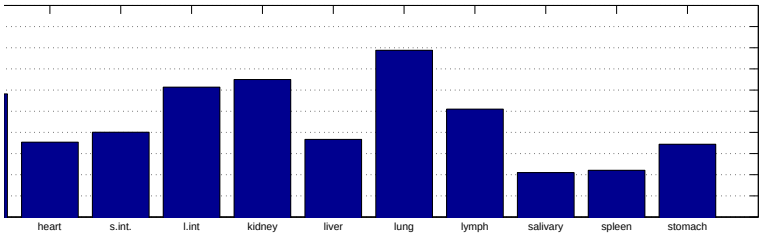

(a) The 16 immune and brain tissues




(b) The 19 developmental, endocrine, and reproductive system tissues




(c) The 15 sensory organs, bone, skeleton, and skin tissues




(d) The 11 remaining organs

Figure 6.7: Number of RIDGEs per tissue

By grouping the tissues into the 7 categories previously described, it was assumed that the RIDGE content would be more similar within a category than between categories, but this was not observed. The average number of RIDGEs per tissue varied for the four figures; the first has $710 \pm 144$, as compared to $581 \pm 213$ for the second, $750 \pm 178$ for the third, and $558 \pm 185$ for the last. The 6 immune related tissues show little variation except for the brown adipose

tissue that only has 321 RIDGEs.

The greatest variance is seen for the endocrine category where both the most RIDGEs, 1047 for the ovary, and the least, 137 for the pancreas tissue, are observed. Another category with little variance are the sensory organs, that all form around 800 RIDGEs. The least coherent category was the 11 remaining organs which is in lie with the naming of this group.

### 6.2.3 Immune system RIDGEs

The adipose tissue is very important for immune response (as for example seen in (Pond, 2005)) and it was known already in 1984 (Kast W. Martin, 1984) that the thymus is important for MHC specificity.

The immune related tissues (adipose, bcells, brown fat, cd4$^+$ tcells, cd8$^+$ tcells, and thymus) are not among the highest scoring tissues with regard to number of observed RIDGEs. Instead ovary, snout epidermis, and frontal cortex are the highest scoring tissues. Thus again showing that RIDGEs are not restricted to the immune system.

559 RIDGEs are observed in both the cd4$^+$ tcells (81% of the 686 observed RIDGEs) and the cd8$^+$ tcells (87% of the 641 observed RIDGEs). The remaining 13 and 19% constitutes tcell specific RIDGEs.

#### 6.2.3.1 RIDGE (Col11a2:H2-DMb2)

This RIDGE is very tissue specific because it is only observed in bone. This RIDGE consists of a group of immune system genes at the 3' end and a group of non-immune system genes at the 5' end, raising the question as to whether the latter act as a loop attachment sites or as a transcription initiator. Although not functionally related to the immune-genes, the combination yields both significant gene and UTR score. Thus implying both sequence duplication events and shared upstream regulatory elements with 18 shared TFBS found.

The RIDGE is 112 kbp long and has a RIDGE activity score of 5.08 which is not significantly high. Col11a2 is not involved in the immune response, but with skeletal development, structural molecule activity, collagen, collagen fibril organisation, phosphate transport, and cell adhesion (The Gene Ontology Consortium, 2000). (Bairoch et al., 2005) Brd2 is not involved with the immune response; but with protein serine/threonine kinase activity and spermatogenisis. (The Gene Ontology Consortium, 2000)

#### 6.2.3.2 RIDGE (Psmb8:H2-Eb1)

The RIDGE (Psmb8:H2-Eb1) is observed in activated macrophages, and in four immune tissues (adipose, cd4$^+$ tcells, cd8$^+$ tcells, and the thymus) in addition to the lymph node (and is thus considered a tissue-specific RIDGE).

This RIDGE was not observed in the time series because H2-Eb1 is not projected. The version (Psmb8:H2-Aa) is activated around 3 hours after treatment, and remained active although with a few exceptions (for primed macrophages at 600 and for viral activated macrophages at 210 and 660) for all three conditions. This shorter version is furthermore found in the adrenal gland, bone, lung, lymph node, and the ovary and therefore not tissue-specific.

#### 6.2.3.3 RIDGE A20 (H2-T24, H2-T23, and H2-L)

- H2-T24, {H2-T23}, and H2-L

RIDGE A20 is activated around 3 hours after treatment and remained active throughout the entire time series. Of the members in RIDGE A20 only the gene H2-L (RIDGE A22) is observed in four of the immnune system tissues (bcells, $cd4^+$ tcells, $cd8^+$ tcells, and the thymus; although absent in the two adipose tissues) as well as in embryo at day 6.5. RIDGE A22 is therefore tissue-specific, whereas H2-T24 is not active in any of the tissues, and H2-T23 not projected.

#### 6.2.3.4 H2-Q8, H2-Q8, H2-Q9, H2-Q10 and Pouf51

- [{H2-Q8}], (H2-Q8), ({H2-Q9}), H2-Q10, and Pou5f1

RIDGE (H2-Q8, H2-Q8, and H2-Q10) is observed in adipose (but lacking in the other five immune tissues) as well as bonemarrow, liver, lymph, and the ovary, and is therefore tissue-specific. The RIDGE version (H2-Q8, H2-Q10, and Pou5f1) is not observed in any tissues. The RIDGE (H2-Q8, H2-Q9, H2-Q10, and Pou5f1) is observed in primed macrophages three hours after treatment, with the exception of time point 720; and at 120 and 180-690 in viral activated macrophages and macrophages that were both primed and viral activated but for the latter also present 30 minutes after activation.

### 6.2.4 Discussion

Macrophage activation, be it from priming with IFN-$\gamma$ or viral activation with mCMV results in a process that continues well beyound 12 hours (as seen both in the time series and marcophage activation datasets). For instance, the RIDGEs observed at time 0 and time 720 are very different in both number and composition. Especially for macrophages that were both primed and viral activated where the most RIDGEs are observed at time point 720 and the fourth least is observed for time 0.

The KIAA1949:GNL1 region as seen in the macrophage activation dataset (5.2.1.1) is not observed in the tissue dataset because only four of the 10 genes are reliable projected for this dataset, and only six for the time series dataset, as compared to the macrophage activation

dataset where seven of the 10 genes are projected. Abcf1 is sparsely expressed (only four time points), whereas both GNL1 and Dhx16 are present in at least 54 tissues. Furthermore both Mrps18b and Dhx16 are housekeeping genes.

- (NM_183282), (Adamts10), [Myo1f], ([{Zfp414}]), and ([Pram1]) (216 kbp)

The RIDGE (Adamts10, Myo1f, and Pram1), containing one of the genes in RIDGE A05, is only observed in bone, whereas the overlapping RIDGE (NM_183282 and Adamts10) is only present in testis. The significant gene score and UTR score for the three-gene RIDGE help validate the choice of RIDGE A06 (March2 and {Rab11b}) in the previous chapter. This RIDGE can not be observed since Rab11b is not projected for either the time series or the tissue dataset.

#### 6.2.4.1 All RIDGEs in the tissue dataset were found in the time series

For the MHC locus all potential RIDGEs found in the tissue dataset were found, at least partially, in the time series.

#### 6.2.4.2 Sparsely expressed RIDGEs that map to the MHC locus

RIDGE A10 (Cdsn and Gtf2h4) and RIDGE A02 (Pknox1 and U2af1) are both only observed in a single condition, the rested, for the macrophage dataset. Furthermore they are both only observed in 2 time points in primed macrophages.

RIDGE (C6orf15 and VARSL) and the elongated (Cdsn, C6orf15, and VARSL) is only observed at time point 60 and 90 respectively. Cdsn is also observed in the macrophage activation dataset; in RIDGE A10 with Gtf2h4. It was found to be de-activated as the immune response was activated. RIDGE A10 is observed in macrophages that were both primed and viral activated at time 0, which is the only time point at which the gene Gtf2h4 is considered active for any of the three conditions.

Another example of a sparsely expressed RIDGE that map to the MHC locus is the three Pknox1 RIDGEs [(1500032D15, 4833413E03, and Pknox1), (Wdr4 and the three previous genes) and (the first three genes and Cbs)]. These are only present at time 0 and 150. The RIDGEs A02 (Pknox1 and U2af1) and A01 are not observed since U2af1 is not projected. For the macrophage activation dataset these two RIDGEs are only observed in macrophages that were both primed and viral activated.

## 6.3 Additional loci

The RIDGE analysis is here extended to include additional genomic loci; the protocadherin locus on chromosome 18, and the immunoglobin loci on chromosomes 6 and 12. Furthermore

random regions and their RIDGE contents are also described.

The MHC locus is 5.8 Mbp long and localise to [30852327 - 36681079] on chromosome 17 and contains 250 genes. It has a gene sparsity of one gene per 23 kbp genomic material and a gene coverage of 44%.

### 6.3.1 The Protocadherin locus on chromosome 18

The protocadherin locus on chromosome 18 is 1.4 Mbp long and contains 52 genes. It has one gene every 26 kbp and with a gene coverage of 46%. Because the locus is only one fifth of the size of the MHC locus and contains one fifth of the number of genes; the gene sparsity and gene coverage are very similar to that of the MHC locus.

Protocadherins are calcium-dependent cell adhesion molecules involved in many biological processes, including cell recognition, cell signaling during embryogenesis, and the formation of neural circuits, and is expressed in the nervous system. (Morishita and Yagi, 2007)

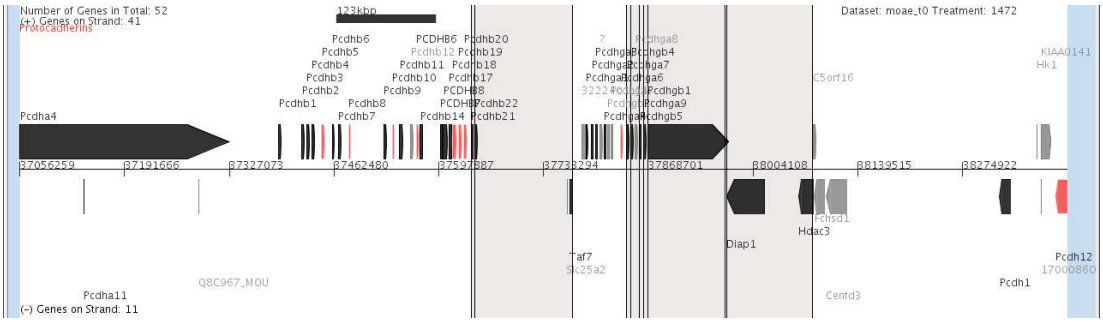



Figure 6.8: RIDGEs found in the Protocadherin locus for the time series dataset. The 11 genes located on the anti-sense strand are shown in *italic*; {Pcdha4}, ({*Pcdha11*}),([{*Q8C967*}]), (Pcdhb1), ({Pcdhb2}), ({Pcdhb3}), (Pcdhb4), (Pcdhb5), ({Pcdhb6}), (Pcdhb7), ({Pcdhb8}), ({Pcdhb9}), ({Pcdhb10}), (Pcdhb11), ([{Pcdhb12}]), ({PCDHB6/Pcdhb13)}), (Pcdhb14), ({PCDHB7/Pcdhb15}), (PCDHB8/Pcdhb16), (Pcdhb17), ({Pcdhb18}), ({Pcdhb19}), (Pcdhb20), (Pcdhb21), (Pcdhb22), ([*Slc25a2*]), *(Taf7*), ([{32222401L}]), ({Pcdhga1}), ({Pcdhga2}), ({Pcdhga3}), ([{?}]), ({Pcdhga4}), ([{Pcdhgb2}]), ([{Pcdhga5}]), ({Pcdhga6}), ({Pcdhga7}), ({Pcdhgb4}), ([{Pcdhga8}]), ({Pcdhgb5}), ({Pcdhga9}), (Pcdhgb1), *Diap1*, *Hdac3*, ([C5orf16]), ([{*Fchsd1*}]), [*Centd3*], (*Pcdh1*), ([{Hk1}]), ([{*1700086O06*}]), ([KIAA0141]*),* and *Pcdh12* (1354 kbp).

Of the 52 genes in the protocadherin locus 39 (75%) are protocadherin genes; representing the entire Pcdhβ family. 9 of the 11 genes on the anti-sense strand are not protocadherins genes (67%), showing a preferential transcriptional orientation. 21 members of the Pcdhγ family map to the same 100 kbp long Ensembl gene identifier; Pcdhgb1.

For the macrophage activation dataset, no RIDGEs are observed in the protocadherin lo-

cus mainly because only 5 genes are projected for this locus. For the time series dataset the RIDGE (Diap1 and Hdac3) is found in addition to 20 RIDGEs in 3 overlapping regions; (Pcdhb1:Pcdhb11), (Pcdhga6:Pcdhgb1), and the (Pcdhb20:Taf7) region. The (Diap1 and Hdac3) RIDGE is also observed in the tissue dataset (cd8$^+$ tcells, prostate, and the thymus) although there elongated with the gene C5orf16 who is not projected for the time series dataset. The single-gene RIDGE (Pcdhgb1) is found in the olfactory bulb and in five of the ten brain tissues; the cerebellum, the dorsal striatum, the frontal cortex, the hippocampus, and the preoptic area.

#### 6.3.1.1 RIDGE (Diap, Hdac3, and C5orf16)

The RIDGE (Diap, Hdac3, and C5orf16), 116 kbp, has a significantly high gene score, but not UTR score. The exon density is 11%. There is a long distance, 46 kbp, in between Diap1 (48 kbp) and Hdac3 (18 kbp), with no missing genes, and a significantly short one between Hdac3 and C5orf16 (4 kbp). Hdac3 is involved in histone deacetylase complex, binding (of DNA and protein), transcription, anti-apoptosis, regulation of progression through mitotic cell cycle, transcription factor binding, chromatin modifications, histone deacetylation, hydrolase activity, and located in the nucleus and the cytoplasm. (The Gene Ontology Consortium, 2000) Diap1 is involved in binding (of actin, receptor, protein, and profilin), sensory perception of sound, cell organisation and biogenesis, Rho GTPase binding, actin cytoskeleton organisation and biogenesis, actin filament polymerisation, and is located in the cytoskeleton. (The Gene Ontology Consortium, 2000) C5orf16 is integral to membrane and located in the membrane (The Gene Ontology Consortium, 2000). Hdac3 and Diap1 both bind protein and other molecules. All three genes have 2 transcripts each. This RIDGE was missing from the MDS dataset because Hdac3 was absent for all four conditions, even though Diap1 was present for all conditions, and C5orf16 had no reliable probe-to-gene projection.

#### 6.3.1.2 RIDGE (Pcdhgb1)

The single-gene RIDGE (Pcdhgb1), 100 kbp, has an exon density of 33% for the RIDGE, which is very high. The gene is involved in binding (of calcium ion and protein), membrane fraction, is integral to plasma membranes, homophilic cell adhesion, and calcium-dependent cell-cell adhesion. It is located in the intracellular junction and in the membrane. (The Gene Ontology Consortium, 2000) Not very surprisingly this long gene has many, 7, transcripts with varying length and exon content.

### 6.3.2 The Immunoglobin locus on chromosome 6

The immunoglobin locus on chromosome 6 is 27.6 Mbp long and contains 350 genes and contains one gene per 79 kbp with a gene density of 33%. This is both sparser and provides

less gene coverage than the MHC locus, but still contains one hundred more genes.

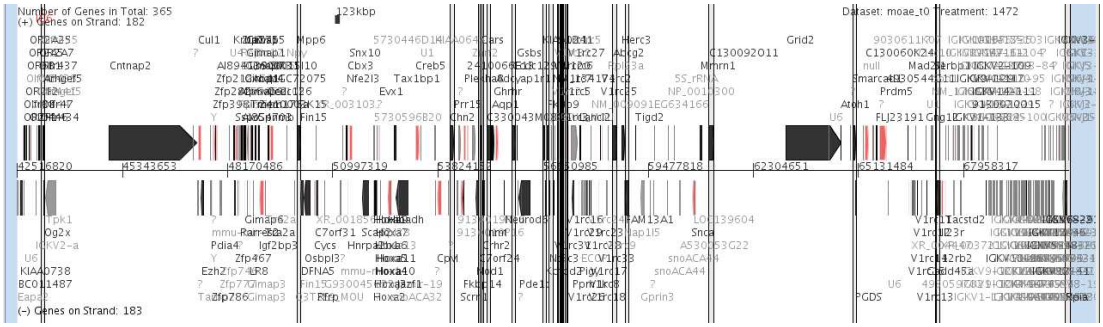


Figure 6.9: The immunoglobulin locus on chromosome 6.

7 RIDGEs are observed for the macrophage activation dataset in the IG locus on chromosome 6; 3 two-gene RIDGEs and 4 single-gene RIDGEs. 3 of these RIDGEs are static and two of these are single-gene RIDGEs. For the time series a total of 60 RIDGEs are observed; 23 non-overlapping RIDGEs and 10 overlapping regions with 37 RIDGEs, where the overlapping RIDGEs had multiple members. For the tissue dataset 5 RIDGEs are observed, of which 4 are single-gene RIDGEs all of which were found in the time series; (Mpp6), (Hibadh), (Creb5), and (Fkbp9), and a two-gene RIDGE (Gimap9 and Repin1).

### 6.3.3 The Immunoglobin locus on chromosome 12

The immunoglobin locus on chromosome 12 contains 147 genes, and is 2.7 Mbp long with one gene every 19 kbp and a gene coverage of 4%. The gene coverage is that low because the total gene length for these 147 genes is only 97 kbp where a gene is only $659 \pm 1268$ bp long. In addition most of the genes in the locus are not reliable projected. Thus no RIDGEs were expected for this locus, and none was observed for the macrophage activation dataset or for the tissue dataset, although five RIDGEs are observed for the time series; (IGHV2-9-1 and IGHV5-17), (C14orf80 and IGHA), (Tmem121, IGHA, IGHE and IGHG2C), (Tmem121:IGHE), and (Vipr2, Zfp386 and IGHV1S12).

Of the 147 genes in the locus, all but 6 are located on the anti-sense strand (and neither of these six are a immunoglobulin genes) thereby showing a preferential orientation of transcription.

### 6.3.4 Random regions

In this section the RIDGE content observed for the macrophage activation dataset for a number of random regions are presented. First the start of chromosome 17 up to the beginning of

the MHC locus is investigated and then the [30852327 - 36681079] region for the other 20 chromosomes (the placement of the MHC locus). All these examples exhibit fewer, and shorter, RIDGEs than those observed for the MHC locus.

#### 6.3.4.1 RIDGEs 5' of the MHC locus

The 5' region of the MHC locus is investigated. This region is 5.3 times the length of the MHC locus (31 Mbp). The region contains 477 genes with one gene every 65 kbp and with a gene coverage of 37%. 15 RIDGEs are observed where about half are static single genes RIDGEs.

#### 6.3.4.2 5.8 Mbp long regions

The MHC locus is located on chromosome 17 at [30852327 - 36681079]. The same physical region is investigated for the other 20 chromosomes, although chromosome Y does not contain any genes for this region.

| Chrom | sparsity | coverage | #genes | | Chrom | sparsity | coverage | #genes |
|---|---|---|---|---|---|---|---|---|
| 1 | 104 | 34 | 56 | | 11 | 104 | 48 | 56 |
| 2 | **53** | 56 | **110** | | 12 | 117 | 30 | 50 |
| 3 | 121 | 23 | 48 | | 13 | 114 | 33 | 51 |
| 4 | 139 | 25 | 42 | | 14 | 93 | 39 | 63 |
| 5 | 54 | 54 | 107 | | 15 | 121 | 41 | 48 |
| 6 | 158 | 39 | **37** | | 16 | 66 | **61** | 88 |
| 7 | 99 | **15** | 59 | | 18 | 68 | 47 | 86 |
| 8 | 142 | 17 | 41 | | 19 | 124 | 26 | 47 |
| 9 | 119 | 20 | 49 | | X | 88 | 22 | 66 |
| 10 | **182** | 22 | 32 | | *17* | *23* | *44* | *250* |

Table 6.2: The sparsity (in kbp), the coverage (in %), and the number of genes per each 5.8 Mbp long chromosome region. The extreme values are shown in bold.

Because it has the highest gene coverage chromsome 16 might have be expected to contain the most RIDGEs, and it does with 9 RIDGEs. Although, the next RIDGE dense chromosome should have been chromosome 2 and not chromosome 3. 5 chromosomes (1, 9, 10, X, and Y) have no RIDGEs and the remaining chromosomes all exhibit 6 or fewer RIDGEs. Furthermore both chromosome 2 and 5 are less sparse than the IG locus on chromosome 6 but exhibit fewer RIDGEs. All of which provide evidence for non-random gene organisation.

The nine RIDGEs on chromosome 16 are 4 overlapping RIDGEs with the members (0610012G03, Ncbp2, and Pak2, Pigx, and Bex6) all of which are static and exhibit significant low RIDGE

activity scores and significant high gene scores. Furthermore the other five RIDGEs in the region are also all static RIDGEs with significant low RIDGE activity scores and significant high gene scores. Therefore these nine RIDGEs are examples of good scoring RIDGEs outside of known genomic loci with members that might be co-regulated and have evolved through recent sequence duplications events.

#### 6.3.4.3 Discussion

From the analysis of the three genomic loci and the 20 random regions it is clear that:

1. RIDGEs are not restricted to immune system genes (although immune system RIDGEs are observed)
2. known genomic loci contain more, and longer, RIDGEs than random regions,
3. RIDGEs outside such loci tended toward single gene RIDGEs unaffected by the condition of study (although examples of good scoring co-regulated RIDGEs are observed),
4. neither region length nor gene content (gene sparcity and gene coverage) can fully explain the observed RIDGEs.

It could be argued that since random regions are less likely to contain RIDGEs, then the randomised genomes will be less likely to contain RIDGEs. By forcing a gene to remain on the same chromosome it increases the probability of the loci genes to remain physically linked in the randomisation process. Also the actual number of observed RIDGEs was atleast 20% more than expected, and the genome is not expected to exhibit this high degree of genetic loci.

# Chapter 7

# Discussion

In the previous chapters, results and discussion have been intertwined in order to present each analysis separately. This final thesis chapter will investigate some of the more significant results and draw an overall conclusion. Finally, further research areas are suggested.

RIDGEs were found to exist in the mouse genome, and examples that consist of functionally related and physically linked genes that are not created by sequence duplication events were observed.

## 7.1 RIDGEs

### 7.1.1 Evolutionary linked units

In the first chapter, five models of operon evolution were presented (1.1.2.4). During this work it was found that some RIDGEs encode proteins belonging to the same family, for example the Proteasome, supporting the natal model of operon evolution. But RIDGE formation may also confer a selective benefit to the individual via decreased cost of protein synthesis (a functional relation again), supporting the co-regulation model. (Lawrence, 1997) Furthermore examples of shared regulatory control for RIDGE members are found, supporting operon evolution through trans-splicing events. (Blumenthal, 1998) The selfish operon model postulates that members are made up of weakly selected functions, but since both immune response and housekeeping RIDGEs are found this does at least not appear to be the mechanism behind all RIDGE formation. (Lawrence, 1997)

### 7.1.2 RIDGE definition

The extensive literature review on loop sizes (as presented in 4.1.1) in combination with the knowledge that loops could contain a single gene (Sumner, 2003) lead to the proposed RIDGE definition. A RIDGE is a sub-genomic region consisting of physically linked genes that are

functionally related, exhibits co-expression and co-regulation, and span around 110 kbp of genomic material.

The way the RIDGE detection algorithm in SORGE was implemented forced all potential RIDGEs to consist of physically linked genes in the form of consecutive neighbors, and to have co-expression in at least one condition. Functional associations between the RIDGE members are seen, such as immune system RIDGEs, housekeeping RIDGEs, and RIDGEs of other functional relations, such as cell cycle genes. Co-regulation of RIDGEs is seen for many RIDGEs in the form of significant *upstream sequence similarity scores* and numerous shared transcription factor binding sites in their 5' regions. Furthermore, the *coding sequence similarity score* associated with the RIDGEs imply that RIDGE members did not arise through recent sequence duplications.

A RIDGE should technically be able to contain silenced genes since the intervening DNA could be considered to loop out. (Hurst et al., 2004) Due to the large quantities of missing data, most notably gene expression data, a RIDGE was in this thesis not allowed to contain silenced gene, as it introduced too much noise. However, genes without a reliable probe-to-gene projection were allowed inside a RIDGE, as these might be either expressed or silenced.

#### 7.1.2.1 RIDGE dimension

RIDGE formation in the MHC locus was independent of the chosen RIDGE dimension (as seen in section 4.3) in the sense that the same members were found in four of the five models. RIDGEs observed for the shorter RIDGE dimension were elongated or combined to make up the RIDGEs found for the longer dimensions, although some new RIDGEs are introduced for each length.

The longest RIDGE dimension tested was too long and did not result in any meaningful RIDGEs. This contradicts the results found by Munkel et al (Munkel and Langowski, 1998) where loops of around 240 kbp were also reported.

### 7.1.3 Consecutive RIDGE analysis

When RIDGEs were defined as 2 or more consecutive active genes, less RIDGEs were seen in the mouse genome than in the randomised genomes. In fact, the latter contained roughly 10% more RIDGEs than the mouse genome. Although when a RIDGE had to have at least 3, 4, or 5 members, then these are more abundant in the mouse genome than in the randomised genomes. This would imply that gene pairs are not occurring more frequently than by chance, although gene triplets do. This contrasts with the results from Cohen et al (Cohen et al., 2000) who found correlated gene pairs and triples, but not quadruplets.

### 7.1.4 Immune system RIDGEs

The majority of the gene expression changes during the immune response was observed after 3 hours (as seen in 6.2.1). While macrophages have an immediate response to activation (Schroder et al., 2004) they also exhibit a high and long response, continuing well beyond 12, and even 24, hours. Evidence for this was observed in the time series dataset with a large discrepancy in RIDGE presence at time 0 and 720, especially for macrophages that were both primed and viral activated.

The immune related tissues (bcells, $cd4^+$ tcells, $cd8^+$ tcells, thymus, adipose, and brown fat) were not among the highest scoring tissues with regard to RIDGE number, although immune system RIDGEs are generally restricted to immune system tissues (with the notable exception of brown fat that did not contain any RIDGEs in the MHC locus). The tissues where the most RIDGEs are found (all from different categories) are ovary, snout epidermis, and frontal cortex. This implies that RIDGEs are not restricted to the immune system.

On the other hand; RIDGEs exhibited non-random chromosome organisation (discussed in 6.1.1) that can not be fully explained by their length, the number of genes they contain, the number of reliable projected genes, gene density, and gene coverage.

### 7.1.5 Housekeeping RIDGEs

A consistent gene order is essential for the assembly of somatically re-arranged genes as seen by the IGs, the TCRs, or the protocadherins. (Wu and Maniatis, 1999) Furthermore linkage, interactions of the products of polymorphic alleles, and clusters of imprinted genes could determine correct gene expression. Where the strongest selection is probably for housekeeping genes since these are both broadly and highly expressed. (Singer et al., 2005) For this study one third of the RIDGEs were found to have housekeeping properties. These RIDGEs contain 1421 non-overlapping genes which equals approximately 5% of the genes in the mouse genome. Coppe et al (Coppe et al., 2006) claimed that 10% of the genome consists of housekeeping genes, implying that half the housekeeping genes fall inside RIDGEs. Although this could be because only half of the genes are represented on the GNF array, so potentially all housekeeping genes fall inside RIDGEs.

#### 7.1.5.1 One housekeeping RIDGE in the MHC locus

Only one housekeeping RIDGE was found inside the MHC locus (and none were really expected); (Notch3 and Rrp1B). This RIDGE is present in 14 tissues, one for each of the seven categories (making it a real housekeeping RIDGE). No housekeeping RIDGEs map to the IG locus on chromosome 12 or the protocadherin locus. Although quite a few single-gene housekeeping RIDGEs fell inside the IG locus on chromosome 6.

### 7.1.6 Functional associations between RIDGE members (and RIDGEs)

Ingenuity was used to suggest functional associations between RIDGE members and RIDGEs. However if they are not known to be functionally related in a network it is not possible to conclude that the genes do not interact, due to the incomplete interaction knowledge. For example according to Ingenuity there are no known functional interactions between the RIDGE members Akap8l, Wiz, and A430107D22. Although according to their GO-terms the first two genes are both associated with nucleic acid binding, zinc ion binding, and metal ion binding, and both locate intracellular.

Functional associations are found between RIDGE members throughout the genome. For instance immune system RIDGEs are found both inside and outside the investigated immune loci, as are non-immune RIDGEs. Most RIDGEs outside of known genomic loci (for example the protocadherin locus) tended toward single-gene RIDGEs unaffected by the conditions of study, implying that the cell line macrophage is specialised towards immune-related genes.

It is pleasing to note that the RIDGEs found within the MHC class II locus were enriched for members in a network underlying cell signaling, immune response, and protein degradation, all in line with known MHC functionality. Furthermore IFN-γ acts as both a hub and a scaffold for this network. (see section 5.3.4)

#### 7.1.6.1 Long-range functional associations

One potential example of long-range influence of physical position on gene expression is RIDGE A15 (5.3.2.2). The individual networks for these RIDGE members could be merged around TNF; a gene located 700 kbp away from Dhx16 on chromosome 17 in mouse.

#### 7.1.6.2 Fully connected networks are seen for RIDGE members

When all RIDGE members for a specific condition were analysed they tended to form fully connected networks in a higher degree than random groups of genes. This could be an artifact of studying the macrophage cell line in combination with invasive immune treatments such as priming with INF-γ and viral activation with mCMV - the genes that are found active, and therefore potentially in a RIDGE are more likely to be associated with the immune response than random genes. Although fully connected networks are also found in some tissue datasets, like the adrenal gland, others form incomplete networks that lack one or more molecules, like the kidney.

### 7.1.7 Time series analysis

For the MHC locus all RIDGEs found in the tissue dataset are also found, at least partially, in the time series. This means that the observed RIDGEs are probably not random occurrence(s)

for one specific biological condition.

Large overlap in RIDGE characteristics between the different time points were observed. RIDGEs were also noted to elongate by adding more members as the cascade progressed. The latter is a possible example of fine tuning within a RIDGE, where some RIDGE members are silenced in certain conditions.

#### 7.1.7.1 Cyclic responses

In the time series dataset (6.2.1), clear resting periods are seen followed by highly activated periods. One such highly active period occur seven hours after treatment and another nine and a half hours after treatment. These periods implie that macrophages have a cyclic response to IFN-γ priming and mCMV viral activation. Further evidence for a cyclic response is found in viral activated macrophage. This erratic response only corresponds to the other two conditions at time 210 and 420, where they behave similarly (although macrophages that are both primed and viral activated generally contain more RIDGEs). The exception to this is the period 240-420 minutes after treatment; a highly active period in the interferon response cascade but a resting period for the response to both priming and viral activation.

The most RIDGEs, for both primed macrophages and viral activated macrophages, are observed five hours after treatment. Macrophages that are both primed and viral activated peak at time 0. This implies a more aggressive immune response when these two invasive treatments are combined. The most RIDGEs, for primed macrophages, are found 630 minutes after treatment. The MHC locus restricted analysis revealed that the most RIDGEs are found in the last three hours, and the least in the first 3 hours. Therefore even if macrophages have a quick response to interferon (Schroder et al., 2004), the longer after activation the greater the response.

### 7.1.8 RIDGE gain, RIDGE loss, static RIDGEs, and RIDGEs in a flux

An important example of RIDGE gain is seen in RIDGE A07; a classical MHC class II locus RIDGE (Psmb8:H2-Eb1), that appeared in activated macrophages. An example of RIDGE loss is RIDGE A10 (Cdsn and Gtf2h4) that is silenced once the macrophage is activated. The latter RIDGE members are associated with skin and hair; functions that are probably less important during an immune response. Another example of re-prioritising is two of the genes that are silenced after activation in the MDS dataset; A430107D22 and Flot1. Both of these genes are involved in signal transduction. A430107D22 is silenced in primed macrophages and Flot1 in viral activation macrophages (although neither gene is silenced in macrophages that are both primed and infected). Flot1 is present at most time points (66 time points), especially for both and primed macrophages. Whereas A430107D22 was more silenced and only present in 50 time points.

Another group of RIDGEs were the overlapping RIDGES A11-A19 in the Ier3:H2-L region. These RIDGE members are in a state of flux with both up and down regulation of members within the same condition.

## 7.2 Further Work

### 7.2.1 Are RIDGEs conserved over evolution?

The work presented in this thesis show that there is indeed higher order chromatin organisation in the form of RIDGE formation, but are RIDGEs found in other eukaryotic species as well? This could be studied by investigating microarray data and chromosomal organisation from additional species, such as human and fruit fly. This would help validate the presence of higher-order chromatin organisation in all eukaryotes.

### 7.2.2 Longer time series with less time in between time points

Here it was seen that the macrophage response to IFN-γ utilised RIDGE formation spanning a 12 hour window. It would also be interesting to see if RIDGE formations are seen in an even higher degree during the first 15-30 minutes after IFN-γ treatment. This because it could be argued that the faster the response to external stimuli need to be, the more benefit an organism would gain from a RIDGE organisation.

### 7.2.3 Biological replicates

The analysis should be repeated with actual whole-genome arrays and preferably with more replicates to remove some of the ambiguity that arise from the determination of active genes.

### 7.2.4 Predictive biology

RIDGE formation could be used to predict what genes will be expressed in a given situation, or the function of an unknown gene, and thus is a step toward predictive biology. For example the housekeeping gene Rrp1B is probably associated with protein binding, superoxide metabolism, and metal ion binding as based on the functional annotation for the RIDGE member Notch3. Furthermore the unknown, missing, RIDGE member is probably a housekeeping gene associated with similar functions.

So further work would be to take the data presented in this study, for example for the Notch3 RIDGE, and make predictions about when, and where, a gene should be active. Predictions could also include functions for an unknown gene. Based on these design an experiment to prove, or refute, the predictions.

## 7.3 Conclusion

This thesis examined the following hypothesis: *There are sub-genomic loci, RIDGEs, in the genome consisting of physically linked genes that are functionally related, and exhibit co-expression and co-regulation.* In this study we observed clusters of genes that are functionally related (both with specific function and housekeeping genes), co-expressed, co-regulated, and fall into genomic structures that fit with the RIDGE model. The data presented in this thesis supports non-random gene order for eukaryotes in line with a number of other genome studies. (Hurst et al., 2004; Cohen et al., 2000; Kruglyak and Tang, 2000; Roy et al., 2002; Boutanaev et al., 2002; Spellman and Rubin, 2002; Lercher et al., 2002; Versteeg et al., 2003; Singer et al., 2005; Caron et al., 2001; Williams and Bowles, 2004) Furthermore, Williams et al (Williams et al., 2002) showed clusters of both housekeeping and immunogenic genes, as was confirmed in this study.

The presence of RIDGEs, as seen in the investigated loci and tissues, supports the hypothesis that there is a connection between chromosome organisation and gene function. The narrow focus, on a well-defined region, in the mouse genome made it possible to map the biology and determine if the found RIDGEs are biological meaningful. RIDGEs were found to not exclusively consist of immune system genes, cell cycle genes, nor of housekeeping genes (although examples of all three categories were found).

To conclude; *there are loci in the mouse genome where physically linked and functionally related genes are co-expressed and co-regulated.*

# Appendix A

# ER diagrams and a JAVA class diagram

## A.1 ER diagram of the project part of the genomic database

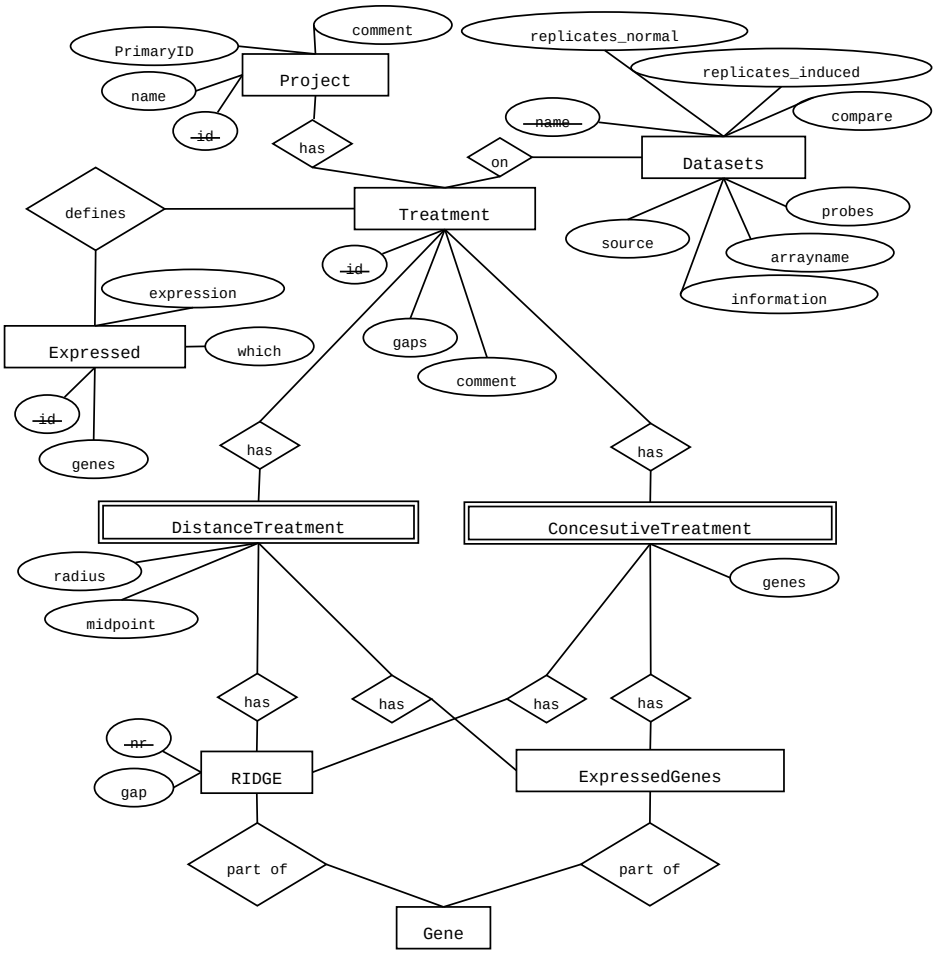



Figure A.1: ER diagram of the project part of the genomic database

## A.2 ER diagram of the bootstrap part of the genomic database

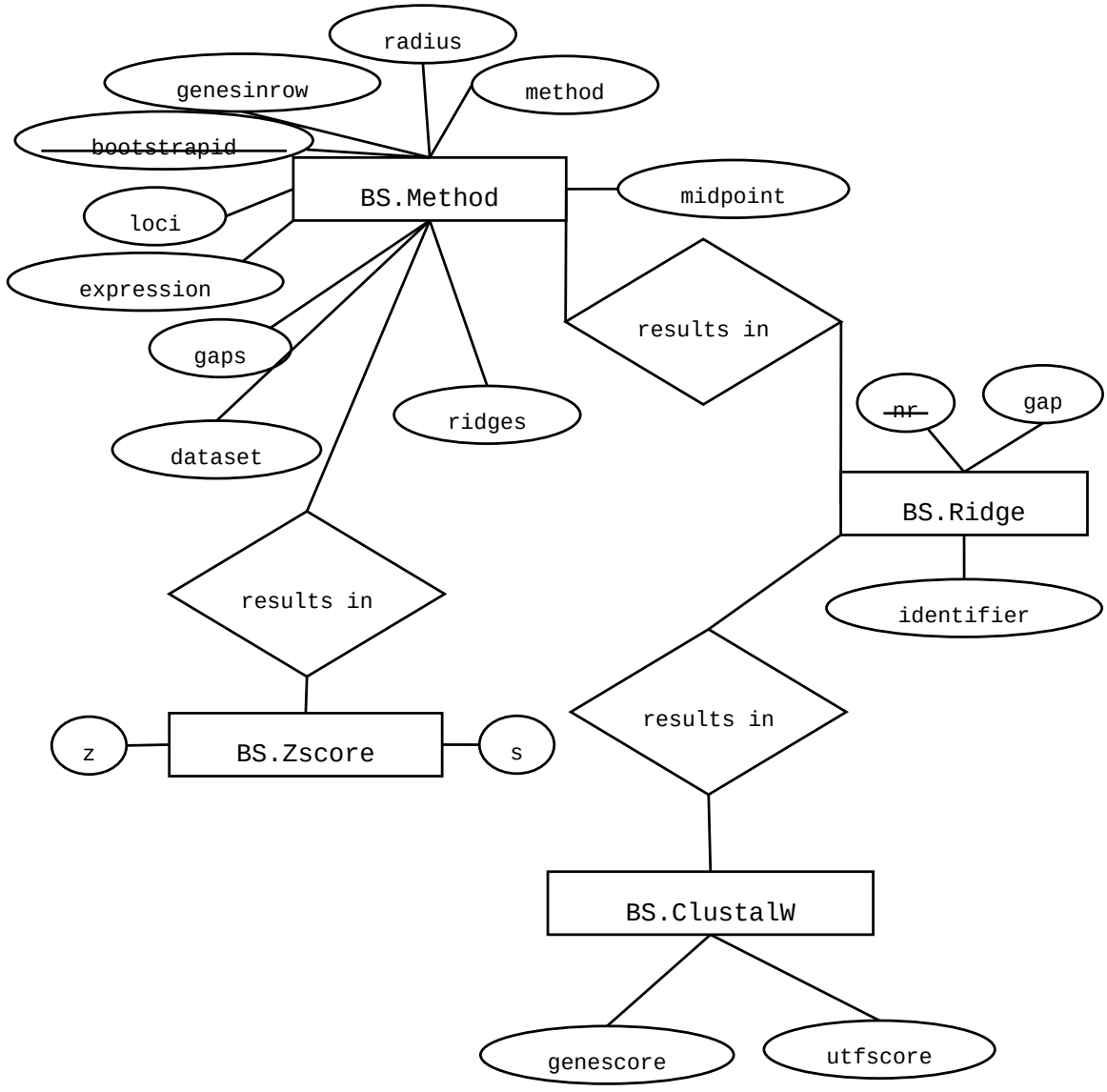



Figure A.2: ER diagram of the bootstrap part of the genomic database

## A.3 ER diagram of the functional annotation database

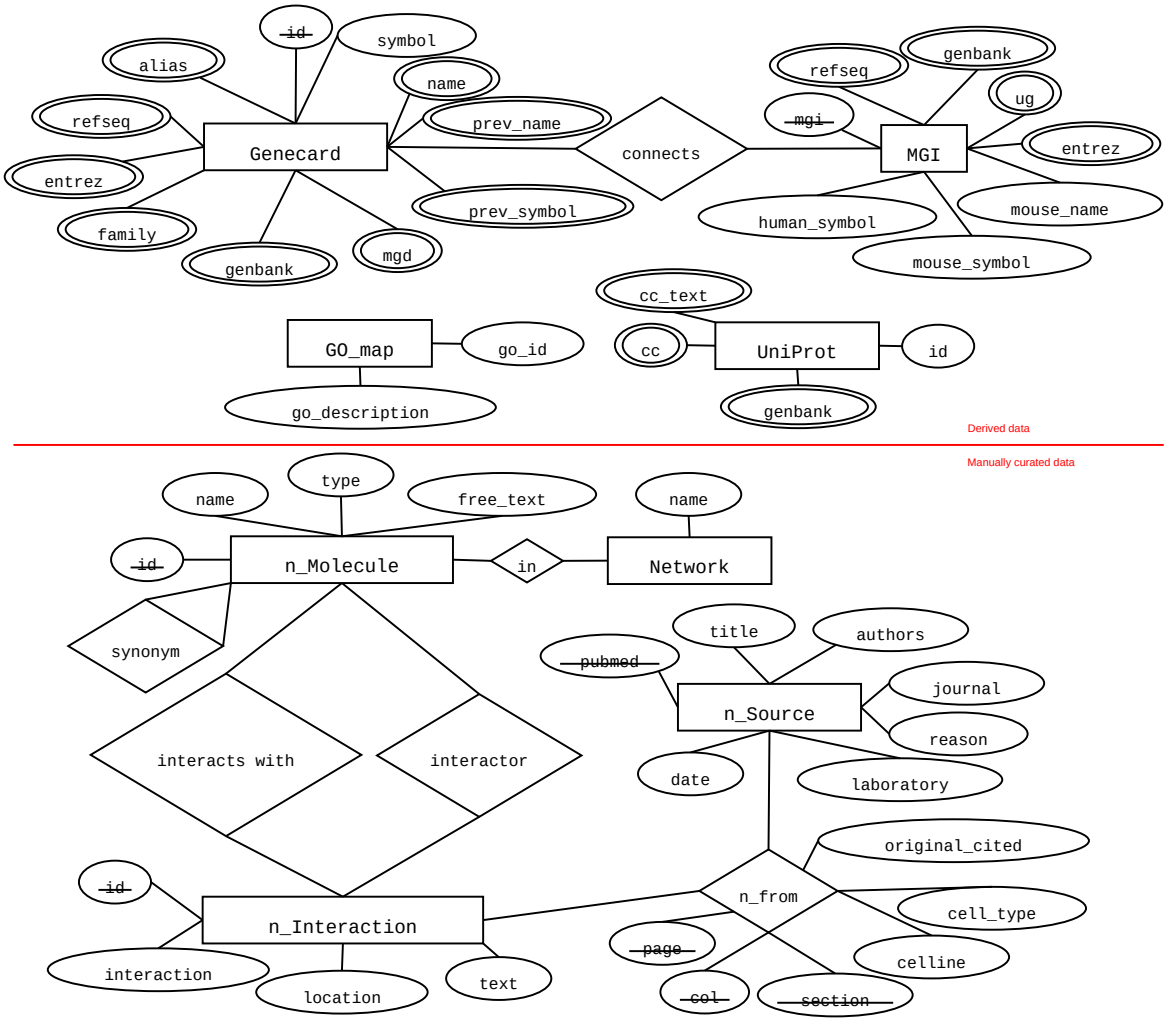



Figure A.3: ER diagram of the functional annotation and molecular interaction database (divided into a derived and a manually curated module).

# Appendix B

# Immune system genes

The 52 articles curated into the molecular interaction database within SORGE DB are the following: (Spilianakis et al., 2003; Masternak et al., 2003; Kretsovali et al., 1998; Zika and Ting, 2005; Karlsson, 2005; Cella et al., 1997; Harding and Neefjes, 2005; Landais et al., 1985; Boss and Jensen, 2003; Ting and Trowsdale, 2002; Aaronson and Horvath, 2002; Hegde et al., 2003; Shortman and Heath, 2001; Huang and Westerlund, 1999; Eberl et al., 1998; Wagle et al., 1999; Cherukuri et al., 2001; Turley et al., 2000; Meraz et al., 1996; Harton and Ting, 2000; Siemasko and Clark, 2001; Muhlethaler-Mottet et al., 1997; Waldburger et al., 2001; Morris et al., 1994b; Perraudeau et al., 2000; Panjwani et al., 1999; Danchin et al., 2004; Lepin et al., 2000; Yu et al., 2005; Ge et al., 2004; Williams et al., 2002; Murray et al., 1999; Wright et al., 1995; Speek et al., 1996; Herberg et al., 1998; Tripodis et al., 1998; Ohta et al., 2002; Rudensky et al., 1991; Kelley et al., 2005; Trowsdale, 2002; Liu and Shaw, 2001; Yang et al., 1998; Yang and Yu, 2000; Meng Kian Tee and Miller, 1995; Alfonso et al., 2001; Aziz et al., 1993; Estess et al., 1986; Huang and Yanda, 2004; Walter et al., 2000, 2001; Benoist et al., 1983)

# Appendix C

# Observed RIDGEs with zero, one, and two gaps

## C.1 RIDGEs in the MHC locus identified by SORGE

### C.1.1 RIDGEs with no silenced genes

| RIDGE identifier: Genes in the RIDGE | control | IFN | mCMV | both |
|---|---|---|---|---|
| **A01**: 1500032D16, Pknox1, U2af1 | PAP | P↓↓ | PAP | PP↑ |
| **A02**: Pknox1, U2af1 | AP | ↓↓ | AP | P↑ |
| **A03**: Brd4 | P | P | P | P |
| **A04**: Akap8l, Wiz, A430107D22 | PPP | PP↓ | ↑P↑ | ↑↓P |
| **A05**: Myo1f, March2 | PP | PP | ↑P | ↑P |
| **A06**: March2, Rab11b | PP | PP | PP | PP |
| **A07**: Psmb8, Tap2, H2-Ab1, H2-Aa, H2-Eb1 | PPPAA | ↑P↑↑↑ | ↑↑↑↑↑ | ↑↑↑↑↑ |
| **A08**: Tap2, H2-Ab1, H2-Aa, H2-Eb1 | PPAA | P↑↑↑ | ↑↑↑↑ | ↑↑↑↑ |
| **A09**: H2-Q2 | P | P | P | P |
| **A10**: Cdsn, Gtf2h4 | PP | AP | AP | AP |
| **A11**: Ier3, Flot1, TUBB, KIAA1949, Dhx16, 2310061I04, Mrps18b | PPPPPPP | PP↓↑↓P↓ | ↓M↓P↓PP | ↓P↓PPP↑ |
| **A12**: Flot1, TUBB, KIAA1949, Dhx16, 2310061I04, Mrps18b | PPPPPP | P↓↑↓P↓ | M↓P↓PP | P↓PPP↑ |
| **A13**: TUBB, KIAA1949, Dhx16, 2310061I04, Mrps18b, Abcf1 | PPPPPP | ↓↑↓P↓↓ | ↓P↓PP↑ | ↓PPP↑P |
| **A14**: KIAA1949, Dhx16, 2310061I04, Mrps18b, Abcf1 | PPPPP | ↑↓P↓↓ | P↓PP↑ | PPP↑P |
| **A15**: KIAA1949, Dhx16, 2310061I04, Mrps18b, Abcf1, GNL1 | PPPPPM | ↑↓P↓↓M | P↓PP↑M | PPP↑PP |
| **A16**: Dhx16, 2310061I04, Mrps18b, Abcf1 | PPPP | ↓P↓↓ | ↓PP↑ | PP↑P |
| **A17**: Dhx16, 2310061I04, Mrps18b, Abcf1, GNL1 | PPPPM | ↓P↓↓M | ↓PP↑M | PP↑PP |
| **A18**: 2310061I04, Mrps18b, Abcf1, GNL1, H2-T24 | PPPMP | P↓↓M↑ | PP↑M↑ | P↑PP↑ |
| **A19**: Mrps18b, Abcf1, GNL1, H2-T24, H2-T23 | PPMPP | ↓↓M↑↑ | P↑M↑↑ | P↑PP↑↑ |
| **A20**: H2-T24, H2-T23, H2-L | PPP | P↑↑ | ↑↑↑ | ↑↑↑ |
| **A211**: H2-T23, H2-L | PP | ↑↑ | ↑↑ | ↑↑ |
| **A222**: H2-L | P | ↑ | ↑ | ↑ |

Table C.1: Observed RIDGEs with no silenced genes The left most column shows the RIDGE identifier and, in genomic order, the genes it contains. For each of the four treatments a gene is flagged as present (P), marginal (M), absent (A), or as having a significant change in expression level from control, i.e. foldchange$>$2 or p$<$0.05, either up $(\uparrow)$ or down $(\downarrow)$, and genes highlighted in red are either marginal or absent. A RIDGE is only present for a specific condition, if all genes are shown in black; for example gene A430107D22 in RIDGE R01 is downregulated in the primed macrophages and thereby goes from present to absent. RIDGEs that are present under all 4 treatments, for example RIDGE R03, is refereed to as static, even though the genes might change. Double lines show where a group of overlapping RIDGEs start, and end.

### C.1.2 RIDGEs with one silenced gene

| RIDGE nr: Genes in the RIDGE | control | IFN | mCMV | both |
|---|---|---|---|---|
| **B01**: Zfp297, Tapbp, Wdr46, H2-K1 | PPMP | PP↓↑ | P↑A↑ | P↑↓↑ |
| **B02**: Tapbp, Wdr46, H2-K1 | PMP | P↓↑ | ↑A↑ | ↑↓↑ |
| **B03**: Brd2, H2-DMa, H2-Dmb2, Psmb9, Tap1, Psmb8, Tap2 | PMAPMPP | ↑M↑↑↑↑P | PMA↑↑↑↑ | P↑↑↑↑↑↑ |
| **B04**: H2-DMa, H2-DMb2, Psmb9, Tap1, Psmb8, Tap2, H2-Ab1 | MAPMPPP | M↑↑↑↑P↑ | MA↑↑↑↑↑ | ↑↑↑↑↑↑↑ |
| **B05**: H2-Dmb2, Psmb9, Tap1, Psmb8, Tap2, H2-Ab1 | APMPPP | ↑↑↑↑P↑ | A↑↑↑↑↑ | ↑↑↑↑↑↑ |
| **B06**: Psmb9, Tap1, Psmb8, Tap2, H2-Ab1 | PMPPP | ↑↑↑P↑ | ↑↑↑↑↑ | ↑↑↑↑↑ |
| **B07**: Psmb9, Tap1, Psmb8, Tap2, H2-Ab1, H2-Aa, H2-Eb1 | PMPPPAA | ↑↑↑P↑↑↑ | ↑↑↑↑↑↑↑ | ↑↑↑↑↑↑↑ |
| **B08**: Tap2, H2-Ab1, H2-Aa, H2-Eb1, H2-Eb2, H2-Ea | PPAAAP | P↑↑↑A↑ | ↑↑↑↑A↑ | ↑↑↑↑A↑ |
| **B09**: H2-Ab1, H2-Aa, H2-Eb1, H2-Eb2, H2-Ea | PAAAP | ↑↑↑A↑ | ↑↑↑A↑ | ↑↑↑A↑ |
| **B10**: ?, Ddah2, Csnk2nb, Bat4 | PAPP | ↓↓PP | ↓APP | ↓APP |

Table C.2: Observed RIDGEs with one silenced gene

When we allow up to a single silenced gene in a RIDGE, the previous described RIDGEs are found in addition to 10 new RIDGEs. These RIDGEs fall into three regions; 7 RIDGEs in the MHC class II locus, 2 in the Zfp297:H2-K1 region, and 1 more RIDGE.
Here genes highlighted in blue are either absent or marginal. Because we now allow a silenced gene, it takes 2, or more, silenced genes to affect the RIDGE (shown in red). For example RIDGE 1197.10 has more than 2 silenced genes and are therefore shown in red, whereas RIDGE 1202.07 only has one silenced gene (shown in blue).

### C.1.3 RIDGEs with two silenced genes

| RIDGE nr: Genes in the RIDGE | control | IFN | mCMV | both |
|---|---|---|---|---|
| **C01**: Rxrb, ?, H2-Oa, Brd2, H2-DMa, H2-DMb2 | AAAPMA | A↓A↑M↑ | MAAPMA | PAAP↑↑ |
| **C02**: Csnk2b, Bat4, Apom, Bat3, Bat2, Aif1, Lst1, Ltb, Tnf | PPAPMPPAP | PP↓↓↓↑↑↑↑ | PP↓↓M↑↑A↓ | PP↓↓M↑↑↑P |

Table C.3: Observed RIDGE with 2 silenced genes

When allowing an additional silenced gene, only 2 new RIDGEs are found. Similar to the previous table, blue genes are either marginal or absent, and 3, or more, silenced genes are highlighted in red.

#### C.1.3.1 RIDGE loss after viral activation

RIDGEs A11 and A12, in the Ier3:H2-L region, are lacking after infection because Flot1 is considered marginal, with the following replicate values; 235.5 (P), 233.1 (A), and 238 (M), e.g.

here 3 replicates is not enough to determine the detection call. The four missing genes for both RIDGEs are (in *italic*): Ier3, Flot1, TUBB, *Mdc1*, *5530401N12Rik*, *Nrm*, 2310014H01Rik, Dhx16, 2310061I04Rik, *C6orf134*, and Mrps18b, i.e. there is a large gap in the middle which is evidence in favor of the the GNL1-based RIDGEs described previously rather than these variants. Lipid raft protein Flotillin-1 is associated with the cytoskeleton and integrin signalling, (Jonathan Kerr, 2006) and is part of the caveolae. The caveolae has at least three functions: 1) mediate the transcytosis of macromolecules, 2) be the site of potocytosis (e.g. transverse the plasma membrane), and it 3) may participate in the relay of extracellular signals to the cells interior by organising signal transduction molecules. (Bickel et al., 1997) The mCMV virus might want to disable either the transport or signaling function of Flot1, but here the gene is marginal.

Ier3 and TUBB are both involved in the cell cycle, cell death (apoptosis), and DNA replication, recombination, and repair. Ier3 is further involved in cellular assembly, organisation, development, growth, and proliferation, tissue development, skeletal and muscular disorders, hematological system development and function, the immune response, gene expression, and is integral to membrane. It is a single-pass type II membrane protein found in lung, testes, uterus, and the membrane, and is part of the IER3 family. TUBB is a known housekeeping gene involved in GTPase activity, structural molecule activity, cellular compromise, function, and maintenance, immunological disease, hematological disease, and binding of MHC class I proteins, GTP, and nucleotide. The gene is found in the cytoplasm, the microtubule, the protein complex, the tubulin, and the cytoskeleton, is expressed in the spleen, the thymus, and the immature brain, and associated with interferon.

1. The first RIDGE, A11 - Ier3, Flot1, TUBB, 2310014H01Rik, Dhx16, 2310061I04Rik, and Mrps18b, has a RIDGE activity score of 2.20 which is not significant. The RIDGE length is 95 kbp, and the density 30% with short ($<$14 kbp) inter-gene distances (except for 27 kbp between TUBB and 2310014H01). The genes have a single transcript each with varying number of exons (2-20). Some functional overlap exists between Ier3 and TUBB. Neither the gene score nor the UTR score are significant, but there are still 16 TFBS shared by the seven genes; AP-1, C/EBP$\alpha$ and $\beta$, c-Fos, c-Jun, f($\alpha$)-f(*ilon*), GR, HES-1, HOXA5, JunD, MyoD, myogenin, NF-1, NF-$\kappa$B, TCF-1, USF-1, and YY1.

2. The second RIDGE, A12, lacks the first gene (Ier3) and has a RIDGE activity score of 2.88 which is not significant. The RIDGE length is 93 kbp, and the density 29%, e.g. still dense. Neither the gene score, nor the UTR score are significant, but these 6 genes share the same 16 TFBS as the previous RIDGE.

# Bibliography


Aaronson, D. S. and Horvath, C. M. A Road Map for Those Who Don't Know JAK-STAT. *Science*, 296(5573):1653–1655, 2002.

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., and Moor, B. D. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucl. Acids. Res.*, 31(6):1753–1764, 2003.

Albiez, H., Cremer, M., Tiberi, C., Vecchio, L., Schermelleh, L., Dittrich, S., Küpper, K., Joffe, B., Thormeyer, T., von Hase, J., Yang, S., Rohr, K., Leonhardt, H., Solovei, I., Cremer, C., Fakanand, S., and Cremer, T. Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. *Chromosome Research*, 14(7):707–733, 2006.

Alfonso, C., Han, J.-O., Williams, G. S., and Karlsson, L. The Impact of H2-DM on Humoral Immune Responses. *J Immunol*, 167(11):6348–6355, 2001.

Alsford, S. and Horn, D. Trypanosomatid histones. *Mol Microbiol*, 53(2):365–72, 2004.

Anderson, K., Hess, K. R., Kapoor, M., Tirrell, S., Courtemanche, J., Wang, B., Wu, Y., Gong, Y., Hortobagyi, G. N., Symmans, W. F., and Pusztai, L. Reproducibility of Gene Expression Signature Based Predictions in Replicate Experiments. *Clin Cancer Res*, 12(6):1721–7, 2006.

Aziz, N., Maxwell, M. M., Jacques, B. S., and Brenner, B. M. Downregulation of Ke 6, a novel gene encoded within the major histocompatibility complex, in murine polycystic kidney disease. *Molecular and Cellular Biology*, 13(3):1847–1853, 1993.

Baerends, R., Smits, W., de Jong, A., Hamoen, L., Kok, J., and Kuipers, O. Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data. *Genome Biol*, 5(5): R37, 2004.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. The Universal Protein Resource (UniProt). *Nucl. Acids Res.*, 33 (suppl_1):D154–159, 2005.

Bartoloni, L. and Antonarakis, S. The human sugar-phosphate/phosphate exchanger family SLC37. *Pflugers Arch*, 447(5):780–3, 2004.

Bednar, J., Horowitz, R., Grigoryev, S., Carruthers, L., Hansen, J., Koster, A., and Woodcock, C. Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc Natl Acad Sci USA*, 95(24): 14173–8, 1998.

Benoist, O., Mathis, D. J., Kanter, M. R., Williams, V. E., and McDevitt, H. O. The murine Ia alpha chains, E alpha and A alpha, show a surprising degree of sequence homology. *PNAS*, 80(2):534–8, 1983.

Bickel, P. E., Scherer, P. E., Schnitzer, J. E., Oh, P., Lisanti, M. P., and Lodish, H. F. Flotillin and epidermal surface antigen define a new family of caeolae-associated integral membrane proteins. *Journal of Biology, Chemistry*, 272(31):13793–802, 1997.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Graf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., and Hubbard, T. J. P. Ensembl 2006. *Nucl. Acids Res.*, 34(suppl_1):D556–561, 2006.

Blanton, J., Gaszner, M., and Schedl, P. Protein:protein interactions and the pairing of boundary elements in vivo. *Genes and Development*, 17(5):664–675, 2003.

Blumenthal, T. Gene clusters and polycistronic transcription in eukaryotes. *BioEssays*, 20(6): 480–487, 1998.

Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M., and Kim, S. K. A global analysis of Caenorhabditis elegans operons. *Nature*, 417(6891):851–4, 2002.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M. R., and Cremer, T. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biology*, 3(5):e157, 2005.

Bortoluzzi, S., Rampoldi, L., Simionati, B., Zimbello, R., Barbon, A., d'Alessi, F., Tiso, N., Pallavicini, A., Toppo, S., Cannata, N., Valle, G., Lanfranchi, G., and Danieli, G. A. A Comprehensive, High-Resolution Genomic Transcript Map of Human Skeletal Muscle. *Genome Res.*, 8(8):817–825, 1998.

Boss, J. M. and Jensen, P. E. Transcriptional regulation of the MHC class II antigen presentation pathway. *Current Opinion in Immunology*, 15(1):105–11, 2003.

Boutanaev, A., Kalmykova, A., Shevelyov, Y., and Nurminsky, D. Large clusters of co-expressed genes in the Drosophila genome. *Nature*, 420(6916):666–9, 2002.

Bowcock, A. M. and Krueger, J. G. Getting under the skin: the immunogenetics of psoriasis. *Nature Reviews Immunology*, 5(9):699–711, 2005.

Brooksbank, C., Camon, E., Harris, M., Magrane, M., Martin, M., Mulder, N., O'Donovan, C., Parkinson, H., Tuli, M., Apweiler, R., Birney, E., Brazma, A., Henrick, K., Lopez, R., Stoesser, G., Stoehr, P., and Cameron, G. The European Bioinformatics Institute's data resources. *NAR*, 31(1):43–50, 2003.

Brooksbank, C., Cameron, G., and Thornton, J. The European Bioinformatics Institute's data resources: towards systems biology. *NAR*, 33(Database issue):D46–53, 2005.

Bulger, M. and Groudine, M. Looping versus linking: toward a model for long-distance gene activation. *Genes and Development*, 13:2465.2477, 1999.

Bystricky, K., Heun, P., Gehlen, L., Langowski, J., and Gasser, S. M. Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *PNAS*, 101(47):16495–16500, 2004.

Capellini, T. D., Di Giacomo, G., Salsi, V., Brendolan, A., Ferretti, E., Srivastava, D., Zappavigna, V., and Selleri, L. Pbx1/Pbx2 requirement for distal limb patterning is mediated by the hierarchical control of Hox gene spatial distribution and Shh expression. *Development*, 133 (11):2263–2273, 2006.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M., van Asperen, R., Boon, K., Voute, P., Heisterkamp, S., van Kampen, A., and Versteeg, R. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, 291(5507):1289–1292, 2001.

Cella, M., Sallusto, F., and Lanzavecchia, A. Origin, maturation and antigen presenting function of dendritic cells. *Current Opinion in Immunology*, 9:10–16, 1997.

Chakalova, L., Debrand, E., Mitchell, J., Osborne, C., and P., F. Replication and transcription: shaping the landscape of the genome. *Nature Reviews Genetics*, 6(9):669–677, 2005.

Chen, H., Lin, R. J., Schiltz, R. L., Chakravarti, D., Nash, A., Nagy, L., Privalsky, M. L., Nakatani, Y., and Evans, R. M. Nuclear Receptor Coactivator ACTR Is a Novel Histone Acetyltransferase and Forms a Multimeric Activation Complex with P/CAF and CBP/p300. *Cell*, 90(3):569–580, 1997.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acid Res*, 31 (13):3497–500, 2003.

Cherukuri, A., Cheng, P. C., and Pierce, S. K. The Role of the CD19/CD21 Complex in B Cell Processing and Presentation of Complement-Tagged Antigens. *J Immunol*, 167(1):163–172, 2001.

Cheung, J., Wilson, M., Zhang, J., Khaja, R., MacDonald, J., Heng, H., Koop, B., and Scherer, S. Recent segmental and gene duplications in the mouse genome. *Genome Biology*, 4(8): R47, 2003.

Clark, S. G., Lu, X., and Horvitz, H. R. The Caenorhabditis elegans Locus lin-15, a Negative Regulator of a Tyrosine Kinase Signaling Pathway, Encodes Two Different Proteins. *Genetics*, 137(4):987–997, 1994.

Cohen, B. A., Mitra, R. D., Hughes, J. D., and Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genetics*, 26(2):183–186, 2000.

Colditz, I. Effects of the immune system on metabolism: implications for production and disease resistance in livestock. *Livestock Production Science*, 75:257–268(12), 2002.

Cook, G., Tomlinson, I., Walter, G., Riethman, H., Carter, N., Buluwela, L., Winter, G., and Rabbitts, T. A map of the human immunoglobulin VH locus completed by analysis of the telomeric region of chromosome 14q. *Nat Genet*, 7(2):162–8, 1994.

Cook, P. A chromomeric model for nuclear and chromosome structure. *J. Cell. Sci.*, 108(9): 2927–2935, 1995.

Coppe, A., Danieli, G. A., and Bortoluzzi, S. REEF: searching REgionally Enriched Features in genomes. *BMC Bioinformatics*, 7(453), 2006.

Cremer, T. and Cremer, C. Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells. *Nature Reviews Genetics*, 2(4):292–301, 2001.

Danchin, E., Vitiello, V., Vienne, A., Richard, O., Gouret, P., McDermott, M. F., and Pontarotti, P. The major histocompatibility complex origin. *Immunol Rev*, 198(1):216–232, 2004.

Davalos, N., Garcia-Vargas, A., Pforr, J., Davalos, I., Picos-Cardenas, V., Garcia-Cruz, D., Kruse, R., Figuera, L., Nothen, M., and Betz, R. A non-sense mutation in the corneodesmosin gene in a Mexican family with hypotrichosis simplex of the scalp. *Br.J.Dermatol.*, 153(6):1216–9, 2005.

de Boer, J., de Wit, J., van Steeg, H., Berg, R. J. W., Morreau, H., Visser, P., Lehmann, A. R., Duran, M., Hoeijmakers, J. H. J., and Weeda, G. A Mouse Model for the Basal Transcription/DNA Repair Syndrome Trichothiodystrophy. *Molecular Cell*, 1(7):981–990, 1998.

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., and Lempicki, R. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4(5):P3, 2003.

Dickinson, P. personal communication, 2007.

Dietzel, S., A, J., Kienle, D., Qu, G., Holtgreve-Grez, H., Eils, R., Munkel, C., Bittner, M., Meltzer, P., Trent, J., and Cremer, T. Separate and variably shaped chromosome arm domains are disclosed by chromosome arm painting in human cell nuclei. *Chromosome Res*, 6(1): 25–33, 1998.

Dorus, S., Vallender, E. J., Evans, P. D., Anderson, J. R., Gilbert, S. L., Mahowald, M., Wyckoff, G. J., Malcom, C. M., and Lahn, B. T. GenoMap, a circular genome data viewer. *Cell*, 119:1027–40, 2004.

Dowell, R., Jokerst, R., Day, A., Eddy, S., and Stein, L. The Distributed Annotation System. *BMC Bioinformatics*, 2(1):7, 2001.

Duret, L. and Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution*, 17(1):68–74, 2000.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

Eberl, Jiang, YU, Schneider, Corradin, and Mach. An anti-CD19 antibody coupled to a tetanus toxin peptide induces efficient Fas ligand (FasL)-mediated cytotoxicity of a transformed human B cell line by specific CD4+ T cells. *Clin Exp Immunol*, 114(2):173–178, 1998.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

Eisenberg, E. and Levanon, E. Y. Human housekeeping genes are compact. *Trends in Genetics*, 19(7):362–5, 2003.

Elefant, F., Su, Y., Liebhaber, S., and Cooke, N. Patterns of histone acetylation suggest dual pathways for gene activation by a bifunctional locus control region. *EMBO J*, 19(24):6814–22, 2000.

Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E., and Blake, J. A. The Mouse Genome Database (MGD): from genes to mice–a community resource for mouse biology. *Nucl. Acids Res.*, 33(suppl_1):D471–475, 2005.

Estess, P., Begovich, A. B., Koo, M., Jones, P. P., and McDevitt, H. O. Sequence Analysis and Structure-Function Correlations of Murine q, k, u, s, and f Haplotype I-Abeta cDNA Clones. *PNAS*, 83(11):3594–3598, 1986.

Farré, D., Roset, R., Huerta, M., Adsuara, J. E., Roselló, L., Albà, M., and Messeguer, X. Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. *Nucleic Acid Res*, 31(13):3651–3653, 2003.

Forsberg, E. and Bresnick, E. Histone acetylation beyond promoters: long-range acetylation patterns in the chromatin world. *Bioessays*, 23(9):820–30, 2001.

Fujii, T., Dracheva, T., Player, A., Chacko, S., Clifford, R., Strausberg, R. L., Buetow, K., Azumi, N., Travis, W. D., and Jen, J. A preliminary transcriptome map of non-small cell lung cancer. *Cancer Research*, 62(12):3340–3346, 2002.

Fukuoka, Y., Inaoka, H., and Kohane, I. S. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. *BMC Genomics*, 5(4), 2004.

Ge, H., Yang, G., Huang, L., Motola, D. L., Pourbahrami, T., and Li, C. Oligomerization and Regulated Proteolytic Processing of Angiopoietin-like Protein 4. *J. Biol. Chem.*, 279(3): 2038–2045, 2004.

Ghai R, Hain T, C. T. GenomeViz: visualizing microbial genomes. *BMC bioinformatics*, 5 (198), 2004.

Goldsby, R. A., Kindt, T. J., Osborne, B. A., and Kuby, J. *Immunology*. W.H. Freeman, 5 edition, 2003.

Harding, C. V. and Neefjes, J. Antigen processing and recognition. *Current Opinion in Immunology*, 17(1):55–57, 2005.

Harton, J. A. and Ting, J. P.-Y. Class II Transactivator: Mastering the Art of Major Histocompatibility Complex Expression. *Mol. Cell. Biol.*, 20(17):6185–6194, 2000.

Hebbes, T., Clayton, A., Thorne, A., and Crane-Robinson, C. Core histone hyperacetylation comaps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. *EMBO*, 13(8):1823–30, 1994.

Hegde, N. R., Chevalier, M. S., and Johnson, D. C. Viral inhibition of MHC class II antigen presentation. *Trends in Immunology*, 24(5):278–285, 2003.

Heise, M. T., Connick, M., and Virgin, H. W. I. Murine Cytomegalovirus Inhibits Interferon gamma-induced Antigen Presentation to CD4 T Cells by Macrophages Via Regulation of Expression of Major Histocompatibility Complex Class II-associated Genes. *J. Exp. Med.*, 187(7):1037–46, 1998.

Herberg, J. A., Beck, S., and Trowsdale, J. TAPASIN, DAXX, RGL2, HKE2 and four new genes (BING 1, 3 to 5) form a dense cluster at the centromeric end of the MHC. *Journal of Molecular Biology*, 277(4):839–57, 1998.

Hodges, E., Krishna, M. T., and Smith, J. L. Diagnostic role of tests for T cell receptor (TCR) genes. *J Clin Pathol*, 56:1–11, 2003.

Hoffmann, R. and Valencia, A. A Gene Network for Navigating the Literature. *Nature Genetics*, 36(664), 2004.

Hoffmann, R. and Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21(suppl 2):ii252–8, 2005.

Horowitz, R., Agard, D., Sedat, J., and Woodcock, C. The three-dimensional architecture of chromatin in situ: electron tomography reveals fibers composed of a continuously variable zig-zag nucleosomal ribbon. *J Cell Biol*, 125(1):1–10, 1994.

Huai, Q., Wang, H., Zhang, W., Colman, R., Robinson, H., and Ke, H. Crystal structure of phosphodiesterase 9 shows orientation variation of inhibitor 3-isobutyl-1-methylxanthine binding. *Proc. Natl. Acad. Sci*, 101(26):9624–9, 2004.

Huang, d. W., Sherman, B., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M., Lane, H., and Lempicki, R. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res*, 1(35):W169–75, 2007.

Huang, L., Tzou, P., and Sternberg, P. The lin-15 locus encodes two negative regulators of Caenorhabditis elegans vulval development. *Mol. Biol. Cell*, 5(4):395–411, 1994.

Huang, X. and Westerlund, L. Phenotypic and Functional Properties of Dendritic Cells Isolated from Human Peripheral Blood in Comparison with Mononuclear Cells and T Cells. *Scandinavian Journal of Immunology*, 49(2):177–83, 1999.

Huang, Y. and Yanda, L. Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*, 20(1):21–28, 2004.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., and Birney, E. Ensembl 2005. *Nucl. Acids Res.*, 33(suppl_1):D447–453, 2005.

Hui, L., Lu, J., Han, Y., and Pilder, S. The mouse T complex gene Tsga2, encoding polypeptides located in the sperm tail and anterior acrosome, maps to a locus associated with sperm motility and sperm-egg interaction abnormalities. *Biol. Reprod.*, 74(4):633–43, 2006.

Hurst, L. D., Williams, E. J. B., and Pál, C. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends in Genetics*, 18(12):604–6, 2002.

Hurst, L. D., Pál, C., and Lercher, M. J. The evolutionary dynamics of gene order. *Nature reviews Genetics*, 5(4):299–310, 2004.

Hurst, L. and Smith, N. Do essential genes evolve slowly? *Curr Biol.*, 9(14):747–50, 1999.

Huynen, M. A., Snel, B., and Bork, P. Inversions and the dynamics of eukaryotic gene order. *Trends in Genetics*, 17(6):304–6, 2001.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(Suppl 1):S233–40, 2002.

Izban, M. and Luse, D. Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *J Biol Chem*, 267 (19):13647–55, 1992.

Jackson, D. A., Dickinson, P., and Cook, P. R. The size of chromatin loops in HeLa cells. *EMBO*, 9(2):567–71, 1990.

Jackson, D. and Cook, P. Transcription occurs at a nucleoskeleton. *EMBO*, 4(4):919–25, 1985.

Jackson, D. and Cook, P. Transcriptionally active minichromosomes are attached transiently in nuclei through transcription units. *J Cell Sci*, 105(Pt 4):1143–50, 1993.

Johannes, L., Arnheiter, H., and Meier, E. Switch in antiviral specificity of a GTPase upon translocation from the cytoplasm to the nucleus. *Journal of virology*, 67(3):1653–57, 1993.

Kanehisa, M. A database for post-genome analysis. *Trends Genet*, 13(9):375–376, 1997.

Kanehisa, M. and Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *NAR*, 28(1): 27–30, 2000.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *NAR*, 34:D354–357, 2006.

Karlsson, L. DM and DO shape the repertoire of peptide-MHC-class-II complexes. *Current Opinion in Immunology*, 17(1):65–70, 2005.

Kast W. Martin, Waal Leo P. de, M. C. J. M. Thymus dictates major histcompatibility complex (MHC) specificity and immune response gene pheenotype of class II MHC-restricted T cells but not of class I MHC-restricted T cells. *J.Exp.Med*, 160:1752–1766, 1984.

Kelley, J., Walter, L., and Trowsdale, J. Comparative genomics of major histocompatibility complexes. *Immunogenetics*, 56(10):683–695, 2005.

Kent, W., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., and Haussler, D. The human genome browser at UCSC. *Genome Biol*, 12(6):996–1006, 2002.

Kerkhoven, R., van Enckevort, F., Boekhorst, J., Molenaar, D., and Siezen, R. Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics*, 20(11):1812–4, 2004.

Kim, J., Chung, H.-J., Park, C. H., Park, W.-Y., and Kim, J. H. ChromoViz: multimodal visualization of gene expression data onto chromosomes using scalable vector graphics. *Bioinformatics*, 20(7):1191–1192, 2004.

Kirov, S., Peng, X., Baker, E., Schmoyer, D., Zhang, B., and Snoddy, J. GeneKeyDB: a lightweight, gene-centric, relational database to support data mining environments. *BMC Bioinformatics*, 6(1):72, 2005.

Kleinnjan, D.-J. and von Heyningen, V. Position effect in human genetic disease. *Human Molecular Genetics*, 7(19):1611–1618, 1998.

Knoch, T. A., Munkel, C., and Langowski, J. New Three-Dimensional Organization ofth Foresight Conference on Molecular Nanotechnology. 1998.

Kreiman, G. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucl. Acids. Res.*, 32(9):2889–2900, 2004.

Kretsovali, A., Agalioti, T., Spilianakis, C., Tzortzakaki, E., Merika, M., and Papamatheakis, J. Involvement of CREB Binding Protein in Expression of Major Histocompatibility Complex Class II Genes via Interaction with the Class II Transactivator. *Mol. Cell. Biol.*, 18(11): 6777–6783, 1998.

Kruglyak, S. and Tang, H. Regulation of adjacent yeast genes. *Trends in Genetics*, 16(1): 109–11, 2000.

Krumlauf, R. Hox genes in vertebrate development. *Cell*, 78:191–201, 1994.

Kutchma, A., Quayum, N., and Jensen, J. GeneSpeed: protein domain organization of the transcriptome. *Nucleic Acid Research*, 35(Database issue):D674–79, 2006.

Laat de Wouter, G. F. Spatial organization of gene expression: the active chromatin hub. *Chromosome Research*, 11:447–459, 2003.

Laemmli, U. Levels of organization of the DNA in eucaryotic chromosomes. *Pharmacological reviews*, 30(4):469–76, 1979.

Landais, D., Matthes, H., Benoist, C., and Mathis, D. A molecular basis for the Ia.2 and Ia.19 antigenic determinants. *Proceedings of the National Acedemy of Sciences of the United States of America*, 82:2930–2934, 1985.

Lanning, D. and Lafuse, W. The mouse p52 subunit of the transcription/DNA repair factor TFIIH is located in the class III region of the H2 complex: cloning and sequence polymorphism. *Immunogenetics*, 49(6):498–504, 1999.

Lawrence, J. G. Selfish operons and speciation by gene transfer. *Trends in Microbiology*, 5(9): 355–359, 1997.

Lee, J. M. and Sonnhammer, E. L. Genomic Gene Clustering Analysis of Pathways in Eukaryotes. *Genome Res.*, 13(5):875–882, 2003.

Lepin, E. J. M., Bastin, J. M., Allan, D. S. J., Roncador, G., Braud, V. M., Mason, D. Y., Merwe, P. A. v. d., McMichael, A. J., Bell, J. I., Powis, S. H., and O'Callaghan, C. A. Functional characterization of HLA-F and binding of HLA-F tetramers to ILT2 and ILT4 receptors. *European Journal of Immunology*, 30(12):3552–3561, 2000.

Lercher, M., Urrutia, A., Pavlicek, A., and Hurst, L. A unification of mosaic structures in the human genome. *Human Molecular Genetics*, 12(19):2411–2415, 2003a.

Lercher, M. J., Blumenthal, T., and Hurst, L. D. Coexpression of Neighboring Genes in Caenorhabditis Elegans Is Mostly Due to Operons and Duplicate Genes. *Genome Res.*, 13(2):238–243, 2003b.

Lercher, M., Urrutia, A., and Hurst, L. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet*, 31(2):180–3, 2002.

Litt, M., Simpson, M., Recillas-Targa, F., Prioleau, M., and Felsenfeld, G. Transitions in histone acetylation reveal boundaries of three separately regulated neighboring loci. *EMBO J*, 20(9):2224–35, 2001.

Liu, Y. and Shaw, S. The human genome: an immuno-centric view of evolutionary strategies. *Trends in Immunology*, 22(5):227–229, 2001.

Loveland, J. VEGA, the genome browser with a difference. *Briefing in bioinformatics*, 6(2): 189–93, 2005.

Lucin, P., Pavic, I., Polic, B., Jonjic, S., and Koszinowski, U. H. Gamma interferon-dependent clearance of cytomegalovirus infection in salivary glands. *Journal of Virology*, 66(4):1977–84, 1992.

Lydyard, P., Whelan, A., and Fanger, M. *Immunology*. Instant Notes. BIOS Scientific Publishers, 2 edition, 2004.

Ma, H., Siegel, A. J., and Berezney, R. Association of Chromosome Territories with the Nuclear Matrix: Disruption of Human Chromosome Territories Correlates with the Release of a Subset of Nuclear Matrix Proteins. *J. Cell Biol.*, 146(3):531–542, 1999.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucl. Acids Res.*, 33(suppl_1):D54–58, 2005.

Marshall, W. Recent advantages in siRNA mediated gene silencing technology, 2005-06-15 2005.

Martin, E. and Hine, R. S., editors. *Dictionary of biology*. Oxford, 4 edition, 2000.

Masternak, K., Peyraud, N., Krawczyk, M., Barras, E., and Reith, W. Chromatin remodeling and extragenic transcription at the MHC class II locus control region. *Nature Immunology*, 4(2):132–137, 2003.

Masternak, K. and Reith, W. Promoter-specific functions of CIITA and the MHC class II enhanceosome in transcriptional activation. *The EMBO Journal*, 21(6):1379–1388, 2002.

May, N. L., Dubaele, S., Santis, L. P. D., Billecocq, A., Bouloy, M., and Egly, J.-M. TFIIH Transcription Factor, a Target for the Rift Valley Hemorrhagic Fever Virus. *Cell*, 116(4): 541–550, 2004.

Mazmanian, S. K., Liu, C. H., Tzianabos, A. O., and Kasper, D. L. An Immunomodulatory Molecule of Symbiotic Bacteria Directs Maturation of the Host Immune System. *Cell*, 122 (1):1–7–18, 2005.

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C.-S., Jan, Y. N., Kenyon, C., Bargmann, C. I., and Li, H. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genetics*, 36(2):197–204, 2004.

McGrath, J. and Wessagowit, V. Human hair abnormalities resulting from inherited desmosome gene mutations. *Keio.J.Med.*, 54(2):72–79, 2005.

McGrath, J. A. Inherited disorders of desmosomes. *Australian Journal of Dermatology*, 46(4): 221–229, 2005.

Meng Kian Tee, Axel A. Thomson, J. B. and Miller, W. L. Sequences Promoting the Transcription of the Human XA Gene Overlapping P450c21A Correctly Predict the Presence of a Novel, Adrenal-Specific, Truncated Form of Tenascin-X. *Genomics*, 28(2):171–78, 1995.

Meraz, M. A., White, J. M., Sheehan, K. C. F., Bach, E. A., Rodig, S. J., Dighe, A. S., Kaplan, D. H., Riley, J. K., Greenlund, A. C., and Campbell, D. Targeted Disruption of the Stat1 Gene in Mice Reveals Unexpected Physiologic Specificity in the JAK-STAT Signaling Pathway. *Cell*, 84(3):431–442, 1996.

Messeguer, X., Escudero, R., Farré, D., Nuñez, O., Martínez, J., and Albà, M. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics*, 18(2):333–334, 2002.

Miller, M. A., Cutter, A. D., Yamamoto, I., Ward, S., and Greenstein, D. Clustered Organization of Reproductive Genes in the C. elegans Genome. *Current Biology*, 14(14):1284–90, 2004.

Moens, C. B. and Selleri, L. Hox cofactors in vertebrate development. *Developmental Biology*, 291(2):193–206, 2006.

Morishita, H. and Yagi, T. Protocadherin family: diversity, structure, and function. *Current Opinion in Cell Biology*, 19(5):584–92, 2007.

Morris, P., Shaman, J., Attaya, M., Amaya, M., Goodman, S., Bergman, C., Monaco, J. J., and Mellins, E. An essential role for HLA-DM in antigen presentation by class II major histocompatibility molecules. *Nature*, 368:551 – 554, 1994a.

Morris, P., Shaman, J., Attaya, M., Amaya, M., Goodman, S., Bergman, C., Monaco, J. J., and Mellins, E. An essential role for HLA-DM in antigen presentation by class II major histocompatibility molecules. *Nature*, 368:551–4, 1994b.

Motoo, K. *Population genetics, molecular evolution, and the neutral theory: Selected papers*. University of Chicago Press, 1994.

Muhlethaler-Mottet, A., Otten, L. A., Steimle, V., and Mach, B. Expression of MHC class II molecules in different cellular and functional compartments is controlled by differential usage of multiple promoters of the transactivator CIITA. *The EMBO Journal*, 16(10):2851–2860, 1997.

Munkel, C. and Langowski, J. Chromosome structure predicted by a polymer model. *Physical Review*, 57(5):5888–5896, 1998.

Munkel, C., Eils, R., Dietzel, S., Zink, D., Mehring, C., Wedemann, G., Cremer, T., and Langowski, J. Compartmentalization of Interphase Chromosomes Observed in Simulation and Experiment. *Journal of Molecular Biology*, 285(3):1053–1065, 1999.

Murray, B. W., Sultmann, H., and Klein, J. Analysis of a 26-kb Region Linked to the Mhc in Zebrafish: Genomic Organization of the Proteasome Component beta/Transporter Associated with Antigen Processing-2 Gene Cluster and Identification of Five New Proteasome beta Subunit Genes. *J Immunol*, 163(5):2657–2666, 1999.

Nicklin, J., Graeme-Cook, K., and R, K. *Microbiology*. BIOS, 2nd edition, 2002.

Niehrs, C. and Pollet, N. Synexpression groups in eukaryotes. *Nature*, 402(6761):483–7, 1999.

Ohta, Y., McKinney, E. C., Criscitiello, M. F., and Flajnik, M. F. Proteasome, Transporter Associated with Antigen Processing, and Class I Genes in the Nurse Shark Ginglymostoma cirratum: Evidence for a Stable Class I Region and MHC Haplotype Lineages. *J Immunol*, 168(2):771–781, 2002.

Okada, T. A. and Comings, D. E. Higher order structure of chromosomes. *Chromosoma*, 72 (1):1–14, 1979.

Okuda, S., Katayama, T., Kawashima, S., Goto, S., and Kanehisa, M. ODB: a database of operons accumulating known operons across multiple genomes. *NAR*, 34(Database issue): D358–62, 2006.

Oldham, W. M. and Hamm, H. E. Heterotrimeric G protein activation by G-protein-coupled receptors. *Nature Reviews Molecular Cell Biology*, 9(1):60–71, 2008.

Oliver, B., Parisi, M., and Clark, D. Gene expression neighbourhoods. *Journal of Biology*, 1 (4):4.1–4.3, 2002.

Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, E. J., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., Bhattacharyya, A., Burd, H., Gardner, W., Hanke, P., Kapatral, V., Mikhailova, N., Vasieva, O., Osterman, A., Vonstein, V., Fonstein, M., Ivanova, N., and Kyrpides, N. The ERGO genome analysis and discovery system. *NAR*, 31 (1):164–71, 2003.

Palstra, R., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and de Laat, W. The beta-globin nuclear compartment in development and erythroid differentiation. *Nature Genetics*, 35(2):190–194, 2003.

Panjwani, N., Akbari, O., Garcia, S., Brazil, M., and Stockinger, B. The HSC73 Molecular Chaperone: Involvement in MHC Class II Antigen Presentation. *J Immunol*, 163(4):1936–1942, 1999.

Pearson Education, I. Lac operon, 2005.

Pedersen, A., Jensen, L., Brunak, S., Staerfeldt, H., and Ussery, D. A DNA structural atlas for Escherichia coli. *J Mol Biol*, 299(4):907–30, 2000.

Perraudeau, M., Taylor, P. R., Stauss, H. J., Lindstedt, R., Bygrave, A. E., Pappin, D. J. C., Ellmerich, S., Whitten, A., Rahman, D., Canas, B., Walport, M. J., Botto, M., and Altmann, D. M. Altered major histocompatibility complex class II peptide loading in H2-O-deficient mice. *European Journal of Immunology*, 30:2871–2880, 2000.

Pollock, J. L., Presti, R. M., Paetzold, S., and IVth, H. W. V. Latent Murine Cytomegalovirus Infection in Macrophages. *Virology*, 227(1):168–79, 1997.

Pond, C. M. Adipose tissue and the immune system. *Prostaglandins, Leuokotriences and essential fatty acids*, 73(1):17–30, 2005.

Pontius, J., Wagner, L., and Schuler, G. UniGene: a unified view of the transcriptome, 2003. URL `http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/handbook/ch21d1.pdf`.

Popkin, D. L. and Virgin, H. W. I. Murine Cytomegalovirus Infection Inhibits Tumor Necrosis Factor Alpha Responses in Primary Macrophages. *J. Virol.*, 77(18):10125–10130, 2003.

Purves, W. K., Sadava, D., Orians, G. H., and Heller, H. C. *Life - the science of biology*. W.H. Freeman, 6 edition, 2001.

Rao, M. S., Kumari, G., Balasundaram, D., Sankaranarayanan, R., and Mahalingam, S. A Novel Lysine-rich Domain and GTP Binding Motifs Regulate the Nucleolar Retention of Human Guanine Nucleotide Binding Protein, GNL3L. *Journal of Molecular Biology*, 364 (4):637–54, 2006.

Richmond, T. J. and Davey, C. A. The structure of DNA in the nucleosome core. *Naute*, 423: 145–150, 2003.

Roy, P., Stuart, J., Lund, J., and Kim, S. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature*, 418(6901):975–9, 2002.

Rudensky, A. Y., Preston-Hurlburt, P., Hong, S.-C., Barlow, A., and Jr, C. A. J. Sequence analysis of peptides bound to MHC class II molecules. *Nature*, 353:622–27, 1991.

Sachs, R., Engh, G., Trask, B., Yokota, H., and Hearst, J. A Random-Walk/Giant-Loop Model for Interphase Chromosomes. *PNAS*, 92(7):2710–2714, 1995.

Sagerström, C. G. pbX Marks the Spot. *Developmental Cell*, 6(6):737–8, 2004.

Saitoh, Y. and Laemmli, U. Metaphase chromosome structure: bands arise from a differential folding path of the highly AT-rich scaffold. *Cell*, 76(4):609–622, 1994.

Sakoda, Y., Hashimoto, D., Asakura, S., Takeuchi, K., Harada, M., Tanimoto, M., and Teshima, T. Donor-derived thymic-dependent T cells cause chronic graft-versus-host disease. *Blood*, 109(4):1756–64, 2007.

Sato, N. and Ehira, S. GenoMap, a circular genome data viewer. *Bioinformatics*, 19(12): 1583–4, 2003.

Schmitt-Egenolf, M., Windemuth, C., Hennies, H. C., Albis-Camps, M., von Engelhardt, B., Wienker, T., Reis, A., Traupe, H., and Blasczyk, R. Comparative association analysis reveals that corneodesmosin is more closely associated with psoriasis than HLA-B in German families. *Tissue Antigens*, 57(5):440–6, 2001.

Schroder, K., Hertzog, P. J., Ravasi, T., and Hume, D. A. Interferon-gamma: an overview of signals, mechanisms and functions. *J Leukoc Biol*, 75(2):163–189, 2004.

Shiina, T., Inoko, H., and Kulski, J. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens*, 64(6):631–49, 2004.

Shortman, K. and Heath, W. R. Immunity or tolerance? That is the question for dendritic cells. *Nature Immunology*, 2(11):988 – 989, 2001.

Siemasko, K. and Clark, M. R. The control and facilitation of MHC class II antigen processing by the BCR. *Current Opinion in Immunology*, 13(1):32–36, 2001.

Singer, G. A. C., Lloyd, A. T., Huminiecki, L. B., and Wolfe, K. H. Clusters of Co-expressed Genes in Mammalian Genomes Are Conserved by Natural Selection. *Mol Biol Evol*, 22(3): 767–775, 2005.

Snoek, M., van Kooij, M., de Groot, K., and van Vugt, H. The mouse p52 subunit of the transcription/DNA repair factor TFIIH is not located in the class III region of the H2 complex, but resides next to a G7a/Bat6 homologue in the telomeric part of the major histocompatibility complex. *Immunogenetics*, 51(2):164–7, 2000.

Speek, M., Barry, F., and Miller, W. Alternate promoters and alternate splicing of human tenascin-X, a gene with 5' and 3' ends buried in other genes. *Hum. Mol. Genet.*, 5(11): 1749–1758, 1996.

Spellman, P. T. and Rubin, G. M. Evidence for large domains of similarily expressed genes in the Drosophila genome. *Journal of Biology*, 1(1):5.1–5.8, 2002.

Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell*, 73(3):521–32, 1993.

Spilianakis, C. and Flavell, R. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nature Immunology*, 5(10):1017–1027, 2004.

Spilianakis, C., Kretsovali, A., Agalioti, T., Makatounakis, T., Thanos, D., and Papamatheakis, J. CIITA regulates transcription onset viaSer5-phosphorylation of RNA Pol II. *The EMBO Journal*, 22(19):5125–5136, 2003.

Stein, L. D. Integrating biological databases. *Nature reviews — Genetics*, 4(5):337–345, 2003.

Stoddart, C. A., Cardin, R. D., Boname, J. M., Manning, W. C., Abenes, G. B., and Mocarski, E. S. Peripheral blood mononuclear phagocytes mediate dissemination of murine cytomegalovirus. *J. Virol.*, 68(10):6243–6253, 1994.

Sumner, A. T. *Chromosomes: organization and function*. Blackwell, 2003.

Sánchez, F., Holm, S. J., Mallbris, L., O'Brien, K. P., and Ståhle, M. STG does not associate with psoriasis in the Swedish population. *Experimental Dermatology*, 13(7):413–8, 2004.

Takada, T., Kumanovics, A., Amadou, C., Yoshino, M., Jones, E., Athanasiou, M., Evans, G., and Fischer, L. K. Species-specific class I gene expansions formed the telomeric 1 mb of the mouse major histocompatibility complex. *Genome Res.*, 13(4):589–600, 2003.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401(6756):921–3, 1999.

Ting, J. P.-Y. and Trowsdale, J. Genetic Control of MHC Class II Expression. *Cell*, 109(2, Supplement 1):S21–S33, 2002.

Tolhuis, B., Palstra, R., Splinter, E., Grosveld, F., and de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, 10:1453–1465, 2002.

Tripodis, N., Mason, R., Humphray, S. J., Davies, A. F., Herberg, J. A., Trowsdale, J., Nizetic, D., Senger, G., and Ragoussis, J. Physical Map of Human 6p21.2-6p21.3: Region Flanking the Centromeric End of the Major Histocompatibility Complex. *Genome Res.*, 8(6):631–643, 1998.

Trowsdale, J. The gentle art of gene arrangement: the meaning of gene clusters. *Genome Biology*, 3(3):comment2002.1 – comment2002.5, 2002.

Turley, S. J., Inaba, K., Garrett, W. S., Ebersold, M., Unternaehrer, J., Steinman, R. M., and Mellman, I. Transport of Peptide-MHC Class II Complexes in Developing Dendritic Cells. *Science*, 288(5465):522–527, 2000.

Ucker, D. and Yamamoto, K. Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates. *J Biol Chem*, 259(12): 7416–20, 1984.

University of Manitoba. III. Chromatin structure, 2005.

Verdugo, R. and Medrano, J. Comparison of gene coverage of mouse oligonucleotide microarray platforms. *BMC Genomics*, 7(1):58, 2006.

Vernikos, G., Gkogkas, C., Promponas, V., and Hamodrakas, S. GeneViTo: visualizing gene-product functional and structural features in genomic datasets. *BMC Bioinformatics*, 4(53), 2003.

Versteeg, G. A., Slobodskaya, O., and Spaan, W. J. M. Transcriptional profiling of acute cytopathic murine hepatitis virus infection in fibroblast-like cells. *J Gen Virol*, 87(7):1961–1975, 2006.

Versteeg, R., van Schaik, B. D., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J., and van Kampen, A. H. The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. *Genome Res.*, 13(9):1998–2004, 2003.

Vogel, J. H., Heydebreck, v. A., Purmann, A., and Sperling, S. Chromosomal clustering of a human transcriptome reveals regulatory background. *BMC Bioinformatics*, 6(1):230, 2005.

Wagle, N. M., Faassen, A. E., Kim, J. H., and Pierce, S. K. Regulation of B Cell Receptor-Mediated MHC Class II Antigen Processing by FcgammaRIIB1. *J Immunol*, 162(5):2732–2740, 1999.

Waldburger, J.-M., Suter, T., Fontana, A., Acha-Orbea, H., and Reith, W. Selective Abrogation of Major Histocompatibility Complex Class II Expression on Extrahematopoietic Cells in Mice Lacking Promoter IV of the Class II Transactivator Gene. *J. Exp. Med.*, 194(4):393–406, 2001.

Walter, W., Lingnau, K., Schmitt, E., Loos, M., and Maeurer, M. MHC class II antigen presentation pathway in murine tumours: tumour evasion from immunosurveillance? *Br J Cancer*, 83(9):1192–1201, 2000.

Walter, W., Scheuer, C., Loos, M., Reichert, T. E., and Maeurer, M. J. H2-Mbeta 1 and H2-Mbeta 2 Heterodimers Equally Promote CLIP Removal in I-Aq Molecules from Autoimmune-prone DBA/1 Mice. *J. Biol. Chem.*, 276(14):11086–11091, 2001.

Wang, P. J., McCarrey, J. R., Yang, F., and Page, D. C. An abundance of X-linked genes expressed in spermatogonia. *Nature Genetics*, 27(4):422–6, 2001.

Weitzman, J. Transcriptional territories in the genome. *Journal of Biology*, 1(2):2:1–2:5, 2002.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.*, 35(suppl_1):D5–12, 2007.

Williams, A., Peh, C., and Elliott, T. The cell biology of MHC class I antigen presentation. *Tissue Antigens*, 59, 2002.

Williams, C., Granner, D., Magnuson, M., and Chalkley, R. Cell specific differences in DNase I hypersensitivity between the two promoters of the rat glucokinase gene. *Biochemical and biophysical research communications*, 215(1):272–9, 1995.

Williams, E. and Hurst, L. Clustering of Tissue-Specific Genes Underlies Much of the Similarity in Rates of Protein Evolution of Linked Genes. *J. of Molecular Evolution*, 54(4):511–8, 2002.

Williams, E. J. and Bowles, D. J. Coexpression of Neighboring Genes in the Genome of Arabidopsis thaliana. *Genome Res.*, 14(6):1060–1067, 2004.

Wright, K., White, L., Kelly, A., Beck, S., Trowsdale, J., and Ting, J. Coordinate regulation of the human TAP1 and LMP2 genes from a shared bidirectional promoter. *J. Exp. Med.*, 181 (4):1459–1471, 1995.

Wu, Q. and Maniatis, T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell*, 97(6):779–90, 1999.

Yamashita, T., Honda, M., Takatori, H., Nishino, R., Hoshino, N., and Kaneko, S. Genome-wide transcriptome mapping analysis identifies organ-specific gene expression patterns along human chromosomes. *Genomics*, 84(5):867–75, 2004.

Yang, Z., Shen, L., Dangel, A., Wu, L.-C., and Yu, C. Four Ubiquitously Expressed Genes, RD SKI2W DOM3Z RP1, Are Present between Complement Component Genes Factor B and C4 in the Class III Region of the HLA. *Genomics*, 53(3):338–347, 1998.

Yang, Z. and Yu, C. Y. Organizations and Gene Duplications of the Human and Mouse MHC Complement Gene Clusters. *Experimental and Clinical Immunogenetics*, 17(1), 2000.

Ye, J., McGinnis, S., and Madden, T. L. BLAST: improvements for better sequence analysis. *NAR*, 34(Web server):W6–W9, 2006.

Yokota, H., van den Engh, G., Hearst, J., Sachs, R., and Trask, B. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. *J Cell Biol*, 130(6):1239–49, 1995.

Yu, X., Burgess, S. C., Ge, H., Wong, K. K., Nassem, R. H., Garry, D. J., Sherry, A. D., Malloy, C. R., Berger, J. P., and Li, C. Inhibition of cardiac lipoprotein utilization by transgenic overexpression of Angptl4 in the heart. *PNAS*, 102(5):1767–1772, 2005.

Yuhki, N., Beck, T., Stephens, R., Nishigaki, Y., Newmann, K., and O'Brien, S. Comparative genome organization of human, murine, and feline MHC class II region. *Genome Res.*, 13 (6A):1169–1179, 2003.

Zakany, J., Kmita, M., Alarcon, P., de la Pompa, J.-L., and Duboule, D. Localized and Transient Transcription of Hox Genes Suggests a Link between Patterning and the Segmentation Clock. *Cell*, 106(2):207–217, 2001.

Zika, E. and Ting, J. P.-Y. Epigenetic control of MHC-II: interplay between CIITA and histone-modifying enzymes. *Current Opinion in Immunology*, 17(1):58–64, 2005.

Zorio, D. A. R., Cheng, N. N., Blumenthal, T., and Spieth, J. Operons as a common form of chromosomal organization in C. elegans. *Nature*, 372(6503):270–2, 2002.

# Bibliography


5M Enterprises Ltd. Porcine reproductive and respiratory syndrom (PRRS) URL www.thepigsite.com/pighealth/article/142/porcine-reproductive-and-respiratory-syndrome-prrs.htm.

Affymetrix Inc. Statistical algorithms reference guide, 2001. URL www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf.

Affymetrix Inc. Statistical algorithms description document, 2002. URL www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

The apache ant project. Welcome, 2007 URL http://ant.apache.org/.

The Apache Software Foundation. Commons-Math: The Apache Commons Mathematics Library, 2008. URL http://jakarta.apache.org/commons/math/index.html.

COPE. CD26, 2008. URL http://www/copewithcytokines.de/cope.cgi?key=CD26.

COPE. Cytokines, 2008. URL http://www/copewithcytokines.de/cope.cgi?key=Cytokines.

COPE. Intercrines, 2008. URL http://www/copewithcytokines.de/cope.cgi?key=Intercrines.

NCBI. CCDS Database URL http://www.ncbi.nlm.nih.gov/projects/CCDS/.

EMBL. Q59FI1 - Vars2l. URL http://marvester.embl.de/marvester/Q59F/Q59FI1.htm.

EMBL. Q8K2X9 - Vars2l. URL http://marvester.embl.de/marvester/Q8K2/Q8K2X9.htm.

EMBL. Q8BIN9 - Vars2l. URL http://marvester.embl.de/marvester/Q8BI/Q8BIN9.htm.

EMBL. Q5SQ95 - Vars2l, a. URL http://marvester.embl.de/marvester/Q5SQ/Q5SQ95.htm.

Engelhardt, P. Manually isolated polytene chromosomes and nuclear envelopes, 1998. URL http://www.lce.hut.fi/~engelhar/Doc/Diss-DiscI.html.

Fortin, I. Domaines proteques du complexe histone aceyltransfease NuA4 implique dans la transciption et le maintien de l'interiteu geee, 2005. URL http://www.theses.ulaval.ca/2005/23265/ch01.html.

Universitesbiblioteket Karolinska institutet. MeSH Tree Location(s) for GTP-binding Proteins. URL http://mesh.kib.ki.se/swemesh/show.swemeshtree.cfm?Mesh_No=D08.811.277.040.330.300\&tool=karo.

Functional Glycomics Gateway. URL http://web.mit.edu/glycomics/consortium/resources/GLYCO_v1GeneList.pdf.

Ge HealthCare. URL http://www.gehealthcare.com/usen/microarrays/docs/genelists/Mouse_HKG_List.txt.

GeneSpring Analysis Platform. GeneSpring Analysis Platform. URL http://www.chem.agilent.com/Scripts/Generic.ASP?lPage=35082\&indcol=Y\&prodcol=Y

GenMapp. URL http://www.genmapp.org.

GeneSpeed. GeneSpeed - Establishing a default expectancy score cutoff. URL http://genespeed.ccf.org/BG_Escore.htm.

Genomics Institute of the Novartis Research Foundation. GNF Genome Informatics Applications & Datasets. URL http://wombat.gnf.org/index.html.

Genomics Institute of the Novartis Research Foundation. Search. URL http://symatlas.gnf.org/SymAtlas.

Ingenuity Systems. Ingenuity Systems. URL http://www.ingenuity.com/.

The Jackson Laboratory. Col1a1. URL http://www.informatics.jax.org.

JUnit.org JUnit.org Resources for Test Driven Development, 2007 URL http://www.junit.org/index.htm.

Jonathan Kerr. Summary of research evidence on ME/CFS and oral presentation compiled by Brame. URL http://64.233.183.104/search?q=cache:t54ABAzglX4J:www.25megroup.org/Campaigning/Gibson\%2520Inquiry\%2520Information/BRAMEEvidenceGibsonInquiryonME.doc+Flot1+cmv+-\%22splice+variants+in+a+CMV+expression+vector\%22&hl=sv&ct=clnk&cd=5.

McClean. Eukaryotic chromosome structure. URL http://www.ndsu.nodak.edu/instruct/mcclean/plsc431/eukarychrom/eukaryo3.htm.

Pierce. Signal Transduction and small GTPase. URL http://www.piercenet.com/Proteomics/browse.cfm?fldID=97D80799-CF9B-4D0D-81EA-5CA7ED33361B.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. GeneCards: encyclopedia for genes, proteins and diseases, 1997. URL http://www.genecards.org/.

The SAX project. About SAX, 2004. URL http://www.saxproject.org/and/orhttp://sourceforge.net/projects/sax/.

Scopes, G. Introduction to the Affymetrix Control Probe Sets. URL http://www.abdn.ac.uk/ims/facilities/microarray/documents/chipcontrols.doc.

Spotfire. Spotfire. URL http://www.spotfire.com.

SuperArray Bioscience Corporation. URL http://www.superarray.com/rt_pcr_product/HTML/PM-HK1B.html.

Twyman, Richard. Model organism: The mouse, 2002. URL `http://genome.wellcome.ac.uk/doc\_wtd020804.html`.

Z-lab. Supplement 2: 451 housekeeping genes from 19 normal human tissue types. URL `http://zlab.bu.edu/HugeIndex/PaperInfo/Supplement\_2-451\_hk\_genes.html`.