

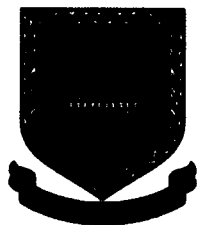
**DEVELOPMENT AND IMPLEMENTATION OF  
ONTOLOGY-BASED SYSTEMS FOR  
MAMMALIAN GENE EXPRESSION  
PROFILING**

**ADÉLE KRUGER**

Thesis presented in fulfilment of the requirements for the Degree  
of *Doctor Philosophiae* at the South African National  
Bioinformatics Institute, Faculty of Natural Sciences, University  
of the Western Cape

August 2009

Advisor: Prof. Winston Hide



**UNIVERSITY of the  
WESTERN CAPE**



**SANBI**

## Keywords

ontology

expression vocabulary

gene expression

cross-species

comparison

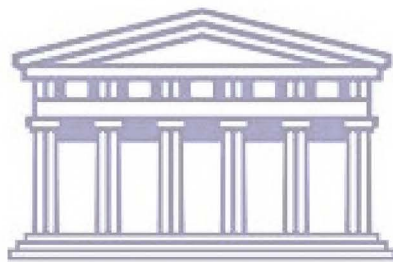
human development

mouse development

cancer/testis

transcription factor

gene regulation



UNIVERSITY *of the*  
WESTERN CAPE



## Abstract

The use of ontologies in the mapping of gene expression events provides an effective and comparable method to determine the expression profile of an entire genome across a large collection of experiments derived from different expression sources. In this dissertation I describe the development of the developmental human and mouse eVOC ontologies and demonstrate the ontologies by identifying genes showing a bias for developmental brain expression in human and mouse, identifying transcription factor complexes, and exploring the mouse orthologs of human cancer/testis genes.

Model organisms represent an important resource for understanding the fundamental aspects of mammalian biology. Mapping of biological phenomena between model organisms is complex and if it is to be meaningful, a simplified representation can be a powerful means for comparison.



The implementation of the ontologies has been illustrated here in two ways. Firstly, the ontologies have been used to illustrate methods to determine clusters of genes showing tissue-restricted expression in humans. The identification of tissue-restricted genes within an organism serves as an indication of the fine-tuning in the regulation of gene expression in a given tissue. Secondly, due to the differences in human and mouse gene expression on a temporal and spatial level, the ontologies were used to identify mouse orthologs of human cancer/testis genes showing cancer/testis characteristics. With the use of model systems such as mouse in the development of gene-targeted drugs in the treatment of disease, it is

important to establish that the expression characteristics and profiles of a drug target in the model system is representative of the characteristics of the target in the system for which it is intended.



UNIVERSITY *of the*  
WESTERN CAPE

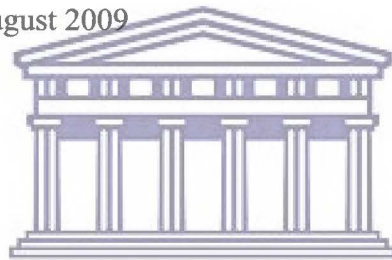
## Declaration

I declare that “Development and implementation of ontology-based systems for mammalian gene expression profiling” is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.



Adèle Kruger

August 2009



UNIVERSITY *of the*  
WESTERN CAPE

## Acknowledgements

I would like to thank my supervisor and mentor, Professor Winston Hide, for his guidance and support throughout this epic journey. His encouragement and work ethic has provided me with the opportunity to attend and present at international conferences, allowing me to establish collaborations with world-renowned scientists in the field. It has been a pleasure and a privilege to work with someone who is so invested in the success of his students.

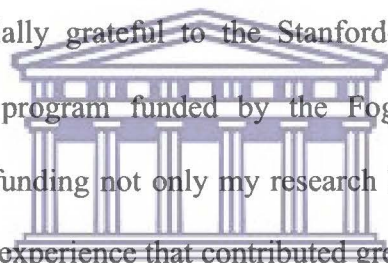
I would also like to thank Oliver Hofmann for his endless patience and support in the research we have conducted together. His insight and advice has played a key role in the completion of this thesis.

To Christopher Maher for his contributions to the cancer/testis research and support in understanding Perl, thank you. Many thanks to my friends and colleagues at SANBI especially Ferial Mullins, Maryam Salie, Judith Jansen, Patricia Josias and Dale Gibbs, who have provided administrative, technical and moral support. Also, a special 'thank you' to Betty Cheng and Russ Altman for their guidance and advice with respect to making a career out of bioinformatics.

In bioinformatics, collaborators play a pivotal role in all research endeavors. I would like to express my gratitude to all those who have contributed to the success of the research presented here. I would especially like to thank Yoshihide Hayashizaki, Piero Carninci and Harukazu Suzuki for their mammoth efforts in the establishment and success of the FANTOM consortium, without which this research would not have been possible. I would also like to extend gratitude and

thanks to all the members of the FANTOM consortium and the RIKEN institute in Japan for their involvement in the provision and analysis of data used in producing this thesis. I thank Duncan Davidson for his feedback during the initial phase of ontology development, Lloyd Old and Andrew Simpson from the Ludwig Institute for Cancer Research (LICR), as well as the members of the Melanoma Research Alliance (MRA) for their contributions to the cancer/testis research.

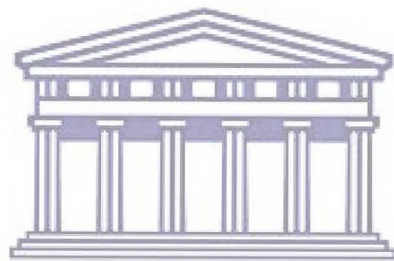
Without funding we would not be able to conduct exciting and cutting-edge research. I would therefore like to take this opportunity to acknowledge the funding agencies and projects that contributed to my research and bioinformatics education. I am especially grateful to the Stanford-South Africa Biomedical Informatics (SSABMI) program funded by the Fogarty International Center (Grant TW-03-008) for funding not only my research but also a research visit to Stanford – an invaluable experience that contributed greatly to my education. The National Bioinformatics Network (NBN) of South Africa and Alternate Transcript Diversity group EU FP programme also funded this research directly. The Medical Research Council (MRC) of South Africa, World Health Organisation (WHO), Oppenheimer Trust and Atlantic Philanthropies funded the projects represented in this thesis and provided financial support for travel. I would also like to acknowledge the funding agencies of our collaborators, RIKEN, FANTOM, LICR and MRA, because without them we would not have any research collaborators or data with which to conduct science.



UNIVERSITY of the  
WESTERN CAPE

On a personal note, I would like to thank my family and friends. Special thanks go to my cousins Johan and Margaret, as well as my aunts Doreen and Mariette for their moral and financial support. I would also like to thank my sister, Marlise, and her husband, Louis-Jacques, for their continuous support. To my best friend, Anrinette, thank you for always believing in me, even when at times I did not. Also, I will be eternally grateful to my mom for understanding the importance of an education and making the sacrifices she did in order for me to be where I am today.

Lastly, I would like to thank my Heavenly Father for providing me with the strength and inspiration necessary to complete such a journey as this.



UNIVERSITY *of the*  
WESTERN CAPE



## Publications arising from this thesis

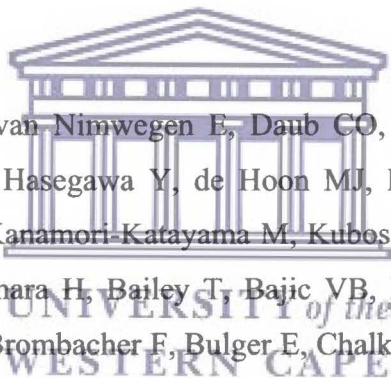
Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y. **The**

**transcriptional landscape of the mammalian genome.** *Science*. 2005. 309(5740):1559-1563.

Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, Yang L, Hofmann O, **Kruger A**, Hide W, Kai C, Kawai J, Hume DA, Carninci P, Hayashizaki Y. **Mice and men: their promoter properties.** *PLoS Genet*. 2006. 2(4):e54.

**Kruger A**, Hofmann O, Carninci P, Hayashizaki Y, Hide W. **Simplified ontologies allowing comparison of developmental mammalian gene expression.** *Genome Biol*. 2007. 8(10):R229.

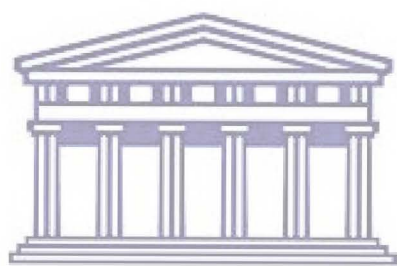
Hofmann O, Caballero OL, Stevenson BJ, Chen YT, Cohen T, Chua R, Maher CA, Panji S, Schaefer U, **Kruger A**, Lehvaslaiho M, Carninci P, Hayashizaki Y, Jongeneel CV, Simpson AJ, Old LJ, Hide W. **Genome-wide analysis of cancer/testis gene expression.** *Proc Natl Acad Sci U S A*. 2008. 105(51):20422-20427.



Suzuki H, Forrest AR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, de Hoon MJ, Katayama S, Schroder K, Carninci P, Tomaru Y, Kanamori-Katayama M, Kubosaki A, Akalin A, Ando Y, Arner E, Asada M, Asahara H, Bailey T, Bajic VB, Bauer D, Beckhouse AG, Bertin N, Bjorkegren J, Brombacher F, Bulger E, Chalk AM, Chiba J, Cloonan N, Dawe A, Dostie J, Engstrom PG, Essack M, Faulkner GJ, Fink JL, Fredman D, Fujimori K, Furuno M, Gojobori T, Gough J, Grimmond SM, Gustafsson M, Hashimoto M, Hashimoto T, Hatakeyama M, Heinzl S, Hide W, Hofmann O, Hornquist M, Huminiecki L, Ikeo K, Imamoto N, Inoue S, Inoue Y, Ishihara R, Iwayanagi T, Jacobsen A, Kaur M, Kawaji H, Kerr MC, Kimura R, Kimura S, Kimura Y, Kitano H, Koga H, Kojima T, Kondo S, Konno T, Krogh A, **Kruger A**, Kumar A, Lenhard B, Lennartsson A, Lindow M, Lizio M, Macpherson C, Maeda N, Maher CA, Maqungo M, Mar J, Matigian NA, Matsuda H, Mattick JS, Meier S, Miyamoto S, Miyamoto-Sato E, Nakabayashi K, Nakachi Y, Nakano M, Nygaard S, Okayama T, Okazaki Y, Okuda-Yabukami H, Orlando V, Otomo J, Pachkov M, Petrovsky N, Plessy C, Quackenbush J, Radovanovic A, Rehli M, Saito R, Sandelin A, Schmeier S, Schonbach C, Schwartz AS, Semple CA, Sera



M, Severin J, Shirahige K, Simons C, St Laurent G, Suzuki M, Suzuki T, Sweet MJ, Taft RJ, Takeda S, Takenaka Y, Tan K, Taylor MS, Teasdale RD, Tegner J, Teichmann S, Valen E, Wahlestedt C, Waki K, Waterhouse A, Wells CA, Winther O, Wu L, Yamaguchi K, Yanagawa H, Yasuda J, Zavolan M, Hume DA, Arakawa T, Fukuda S, Imamura K, Kai C, Kaiho A, Kawashima T, Kawazu C, Kitazume Y, Kojima M, Miura H, Murakami K, Murata M, Ninomiya N, Nishiyori H, Noma S, Ogawa C, Sano T, Simon C, Tagami M, Takahashi Y, Kawai J, Hayashizaki Y. **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nat Genet.* 2009. 41(5):553-562.



UNIVERSITY *of the*  
WESTERN CAPE

# Table of Contents


Keywords	ii
Abstract	iii
Declaration	v
Acknowledgements	vi
Publications arising from this thesis	ix
Table of Contents	xii
List of Figures	xiii
List of Tables	xiv
List of Appendices	xvi
Abbreviations	xviii
Preface	xx
Chapter 1: Simplified ontologies allowing comparison of developmental mammalian gene expression	1
Chapter 2: Expression profiling reveals tissue-restricted transcription factor complexes	32
Chapter 3: Mouse gene expression analysis of cancer/testis orthologs restricts candidates for cancer therapy	55
Conclusions	70
Afterword	75
References	79

# List of Figures

## Chapter 1

- Figure 1: Venn diagram illustrating the integration of mouse and human ontologies represented by the eVOC system. 15
- Figure 2: Screenshot of the Mouse Development ontology, visualized in COBrA. 17
- Figure 3: Screenshot of the individual Theiler Stage 13 ontology, visualized in COBrA. 18
- Figure 4: Diagram illustrating the sets of genes analyzed for developmental brain expression bias. 25

## Chapter 2

- 
- Figure 1a: Illustration of genes clustering together based on correlated co-expression. 46
- Figure 1b: Illustration of genes clustering together based on correlated co-expression. 47

## Chapter 3

- Figure 1: Flow-diagram representing the categorisation of mouse genes into cancer/testis categories. 63
- Figure 2: Visualisation of the gene expression profile of 63 mouse orthologs. 66

# List of Tables

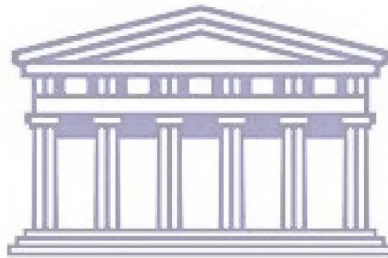
## Preface

Table 1: A list of ontologies available from the Open Biomedical Ontologies (OBO) Foundry.	xxii
--	------

## Chapter 1

Table 1: Statistics of the individual developmental eVOC ontologies, representing the alignment between human and mouse stages.	19
---	----

Table 2: Genes showing developmental expression bias in human and mouse brain.	26
--	----



## Chapter 2

Table 1: A list of the 145 genes expressed in less than 25% of all tissues.	42
---	----

Table 2a: The top five physiological system development and functions over-represented by genes showing restricted expression.	49
--	----

Table 2b: The top five diseases and disorders associated with the genes showing restricted expression in less than 25% of all tissues.	51
--	----

Table 3: A list of canonical pathways over-represented by genes showing restricted expression in less than 25% of all tissues.	52
--	----

## Chapter 3

Table 1: Classification of categories for cancer/testis genes. 64

Table 2: Gene identifiers and symbols of mouse genes showing testis-restricted, testis/brain-restricted or testis-selective expression, along with their human orthologs. 67



UNIVERSITY *of the*  
WESTERN CAPE

# List of Appendices

## Chapter 1

- Appendix I: Transcriptional landscape of the mammalian genome,  
*Science*. 2005. 309(5740):1559-1563. 84
- Appendix II: Mice and men: their promoter properties. *PLoS Genet*.  
2006. 2(4):e54. 89
- Appendix III: Correlation coefficients of genes showing biased  
expression for the developmental brain in human and mouse 102
- Appendix IV: Expression profile of genes showing biased expression  
for the developmental brain in human and mouse 106
- Appendix V: The individual mouse developmental ontologies 114
- Appendix VI: The merged mouse developmental ontologies 154



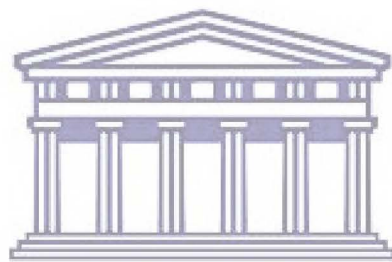
UNIVERSITY of the  
WESTERN CAPE

## Chapter 2

- Appendix VIIa: The transcriptional network that controls growth arrest  
and differentiation in a human myeloid leukemia cell line.  
*Nat Genet*. 2009. 41(5):553-562. 161
- Appendix VIIb: Clusters of genes from Illumina microarray expression  
experiment with early, mid and late response characteristics 171
- Appendix VIII: Expression profile of transcription factors showing  
tissue restriction 176

## Chapter 3

- Appendix IX: Genome-wide analysis of cancer/testis gene expression.  
*Proc Natl Acad Sci U S A.* 2008. 105(51):20422-20427. 183
- Appendix X: Manual curation steps applied in filtering the expression  
array generated for the investigation of 63 potential mouse  
cancer/testis genes 189
- Appendix XI: Expression profile of mouse orthologs of human  
cancer/testis genes 191



UNIVERSITY *of the*  
WESTERN CAPE

## Abbreviations

CAGE – Cap Analysis of Gene Expression

CGAP – Cancer Genome Anatomy Project

CT – cancer/testis

DAG – Directed Acyclic Graph

EMAP – Edinburgh Mouse Atlas Project

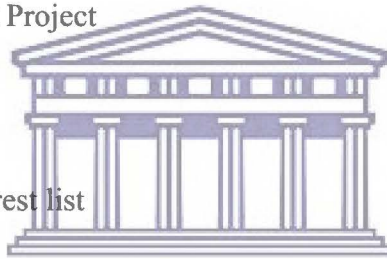
EST – Expressed Sequence Tag

FMA – Foundational Model of Anatomy

GNP – Genome Network Project

GO – Gene Ontology

GOI-list – Genes Of Interest list



HUMAT – Edinburgh Human Developmental Anatomy

LPS – lipopolysaccharide

MA – Adult Mouse Anatomy

MGED – Microarray Gene Expression Data Society

MGI – Mouse Genome Informatics

MPSS – Massively Parallel Signature Sequencing

NCBI – National Center for Biotechnology Information

OBO – Open Biomedical Ontologies



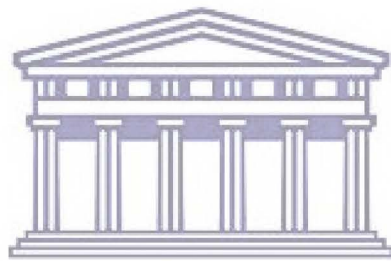
OMIM – Online Mendelian Inheritance in Man

PMA – Phorbol Myristate Acetate

SAEL – SOFG Anatomy Entry List

SAGE – Serial Analysis of Gene Expression

TSS – Transcription Start Site



UNIVERSITY *of the*  
WESTERN CAPE

## Preface

In the post-genomic era, much of the focus of research has shifted from identifying each gene in the human genome, to creating a catalogue of genes listing their corresponding function, regulatory potential, expression profile and disease involvement.

Each cell in an organism contains a complete copy of its genome, thereby providing the expression potential of the organism. Since cells do not simultaneously express all genes in the genome, it is important to determine the location and timing of each gene expression event. This expression profiling can lead to the identification of genes biased in their expression for the developmental program or diseases such as cancer. The identification of genes whose expression is biased for tumorigenic tissues provides the context for the development of drugs or vaccines in the treatment of cancer. The significance of this knowledge is also evident when comparing two species whose genomes show considerable overlap. For example, an orthologous gene may be expressed in both human and mouse but will not necessarily share the same expression profile in both species. Therefore, knowing when and where a gene is expressed is of great importance in drug discovery for disease treatment and understanding the relationship between human genes and their counterparts in the model organisms.

A popular technique used to determine the expression status of a cell is to create a cDNA library from which expressed sequence tags are derived. An expressed sequence tag (EST) is a 200-800 nucleotide sequence from a cDNA clone. An

EST is generated randomly and represents a segment of an mRNA molecule (Adams et al., 1991; Nagaraj et al., 2007). The source of ESTs, namely mRNA, enables these tags to provide a view of the expression state of a cell by identifying the mRNA being expressed in a particular cell at any given time.

Although ESTs provide insights into many biological phenomena such as gene discovery, alternative transcript identification and genome annotation (Nagaraj et al., 2007), the EST transcripts are generated by single-pass sequencing and are therefore very susceptible to errors. The advantage of using ESTs in exploring cellular gene expression lies in their low complexity and cost-effectiveness. Since the use of any technology is dictated by its financial impact, ESTs will continue to be a popular low-cost method among researchers as the current, high-impact sequencing methods become more established.



With the continuous generation of genome-scale data, it is imperative that the biological data be annotated in such a way that it is possible to adequately share and compare data from different biological sources, experiments or laboratories. Since 2000 (Stevens et al., 2000), ontologies have become an accepted method in bioinformatics with which to describe experimental tissue sources and gene expression data. Table 1 lists the 26 anatomical ontologies available from the Open Biomedical Ontology (OBO) Foundry (Smith et al., 2007) as of August 2009. The OBO Foundry provides a library of reference ontologies for the biomedical domain. Strict requirements need to be met for an ontology to be endorsed by the OBO Foundry such as providing a definition for every term within the ontology. Since the implementation of the OBO requirements, the

**Table 1**

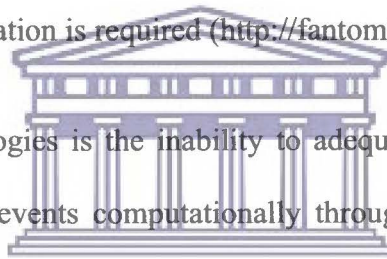
**A list of ontologies available from the Open Biomedical Ontologies (OBO) Foundry. The eVOC ontology is not officially distributed via the OBO foundry, but is included here to give context.**

Ontology	Namespace
Common Anatomy Reference Ontology	CARO
Subcellular anatomy ontology	SAO
Teleost anatomy and development	TAO
C. elegans gross anatomy	WBbt
Spider Ontology	SPD
Mouse adult gross anatomy	MA
Mouse gross anatomy and development	EMAP
Amphibian gross anatomy	AAO
Drosophila gross anatomy	FBbt
Fungal gross anatomy	FAO
Cellular component	GO
Xenopus anatomy and development	XAO
Plant growth and developmental stage	PO
Plant structure	PO
Spatial Ontology	BSPO
C. elegans development	WBls
Mosquito gross anatomy	TGMA
Drosophila development	FBdv
Human developmental anatomy, timed version	EHDA
Dictyostelium discoideum anatomy	DDANAT
Zebrafish anatomy and development	ZFA
Tick gross anatomy	TADS
Foundational Model of Anatomy (subset)	FMA
Medaka fish anatomy and development	MFO
Cell type	CL
Human developmental anatomy, abstract version	EHDA
eVOC Expression vocabulary	eVOC



eVOC ontology is no longer part of the OBO distribution as it does not provide definitions for all its terms. It is an important aim of the project to be included in the OBO distribution and further curation of the ontologies will ensure this.

An ontology is a hierarchical vocabulary used to describe a particular domain, and consists of parent and child terms defined by relationships between them. The most well-known ontology is the Gene Ontology (Ashburner et al., 2000) which describes three domains: the cellular component, molecular function and biological process of an organism. Ontologies are used by most database systems where a user is able to select a search term from a drop-down menu to select, for example the FANTOM3 CAGE Basic Viewer where the user selects the tissue for which expression information is required (<http://fantom3.gsc.riken.jp/>).



The problem with ontologies is the inability to adequately compare human and mouse gene expression events computationally through ontologies due to their individual structures and inherent complexities. An effective tool to enable the ontological comparison between human and mouse will enable the direct inter-species comparison of gene expression events, providing insight into the differences and similarities between the species – an integral aspect of model organism biology.

Model organisms are an important part of biological research because they allow researchers to perform experiments that would be either unethical or fatal if performed on humans. For example, it is considered unethical to genetically modify a human embryo by creating a knock-out of a particular gene purely to determine a possible function for that gene. Model organisms therefore allow us

to study genes *in vivo*, they allow us to test experimental drugs for efficacy and lethality, and they enable us to explore gene expression events throughout the life-span of the organism since its gestation and developmental periods are typically on a scale of days and weeks rather than months and years. The laboratory mouse is a particularly good model for studying cancer because mice have a high tumour incidence, are cheap and easy to handle, can be inbred to eliminate genetic variation effects, and many may be treated at a time to provide replicate data. However, in order for model organism experiments to be informative, it is imperative that we know and understand the similarities and differences between the models and humans. A robust system for comparing human and mouse biology and expression data is therefore critical.

This dissertation describes the development and implementation of an ontology-based system as a consistent approach to gene discovery. The processes required to successfully develop and apply a set of ontologies are to:

- 1) develop a set of ontologies;
- 2) map data to the ontologies by using them to annotate expression data;  
and
- 3) query the system to answer specific questions regarding the data.

Chapter 1 describes the development of a mouse ontology that conforms to the structure of an established human ontology to provide a tool to compare biological aspects of the two species. Both the mouse and human ontologies are also further developed to include the ontological representation of the developing mouse and human, enabling the alignment of mouse and human anatomical

structures for the annotation of expression events. In addition to developing the ontologies, this chapter also describes using the ontologies to annotate 8 852 human and 1 210 mouse cDNA libraries obtained from the Cancer Genome Anatomy Project (CGAP) as an initial dataset with which to illustrate the use of the ontologies.

The remaining two chapters describe how the ontologies developed in Chapter 1 are used in two major collaborations. Both chapters describe two aspects of each collaboration, namely a publication resulting from the collaborative efforts of all the members of the collaboration and an independent study I performed within each collaboration that is unpublished. I therefore, for each chapter, briefly describe my role in the collaboration and the work I performed that resulted in the publications, and thereafter describe in detail the unpublished analyses.

Chapter 2 describes how the ontologies developed in Chapter 1 are used to determine the expression profile of human transcription factors. The investigation of the expression profile enables the identification of transcription factor complexes that show tissue-restricted expression patterns.

The analysis presented as Chapter 3 uses the ontologies described in Chapter 1 to explore the expression profile of the mouse orthologs of human cancer/testis genes with the aim of comparing the human and mouse expression profiles of these genes.

# Chapter 1

## **Simplified ontologies allowing comparison of developmental mammalian gene expression**

### **1.1 Summary**

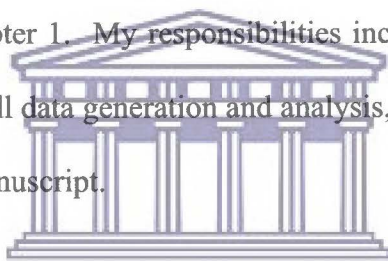
The concept of creating a developmental mouse ontology that is structured in the same way as the existing human eVOC ontologies was suggested as a viable approach while establishing a collaboration as part of the FANTOM consortium - a collaborative effort by many international laboratories with the aim to map out the transcriptional landscape of mouse and human. I was responsible for developing and applying the method of ontology generation for both the mouse and human developmental ontologies. I was also responsible for collecting and annotating the mouse and human CGAP cDNA libraries that have been mapped to the ontologies, as well as the data provided by the FANTOM3 project. The ontologies that I developed, along with the FANTOM data that I mapped to it, were incorporated into the FANTOM CAGE databases (CAGE Basic Viewer and CAGE Analysis Viewer) available online (<http://fantom3.gsc.riken.jp/>).

The FANTOM3 project culminated in a main publication in Science (of which I was co-author (Carninci et al., 2005)) as well as many satellite papers in PLoS Genetics – including a paper which I co-authored (Bajic et al., 2006). For ‘The transcriptional landscape of the mammalian genome’ published in Science (Appendix I), I was responsible for the development of the ontologies which were



used to annotate the expression data used in the paper. In the PLoS Genetics paper, ‘Mice and men: their promoter properties’ (Appendix II), the aim was to classify transcription start sites (TSS) based on the GC content of the 5’ upstream region of each gene. I used the ontology system described in this chapter to provide the expression information for the dataset used in the paper, which shows enrichment of certain tissue categories in each of the four TSS categories identified (Table 6 of Appendix II). The methods and results for both analyses are described in detail in the publications appended.

In addition to developing the ontologies, I was responsible for preparing the manuscript describing the development and application of these ontologies, which is presented here as Chapter 1. My responsibilities included the development of the manuscript concept, all data generation and analysis, as well as the preparation and submission of the manuscript.



Dr Yoshihide Hayashizaki and Dr Piero Carninci provided the request of the developmental ontologies as well as access to the FANTOM3 data. Dr Oliver Hofmann and Dr Winston Hide provided guidance regarding ontology development and application, and oversaw the production of the manuscript.

## 1.2 Aim

The aim of the work presented in this chapter is to develop an ontology system that enables the comparison of human and mouse anatomy throughout development. The use of the ontologies in the annotation of human and mouse

gene expression data provides a means to accurately compare gene expression between human and mouse, thereby identifying similar and unique gene expression patterns between the two species.

## **1.3 Background**

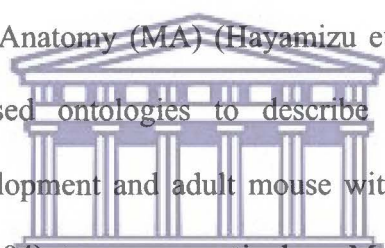
### **1.3.1 Ontologies and gene expression**

Biological investigation into mammalian biology employs standardized methods of data annotation by consortia such as MGED (Microarray Gene Expression Data Society) and CGAP (Cancer Genome Anatomy Project) or collaborative groups such as the Genome Network Project group at the Genome Sciences Centre at RIKEN, Japan (<http://gsc.riken.go.jp/indexE.html>). Data generated by these consortia include microarray, CAGE (Cap Analysis of Gene Expression), SAGE (Serial Analysis of Gene Expression) and MPSS (Massively Parallel Signature Sequencing) as well as cDNA and EST (Expressed Sequence Tags) libraries. The diversity of data types offers the opportunity to capture several views on concurrent biological events, but without standardization between these platforms and data types information is lost, reducing the value of comparison between systems. The terminology used to describe data provides a means for the integration of different data types such as EST or CAGE.

An ontology is a commonly used method of standardization in biology. It is often defined as a formal description of entities and the relationships between them, providing a standard vocabulary for the description and representation of terms in

a particular domain (Bard and Winter, 2001; Gkoutos et al., 2005). Given a need and obvious value in comparison of gene expression between species, anatomical systems and developmental states, we have set out to discover the potential and applicability of such an approach to compare mouse and human systems.

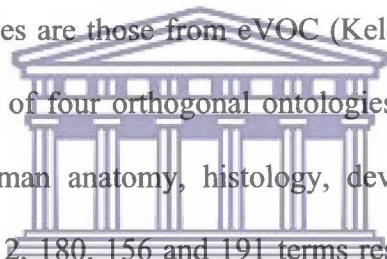
Many anatomical and developmental ontologies have been created, each focusing on their intended organisms. As many as 62 ontologies describing biological and medical aspects of a range of organisms can be obtained from the Open Biomedical Ontologies (OBO) website (<http://www.obofoundry.org/>), a system set up to provide well-structured controlled vocabularies of different domains in a single website. The Edinburgh Mouse Atlas Project (EMAP) (Baldock et al., 2003) and Adult Mouse Anatomy (MA) (Hayamizu et al., 2005) ontologies are the most commonly used ontologies to describe mouse gene expression, representing mouse development and adult mouse with 13 730 (October, 2005) and 7 702 (October, 2004) terms respectively. Mouse Genome Informatics (MGI), the most comprehensive mouse resource available, uses both ontologies. Human gene expression however, can be represented as developmental and adult ontologies by the Edinburgh Human Developmental Anatomy (HUMAT) ontology (Hunter et al., 2003) consisting of 8 316 terms (October, 2005) and the mammalian Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) consisting of more than 110 000 terms (January, 2002). Selected terms from the above ontologies have been used to create a cross-species list of terms known as the SOFG Anatomy Entry List (SAEL) (Parkinson et al., 2004). Although these ontologies more than adequately describe the anatomical structures of the developing organism, with the exception of SAEL, they are structured as Directed



UNIVERSITY of the  
WESTERN CAPE

Acyclic Graphs (DAG), defined as a hierarchy where each term may have more than one parent term (Hayamizu et al., 2005). The DAG structure adds to the inherent complexity of the ontologies, hampering efforts to align them between two species, making the process of a comparative study of gene expression events a challenge.

Efforts are being implemented in order to simplify ontologies for gene expression annotation. The Gene Ontology (GO) Consortium's GO slim (Martin et al., 2004) contains less than 1% of terms in the GO ontologies. GO slim is intended to provide a broad categorization of cDNA libraries or microarray data when the fine-grained resolution of the original GO ontologies are not required. Another set of simplified ontologies are those from eVOC (Kelso et al., 2003). The core eVOC ontologies consist of four orthogonal ontologies with a strict hierarchical structure to describe human anatomy, histology, development and pathology, currently consisting of 512, 180, 156 and 191 terms respectively (August, 2006). The aim of the eVOC project is to provide a standardized, simplified representation of gene expression, unifying different types of gene expression data and increasing the power of gene expression queries. The simplified representation achieved by the eVOC ontologies is due to the implementation of multiple orthogonal ontologies with a lower level of granularity than its counterparts.

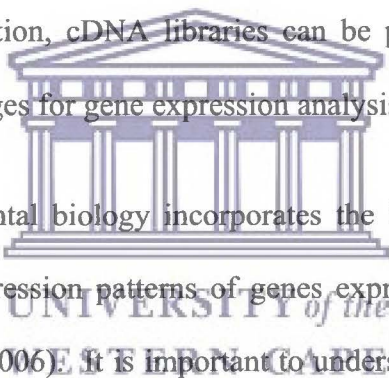


UNIVERSITY of the  
WESTERN CAPE



### 1.3.2 Mammalian development

The laboratory mouse is being used as a model organism to study the biology of mammals (Marra et al., 1999). The expectation is that these studies will provide insight into the developmental and disease biology of humans, coloured by the finding that 99% of the 25 000 – 30 000 mouse genes may have a human ortholog (only 1% of mouse genes do not have a human ortholog) and at least 80% of mouse genes are 1:1 orthologs where the mouse sequence is the best match to the human sequence and vice versa (Waterston et al., 2002). Given the similarity between the two species, it is possible to perform functional experiments on mouse and transfer any knowledge obtained to enhance our understanding of human biology. In addition, cDNA libraries can be prepared from very early mouse developmental stages for gene expression analysis.



The study of developmental biology incorporates the identification of both the temporal and spatial expression patterns of genes expressed in the embryo and fetus (Magdaleno et al., 2006). It is important to understand developmental gene expression because many genetic disorders originate during this period (Lindsay and Copp, 2005). Similarities in behavior and expression profiles between cancer cells and embryonic stem cells (Kho et al., 2004) also fuel the need to investigate developmental biology.

Using mice as model organisms in research requires the need for comparison of resulting data and provides a means to compare mouse data to humans (Lindsay and Copp, 2005). The cross-species comparison of human and mouse gene expression data can highlight fundamental differences between the two species

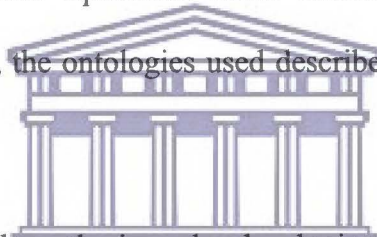
such as greater olfactory and immune capabilities, impacting on areas as diverse as the effectiveness of therapeutic strategies in the treatment of cystic fibrosis or Alzheimer's to the elucidation of the components such as tail, fur and whiskers that determine species. Using ontology-annotated gene expression events to compare across species provides a structured and accurate means of identifying identical gene expression context between the species, particularly if the annotation of each species differs in granularity.

### 1.3.3 Cross-species gene expression comparison

Function of most human genes has been inferred from model organism studies, based on the transitive assumption that genes sharing sequence similarity also share function when conserved across species (Zhou and Gibson, 2004). The same principle can be applied to gene regulation. The first step is to find not only the orthologs, but the commonly expressed orthologs. We predict that although two genes are orthologous between human and mouse, their expression patterns differ on the temporal and spatial level, indicating that their regulation may differ between the two species.

The terminology currently used to annotate human and mouse gene expression can be ambiguous (Eilbeck et al., 2005) among species since one term may be used to describe many different structures or one structure may be defined by more than one term, which is a result of different ontologies being used to annotate different species. The way in which we circumvented this issue is to

effectively map the ontology terms across species by using the same terminology for each species. This adaptation allows the integration of human and mouse ontologies as well as the comparison of the data it is used to annotate – a feature not possible with current ontologies. Although the EMAP, MA, HUMAT and FMA ontologies describe the anatomical structures throughout the development of the mouse and human, their complexities complicate the alignment of the anatomy between the two species. With the alignment of terms between a mouse and human ontology, the data mapped to each term becomes comparable, allowing efficient and accurate comparison of mammalian gene expression. A SAEL-related project, XSPAN (Dennis et al., 2003), is aimed at providing a web tool to enable users to find equivalent terms between ontologies of different species. Although useful, the ontologies used describe only spatial anatomy and are not temporal.



UNIVERSITY of the  
WESTERN CAPE

We have attempted to address the issue by developing simplified ontologies that allow the comparison of gene expression between human and mouse on a temporal and spatial level. The distribution of human and mouse anatomy terms across development match the structure of the human adult ontologies that form the core of the eVOC system.

Due to the ambiguous annotation of current gene expression data between human and mouse, and the lack of data mappings accompanying the available ontologies, the ontologies presented here have been developed in concert with semi-automatic mapping and curation of 8 852 human and 1 210 mouse cDNA libraries. We have therefore created a resource of simplified, standardized gene expression enabling

cross-species comparison of gene expression between mammalian species that is publicly available.

## 1.4 Materials and methods

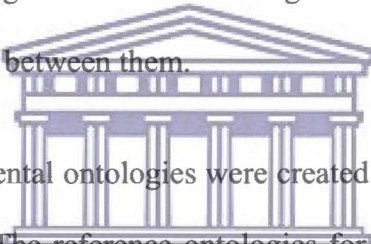
### 1.4.1 Ontology development

The ontologies were constructed using the COBrA (Aitken et al., 2005) and DAG-edit (<http://www.geneontology.org/GO.tools.shtml#dagedit>) ontology editors. Each term has a unique accession identifier with 'EVM' as the namespace for mouse and 'EV' for human, followed by seven numbers. This is consistent with the rules defined by the GO consortium (Ashburner et al., 2000).

Using the human adult eVOC anatomical system ontology as a template, terms from the Theiler stage 26 (mouse developmental stage immediately prior to birth) section of the EMAP ontology were inserted to create the Theiler stage 26 developmental eVOC mouse ontology. Proceeding from Theiler stage 26 to Theiler stage 1, each stage was used as a template for the next stage and any term not occurring at that specific stage, using EMAP as reference, was removed. Similarly, if a term occurred in EMAP that was not present in the previous stage, it was added to the ontology. The result is a set of 26 ontologies, one for each Theiler stage of mouse development, with many terms appearing and disappearing throughout the ontologies according to changes of anatomy during mouse development.



The Theiler stage 28 (adult mouse) ontology was constructed in the same way as the developmental ontologies, using the MA ontology as a reference. A previously not available Theiler stage 27 ontology was developed by comparing Theiler stage 26 and Theiler stage 28. Any terms that differed between the two stages were manually curated and included or removed in Theiler stage 27 as needed. The Theiler stage 27 ontology therefore represents all immature, post-natal anatomical structures. Theiler stage 28 ontology terms have been mapped to the adult human eVOC terms by using the human eVOC accession identifiers as database cross-references in the mouse ontology. Similarly, the EMAP accession number for each term was mapped to the developmental mouse ontologies. The result is a set of 28 ontologies that are an untangled form of the EMAP and MA ontologies, with mappings between them.



A set of human developmental ontologies were created by using the same method as was used for mouse. The reference ontologies for human development were the HUMAT ontologies, which describes the first 23 Carnegie stages of development, classified according to morphological characteristics.

The 28 mouse and 23 human ontologies were merged into two ontologies – one for mouse and one for human. Each merged ontology (named Mouse Development and Human Development) contains all terms present in the individual ontologies. A Theiler Stage ontology was created for mouse, which contains all 28 Theiler stages categorized into embryo, fetus or adult. The existing eVOC Development Stage ontology serves as the human equivalent of the mouse Theiler Stage ontology. The Mouse Development, Human

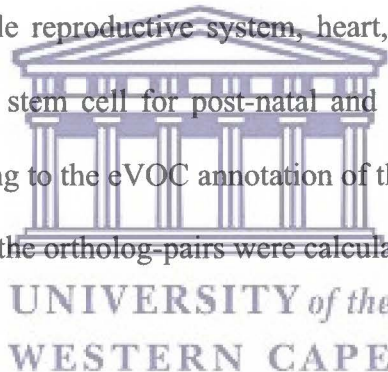
Development, Theiler Stage and the existing Development Stage ontologies form the core of the Developmental eVOC ontologies.

#### 1.4.2 Data mapping

Mouse and human cDNA libraries were obtained from the publicly available CGAP resource (January, 2006) and mapped (semi-automated) to the entire set of eVOC ontologies. The eVOC ontologies consist of Anatomical System, Cell Type, Developmental Stage, Pathology, Associated With, Treatment, Tissue Preparation, Experimental Technique, Pooling and Microarray Platform. The 'age' annotation of the mouse CGAP libraries were manually checked against the Gene Expression Database (version 3.41; December, 2005) (Hill et al., 2004) to determine the Theiler stage of each library. Due to the lack of a resource providing the Carnegie stage annotation for cDNA libraries, the human cDNA libraries were annotated according to the age annotation originally provided by CGAP. Genes associated with each mouse and human cDNA library were obtained from NCBI's UniGene (March, 2006) (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>). A list of human-mouse orthologs were obtained from HomoloGene (build 53) (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene>).

### 1.4.3 Data mining

The genes were filtered according to the presence or absence of expression evidence and homology. A gene passed the selection criteria if it has an ortholog and if both genes in the ortholog pair have eVOC-annotated expression. According to eVOC annotation, genes were categorized into those that showed expression in normal adult brain and those expressed in normal developmental brain, many genes appearing in more than one category. Genes expressed in normal adult brain were subtracted from those with expression in normal developmental brain to establish genes whose expression in the brain occurs only during development. The expression profiles of the developmentally-biased genes annotated to female reproductive system, heart, kidney, liver, lung, male reproductive system and stem cell for post-natal and developmental expression were determined according to the eVOC annotation of the cDNA libraries, and the correlation coefficient of the ortholog-pairs were calculated.



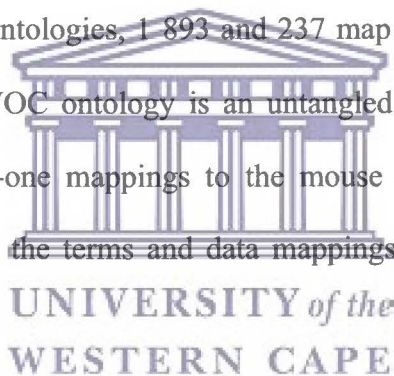
## 1.5 Results and discussion

### 1.5.1 Ontology development

The ontologies were originally created to accommodate requests by the FANTOM3 consortium (Carninci et al., 2005) for a simple mouse ontology that could be used in alignment to the human eVOC ontologies. The FANTOM3 project was a collaborative effort by many international laboratories to analyze the mouse and human transcriptome. The aim was to generate a transcriptional

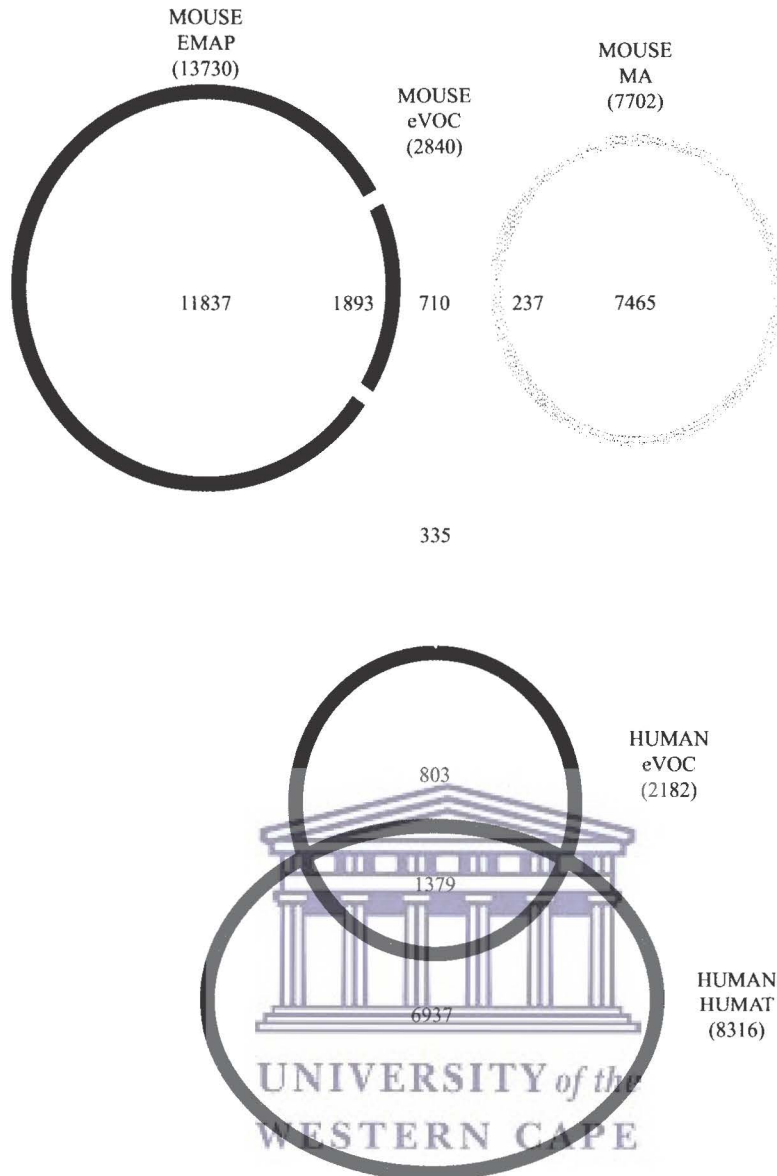
in the developmental eVOC ontologies to ensure interoperability between external ontologies and eVOC. Terms from the mouse have also been mapped to those from human to enable cross-species comparison of the data mapped.

The integration of the ontologies is described in Figure 1, where ‘Mouse eVOC’ refers to the individual mouse ontologies and ‘Human eVOC’ refers to the individual human ontologies (including the adult human ontology). The EMAP and MA ontologies represent mouse pre- and post-natal developmental anatomical structures, respectively, and therefore exhibit no commonality. The mouse developmental eVOC ontologies integrate the two ontologies by containing terms from, and mappings to, both the EMAP and MA ontologies. Of the 2 840 terms in the individual mouse ontologies, 1 893 and 237 map to EMAP and MA. The human developmental eVOC ontology is an untangled version of the HUMAT ontology and has one-to-one mappings to the mouse developmental ontology, providing a link between the terms and data mappings between the mouse and human ontologies.



The presence of species-specific anatomical structures posed a challenge when aligning the mouse and human terms. An obvious example is the presence of a tail in mouse but not in human. We decided that there would simply be no mapping between the two terms. Further challenges involved structures such as paw and hand. The two terms cannot be made identical because it is incorrect to refer to the anterior appendage of a mouse as a hand. However, due to the fact that the mouse paw and human hand share functional similarities, the two terms are not identical, but are mapped to each other based on functional equivalence.





**Figure 1**

**Venn diagram illustrating the integration of mouse and human ontologies represented by the eVOC system. The total number of terms in each ontology is in parentheses. The numbers in each set are the number of terms in the intersection represented by that set. 'Mouse eVOC' represents the 28 individual mouse ontologies and 'Human eVOC' represents the 23 individual human and adult ontologies; therefore, the numbers in parentheses refer to the total number of terms in all the eVOC ontologies for each species. The intersection of the Mouse eVOC with the EMAP and MA ontologies represents the number of terms in Mouse eVOC that have database cross-references to EMAP and MA. Similarly, the intersection of the Human eVOC and HUMAT sets represents the number of Human eVOC terms that map to HUMAT terms. The number within the arrows represents the number of mapped human and mouse eVOC terms.**



In order to provide simplified ontologies, the 28 mouse and 23 human ontologies were merged to create two ontologies – one for each species. In addition, a Theiler Stage ontology was created that represents the Theiler stages of mouse development. The human stage ontology is represented by the current eVOC Development Stage. A cross-product of two terms (one from the merged and one from the stage ontology) for a species can therefore represent any anatomical structure at any stage of development.

The relationship between the Developmental Mouse and individual ontologies is illustrated in Figure 2, where the term ‘brain’ is mapped to 12 terms in the individual ontologies and therefore occurs in 12 of the 28 Theiler stages. All terms in the individual ontologies that are derived from EMAP or MA for mouse, and HUMAT for human are mapped to the corresponding term by adding the term’s accession from the external ontology as a database cross-reference in the eVOC ontologies. Figure 3 shows that the database cross-reference is the accession of the EMAP term, indicating that ‘intestine’ of the ‘Theiler stage 13’ ontology is equivalent to the term represented by ‘EMAP:600’. This feature allows cross-communication, and thereby integration, of the EMAP, MA, HUMAT and eVOC ontologies.

The ontologies presented here are simplified versions of existing human and mouse developmental and adult ontologies, containing 1 670 and 2 840 terms respectively. Table 1 shows the number of terms and database cross-references for the individual mouse and human ontologies. The Theiler Stage 4 ontology contains 12 terms and has 9 mappings to the EMAP ontology. The mouse and

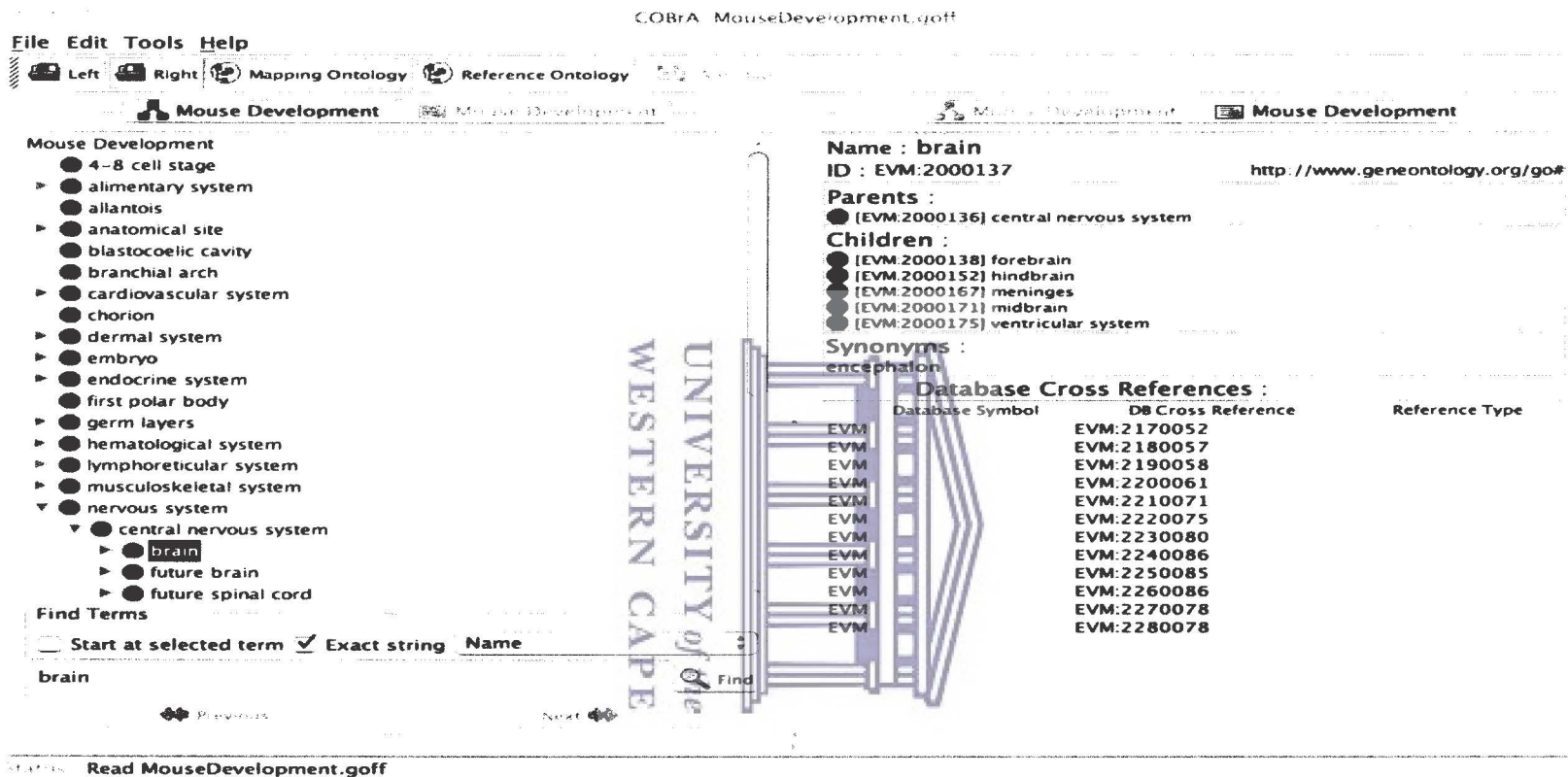


Figure 2

Screenshot of the Mouse Development ontology, visualised in COBRA. The left panel shows the hierarchy of the ontology, with 'brain' as the highlighted term. The right panel lists the 12 database cross-references mapped to 'brain', representing the accession of 'brain' in each of the 12 individual ontologies.

COBrA TS13.goff

File Edit Tools Help

Left Right Mapping Ontology Reference Ontology

eVOC Mouse Ontologies

eVOC Mouse Ontologies

eVOC Mouse Ontologies

eVOC Mouse Ontologies

- Theiler Stage 13
  - alimentary system
    - diverticulum
    - gastrointestinal tract
      - **intestine**
      - mesentery
  - ▶ ● anatomical site
  - ▶ ● branchial arch
  - ▶ ● cardiovascular system
  - ▶ ● endocrine system
  - ▶ ● germ layers
  - ▶ ● hematological system
  - ▶ ● nervous system
  - ▶ ● primitive streak
  - ▶ ● umbilical cord
  - ▶ ● unclassifiable
  - ▶ ● urogenital system

Name : intestine  
ID : EVM:2130003 <http://www.geneontology.org/go#>  
Parents :  
● [EVM:2130057] gastrointestinal tract  
Synonyms :  
gut

Database Cross References :


Database Symbol	DB Cross Reference	Reference Type
EMAP	EMAP:600	

Find Terms

Start at selected term  Exact string Name

intestine

Previous Next



Status: Read TS13.goff

Figure 3

Screenshot of the individual Theiler Stage 13 ontology, visualised in COBrA. The left panel displays the ontology with terms of anatomical structures occurring only in Theiler stage 13 of mouse development. The right panel lists the accession of the equivalent term in the external ontology as a database cross-reference.

**Table 1**

**Statistics of the individual developmental eVOC ontologies, representing the alignment between human and mouse stages. The first three columns display the individual mouse ontologies, the number of terms in each ontology, and the number of external references of each. The last three columns display the individual human ontologies, the number of terms, and the number of external references of each. The external references refer to the EMAP and MA ontologies for mouse, and to HUMAT for human. The alignment of the rows between the mouse and human ontologies represents the alignment of the Theiler and Carnegie stages of development based on morphological similarities. For example, the Theiler Stage 4 ontology contains 12 terms and has 9 mappings to the EMAP ontology. Mouse Theiler Stage 4 is equivalent to human Carnegie Stage 3. The Carnegie Stage 3 ontology contains 13 terms and has 11 mappings to terms from the HUMAT ontology.**

Theiler Stage	Mouse Terms	External Reference	Carnegie Stage	Human Terms	External Reference
1	6	4	1	5	4
2	5	3	2	5	4
3	6	4			
4	12	9	3	13	11
5	9	6			
6	10	7	4	10	8
7	11				
8	12	10	5a	10	8
			5b	11	10
			5c	9	8
9	14	14	6a	14	16
			6b	19	18
10	14	18	7	20	17
11	32	29	8	22	19
12	56	63	9	52	54
13	55	64	10	60	80
14	67	85	11	72	92
15	80	109	12	80	98

Theiler Stage	Mouse Terms	External Reference	Carnegie Stage	Human Terms	External Reference
16	93	128	13	103	131
17	103	137	14	122	149
18	116	155	15	131	165
19	134	173	16	155	178
20	157	171	17	170	184
21	193	239	18	188	223
			19	199	237
22	209	299	20	200	237
23	216	303			
24	226	316			
25	234	339			
26	238	348			
27	266	0			
28	266	246	adult	512	
<b>TOTAL</b>	<b>2840</b>	<b>3288</b>	<b>TOTAL</b>	<b>2049</b>	<b>1951</b>


  
 UNIVERSITY *of the*  
 WESTERN CAPE

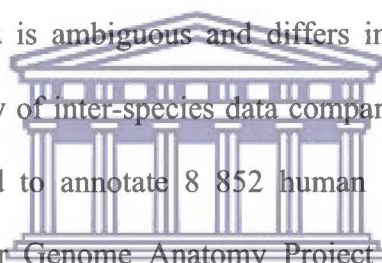


human stages have been aligned in the table and therefore shows that mouse Theiler Stage 4 is equivalent to human Carnegie Stage 3, based on morphological similarities during development (<http://www.ana.ed.ac.uk/anatomy/database/humat/MouseComp.html>). The Carnegie Stage 3 ontology contains 13 terms and has 11 mappings to the HUMAT ontology. The difference in the number of ontology terms and external references is attributed to the addition of terms to maintain the standard structure of the eVOC system. In this example, the term 'germ layers' is in the eVOC ontologies, but not in the EMAP or HUMAT ontologies. Many eVOC terms are mapped to more than one term in the external referencing ontology as an artifact of the simplification of the ontologies, resulting in a one-to-many relationship between eVOC and its reference ontology. For example, 'myocardium' at Theiler Stage 12 in the eVOC ontologies is mapped to five EMAP identifiers. Each EMAP identifier references a cardiac muscle, but at a different location. eVOC does not distinguish between cardiac muscle of the common atrial chamber (EMAP:337) and cardiac muscle of the rostral half of the bulbus cordis (EMAP:330). Compared to their counterparts, the Developmental eVOC ontologies represent 22% of both the human HUMAT and mouse EMAP ontologies, with the only relationship between the terms being 'IS\_A'. Note that relationships within the eVOC ontologies only indicate an association between parent and child term and do not systematically distinguish between is\_a or part\_of relationships. As eVOC moves to adopt relationship types from the OBO Relation Ontology (Smith et al., 2005) relations will be reviewed and curated. Using a principle of data-driven development, eVOC terms are

added at an annotator's request, resulting in a dynamic vocabulary describing gene expression.

### 1.5.2 Data mapping

The resources providing ontologies to annotate gene expression do not always provide the data itself. In order to obtain mouse and human data, one would have to search separate databases for each species. An example of this would be searching MGI for mouse gene expression data, and ArrayExpress for human. Apart from having to access different databases to obtain data, the terminology used to describe the data is ambiguous and differs in the level of granularity, impacting on the accuracy of inter-species data comparison. The ontology terms have therefore been used to annotate 8 852 human and 1 210 mouse cDNA libraries from the Cancer Genome Anatomy Project (CGAP) (January, 2006) (<http://cgap.nci.nih.gov/>).



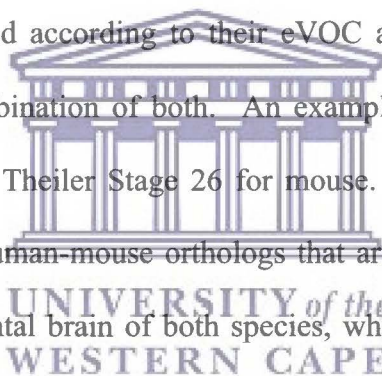
UNIVERSITY of the  
WESTERN CAPE

The mapping process revealed inconsistencies in the annotation of the human and mouse CGAP cDNA libraries, requiring manual intervention and emphasizing the need for a standardized annotation. All genes associated with the libraries have been extracted by association through UniGene (March, 2006). A gene was considered to be associated with a cDNA library if at least one EST was evident for the gene in a particular library. The result is a set of 21 152 human and 24 047 mouse genes from UniGene that are represented by CGAP cDNA libraries and annotated with eVOC terms, and represent the set of human and mouse genes for

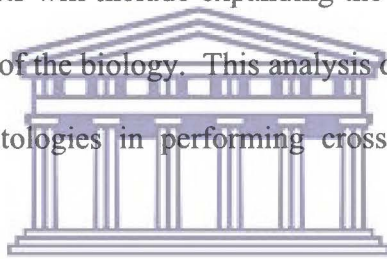
which there is expression evidence. CGAP represents an ascertainment bias where there is a strong over-representation for cancer genes, and therefore future efforts for this research will include obtaining a well-represented, evenly distributed dataset of human and mouse gene expression. The list of human and mouse orthologs were extracted from HomoloGene to represent the 16 324 human-mouse orthologs. Two genes were considered to be orthologs if they shared the same HomoloGene group identifier (March, 2006).

### 1.5.3 Data mining

Genes may be categorized according to their eVOC annotation on a spatial or temporal level, or a combination of both. An example of this would be genes expressed in the heart at Theiler Stage 26 for mouse. For the purposes of this study, we searched for human-mouse orthologs that are expressed in the normal postnatal and developmental brain of both species, where a gene is classified as normal if it's originating library was annotated as 'normal'. Research involving gene expression of the brain aims at identifying causes of psychological and neurological diseases, many of these diseases originating during development. With the use of mice as model organisms in this kind of research, it is important to identify genes which are co-expressed in human and mouse on the temporal and spatial level. The results of our analysis show that of the available 16 324 human-mouse orthologs, 14 434 can be found in CGAP libraries for both human and mouse. When looking at brain gene expression, we could segregate genes according to their spatial and temporal expression patterns. We found that of all



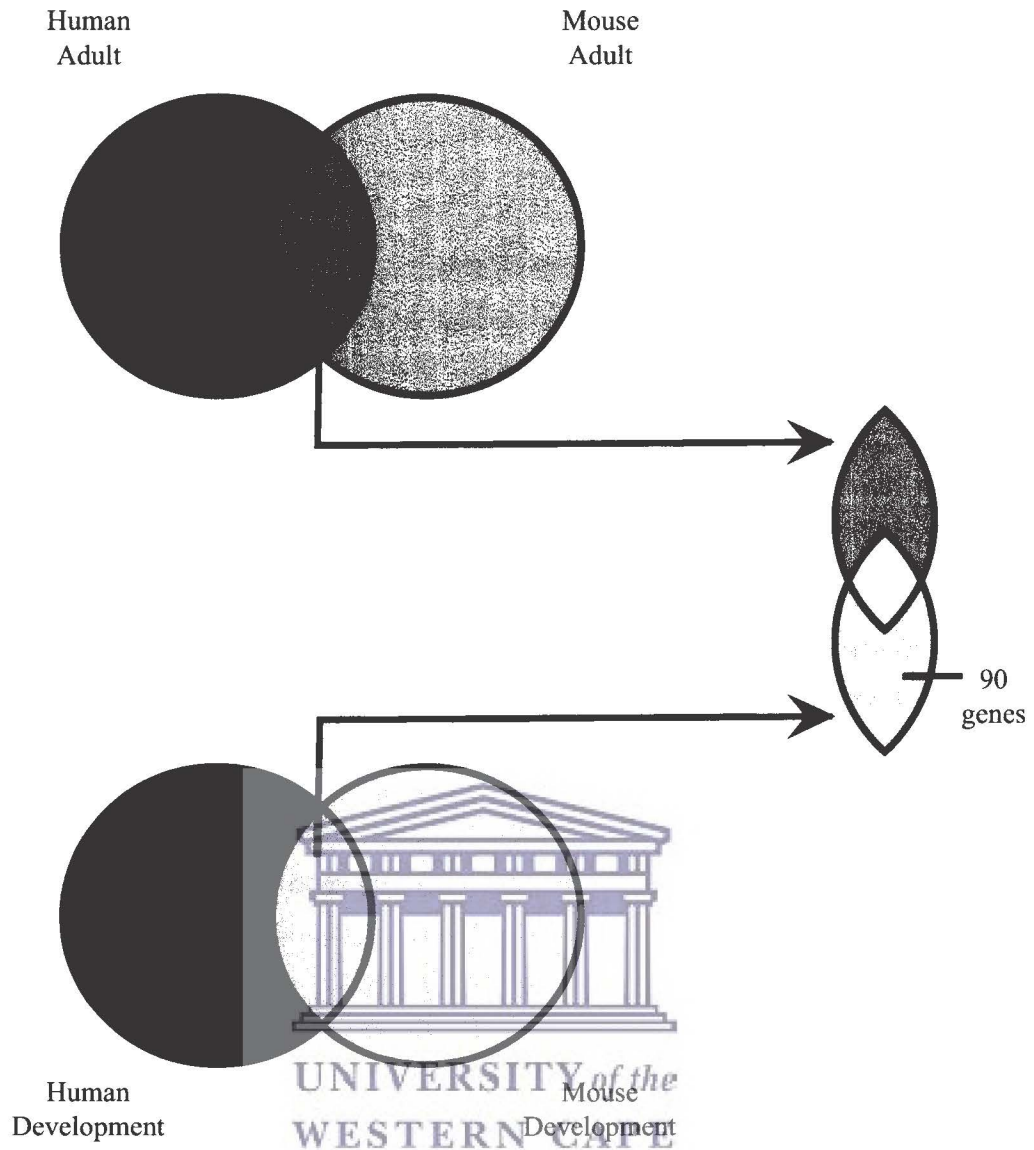
the orthologs expressed in the brain, 10 980 genes were expressed in the post-natal brain of both species whereas 1 692 genes were expressed in the developing brain of both species. Of these two sets of genes, 90 genes were found to have biased expression for developmental brain (Table 2) where developmentally biased genes are those that are expressed during development and not the post-natal organism in either human, mouse or both species (see Figure 4 for illustration). It is important to note that only genes whose orthologs also have expression evidence were considered for analysis. This small number of genes found to be biased for expression during brain development in both species may be a result of data-bias due to the difficulty involved in accessing developmental libraries. Our future efforts will include expanding the data platforms to provide data that is representative of the biology. This analysis does however demonstrate the usefulness of the ontologies in performing cross-species gene expression analyses.



UNIVERSITY of the  
WESTERN CAPE

The Gene Ontology (GO) categories that are highly associated with the 90 genes biased for developmental brain expression were extracted with the use of the DAVID bioinformatics resource (Dennis et al., 2003). The human representatives of the human-mouse orthologs cluster with GO terms such as ‘nervous system development’ and ‘cell differentiation’, suggesting a shared role for development of the mammalian brain, and therefore may be potential targets for the analysis in neurological diseases. Given the existence of ascertainment bias on these kinds of data, it was still surprising to see how many genes passed the stringent selection criteria. Searching the Online Mendelian Inheritance of Man (OMIM) database





**Figure 4**

**Diagram illustrating the sets of genes analysed for developmental brain expression bias. Genes for human and mouse grouped together if they are expressed in post-natal or developmental brain, respectively. The intersection between the human and mouse developmental brain genes represent those genes showing common expression in the two species. Subtracting genes commonly expressed in human and mouse post-natal brain determines those genes that show developmental restriction in either human, mouse or both species.**



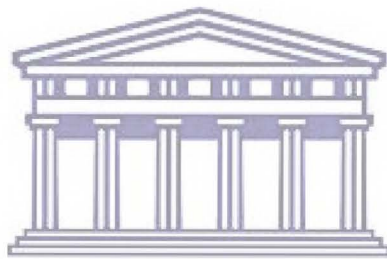
**Table 2**

**Genes showing developmental expression bias in human and mouse brain. The table lists the HomoloGene group identifier, Entrez Gene identifier and gene symbol of the 90 human-mouse orthologs found to have an expression bias towards the embryonic and fetal stages of brain development, without expression during postnatal development. Genes were only considered for analysis if they have an ortholog, and if the ortholog also has expression evidence based on eVOC annotation.**

HomoloGene group identifier	Human Entrez Gene ID	Human Entrez Gene Symbol	Mouse Entrez Gene ID	Mouse Entrez Gene Symbol
32	435	ASL	109900	Asl
268	5805	PTS	19286	Pts
413	353	APRT	11821	Aprt
1028	1606	DGKA	13139	Dgka
1290	9275	BCL7B	12054	Bcl7b
1330	857	CAV1	12389	Cav1
1368	1054	CEBPG	12611	Cebpg
1871	4760	NEUROD1	18012	Neurod1
1933	5050	PAFAH1B3	18476	Pafah1b3
2212	6182	MRPL12	56282	Mrpl12
2593	7913	DEK	110052	Dek
2880	8835	SOCS2	216233	Socs2
3476	9197	SLC33A1	11416	Slc33a1
4397	8971	H1FX	243529	H1fx
4983	10991	SLC38A3	76257	Slc38a3
6535	11062	DUS4L	71916	Dus4l
7199	11054	OGFR	72075	Ogfr
7291	10683	DLL3	13389	Dll3
7500	5806	PTX3	19288	Ptx3
7516	389075	RESP18	19711	Resp18
7667	1154	CISH	12700	Cish
7717	24147	FJX1	14221	Fjx1
7922	6150	MRPL23	19935	Mrpl23
9120	25851	DKFZP434B0335	70381	2210010N04Rik
9355	51637	C14orf166	68045	2700060E02Rik
9813	55627	FLJ20297	77626	4122402O22Rik
10026	55172	C14orf104	109065	1110034A24Rik
10494	58516	FAM60A	56306	Tera
10518	84273	C4orf14	56412	2610024G14Rik
10663	57171	DOLPP1	57170	Dolpp1
10695	57120	GOPC	94221	Gopc
10774	57045	TWSG1	65960	Twsg1
11653	79730	FLJ14001	70918	4921525L17Rik
11920	84303	CHCHD6	66098	Chchd6

HomoloGene group identifier	Human Entrez Gene Gene ID	Human Entrez Gene Symbol	Mouse Entrez Gene ID	Mouse Entrez Gene Symbol
11980	84262	MGC10911	66506	1810042K04Rik
12021	84557	MAP1LC3A	66734	Map1lc3a
12418	124056	NOXO1	71893	Noxo1
12444	84902	FLJ14640	72140	2610507L03Rik
12993	84217	ZMYND12	332934	Zmynd12
14128	91107	TRIM47	217333	Trim47
14157	90416	CCDC32	269336	Ccdc32
14180	115294	PCMTD1	319263	Pcmtd1
14667	113510	HEL308	191578	Hel308
15843	79591	C10orf76	71617	9130011E15Rik
16890	399664	RKHD1	237400	Rkhd1
17078	387914	TMEM46	219134	Tmem46
17523	115290	FBXO17	50760	Fbxo17
18123	140730	RIMS4	241770	Rims4
18833	143678	LOC143678	75641	1700029I15Rik
18903	440193	KIAA1509	68339	0610010D24Rik
19028	146167	LOC146167	234788	Gm587
20549	4324	MMP15	17388	Mmp15
21334	10912	GADD45G	23882	Gadd45g
22818	29850	TRPM5	56843	Trpm5
24848	266629	SEC14L3	380683	RP23-81P12.8
26702	93109	TMEM44	224090	Tmem44
27813	84865	FLJ14397	243510	A230058J24Rik
31656	27000	ZRF1	22791	Dnajc2
32293	51018	CGI-115	67223	2810430M08Rik
32331	51776	ZAK	65964	B230120H23Rik
32546	64410	KLHL25	207952	Klhl25
32633	136647	C7orf11	66308	2810021B07Rik
35002	93082	LINCR	214854	Lincr
37917	1293	COL6A3	12835	Col6a3
40668	9646	SH2BP1	22083	Sh2bp1
40859	27166	PX19	66494	2610524G07Rik
41703	118881	COMTD1	69156	Comtd1
45198	65117	FLJ11021	208606	1500011J06Rik
45867	139189	DGKK	331374	Dgkk
46116	401399	LOC401399	101359	D330027H18Rik
49899	143282	C10orf13	72514	2610306H15Rik
49970	83879	CDCA7	66953	Cdca7
55434	1289	COL5A1	12831	Col5a1
55599	669	BPGM	12183	Bpgm
55918	6882	TAF11	68776	Taf11
56005	6328	SCN3A	20269	Scn3a
56571	26503	SLC17A5	235504	Slc17a5
56774	54751	FBLIM1	74202	Fblim1

HomoloGene group identifier	Human Entrez Gene ID	Human Entrez Gene Symbol	Mouse Entrez Gene ID	Mouse Entrez Gene Symbol
64353	126374	WTIP	101543	Wtip
65280	286128	ZFP41	22701	Zfp41
65318	23361	ZNF629	320683	Zfp629
65328	7559	ZNF12	231866	Zfp12
68420	9559	VPS26A	30930	Vps26
68934	57016	AKR1B10	14187	Akr1b8
68973	1663	DDX11	320209	Ddx11
68998	170302	ARX	11878	Arx
78698	387876	LOC387876	380653	Gm872
81871	56751	BARHL1	54422	Barhl1
82250	150678	MYEOV2	66915	Myeov2
84799	22835	ZFP30	22693	Zfp30

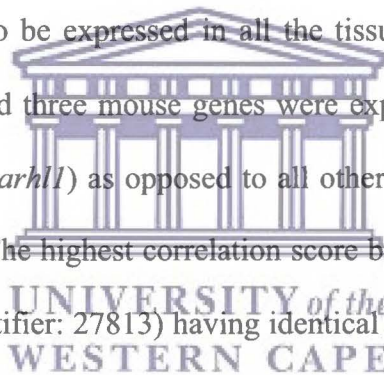


UNIVERSITY *of the*  
WESTERN CAPE



implicated some of the 90 genes, such as *GOPC*, *ARX* and *DEK*, in diseases such as astrocytoma, lissencephaly and leukemia.

To assess the similarity in expression across major human and mouse tissues other than brain, the expression profiles of the 90 genes with bias for developmental expression were determined for developmental and adult expression in the following tissues: female reproductive system, heart, kidney, liver, lung, male reproductive system and stem cell. These tissues were chosen based on the availability of data for each tissue in the developmental and adult categories. For each ortholog-pair, we determined the correlation between their expression profiles (see Appendix III). We found that, according to the cDNA libraries, one mouse gene was found to be expressed in all the tissues in both post-natal and development (*Twsg1*), and three mouse genes were expressed only in the mouse brain (*Resp18*, *Gm872*, *Barhl1*) as opposed to all other tissues (see Appendix IV for expression profile). The highest correlation score between an ortholog-pair is 0.646 (HomoloGene identifier: 27813) having identical expression profiles during development (expressed in liver and stem cell), but differing during post-natal expression (expression in mouse heart, kidney and stem cell but not in their human counterparts). The correlations observed suggest that the expression profiles of orthologs across these major tissues are only partially conserved between human and mouse. This finding strengthens our understanding of orthologous gene expression in that although two genes are orthologs, they do not share temporal and spatial expression patterns and therefore probably do not share a majority of their regulatory modules (Odom et al., 2007).



Developmental gene expression may be subdivided into embryonic and fetal expression which in turn may be categorized further according to the Theiler and Carnegie stages for mouse and human, allowing a high-resolution investigation of gene expression profiles between the two species. This stage by-stage expression profile for human and mouse will allow investigation into common regulatory elements of co-developmentally expressed genes and give new insight into the characterization of the normal mammalian developmental program.

## 1.6 Conclusions

The developmental mouse ontologies were developed in collaboration with the FANTOM3 consortium to have the same structure and format as the existing human eVOC ontologies to enable the comparison of developmental expression data between human and mouse. The developmental ontologies have been constructed by integrating the Edinburgh Mouse Atlas Project, Mouse Anatomy, the developmental Human Anatomy and the human adult eVOC ontologies. The re-organization of existing ontological systems under a uniform format allows the consistent integration and querying of expression data from both human and mouse databases, creating a cross-species query platform with one-to-one mappings between terms within the human and mouse ontologies.

The ontologies have been used to map human and mouse gene expression events, and can be used to identify differential gene expression profiles between the two species. In future, the ontologies presented here will be used to investigate the



transcriptional regulation of genes according to their characteristics based on developmental stage, tissue and pathological expression profiles, providing insight into the mechanisms involved in the differential regulation of genes across mammalian development.

## **1.7 Availability**

The mouse eVOC ontologies, their mappings and the datasets referred to in this manuscript are available under a FreeBSD-style license at the eVOC website (<http://www.evocontology.org>) and are appended here as Appendix V and VI.



## Chapter 2

### Expression profiling reveals tissue-restricted transcription factor complexes

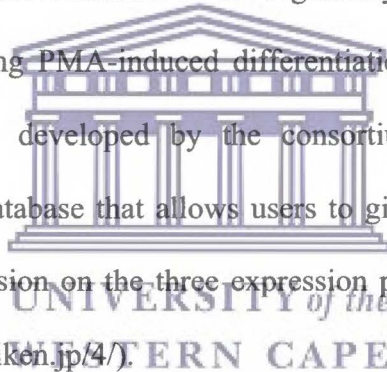
#### 2.1 Summary

The study presented in this chapter formed part of a major effort by the Genome Network Project (GNP) aimed at understanding the transcriptional networks involved in the growth arrest and differentiation in mammalian cells, using THP-1 cells (Human acute monocytic leukemia cell line) as a model system. My involvement in the project was two-fold:

1. Assist in analysing the response of 1 805 transcription factors from THP-1-derived macrophage cells to LPS stimulation over a range of time-points; and
2. Investigate the tissue expression profiles of 1 805 transcription factors under investigation.

In (1) above, THP-1 cells were induced to differentiate into macrophages by adding phorbol myristate acetate (PMA). After 96 hours, an immune response was induced by adding lipopolysaccharide (LPS) and the effect on transcription was monitored over a time-series of 0.5h, 1h, 2h, 3h, 4h, 8h, 10h, 12h, 18h and 24h. For each time-point, expression data was generated on three platforms: Illumina microarray, CAGE tags (cap analysis of gene expression) and qRT-PCR. I was part of the group that used the expression data from the Illumina platform to

determine which genes were up- and down-regulated during the early (0.5h, 1h, 2h, 3h), middle (4h, 8h, 10h) and late (12h, 18h, 24h) response to LPS stimulation. The results of this analysis formed the basis of the paper ‘The transcriptional network that controls growth and differentiation in a human myeloid leukemia cell line’ published in Nature Genetics by the GNP (Suzuki et al., 2009), wherein I am listed as co-author due to my involvement in the analysis. The publication is appended as Appendix VIIa. My analysis method and interpretation that contributed to the publication is appended as Appendix VIIb. The analysis yielded the categorisation of 193 genes into 10 categories according to their level of expression across ten time-points. The categorisation of these genes contributed to the identification of the regulatory motifs whose activity is significantly altered during PMA-induced differentiation. In addition, the data and computational tools developed by the consortium members have been collated into an online database that allows users to give a gene as input and is provided with its expression on the three expression platforms across the time-series (<http://fantom.gsc.riken.jp/4/>).



In (2) above, I used the ontologies and mappings described in Chapter 1 to determine the tissue expression profiles of the list of transcription factors under investigation by the GNP (1 805 genes). The list of genes for which an expression profile was required was provided to me by the GNP. I was responsible for the development, implementation and interpretation of the analysis, which is presented here as Chapter 2. The results of this analysis were provided to the GNP to assist in the interpretation and discussion of the results presented in the publication.

## 2.2 Aim

The aim of this chapter is to use the Developmental eVOC system to illustrate the identification of tissue-restricted, co-expressing transcription factors. The identification of co-expressing genes gives insight into the regulation of genes specific to a particular cell type or disease.

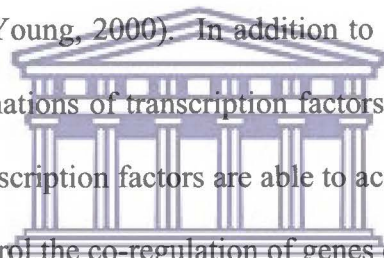
## 2.3 Background

Each gene in a cell has a spatial and temporal fate whereby it is only expressed in certain tissues at defined times throughout the life span of the organism. The exact timing of gene expression is a tightly controlled process (Dymlacht, 1997) and a slight deviation in this process causes aberrant gene expression that could lead to disease or a cell following an inappropriate developmental path. The origin of many diseases such as cancer (Liao et al., 2009), Alzheimer's (de la Monte et al., 1995) and multiple sclerosis (Sato et al., 2007) can be attributed to aberrant gene expression, making this process a topic of much investigation. In order to understand how the uncontrolled regulation of gene expression causes disease, it is important to understand how normal gene expression events are regulated within the cell.

Transcription factors are sequence-specific DNA-binding proteins forming the regulatory machinery responsible for the differential gene expression,

development and regulation of cellular processes in an organism. Transcription factors function by binding to a promoter sequence in the upstream, untranslated region of a gene, allowing RNA polymerase II to bind and initiate transcription (Nikolov and Burley, 1997).

It is widely accepted that transcription factors function in complexes (Sandelin et al., 2007) rather than individually. The activation of transcription is greatly influenced by the composition of these transcription factor complexes where the presence or absence of even one transcription factor can alter the ability of the complex to activate transcription (Reid et al., 2009). This sensitive transcriptional switch therefore affects the regulation of gene expression on a spatial and temporal level (Lee and Young, 2000). In addition to one gene being controlled by many different combinations of transcription factors, it is also known that any given combination of transcription factors are able to activate more than one gene, providing a means to control the co-regulation of genes (Reid et al., 2009).



UNIVERSITY of the  
WESTERN CAPE

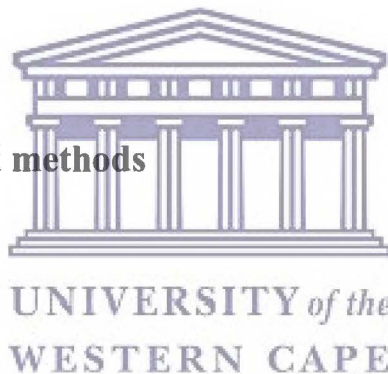
The efficiency of transcription factors are also variable, with some having a high DNA-binding affinity and others having low affinity, creating a mechanism whereby the cell can control the number of mRNA molecules transcribed from a gene. In addition, it is suggested that ubiquitously expressed transcription factors control a broad set of genes that are then fine-tuned by tissue-specific transcription factors (Vaquerizas et al., 2009). Regulation of gene expression by transcription factors is therefore greatly influenced by their tissue expression profiles as well as their involvement in transcription factor complexes.



Conventional expression profiling experiments focus on a few individual genes of interest. With the discovery of high-throughput technologies, it has become increasingly apparent that genes should be analysed within their genomic context. Since transcription factors function as groups or complexes, it is necessary that our investigations of gene expression events reflect this. The aim of this study is to identify tissue-restricted transcription factor complexes based on the co-expression of 1 805 transcription factors. The rationale behind this is that the identification of transcription factors responsible for tissue-specific expression of a particular gene may be investigated across different pathological states, thereby giving insight into the genes responsible for the disease in question.

## **2.4 Materials and methods**

### **2.4.1 Data generation**



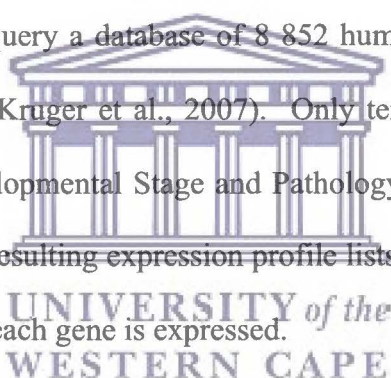
The members of the Genome Network Project (GNP), for which this study was conducted, compiled a list of human transcription factors for analysis, hereafter referred to as the Genes Of Interest list (GOI-list) (March, 2007). The genes in this list originally contained all 2 353 known human transcription factors based on qRT-PCR experiments. Manual curation of the GOI-list resulted in 1 805 transcription factors that conform to the following criteria:

- a) has a DNA-binding domain;

- b) shows evidence of nuclear localization according to LOCATE (Sprenger et al., 2008); and
- c) is annotated as a transcriptional regulator according to the Gene Ontology database (Ashburner et al., 2000).

A transcription factor was excluded from the GOI-list if there was strong evidence supporting localisation outside of the nucleus.

To generate expression profiles for each of the genes in the GOI-list, their Entrez Gene identifiers were obtained from the National Center for Biotechnology Information (NCBI) UniGene database (March, 2009) (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>). The Entrez Gene identifiers were used to query a database of 8 852 human cDNA libraries in the eVOC ontology system (Kruger et al., 2007). Only terms from the Anatomical System, Cell Type, Developmental Stage and Pathology ontologies were used to annotate the genes. The resulting expression profile lists the annotations of all the cDNA libraries in which each gene is expressed.

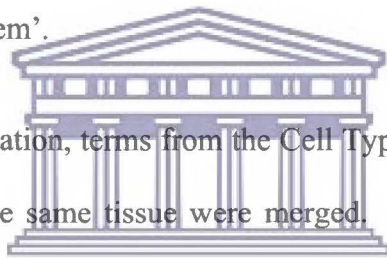


#### **2.4.2 Pseudoarray generation and expression filtering**

The gene expression profiles were converted into a binary pseudoarray by listing the genes in the first column and all annotations in the first row of a table. If a gene is annotated with a term, the value in the array corresponding with that gene and term is '1'. Similarly, if a gene is not annotated with a term, the value in the

array is '0', creating a binary code for presence ('1') and absence ('0') of expression of a gene across a list of tissues represented by ontology terms.

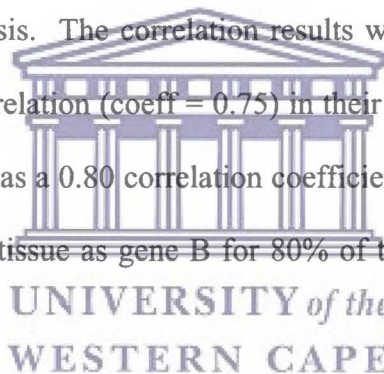
The pseudoarray was filtered for annotations resulting from cDNA libraries derived from normal tissues. A library is considered to be from normal tissue only if the annotation explicitly states 'normal'. Annotations were discarded where the originating tissue samples were pooled or if the Anatomical System term was 'unclassifiable', indicating the sample was from an unknown tissue type. In addition, the developmental stage information was removed and identical terms from different stages were merged. Terms were collated if they were located on the same branch of a hierarchy, eg. ovary and uterus were collated and renamed 'female reproductive system'.



To avoid redundant annotation, terms from the Cell Type and Anatomical System ontologies referring to the same tissue were merged. The terms 'macrophage', 'lymphocyte' and 'bone marrow' were merged with 'blood', 'lymph' and 'bone', respectively. Due to ubiquitous expression, all terms relating to 'brain' were removed, and the following terms were collated as 'other': adipose tissue, auditory apparatus, bladder, cartilage, gall bladder, gastrointestinal tract, larynx, muscle, omentum, oral cavity, pharynx, skeletal muscle, skin, spinal cord, synovium, tonsil, umbilical cord and visual apparatus. In order to explore tissue-restricted expression, genes were further filtered based on the number of terms to which they are annotated. Only genes expressed in less than 25% of tissues were used for further analysis.

### 2.4.3 Expression clustering

To determine genes exhibiting similar expression patterns, the correlation coefficient of each gene pair was calculated. A correlation coefficient describes the strength of a linear relationship between two variables and has a value between '-1' (negatively correlated) and '1' (positively correlated). The correlation coefficients were calculated computationally by means of the *numpy* module of the Python scripting language. Genes showing no correlation in their expression have a correlation coefficient '0' and genes whose expression are perfectly correlated have a correlation coefficient '1'. Since the aim of the study was to find co-expressing transcription factors, negatively correlated genes were not included in the analysis. The correlation results were filtered for gene pairs showing at least 75% correlation (coeff = 0.75) in their expression. For example, if a gene pair (A and B) has a 0.80 correlation coefficient, it indicates that gene A is expressed in the same tissue as gene B for 80% of the time, indicating a high degree of co-expression.

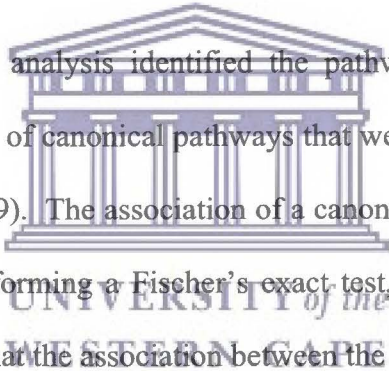


Genes were defined as clustering together in a network if a node (gene) is connected to another node (corresponding gene pair) by an edge (correlation coefficient  $\geq 0.75$ ). The nodes and edges resulting from the expression correlation calculations were visualised using the Cytoscape network and visualisation tool (Shannon et al., 2003).



#### 2.4.4 Functional analysis

The list of tissue-restricted genes was analysed through the use of Ingenuity Pathway Analysis (IPA) version 7.5 (<http://www.ingenuity.com>). The set of genes was uploaded into the application as a list of Entrez Gene identifiers. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base. The Functional Analysis component of the application identified the biological functions and diseases that were most significant to the data set. A Fischer's exact test was used to calculate a p-value determining the probability that each biological function and disease assigned to that data set is random.



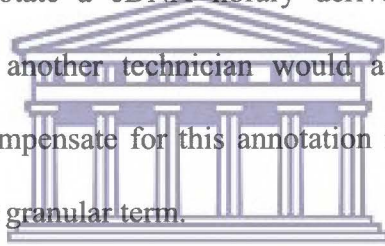
The Canonical Pathways analysis identified the pathways from the Ingenuity Pathways Analysis library of canonical pathways that were most significant to the data set (as at August 2009). The association of a canonical pathway and the data set was measured by performing a Fischer's exact test, calculating a p-value to illustrate the probability that the association between the pathway and genes in the data set is due to chance.



## 2.5 Results and discussion

### 2.5.1 Data generation and expression profiling

Of the 1 805 genes in the TF-list, 60 genes were not represented by the cDNA libraries in the eVOC ontology system. The remaining 1 745 genes were represented by 239 unique annotation tuples, where a tuple is a list of four terms (one from each ontology) representing a cDNA library. For example, the tuple representing a cDNA library obtained from the epithelial cells of a normal fetal kidney is 'kidney|epithelial cell|fetus|normal'. Due to the hierarchical nature of an ontology, libraries are often annotated with differing granularity. For example, one technician may annotate a cDNA library derived from hippocampus as 'hippocampus', whereas another technician would annotate the same cDNA library as 'brain'. To compensate for this annotation inconsistency, terms were merged to reflect the least granular term.



UNIVERSITY of the  
WESTERN CAPE

The merging and removal of terms resulted in 1 734 genes represented by 21 ontology terms. To determine which genes showed tissue-restricted expression, the genes were further filtered based on the number of tissues in which they are expressed. Table 1 lists the 145 genes that are expressed in less than 25% of the tissues represented by the 21 ontology terms. It should be noted that, as with most analyses, the results obtained here might be subjected to a data bias. Since only one expression source (namely ESTs) is used, it is possible that the expression of certain genes were not captured. Although the focus of this study is the development of a method to determine tissue-restricted expression factors, the

**Table 1**

**A list of the 145 genes expressed in less than 25% of all tissues. The table consists of two panels, each listing the Entrez gene identifier and gene symbol for the human transcription factors showing tissue-restricted expression.**

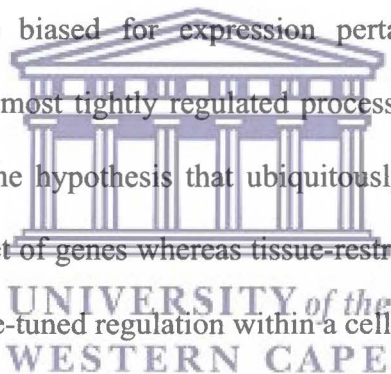
GeneID	GeneSymbol	GeneID	GeneSymbol
326	AIRE	8345	HIST1H2BH
430	ASCL2	8820	HESX1
579	BAPX1	8970	HIST1H2BJ
668	FOXL2	9970	NR1I3
1032	CDKN2D	10215	OLIG2
1053	CEBPE	10655	DMRT2
1745	DLX1	10794	ZNF272
1746	DLX2	11077	HSF2BP
1748	DLX4	11281	POU6F2
1761	DMRT1	25806	VAX2
1961	EGR4	26038	CHD5
1993	ELAVL2	26108	PYGO1
2016	EMX1	26468	LHX6
2020	EN2	27023	FOXB1
2103	ESRRB	27164	SALL3
2118	ETV4	27288	HNRNPG-T
2294	FOXF1	27439	CECR6
2295	FOXF2	30009	TBX21
2297	FOXD1	30012	TLX3
2302	FOXJ1	50805	IRX4
2304	FOXE1	51022	GLRX2
2306	FOXD2	51402	LWI
2623	GATA1	51450	PRRX2
2672	GFI1	54626	HES2
3007	HIST1H1D	55552	HSZFP36
3008	HIST1H1E	55659	ZNF416
3009	HIST1H1B	56938	ARNTL2
3110	HLXB9	56978	PRDM8
3198	HOXA1	57116	ZNF695
3205	HOXA9	57332	CBX8
3207	HOXA11	57343	ZNF304
3209	HOXA13	57801	HES4
3231	HOXD1	58495	OVOL2
3234	HOXD8	60529	ALX4
3642	INSM1	63978	PRDM14
3975	LHX1	79192	IRX1
4210	MEFV	79722	FLJ11795
4656	MYOG	79816	TLE6
4796	NFKBIL2	79862	ZNF669
4821	NKX2-2	80032	ZNF556

GeneID	GeneSymbol	GeneID	GeneSymbol
4861	NPAS1	84127	RUNDC2A
4901	NRL	84911	ZNF382
5013	OTX1	85409	NKD2
5076	PAX2	85446	ZFHX2
5077	PAX3	89870	TRIM15
5079	PAX5	90649	ZNF486
5081	PAX7	94039	ZNF101
5453	POU3F1	94234	FOXQ1
5454	POU3F2	116448	OLIG1
5455	POU3F3	126295	LOC126295
5462	POU5F1P1	129025	SUHW1
5992	RFX4	136051	DKFZp762I137
6474	SHOX2	138474	TAF1L
6493	SIM2	140883	SUHW2
6496	SIX3	142689	ASB12
6664	SOX11	146434	ZNF597
6689	SPIB	148268	ZNF570
6877	TAF5	148979	GLIS1
6899	TBX1	161253	FLJ38964
6913	TBX15	162979	ZNF342
7023	TFAP4	163059	ZNF433
7161	TP73	163071	ZNF114
7291	TWIST1	170302	ARX
7310	U2AF1L1	171392	ZNF675
7546	ZIC2	221527	ZBTB12
7621	ZNF70	245806	VGLL2
7673	ZNF222	253738	EBF3
7675	ZNF121	283078	MKX
7710	ZNF154	285676	ZNF454
7768	ZNF225	339416	ANKRD45
8092	CART1	339488	TFAP2E
8193	DPF1	341405	ANKRD33
8320	EOMES		

addition of data sources such as CAGE, MPSS and SAGE will dramatically increase the quality of the results.

Not surprisingly, more than 80% of the restricted genes are regulators of gene expression according to their Gene Ontology annotations. In addition, a small percentage of the restricted genes are involved in immune system development (*BAPX1*, *TBX21* and *SPIB*), embryonic development (*EOMES*, *OTX1*, *BAPX1*, *FOXE1*, *HOXD8*, *SIM2*, *FOXF1*, *LHX1*, *VAX2*, *FOXF2*, *TRIM15*, *GF11*, *ASCL2*, *FOXL2*, *TBX1* and *ZIC2*) and cell fate specification (*NKX2-2*, *TLX3* and *GF11*).

The pseudoarray illustrating the expression profiles of these genes is represented by Appendix VIII. It is interesting to note that these genes showing tissue-restricted expression are biased for expression pertaining to developmental processes – probably the most tightly regulated processes in an organism. This observation strengthens the hypothesis that ubiquitously expressed transcription factors regulate a broad set of genes whereas tissue-restricted transcription factors are responsible for the fine-tuned regulation within a cell.



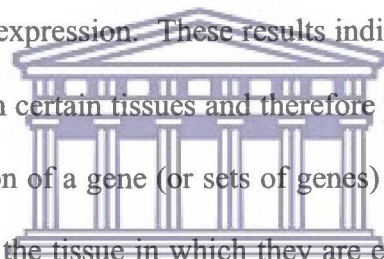
### **2.5.2 Expression clustering**

The current knowledge of transcription factor function suggests that they function as protein complexes, indicating that the functional and expression profiling of a single transcription factor is unuseful. In order to determine how transcription factors regulate gene expression, it is important to determine which transcription factors function together. The correlations of gene expression profiles were



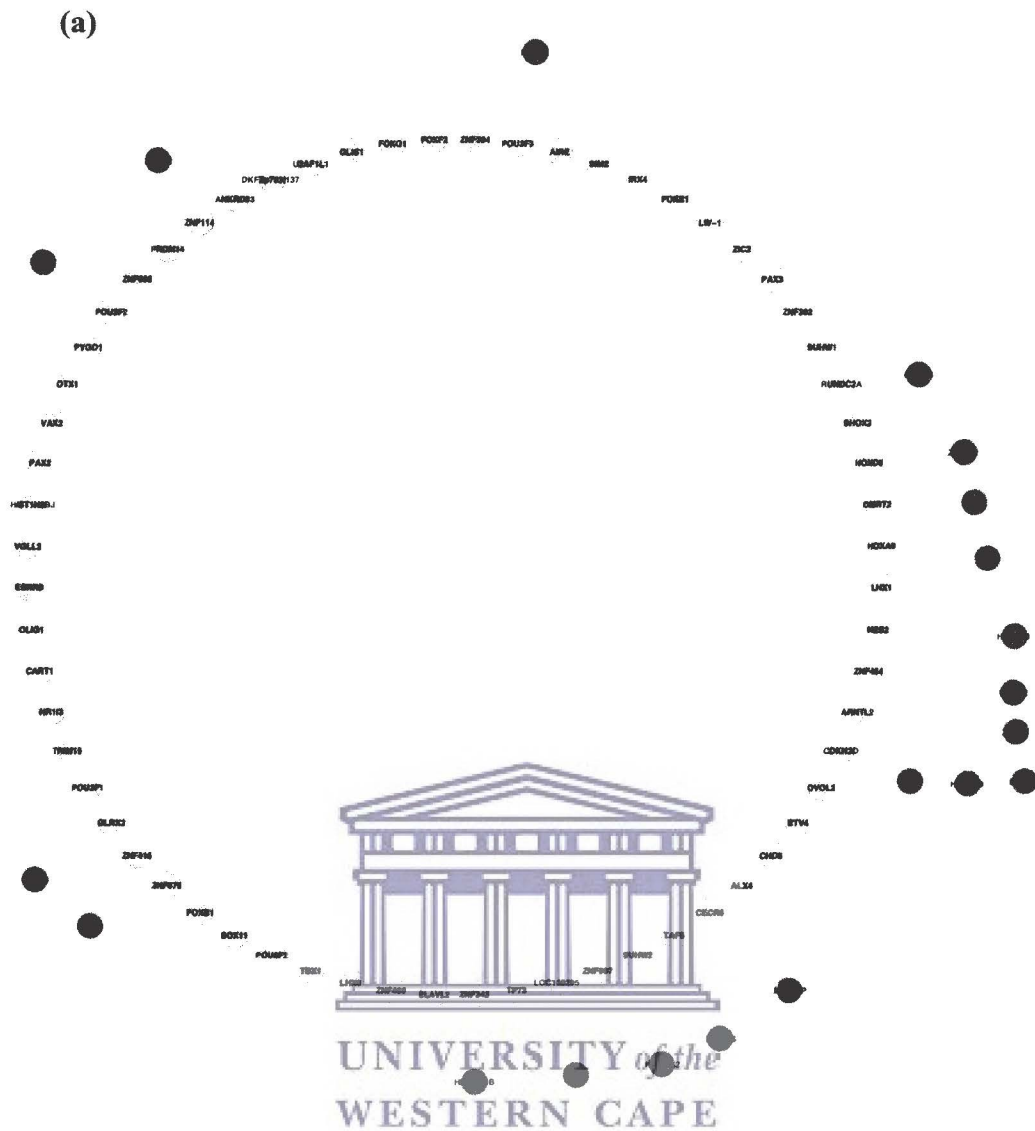
determined in order to assess which genes co-express across a range of tissues. The co-expression of transcription factors implicates their involvement in the co-regulation of their target genes, providing the basis for further functional studies.

A moderate correlation cutoff of 75% resulted in 112 genes represented by 8 gene clusters. Genes clustered together if there was at least one edge (correlation coefficient  $\geq 0.75$ ) between two genes. Not surprisingly, the results show one large gene cluster (Figure 1a) with a few smaller clusters (Figure 1b). Investigations of the annotations of the genes in Figure 1b reveal a few clusters (3, 4 and 5) that exhibit tissue-restricted expression for female reproductive system, male reproductive system and stem cell, respectively. In addition, clusters 6, 7 and 8 show tissue-biased expression. These results indicate that the genes in each cluster are co-expressed in certain tissues and therefore possibly function as a unit to activate the transcription of a gene (or sets of genes) responsible for the tissue-specific characteristics of the tissue in which they are expressed. For example, it is feasible that because the genes in cluster 5 (*DLX2*, *BAPX1* and *ZBTB12*) co-express only in the stem cell population that these transcription factors may be responsible for regulating the genes that define stemness (self-renewal, chemoresistance, pluripotency). Since we see transcription factors biased for expression in tissues that have developmental functions (female reproductive system, male reproductive system and stem cell), we can intuitively predict that the corresponding transcription factors play a role in the regulation of the development of the cell. It is even possible, given the tissues in which these genes are restricted, that they regulate the stem cell state of a cell since the male and female reproductive system has stem cell-containing tissues. The tissues



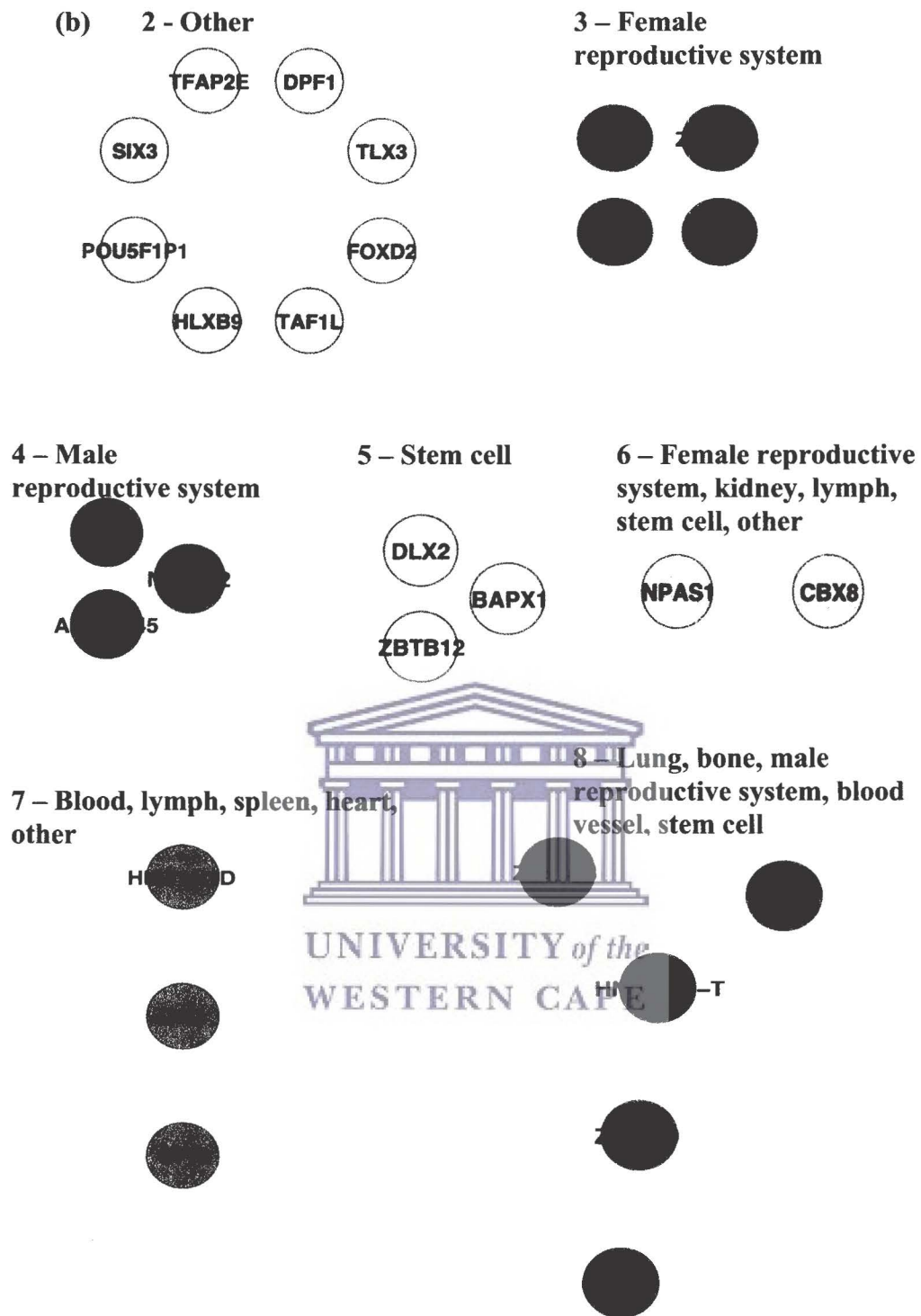
UNIVERSITY of the  
WESTERN CAPE





**Figure 1a**

**Illustration of genes clustering together based on correlated co-expression. All gene clusters represent the sets of genes that cluster together based on a correlation coefficient larger than 0.75.**

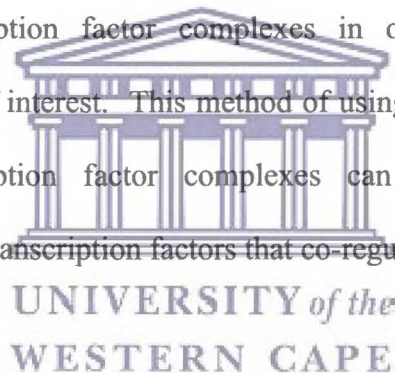


**Figure 1b**

**Illustration of genes clustering together based on correlated co-expression. All gene clusters represent the sets of genes that cluster together based on a correlation coefficient larger than 0.75. Clusters 2 – 8 represents genes and tissues for which there is biased expression.**

represented by the tissue-biased clusters (lung, bone, kidney, heart, lymph and blood) also have a stem cell niche with cells progressing through a defined cell lineage.

Although the above statements require experimental validation, what we see here is the identification of several complexes of transcription factors that show an expression bias towards certain tissues and therefore possibly interact with each other to combinatorially regulate a defined set of target genes. It is possible that the addition or omission of even one transcription factor in a complex may alter the regulation of a gene not only quantitatively, but also on a temporal and spatial level. It is for this reason that it is important for researchers to determine the composition of transcription factor complexes in order to understand the regulation of any gene of interest. This method of using ontologies to determine tissue-restricted transcription factor complexes can therefore be used to computationally predict transcription factors that co-regulate a set of genes.



### **2.5.3 Functional analysis**

A functional analysis of a list of genes reveals processes with which the genes are associated, thereby giving insight into the processes governing a particular cell type or state. The functional analysis of the 145 transcription factors that exhibit a restricted expression profile suggests a functional bias towards developmental processes. Table 2a lists the top five physiological functions associated with the restricted gene set, showing a significant enrichment for the development of

**Table 2a**

**The top five physiological system development and functions over-represented by genes showing restricted expression.**

Physiological System Development and Function	P-value
Organ development	4.73E-15 - 1.57E-02
Nervous System Development and Function	1.33E-10 - 2.34E-02
Lymphoid Tissue Structure and Development	3.60E-07 - 2.12E-02
Digestive System Development and Function	1.83E-04 - 1.83E-04
Organismal Development	2.92E-04 - 2.92E-04



UNIVERSITY *of the*  
WESTERN CAPE

organs and the organism as a whole. Investigation into the top five diseases associated with the data set shows that cancer is significantly over-represented (Table 2b). In addition, analysis of the canonical pathways suggests the Sonic Hedgehog Signaling pathway as the most significantly over-represented pathway by the data set (Table 3) with a p-value of  $1.99 \times 10^{-01}$ . Although the p-value presented here does not fall below the accepted 0.005, it does support the findings presented in 2.5.2. The p-value obtained from enrichment analyses is influenced by the size of the gene list being investigated, where a larger gene list will have a higher statistical power resulting in more significant p-values. Even so, the order of enriched terms will remain fairly stable regardless of the size of the gene list, provided the lists of different sizes are being sampled from the same data set (Huang da et al., 2009). We can therefore argue that the Hedgehog pathway is significantly over-represented even though a high p-value is obtained, since it is most likely a result of having a small gene list. The Hedgehog pathway is a key regulator of embryonic development and is highly conserved from insects to mammals. Altered Hedgehog pathway activity can lead to certain cancers such as basal cell carcinoma. There is also increasing evidence that this pathway is involved in regulating adult stem cells (Bhardwaj et al., 2001) and over-representation of this pathway is associated with proliferation and development (Kenney et al., 2003).

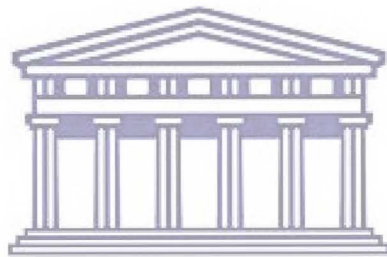
The over-representation of developmental functions, diseases and canonical pathways in the data set is strong evidence that the transcription factors showing a tissue-restricted expression bias are those factors that are responsible for the fine-tuning of the regulation of developmental gene expression. These tissue-restricted



**Table 2b**

**The top five diseases and disorders associated with the genes showing restricted expression in less than 25% of all tissues.**

<b>Diseases and Disorders</b>	<b>P-value</b>
Developmental Disorder	2.99E-03 - 3.88E-02
Antimicrobial Response	7.87E-03 - 7.87E-03
Cancer	7.87E-03 - 3.88E-02
Dermatological Diseases and Conditions	7.87E-03 - 3.88E-02
Endocrine System Disorders	7.87E-03 - 7.87E-03

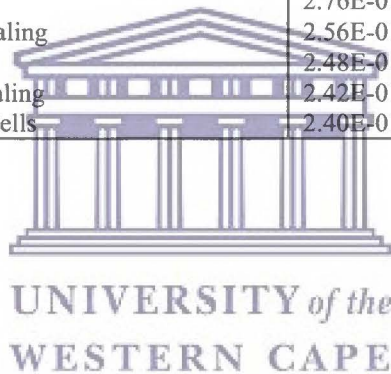


UNIVERSITY *of the*  
WESTERN CAPE

**Table 3**

**A list of canonical pathways over-represented by genes showing restricted expression in less than 25% of all tissues.**

<b>Ingenuity Canonical Pathways</b>	<b>-Log(P-value)</b>
Sonic Hedgehog Signaling	7.02E-01
Estrogen Receptor Signaling	6.92E-01
Allograft Rejection Signaling	6.16E-01
T Helper Cell Differentiation	5.95E-01
Autoimmune Thyroid Disease Signaling	5.95E-01
Graft-versus-Host Disease Signaling	5.76E-01
Dendritic Cell Maturation	5.07E-01
ATM Signaling	4.79E-01
TREM1 Signaling	4.58E-01
Basal Cell Carcinoma Signaling	4.11E-01
PXR/RXR Activation	4.06E-01
Caveolar-mediated Endocytosis	3.71E-01
CTLA4 Signaling in Cytotoxic T Lymphocytes	3.24E-01
Melanocyte Development and Pigmentation Signaling	3.13E-01
Virus Entry via Endocytic Pathways	3.10E-01
p53 Signaling	3.02E-01
Glioma Signaling	2.76E-01
Type I Diabetes Mellitus Signaling	2.56E-01
14-3-3-mediated Signaling	2.48E-01
Glucocorticoid Receptor Signaling	2.42E-01
CD28 Signaling in T Helper Cells	2.40E-01



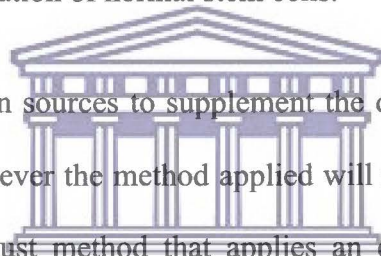
transcription factors may therefore also be implicated in the development of cancers and developmental disorders originating from a dysregulation of genes in a cell.

## 2.6 Conclusions

This study explored the expression profiles of a list of transcription factors known to localise in the nucleus. The aim of the study was to determine which transcription factors show tissue-restricted expression. The use of an ontology-based system enabled the identification of 145 transcription factors whose expression was limited to less than 25% of the 21 tissues represented by the dataset. Investigation of the results revealed that the tissue-restricted transcription factors are involved in developmental processes such as immune system development, embryonic development and cell fate specification. The Sonic Hedgehog Signaling pathway was the most significantly over-represented pathway in the data set, providing further evidence of a significant role of these genes in the development of an organism. In addition, the tissues in which the transcription factors showed biased expression are those tissues in which cells are continuously re-generating, indicating that these transcription factors may play a crucial role in the regulation of the progression of a cell down a defined cell lineage.

It is becoming increasingly apparent that transcription factors do not function individually, but rather as complexes. The identification of co-expressing

transcription factors will therefore be able to make an initial identification of transcription factor complexes. Clustering tissue-restricted genes based on a 75% correlation of their expression enabled the identification of 3 transcription factor complexes showing tissue-restricted (expressed in one tissue only) expression and 3 complexes showing tissue-biased (expressed in a limited number of tissues) expression patterns. The three clusters showing tissue-restricted expression represent the male and female reproductive systems as well as stem cells. We have therefore potentially identified transcription factor complexes that are involved in the regulation of the development of the cell and further investigation of the transcription factors represented by these clusters may contribute to the understanding of the regulation of normal stem cells.



The addition of expression sources to supplement the dataset used here will add quality to the results, however the method applied will not be affected. We have therefore described a robust method that applies an ontology-based system to enable the identification of transcription factor complexes that may be used to identify transcription factor complexes that function in specific tissues thereby enhancing the understanding of the regulatory potential of genes of interest.

## Chapter 3

### **Mouse gene expression analysis of cancer/testis orthologs restricts candidates for cancer therapy.**

#### **3.1 Summary**

The work presented in this chapter was conducted as part of a project aimed at characterising cancer/testis genes in human and mouse. The overall objectives of the project are fourfold:

1. Characterise, and possibly re-classify, all known human cancer/testis genes;
2. Identify novel human cancer/testis genes by means of expression profiling;
3. Identify which cancer/testis genes are most suited for developing cancer drugs or vaccines; and
4. Identify mouse cancer/testis genes to use as a model system for cancer drug and vaccine development.

Objectives (1) and (2) resulted in a publication (Hofmann et al., 2008), wherein my contribution was to:

- a) use the ontologies presented in Chapter 1 to annotate a list of human cancer/testis genes and their mouse orthologs; and
- b) maintain and implement the data-generation pipeline developed by Dr Christopher Maher and Dr Oliver Hofmann.



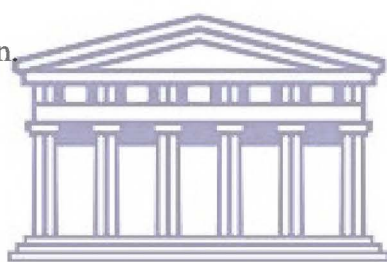
The mouse expression information in (a) was not used in the publication due to the observation that the expression profiles of the orthologs did not conform to expected cancer/testis criteria and further investigation was required (subsequently resulting in this chapter). The human expression information was merged with expression data derived from MPSS, qRT-PCR and CAGE expression data in order to perform a multi-platform expression analysis in the attempt to re-classify human cancer/testis genes. The pipeline in (b) is a sequence of computer scripts coded in Perl, which requires raw CAGE sequence information (Kodzius et al., 2006) as input. CAGE tags are short 10-12bp fragments derived from the 5' coding region of an mRNA and, when mapped to the genome, accurately identifies the point of transcription initiation (transcription start site – TSS). The pipeline orders the CAGE tags according to chromosome and strand, and subsequently clusters the tags to provide quantitative evidence for transcription initiation. When annotated according to the ontology-based system described in Chapter 1, this information provides tissue-based transcription initiation events. When combined with the cDNA library information from the eVOC system as well as qRT-PCR and MPSS data, a genome-wide analysis identified genes whose expression profile classifies them as cancer/testis genes, thereby identifying novel CT genes in human. This work is discussed in detail in ‘Genome-wide analysis of cancer/testis gene expression’ published in PNAS (Hofmann et al., 2008), which is appended as Appendix IX.

This chapter describes objective (4), where my role was to develop, implement and interpret the analysis. The results of this study will be used to make informed decisions regarding the use of mouse as model system for investigation of

cancer/testis genes, and to further understand the relationship between human and mouse cancer/testis orthologs.

### 3.2 Aim

The aim of the analysis presented here is to determine whether the mouse orthologs of the human cancer/testis (CT) gene set exhibits CT characteristics. Since CT genes are a target for gene-based cancer drug therapy, and the development of these drugs includes efficacy and toxicity trials in mouse, it is important to identify human target genes whose mouse counterpart show the same tissue-restricted expression.



### 3.3 Introduction

UNIVERSITY *of the*  
WESTERN CAPE

Cancer is a disease characterised by the uncontrolled growth of cells in any of a variety of tissues such as breast, prostate, lung, liver and pancreas (Jemal et al., 2008). Cancer is an invasive disease and can migrate to different parts of the body. Although there are hundreds of cancer types, they typically fall into one of five categories (leukemia, sarcoma, carcinoma, lymphoma/myeloma, and central nervous system cancers), depending on their tissue of origin. Leukemia is cancer that originates in the bone marrow where blood is formed, resulting in the production of a large number of abnormal blood cells. The sarcoma cancers develop in the connective and supportive tissues such as bone, muscle or fat.

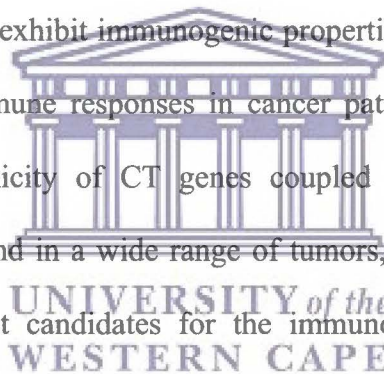
Carcinoma is referred to cancer originating in the skin or in the tissue lining the internal organs. The lymphoma and myeloma cancers originate in the immune system, whereas the central nervous system cancers develop in the brain and spinal cord (<http://www.cancer.gov/cancertopics/what-is-cancer>). In addition, cancers may be classified as either benign (non-metastasizing, non-invasive, non-aggressive) or malignant (metastasizing, invasive, aggressive) tumors, the latter being the most cause of concern.

In 2004, cancer was responsible for the deaths of 7.4 million people worldwide and it is estimated that this figure will rise to 12 million in the year 2030 (<http://www.who.int/en/>). The exact origin of cancer is the topic of much research, however the consensus is that tumorigenic cells have altered genomes compared to normal cells, resulting in aberrant gene expression, function and cellular growth (Bos, 1989). The two main theories for the origin of cancer are the clonal evolution model and the cancer stem cell theory (Gil et al., 2008). The clonal evolution model suggests that a cell acquires a series of mutations during the process of cell division. The cancer stem cell model states that only stem cells proliferate enough times to accumulate cancer-causing mutations and that it is these cells that gives rise to tumors. The cancer stem cell population is a subset of the tumor that possesses the self-renewal and multipotent qualities of normal stem cells.

The cancer stem cell theory suggests that if the cancer stem cell population is not removed from the tumor, the patient will experience a tumor relapse. Conventional cancer therapy includes surgery to excise the tumor followed by

chemo- or radiation-therapy to kill all replicating cells. Since cancer stem cells exhibit intolerance to chemotherapy (Gil et al., 2008) these conventional therapies are not only invasive but potentially ineffective as well. Current research focusing on cancer therapy is therefore aimed at identifying genes expressed specifically in tumors and not in normal tissues, enabling the production of drugs or vaccines to target cells that have become tumorigenic.

Cancer/testis (CT) genes are a group of genes whose expression has been observed in a variety of different tumors (Chitale et al., 2005). However, when observed in normal tissues, the expression of CT genes is limited to the immunoprivileged tissues of testis, ovary and/or placenta (Cho et al., 2006). In addition, many CT genes exhibit immunogenic properties, enabling them to elicit cellular and humoral immune responses in cancer patients (Atanackovic et al., 2006). The immunogenicity of CT genes coupled with their expression in immunoprivileged sites and in a wide range of tumors, allows these genes to be considered as drug target candidates for the immunotherapeutic treatment of cancer.



As with many pharmaceutical products, the process of creating drug targets requires the use of model systems in which to test drugs before being declared fit for clinical trials. Although the mouse is a common model system for studying biological reactions to chemical additives, it is not guaranteed that the human response will be identical. Orthologous genes may be expressed in both human and mouse, but due to different regulators their expression does not necessarily occur on the same temporal and spatial level (discussed in Chapter 1), affecting



their eventual function. For this reason it is important to identify mouse CT genes and to understand their relationship to human orthologs for the development of drug targets for cancer therapy.

### **3.4 Materials and methods**

#### **3.4.1 Data selection and generation**

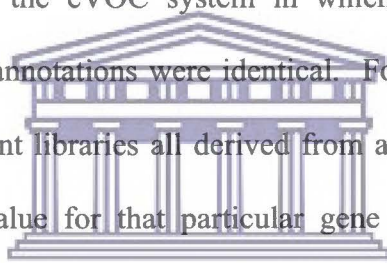
A list of 181 human cancer/testis (CT) genes was obtained from the CT Antigen Database (April, 2009) (<http://www.cta.lncc.br>). The mouse orthologs of the human CT genes were obtained by matching HomoloGene identifiers (as presented in Chapter 1) resulting in only 70 mouse genes. Information for the generation of gene expression profiles of the mouse orthologs was extracted from 1 210 cDNA libraries in the eVOC system (Chapter 1). A gene was annotated with the anatomical, cellular, developmental and pathological terms associated with a library if the gene was found to be expressed in that particular library. In the cases where anatomical terms were not available, terms relating to cell type were used.

Only libraries that were annotated as having normal pathology were categorised as 'normal', whereas all other libraries not explicitly annotated as such were categorised as 'unclassifiable' in terms of pathology. Libraries comprising of more than one sample were excluded from the analysis unless all the samples were obtained from the same anatomical structure under identical pathological conditions.



### 3.4.2 Expression profiling

The expression information generated in 3.3.1 was organised in the form of an array. An expression array consists of a list of genes in the first column of a table, with the first row consisting of all possible annotations from the expression sources. The annotations are a combination of developmental stage, pathology and anatomical structure (or cell type) for each library used. For example, an annotation for a cDNA library obtained from the normal heart of an adult mouse would be 'adult|normal|heart'. The values for the array were based on the number of cDNA libraries from the eVOC system in which a gene was expressed, summing libraries if the annotations were identical. For example, if a gene was expressed in three different libraries all derived from a normal heart of an adult mouse, the expression value for that particular gene with 'adult|normal|heart' annotation would be 3.



UNIVERSITY of the  
WESTERN CAPE

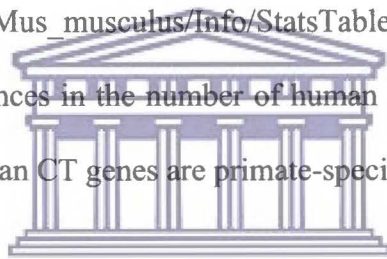
The expression array was subsequently filtered to disregard developmental stage information, remove annotations where the pathology was neither cancer nor normal, and merge terms related in terms of hierarchical structure. Appendix X lists the manual filtering steps performed on the data. A total of 7 genes were not represented by the data and were subsequently removed from the analysis.

Based on the expression profiles derived, genes were classified into three categories: (i) testis-restricted; (ii) testis/brain-restricted; (iii) testis-selective (see

Table 1 for classification and Figure 1 for a flow-diagram describing the categorisation process).

### 3.5 Results and discussion

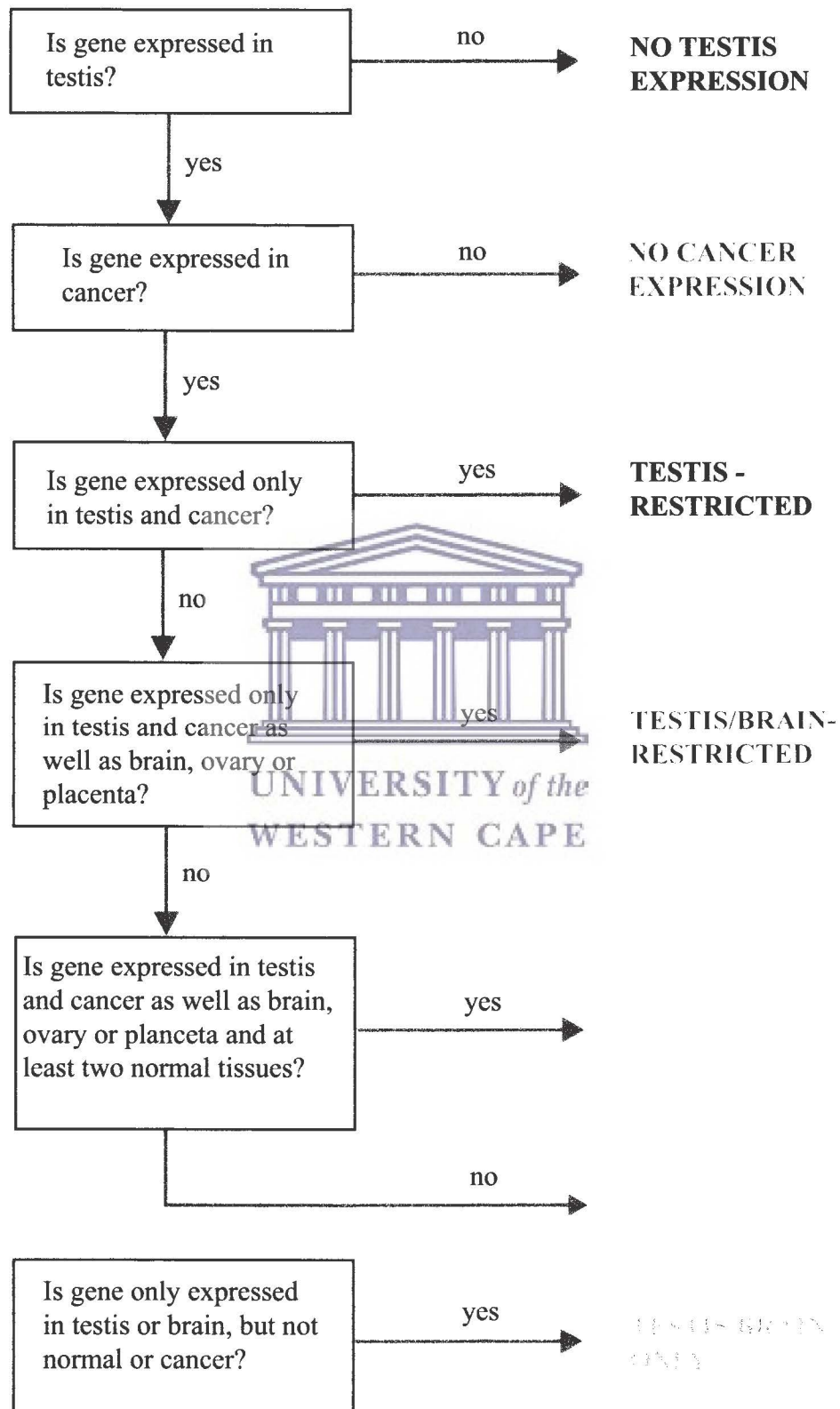
Of the 181 human CT genes, only 70 have mouse orthologs according to the HomoloGene database (April, 2009) (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene>). Although 80 - 99% of mouse genes have human orthologs (discussed in Chapter 1), these percentages still represent between 300 – 6 000 of the estimated 30 000 genes in the mouse genome (NCBI m37, Apr 2007) ([http://www.ensembl.org/Mus\\_musculus/Info/StatsTable](http://www.ensembl.org/Mus_musculus/Info/StatsTable)), thereby easily accounting for the differences in the number of human and mouse CT genes. In addition, many of the human CT genes are primate-specific.



The data filtering process involved removing annotations where the pathology is unclassifiable as well as disregarding developmental stage information. The filtering process is important as it discards genes whose origin is unknown and their expression can therefore not be specifically designated as ‘normal’ or ‘cancer’. The developmental stage information is discarded because there is simply not enough data for each developmental stage to be a category on its own. Terms such as cerebellum and brain that are related in the eVOC hierarchy were merged to reflect the least granular term, resulting in 63 genes represented by 76 unique annotations consisting of 58 normal- and 18 cancer-related annotations. Unfortunately, the filtering of data resulted in 4 genes being excluded from the

**Figure 1**

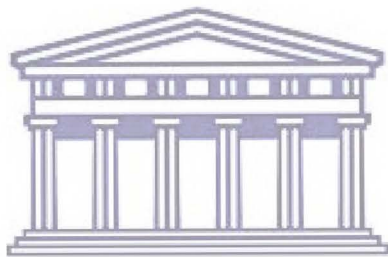
**Flow-diagram representing the categorisation of mouse genes into cancer/testis categories.**



**Table 1**

**Classification categories for cancer/testis genes. Testis- and testis/brain-restricted genes are those biased for expression in immunoprivileged tissues.**

<b>Category</b>	<b>Classification</b>
Testis-restricted	expression in cancer and testis only
Testis/brain- restricted	expression in cancer, testis, placenta, ovary and brain-regions only
Testis-selective	expression in cancer, testis and two other tissues



UNIVERSITY *of the*  
WESTERN CAPE

analysis since they did not have any expression evidence in the remaining cDNA libraries. Although this process results in a loss of data, it increases the confidence of the remaining genes in that they have definite expression in 'normal' and 'cancer' tissues.

The resulting expression profile showed that 4 of the 70 genes were not found to be expressed in a testis library at all (*Il13ra2*, *Ccdc36*, *Otoa* and *Magea8*). There were 0 genes categorised as testis-restricted, 2 classified as testis/brain-restricted (*Syce1* and *Tssk6*) and 7 classified as testis-selective (*Morc1*, *Spa17*, *Dkk11*, *Plac1*, *Piwil2*, *Ly6k* and *Ssxb2*). In addition, there were 17 genes expressed in testis, brain, ovary or placenta but not in normal or cancer tissues. Because these genes are not expressed in cancer, they are not classified as cancer/testis genes.

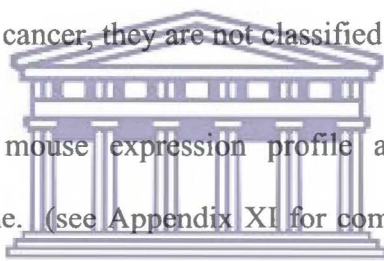


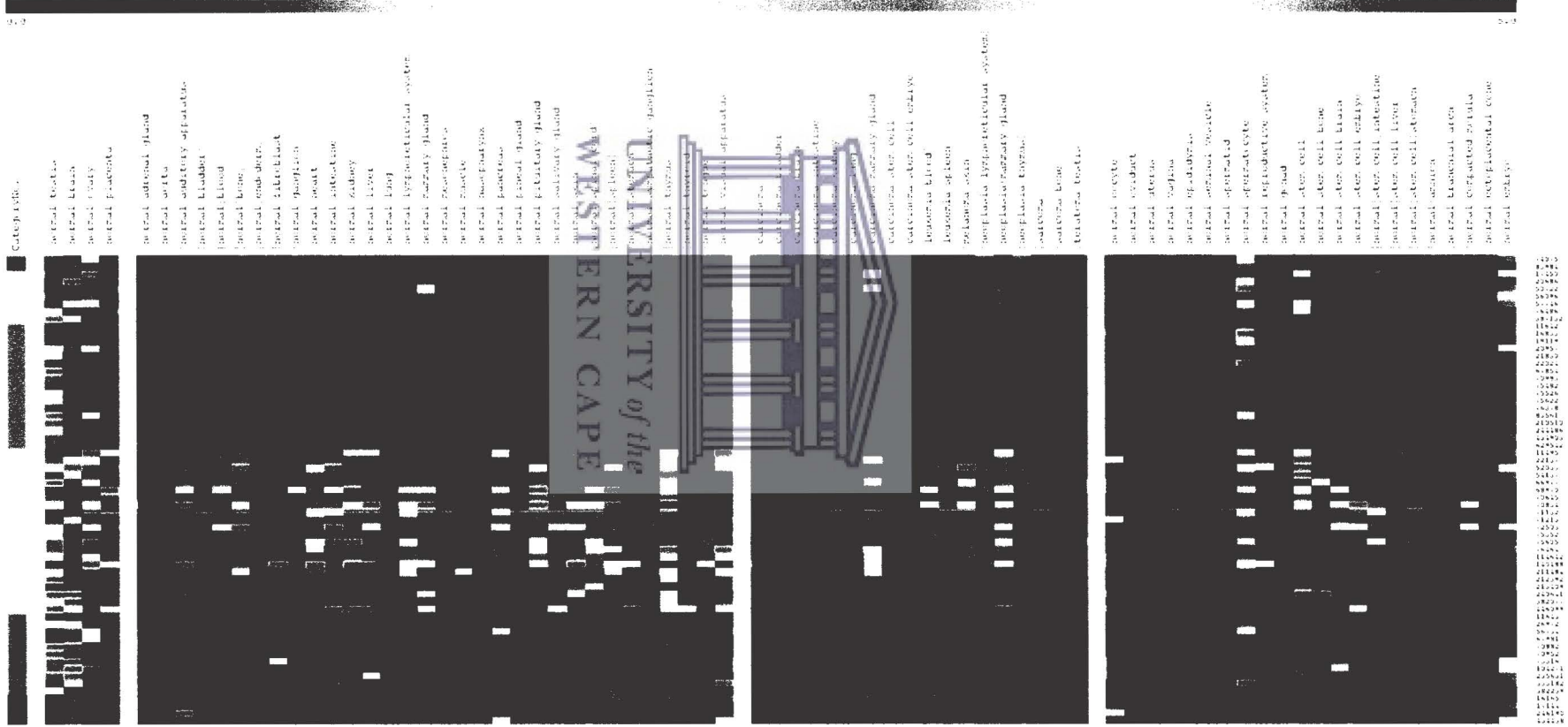
Figure 2 illustrates the mouse expression profile as well as the resulting categorisation of each gene. (see Appendix XI for complete expression profile). The first panel of Figure 2 (Category No.) represents the CT category each gene was categorised as. The second panel represents normal testis, brain, ovary and placenta expression. The third and fourth panels represent normal and cancer expression, respectively. The fifth panel represents expression derived from normal tissues relating to the reproductive system (eg. oocyte and spermatocyte) and stem cells, and were not included in the CT categorisation process. Table 2 provides the testis-restricted, testis/brain-restricted and testis-selective genes along with their human orthologs.

The results are inevitably subject to data bias since the data set is derived from one data type from a single origin and it is therefore possible that some genes are



Figure 2

Visualisation of the gene expression profile of 63 mouse orthologs. The coloured blocks within the array refer to the number of cDNA libraries a gene is expressed in (0 = black; 5 = red). Genes are ordered from top to bottom according to their CT classification (testis/brain-restricted = red; no testis expression = black).



**Table 2**

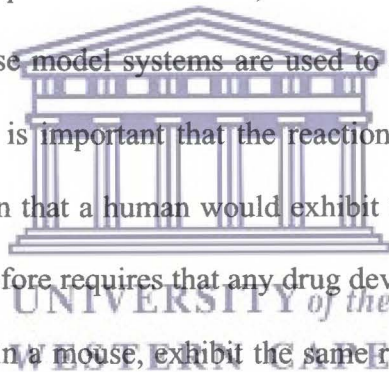
**Gene identifiers and symbols of mouse genes showing testis-restricted, testis/brain-restricted or testis-selective expression, along with their human orthologs.**

Mouse			Human	
GeneID	GeneSymbol	MouseCTcategory	GeneID	GeneSymbol
74075	Syce1	testis/brain	93426	SYCE1
83984	Tssk6	testis/brain	83983	TSSK6
17450	Morc1	testis-selective	27136	MORC1
20686	Spa17	testis-selective	53340	SPA17
50722	Dkk11	testis-selective	27120	DKKL1
56096	Plac1	testis-selective	10761	PLAC1
57746	Piwil2	testis-selective	55124	PIWIL2
76486	Ly6k	testis-selective	54742	LY6K
387132	Ssxb2	testis-selective	6756	SSX1



UNIVERSITY of the  
WESTERN CAPE

more likely to be included in the data set than others. The way in which to minimise the effects of data bias would be to include more data types from different sources. Although it is not presented here, the addition of data sources to the ontology system is strongly suggested. We can therefore not definitively conclude that the genes listed above are never expressed in testis or cancer and testis only. We can, however, illustrate that (a) there is evidence that these genes may not be expressed in testis and therefore possibly not classify as CT genes, and (b) genes that are considered testis-restricted in humans are showing a less-restrictive expression profile when expressed in mouse, which was the purpose of this study. We have therefore assessed the expression profiles of mouse genes whose orthologs, when expressed in humans, show a testis-restricted or testis-biased expression. Because model systems are used to determine the safety and efficacy of a trial drug, it is important that the reaction exhibited by the mouse closely reflects the reaction that a human would exhibit to the same drug. Gene-targeted drug therapy therefore requires that any drug developed to target a human gene should, when tested in a mouse, exhibit the same required response. When an ortholog does not show the same expression pattern in both human and mouse, there is a high probability that the gene performs a different function in each species. It is for this reason that we have set out to determine the expression profile of the mouse orthologs of the human CT genes and we have identified only 7 mouse genes whose expression profile characterises them as potential CT genes and therefore potential candidates for the development of gene-targeted drug therapies in mouse for eventual application in humans. In order for this work to make the transition from hypothetical to actual drug therapy, drugs may



be developed to specifically target the genes highlighted in this study. The ability for a drug to identify, target and destroy a cell expressing a gene characteristic of cancer and no other normal tissue will result in a non-invasive and highly effective means of treating and eradicating cancer.

### 3.6 Conclusions

The answer to effective cancer therapy lies in the ability to distinguish cancer from normal cells. The cancer/testis genes have proven to be promising candidates for drug targeted therapy due to their immunoprivileged properties. Despite the obvious importance of the cancer/testis genes in cancer therapy, these genes are not well characterised and therefore poorly understood. The use of a model system such as mouse provides an effective way to advance our knowledge of the cancer/testis genes. The problem however, is that it has been shown that the temporal and spatial gene expression of human and mouse orthologs differs greatly, emphasising the need to identify mouse CT gene orthologs. The analysis presented here highlights that the mouse orthologs of human CT genes are not necessarily CT genes themselves, and identifies only 7 mouse genes showing CT gene characteristics and have human CT counterparts. These findings provide realistic targets for drug-targeted cancer therapy and deeper characterization because they have, as a result of expression profiling, been identified as genes that potentially perform the same function due to identical expression and will therefore exhibit the same responses to chemical stimuli.



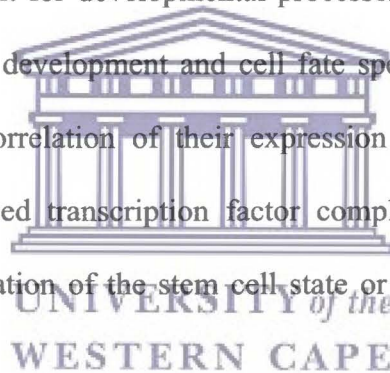
## Conclusions

I have demonstrated the need for an effective way to annotate expression sources such as cDNA libraries in order to allow the universal and computational comparison of the annotated data. The need for the comparison of data is not only limited to data derived from different laboratories, but also data derived from different species. I have addressed the issue of data comparison by developing a set of ontologies that describe human and mouse development. The ontologies are aligned not only between the two species, but also to other available ontologies, allowing the use of computational methods to compare human and mouse gene expression data across a range of sources. In addition, I have used the ontologies to annotate a set of 8 852 human and 1 210 mouse cDNA libraries as an initial dataset to showcase the ontologies.

The use of the ontologies has been demonstrated in several ways. Firstly, the ontologies have been used to compare the expression of human and mouse genes in the developing brain. It was found that of the 16 324 possible human-mouse orthologs, only 90 genes were expressed in the developing brain of both human and mouse. This finding highlights the differences in the temporal and spatial expression patterns of orthologous genes between the two species. I emphasise here that when using model organisms to study the behaviour of genes with the intention of inferring structural and functional information, it is important to establish that the genes of interest have similar spatial and temporal expression profiles in both species under investigation.



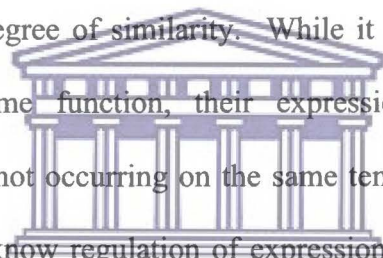
Secondly, the ontologies have been used to determine clusters of tissue-restricted transcription factors. A single gene may be expressed as several different transcripts in different tissues or under different conditions depending on the transcription factors binding to the promoter region of that gene. In addition, it has been found that transcription factors function in complexes and the composition of the transcription factor complexes differ between tissues as well as disease states. The identification of tissue-restricted transcription factors may therefore provide insight into the tissue- or disease-specific regulation of genes. The results from this analysis identified 145 human transcription factors showing a tissue-restricted expression pattern. Investigation into known functions of these genes revealed enrichment for developmental processes such as immune system development, embryonic development and cell fate specification. Clustering of these genes based on correlation of their expression profiles revealed tissue-restricted and tissue-biased transcription factor complexes that are potentially responsible for the regulation of the stem cell state or lineage differentiation of cells.



Lastly, the ontologies have been used to compare the expression profiles of a set of human cancer/testis genes in mouse. Of the 181 known human cancer/testis genes, only 70 have a mouse ortholog according to the HomoloGene database. Of these 70 mouse orthologs, only 63 have expression evidence in the system used. The human cancer/testis genes have been selected based on their biased expression for either testis and cancer, or testis, brain and cancer. The investigation of the 63 mouse orthologs show that 4 genes are not expressed in the testis at all and only 2 and 7 genes showed testis/brain-restricted and testis-

selective expression, respectively. Since the cancer/testis genes are considered extremely good candidates for the development of cancer drugs and vaccines, these findings emphasise the need to consider spatial and temporal differences in gene expression between human and model organisms when using the model organism to investigate the reaction of a set of genes to a drug or vaccine. This analysis also emphasises that mouse genes whose human orthologs are cancer/testis genes, are not necessarily cancer/testis genes themselves.

Each of the studies presented here have provided evidence that many human and mouse orthologs differ in their spatial as well as temporal expression. This would lead one to question whether the genes are truly orthologs even though their sequences have a high degree of similarity. While it is true that two orthologs once performed the same function, their expression clearly has different consequences when it is not occurring on the same temporal and spatial level in both species. Since we know regulation of expression determines the timing of gene expression, it is obvious that the differences between human and mouse is not limited to those genes without any counterparts in the opposite species, but also include those orthologs whose transcriptional regulators differ between the two species. As discussed previously, transcription factors function in complexes and omission or substitution of even one transcription factor in a complex can change the timing of expression of a single gene. It is this quality of transcriptional regulation that allows even a 1% difference in genetic composition to determine the difference between the mouse and human phenotype.



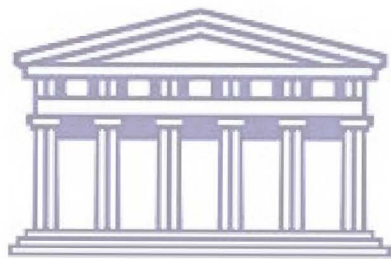
UNIVERSITY of the  
WESTERN CAPE

Our need to find cures for life-threatening diseases such as cancer is a major driving force behind biological research and with the advances of modern medicine we are in a position to develop non-invasive gene-targeted drug therapy. Due to the advantages of using mouse as a model system, the development of most drugs inevitably involves injecting a mouse with a drug to test its efficacy and toxicity. Since gene-targeted drugs aim to identify a specific gene in humans, one would expect the drug to target the same gene in the mouse in which the drug is being tested. It is therefore important to determine if the gene in question is indeed expressed in the mouse in identical tissues and developmental stages as its human counterpart.

Given the importance of the regulation of gene expression timing and the comparison thereof between human and mouse, it is therefore imperative to accurately document a gene's expression profile based on tissue, disease and developmental stage and the work presented here provides a method to address this. It is noted that the analyses presented here used a single source of expression, namely cDNA libraries. While the addition of other expression sources such as microarray, SAGE and CAGE experiments may alter the findings, the methods still apply. I have therefore developed a robust method with which to investigate aspects of mammalian gene expression, which is illustrated here in several ways.

Bioinformatics is, without a doubt, a collaborative science where your data resources are dependent on publically available data as well as that of your collaborators. It is therefore inevitable that your data will be slightly biased in

many ways, which is why it is important to keep in consideration two aspects of this field. Firstly, the integrity of your analysis and subsequent results are directly correlated with the quality, quantity and granularity of your input data. Secondly, any computational expression results or predictions need to be experimentally confirmed in a laboratory.



UNIVERSITY *of the*  
WESTERN CAPE



## Afterword

### Examination questions and answers

**1. In the first sentence of the preface you bring up the term “post-genomic”. Would you not like to argue that we are not in the post-genomic era, but rather right in the smack middle of the genomic era? Is it not premature to speak of the “post-genome”?**

- In this context, the term ‘post-genomic’ refers to the fact that we have passed the point where we have decoded the genome. Whole genomes are being sequenced on a daily basis in laboratories around the world and it is no longer the major bottleneck in genomics. Our challenge now is to interpret the genome by determining the function as well as regulation of all genes and the networks they are involved in.

UNIVERSITY of the

**2. What effect do you think “next-generation” technology will have on gene expression analysis and annotation in general?**

- The ‘next-generation’ technologies enable the sequencing of genes on a much larger scale and at a faster rate than before. While this provides more data for gene expression analysis at higher accuracy, it requires effective data management strategies. Unfortunately, the annotation is not a tightly controlled aspect of data generation and it is my opinion that with the increase in the speed at which data can be generated that this process will be neglected. In order



for us to exploit data to its full potential it should be a requirement that all data submitted to public venues be annotated according to a strict set of rules involving the use of ontologies.

### 3. What are annotations?

- An annotation is a 'label' associated with a particular object with the purpose of describing that object. Data annotations are therefore a set of words used by the researcher generating the data to describe it. A gene will, for example, be annotated according to the tissue from which it was sequenced, such as 'lung' or 'liver'. The more annotations associated with the gene, the more descriptive it becomes (such as annotating the gene according to the developmental stage or pathological state of the originating tissue). Because annotations are assigned by different individuals who would not necessarily annotate a tissue with the same level of detail, all annotations are effectively open to interpretation and prone to errors.

### 4. What is the difference between orthologs and paralogs?

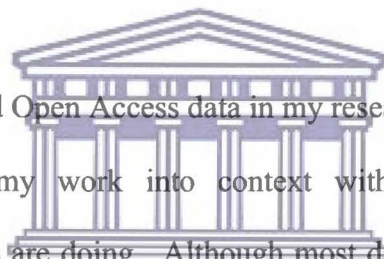
- Orthologs are genes in different species whose sequences diverged during speciation. Paralogs are genes that originated in the same species as a duplication event and the sequences of the two genes subsequently diverged. Orthologs are therefore genes separated by speciation whereas paralogs are genes separated by a duplication event.

**5. What is wrong with this statement: “These two genes are 90% homologous”?**

- Homology refers to two sequences having common ancestry and cannot be quantified. When comparing the composition of two sequences, a percentage is a degree of their SIMILARITY.

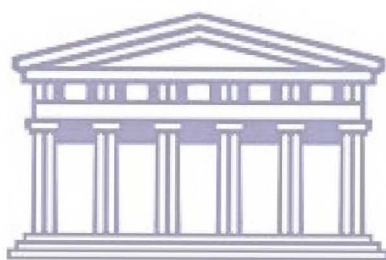
**6. How has Open Access affected your field of research? (has it?). What should the community do differently to make this kind of data more useful? Are there some requirements on data annotation that would make this more useful? If you could change one thing that was done in the past that would have made your work more useful, what would it be?**

- I have used Open Access data in my research and it has enabled me to place my work into context with respect to what other researchers are doing. Although most data is freely-available it is not easily understandable – almost as if it is just dumped into a database because it is a requirement for publication. Adequate descriptions of Open Access data would therefore make it more valuable. One of the stumbling-blocks of my research was the lack of accurate annotation of the data that is provided in public databases, which forced me to discard most of the data anyway (for example cDNA libraries annotated as ‘unclassifiable’ on the anatomical, developmental and pathological level are useless). In hindsight, making an effort to resolve annotations such as ‘unclassifiable’ would have increased the size and value of the data



UNIVERSITY of the  
WESTERN CAPE

set used in all my analyses. This would have required contacting the researcher producing each cDNA library and would be extremely time-consuming. In terms of publications, I was limited to the subscriptions of my host institution and Open Access journals. I found that much of the literature required in my research was not freely-available and therefore inaccessible to me.



UNIVERSITY *of the*  
WESTERN CAPE

## References

- Adams MD, Kelley JM, Gocayne JD, *et al.* (1991). **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science*. 252(5013):1651-1656.
- Aitken S, Korf R, Webber B, Bard J. (2005). **COBrA: a bio-ontology editor.** *Bioinformatics*. 21(6):825-826.
- Ashburner M, Ball CA, Blake JA, *et al.* (2000). **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet*. 25(1):25-29.
- Atanackovic D, Blum I, Cao Y, *et al.* (2006). **Expression of cancer-testis antigens as possible targets for antigen-specific immunotherapy in head and neck squamous cell carcinoma.** *Cancer Biol Ther*. 5(9):1218-1225.
- Bajic VB, Tan SL, Christoffels A, *et al.* (2006). **Mice and men: their promoter properties.** *PLoS Genet*. 2(4):e54.
- Baldock RA, Bard JB, Burger A, *et al.* (2003). **EMAP and EMAGE: a framework for understanding spatially organized data.** *Neuroinformatics*. 1(4):309-325.
- Bard J, Winter R. (2001). **Ontologies of developmental anatomy: their current and future roles.** *Brief Bioinform*. 2(3):289-299.
- Bhardwaj G, Murdoch B, Wu D, *et al.* (2001). **Sonic hedgehog induces the proliferation of primitive human hematopoietic cells via BMP regulation.** *Nat Immunol*. 2(2):172-180.
- Bos JL. (1989). **ras oncogenes in human cancer: a review.** *Cancer Res*. 49(17):4682-4689.
- Carninci P, Kasukawa T, Katayama S, *et al.* (2005). **The transcriptional landscape of the mammalian genome.** *Science*. 309(5740):1559-1563.
- Chitale DA, Jungbluth AA, Marshall DS, *et al.* (2005). **Expression of cancer-testis antigens in endometrial carcinomas using a tissue microarray.** *Mod Pathol*. 18(1):119-126.
- Cho HJ, Caballero OL, Gnjatic S, *et al.* (2006). **Physical interaction of two cancer-testis antigens, MAGE-C1 (CT7) and NY-ESO-1 (CT6).** *Cancer Immun*. 6(12).
- de la Monte SM, Ng SC, Hsu DW. (1995). **Aberrant GAP-43 gene expression in Alzheimer's disease.** *Am J Pathol*. 147(4):934-946.
- Dennis GJ, Sherman BT, Hosack DA, *et al.* (2003). **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol*. 4(5):P3.



Dynlacht BD. (1997). **Regulation of transcription by proteins that control the cell cycle.** *Nature*. 389(6647):149-152.

Eilbeck K, Lewis SE, Mungall CJ, *et al.* (2005). **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biol.* 6(5):R44.

Gil J, Stembalska A, Pesz KA, Sasiadek MM. (2008). **Cancer stem cells: the theory and perspectives in cancer therapy.** *J Appl Genet.* 49(2):193-199.

Gkoutos GV, Green EC, Mallon AM, Hancock JM, Davidson D. (2005). **Using ontologies to describe mouse phenotypes.** *Genome Biol.* 6(1):R8.

Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. (2005). **The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data.** *Genome Biol.* 6(3):R29.

Hill DP, Begley DA, Finger JH, *et al.* (2004). **The mouse Gene Expression Database (GXD): updates and enhancements.** *Nucleic Acids Res.* 32(Database issue):D568-71.

Hofmann O, Caballero OL, Stevenson BJ, *et al.* (2008). **Genome-wide analysis of cancer/testis gene expression.** *Proc Natl Acad Sci U S A.* 105(51):20422-20427.

The Cancer Genome Anatomy Project. <http://cgap.nci.nih.gov/>

FANTOM 4. <http://fantom.gsc.riken.jp/4/>

FANTOM3::Databases. <http://fantom3.gsc.riken.jp/>

RIKEN Genomic Sciences Centre. <http://gsc.riken.go.jp/indexE.html>

EHDA: Human versus mouse development stage comparison.  
<http://www.ana.ed.ac.uk/anatomy/database/humat/MouseComp.html>

National Cancer Institute. <http://www.cancer.gov/cancertopics/what-is-cancer>

Cancer Testis Antigen Database. <http://www.cta.lncc.br>

Ensembl. [http://www.ensembl.org/Mus\\_musculus/Info/StatsTable](http://www.ensembl.org/Mus_musculus/Info/StatsTable)

eVOC ontology. <http://www.evoontology.org>

DAG-edit. <http://www.geneontology.org/GO.tools.shtml#dagedit>

Ingenuity (R) Systems. <http://www.ingenuity.com>

NCBI HomoloGene.

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene>

NCBI UniGene. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>

The Open Biomedical Ontologies. <http://www.obofoundry.org/>



World Health Organization. <http://www.who.int/en/>

Huang da W, Sherman BT, Lempicki RA. (2009). **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc.* 4(1):44-57.

Hunter A, Kaufman MH, McKay A, *et al.* (2003). **An ontology of human developmental anatomy.** *J Anat.* 203(4):347-355.

Jemal A, Siegel R, Ward E, *et al.* (2008). **Cancer statistics, 2008.** *CA Cancer J Clin.* 58(2):71-96.

Kelso J, Visagie J, Theiler G, *et al.* (2003). **eVOC: a controlled vocabulary for unifying gene expression data.** *Genome Res.* 13(6A):1222-1230.

Kenney AM, Cole MD, Rowitch DH. (2003). **Nmyc upregulation by sonic hedgehog signaling promotes proliferation in developing cerebellar granule neuron precursors.** *Development.* 130(1):15-28.

Kho AT, Zhao Q, Cai Z, *et al.* (2004). **Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers.** *Genes Dev.* 18(6):629-640.

Kodzius R, Kojima M, Nishiyori H, *et al.* (2006). **CAGE: cap analysis of gene expression.** *Nat Methods.* 3(3):211-222.

Kruger A, Hofmann O, Carninci P, Hayashizaki Y, Hide W. (2007). **Simplified ontologies allowing comparison of developmental mammalian gene expression.** *Genome Biol.* 8(10):R229.

Lee TI, Young RA. (2000). **Transcription of eukaryotic protein-coding genes.** *Annu Rev Genet.* 34(77-137).

Liao X, Siu MK, Au CW, *et al.* (2009). **Aberrant activation of hedgehog signaling pathway contributes to endometrial carcinogenesis through beta-catenin.** *Mod Pathol.*

Lindsay S, Copp AJ. (2005). **MRC-Wellcome Trust Human Developmental Biology Resource: enabling studies of human developmental gene expression.** *Trends Genet.* 21(11):586-590.

Magdaleno S, Jensen P, Brumwell CL, *et al.* (2006). **BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system.** *PLoS Biol.* 4(4):e86.

Marra M, Hillier L, Kucaba T, *et al.* (1999). **An encyclopedia of mouse genes.** *Nat Genet.* 21(2):191-194.

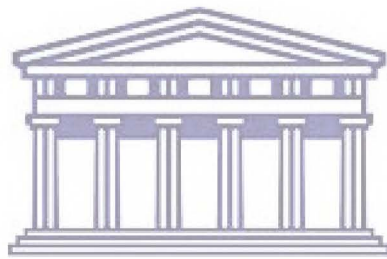
Martin D, Brun C, Remy E, *et al.* (2004). **GOToolBox: functional analysis of gene datasets based on Gene Ontology.** *Genome Biol.* 5(12):R101.

- Nagaraj SH, Gasser RB, Ranganathan S. (2007). **A hitchhiker's guide to expressed sequence tag (EST) analysis.** *Brief Bioinform.* 8(1):6-21.
- Nikolov DB, Burley SK. (1997). **RNA polymerase II transcription initiation: a structural view.** *Proc Natl Acad Sci U S A.* 94(1):15-22.
- Odom DT, Dowell RD, Jacobsen ES, *et al.* (2007). **Tissue-specific transcriptional regulation has diverged significantly between human and mouse.** *Nat Genet.* 39(6):730-732.
- Parkinson H, Aitken S, Baldock RA, *et al.* (2004). **The SOFG anatomy entry list (SAEL): an annotation tool for functional genomics data.** *Comparative and Functional Genomics.* 5(6-7):521-527.
- Reid JE, Ott S, Wernisch L. (2009). **Transcriptional programs: Modelling higher order structure in transcriptional control.** *BMC Bioinformatics.* 10(1):218.
- Rosse C, Mejino JLJ. (2003). **A reference ontology for biomedical informatics: the Foundational Model of Anatomy.** *J Biomed Inform.* 36(6):478-500.
- Sandelin A, Carninci P, Lenhard B, *et al.* (2007). **Mammalian RNA polymerase II core promoters: insights from genome-wide studies.** *Nat Rev Genet.* 8(6):424-436.
- Sato J, Illes Z, Peterfalvi A, *et al.* (2007). **Aberrant transcriptional regulatory network in T cells of multiple sclerosis.** *Neurosci Lett.* 422(1):30-33.
- Shannon P, Markiel A, Ozier O, *et al.* (2003). **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 13(11):2498-2504.
- Smith B, Ceusters W, Klagges B, *et al.* (2005). **Relations in biomedical ontologies.** *Genome Biol.* 6(5):R46.
- Smith B, Ashburner M, Rosse C, *et al.* (2007). **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotechnol.* 25(11):1251-1255.
- Sprenger J, Lynn Fink J, Karunaratne S, *et al.* (2008). **LOCATE: a mammalian protein subcellular localization database.** *Nucleic Acids Res.* 36(Database issue):D230-3.
- Stevens R, Goble CA, Bechhofer S. (2000). **Ontology-based knowledge representation for bioinformatics.** *Brief Bioinform.* 1(4):398-414.
- Suzuki H, Forrest AR, van Nimwegen E, *et al.* (2009). **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nat Genet.* 41(5):553-562.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. (2009). **A census of human transcription factors: function, expression and evolution.** *Nat Rev Genet.* 10(4):252-263.

Waterston RH, Lindblad-Toh K, Birney E, *et al.* (2002). **Initial sequencing and comparative analysis of the mouse genome.** *Nature.* 420(6915):520-562.

Zhou XJ, Gibson G. (2004). **Cross-species comparison of genome-wide expression patterns.** *Genome Biol.* 5(7):232.



UNIVERSITY *of the*  
WESTERN CAPE



# Appendix I Transcriptional landscape of the mammalian genome, *Science*.

## REPORTS

only LKS stations in NH), are fully consistent with this assumption, particularly for the tropical stations. In the extratropics there are only four daytime-only stations so the MSU test is less meaningful, but the two independent estimates do agree within 0.03°C per decade.

To illustrate the importance of the heating bias, we have computed its impact  $\delta_{\text{sol}}$  on the trends at LKS stations. The LKS  $f$  factors, unhomogenized trends, and trends adjusted only for solar heating are given for the middle troposphere and lower stratosphere in Table 2. In the stratosphere, our  $\delta_{\text{sol}}$  is similar to the total adjustments by LKS and others, with trends moving closer to those from MSU (73). At the tropical tropopause (of relevance to stratospheric water vapor),  $\delta_{\text{sol}}$  is somewhat smaller than LKS's. In the troposphere, however,  $\delta_{\text{sol}}$  is much larger than previous adjustments. Indeed, the tropical trend with this adjustment (0.14°C per decade over 1979 to 1997) would be consistent with model simulations driven by observed surface warming, which was not true previously (1). One independent indication that the solar-adjusted trends should be more accurate is their consistency across latitude belts: for the period 1979 to 1997, the spread of values fell by 70% in the lower stratosphere and 25% in the troposphere.

Though this is encouraging, our confidence in these nighttime trends is still limited given that other radiosonde errors have not been addressed. SH trends from 1958 to 1997 seem unrealistically high in the troposphere, especially with the  $\delta_{\text{sol}}$  adjustment, although this belt has by far the worst sampling. Previous homogenization efforts typically produced small changes to mean tropospheric trends, which could mean that other error trends cancel out  $\delta_{\text{sol}}$  in the troposphere. In our judgment, however, such fortuitous cancellation of independent errors is unlikely compared to the possibility that most solar artifacts were previously either missed or their removal negated by other, inaccurate adjustments. To be detected easily, a shift must be large and abrupt, but  $\delta_{\text{sol}}$  was spread out over so many stations (79% of stations during 1979 to 1997 and 90% during 1959 to 1997 experienced  $\Delta T$  trends significant at 95% level), at such modest levels, and of sufficient frequency at many stations that many may have been undetectable. Most important, jumps in the difference between daytime and nighttime monthly means would be detectable at only a few tropical stations because most lack sufficient nighttime data. In any case, we conclude that carefully extracted diurnal temperature variations can be a valuable troubleshooting diagnostic for climate records, and that the uncertainty in late-20th century radiosonde trends is large enough to accommodate the reported surface warming.

### References and Notes

1. B. D. Santer et al., *Science* **309**, 1551 (2005); published online 11 August 2005 (10.1126/science.1114867).
2. J. K. Angell, *J. Clim.* **16**, 2288 (2003).
3. J. R. Lanzante, S. A. Klein, D. J. Seidel, *J. Clim.* **16**, 241 (2003).
4. D. E. Parker et al., *Geophys. Res. Lett.* **24**, 1499 (1997).
5. P. W. Thorne et al., *J. Geophys. Res.*, in press.
6. D. H. Douglass, B. D. Pearson, S. F. Singer, P. C. Knappenberger, P. J. Michaels, *Geophys. Res. Lett.* **31**, L13207 (2004).
7. D. J. Gaffen et al., *Science* **287**, 1242 (2000).
8. D. E. Parker, D. I. Cox, *Int. J. Climatol.* **15**, 473 (1995).
9. M. Free, D. J. Seidel, *J. Geophys. Res.* **110**, D07101 (2005).
10. J. K. Luers, R. E. Eskridge, *J. Appl. Meteorol.* **34**, 1241 (1995).
11. I. Durre, T. C. Peterson, R. S. Vose, *J. Clim.* **15**, 1335 (2002).
12. L. Haimberger, "Homogenization of radiosonde temperature time series using ERA-40 analysis feedback information," Tech. Rep. European Center for Medium Range Weather Forecasting (2005), ERA-40 Project Report Series 23.
13. D. J. Seidel et al., *J. Clim.* **17**, 2225 (2004).
14. P. R. Krishnaiah, B. Q. Miao, *Handbook of Statistics*, P. R. Krishnaiah, C. R. Rao, Eds. (Elsevier, New York, 1988), vol. 7.
15. M. Free et al., *Bull. Am. Meteorol. Soc.* **83**, 891 (2002).
16. W. J. Randel, F. Wu, in preparation.
17. D. J. Seidel, M. Free, J. Wang, *J. Geophys. Res.* **110**, D09102 (2005).
18. A. Dai, K. E. Trenberth, T. R. Karl, *J. Clim.* **12**, 2451 (1999).

19. S. Chapman, R. S. Lindzen, *Atmospheric Tides* (D. Reidel, Norwell, MA, 1970).
20. D. R. Easterling et al., *Science* **277**, 364 (1997).
21. D. J. Gaffen, R. J. Ross, *J. Clim.* **12**, 811 (1999).
22. W. J. Randel et al., *Science* **285**, 1689 (1999).
23. K. M. Liou, T. Sasamori, *J. Atmos. Sci.* **32**, 2166 (1975).
24. R. E. Eskridge et al., *Bull. Am. Meteorol. Soc.* **76**, 1759 (1995).
25. H. Riehl, *Tropical Meteorology* (McGraw Hill, New York, 1954).
26. S. C. Sherwood, *Geophys. Res. Lett.* **27**, 3525 (2000).
27. J. R. Christy, R. W. Spencer, W. B. Norris, W. D. Braswell, D. E. Parker, *J. Atmos. Oceanic Technol.* **20**, 613 (2003).
28. T. Sasamori, J. London, *J. Atmos. Sci.* **23**, 543 (1966).
29. Data files and further information on methods, uncertainty, and interpretation of our results are available as supporting material on Science Online.
30. S.C.S. thanks J. Risbey and K. Braganza for useful discussions. This work was supported by the National Oceanic and Atmospheric Administration Climate and Global Change Program award NA03OAR4310153, and by NSF ATM-0134893.

### Supporting Online Material

[www.sciencemag.org/cgi/content/full/1115640/DC1](http://www.sciencemag.org/cgi/content/full/1115640/DC1)

Methods

SOM Text

Data files

References and Notes

2 June 2005; accepted 27 July 2005

Published online 11 August 2005;

10.1126/science.1115640

include this information when citing this paper.

## The Transcriptional Landscape of the Mammalian Genome

The FANTOM Consortium\* and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group)\*

This study describes comprehensive polling of transcription start and termination sites and analysis of previously unidentified full-length complementary DNAs derived from the mouse genome. We identify the 5' and 3' boundaries of 181,047 transcripts with extensive variation in transcripts arising from alternative promoter usage, splicing, and polyadenylation. There are 16,247 new mouse protein-coding transcripts, including 5154 encoding previously unidentified proteins. Genomic mapping of the transcriptome reveals transcriptional forests, with overlapping transcription on both strands, separated by deserts in which few transcripts are observed. The data provide a comprehensive platform for the comparative analysis of mammalian transcriptional regulation in differentiation and development.

The production of RNA from genomic DNA is directed by sequences that determine the start and end of transcripts and splicing into mature RNAs. We refer to the pattern of transcription control signals, and the transcripts they generate, as the transcriptional landscape. To describe the transcriptional landscape of the mammalian genome, we combined full-length cDNA isolation (1) and 5'- and 3'-end sequencing of cloned cDNAs, with new cap-analysis gene expression (CAGE) and gene identification signature (GIS) and gene signature cloning (GSC) ditag technologies for the identification of RNA and mRNA sequences corresponding to transcription initi-

ation and termination sites (2, 3). A detailed description of the data sets generated, mapping strategies, and depth of coverage of the mouse transcriptome is provided in supporting online material (SOM) text 1 (Tables 1 and 2). We have identified paired initiation and termination sites, the boundaries of independent transcripts, for 181,047 independent transcripts in the transcriptome (Table 3). In total, we found 1.32 5' start sites for each 3' end and 1.83 3' ends for each 5' end (table S1). Based on these data, the number of transcripts is at least one order of magnitude larger than the estimated 22,000 "genes" in the mouse genome (4) (SOM text 1), and the



REPORTS

large majority of transcriptional units have alternative promoters and polyadenylation sites. The use of genome tiling arrays (5-7) in humans has also implied that the number of transcripts encoded by the genome is at least 10 times as great as the number of "genes." To extend the mouse data, two HepG2 CAGE libraries, one constructed with random primers and the other with oligo-dT primers, were combined to produce 1,000,000 CAGE tags. Mapping of these tags to the human genome identified the likely promoters and transcriptional starting site (TSS) of many of the gene models identified by tiling array, also called transfrags (5), and clearly indicates that the same level of transcriptional diversity occurs in humans as in mice (table S2).

The mapping of ends of transcripts can be used to identify the genomic span of the primary transcript. Figure 1A shows length distributions of the predicted genomic regions spanned by mouse cDNAs showing a bimodal distribution and compares them with one peak for unspliced and another for spliced RNAs. At the upper end of the distribution are candidate mega transcripts (transcripts originating from genomic regions in the order of millions of base pairs). For example, we located six pairs of genome signature cloning (GSC) ditags to RIKEN clone ID 9330159J16 and corresponding RIKEN expressed sequence tags (ESTs). This clone encodes for a previously unidentified large

transcript that is similar to a protein tyrosine phosphatase, receptor type D (accession no. BC086654), the genomic structure of which has not been previously reported (8). The predicted mRNA is 2475 base pairs in length but spans a genomic region of 2.2 megabases (Mb).

We previously coined the term transcriptional units (TUs), which groups mRNAs that share at least one nucleotide and have the same genomic location and orientation (9). However, TU fusions can join unrelated and differently annotated transcripts (SOM text 2). Therefore, we define a transcriptional framework (TK) as grouping transcripts that share common expressed regions as well

The FANTOM Consortium:

P. Carninci,† T. Kasukawa, S. Katayama, J. Gough,† M. C. Frith,† N. Maeda, R. Oyama, T. Ravasi,† B. Lenhard,† C. Wells,† R. Kodrus, K. Shimokawa, V. B. Bajic,† S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wirming, V. Adkins, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojorori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminecki, M. Iacono, K. Ikeyo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. T. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Ljuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakaguchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schönbach, K. Sekiguchi, C. A. M. Semplic, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamashita, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult,† S. M. Grimmond, R. D. Teasdale, E. T. Liu,† V. Brusic, J. Quackenbush,† C. Wahlestedt,† J. S. Mattick,† D. A. Hume†

RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group):

C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki,† J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashima, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessey, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawaji

General Organizer:

Y. Hayashizaki††

\*Affiliations can be found on Science Online (available at [www.sciencemag.org/cgi/content/full/309/5740/1559/DC1](http://www.sciencemag.org/cgi/content/full/309/5740/1559/DC1)).

†These authors are core authorship members.

‡To whom correspondence should be addressed.

E-mail: yoshide@gsc.riken.jp

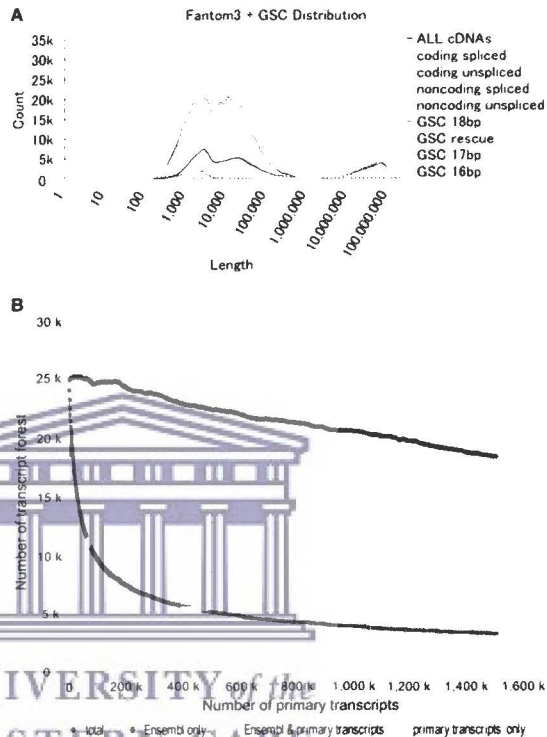


Fig. 1. Genome-transcriptome relation. (A) Genome span covered by full-length cDNA and GIS/GSC ditags shows similar distribution with two main peaks. Ditags mapping follows the same distribution profile at various mapping thresholds, with a minimum around 2 to 2.5 Mb. Mapping events above this genomic span are nonspecific. Count displays the number of events in the size interval. (B) Asymptotic unit collapse. Due to extensive overlap of the genome, transcripts overlap to the extent that they collapse to a few GFs. Simulating addition of ditags shows the collapsing rate of the known annotated genes into 9976 elements only. Primary transcripts only, GFs identified by GSC ditags only; Ensembl only, GFs produced by the 3332 Ensembl-only annotated transcripts; total, the total number of GFs.

as splicing events, TSS, or termination events (SOM text 1).

TKs can be clustered together into transcript forests (TFs), genomic regions that are transcribed on either strand without gaps. TFs encompass 62.5% of the genome (table S1) and are separated by regions

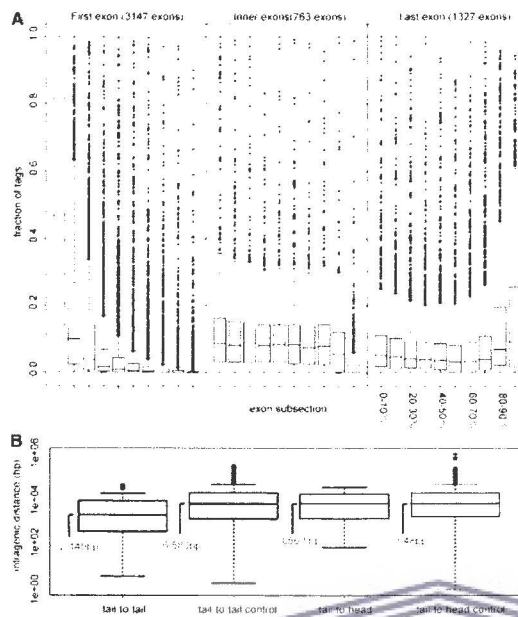
devoid of transcription, or transcription deserts. With the inclusion of GSC tags in addition to full-length cDNA and paired EST sequences, the estimated total number of transcript forests is 18,461, which will collapse further with increasing depth of coverage (Fig. 1B).

The approach used to isolate full-length cDNAs, based on library subtraction and previously unidentified 5'3' end selection before full-insert sequencing, was weighted toward identification of representative transcripts. Nevertheless, 78,393 different splicing variants were identified, such that 65% of TUs contain multiple splice variants (Table 2), an increase from our previous estimate (41%) (9). This is still expected to be an underestimate, and new approaches will be necessary for a full evaluation of exon diversity (10).

Transcript diversity also arises through alternative termination. Little is known about sequence motifs that control alternative polyadenylation. We identified 27 motif families with six or more nucleotides that were statistically overrepresented within 120 base pairs of the polyadenylation site of individual transcripts in our data set. These motifs represent candidate modulators of polyadenylation site for eight unconventional alternative polyadenylation signals (1) (table S3). In addition, we found a widespread motif family with sequence TTGTTT, which was associated with both the canonical (AAUAAA and AUUAAA) and unconventional signals (1, 11).

Gene names of 56,722 transcripts that were protein coding were assigned according to annotation rules (9, 12). Their encoded protein sequences were combined with the publicly available proteins supported by cDNA sequences (8). This generated a nonredundant set of 51,135 proteins with experimental evidence [isoform protein set (IPS)], 36,166 of which are complete (complete IPS). By comparison, the mammalian gene collection (<http://mgc.nci.nih.gov>) has cloned, as of July 2005, only ~16,700 transcripts (11,514 nonredundant). In the FANTOM3 data set, 16,274 protein sequences are newly described. Their splice variants were grouped together into 19,313 TKs. For 9002 of these, a previously known sequence maps to the same TK (locus), but 4311 clusters (5154 different proteins) map to new TKs (SOM text 3).

There are a total of 32,129 protein-coding TKs on the genome, of which 19,197 have only a single protein splice form, although 2525 of those do have an alternative noncoding splice variant. The SUPERFAMILY analysis of structural classification of protein database (SCOP) domain architectures (13) was carried out for each sequence. Of the 12,932 TKs that show variation in splicing, 8365 showed variation in SCOP domain prediction. Of the 12,932 variable TKs, 2392 produce proteins with different observed contents of InterPro entries. More than two alternatives were observed in 439 of the 2392 InterPro-variable TKs. Thus, in the majority of variable loci, splicing controls some aspect of domain content or organization. To seek evidence for such an impact in specific sets of regulatory proteins, we compared a representative protein set



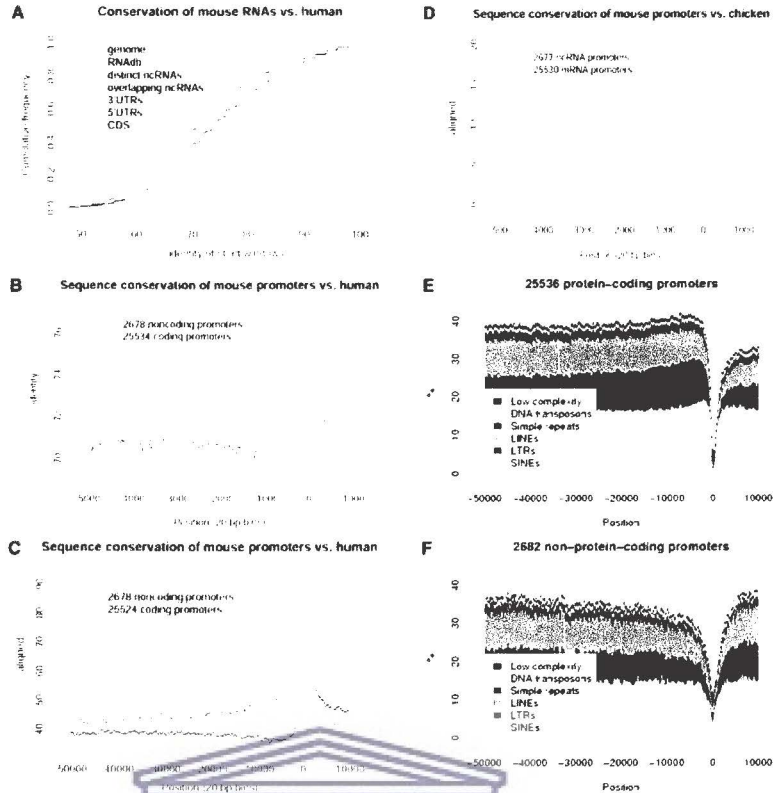
**Fig. 2.** Transcription originating in 3'UTRs. (A) For each analyzed exon, the fraction of tags mapped to 10 equally large subsections of the exon was calculated. (Left) CAGE tags mapping to the first exon are prevalently located in the first part of the exon. (Middle) CAGE tags mapping to internal exons are uniformly distributed. (Right) Last exons show a distinct overrepresentation of CAGE tags mapping close to the 3' end. (B) Distance to the closest downstream gene for the set of highly expressed TUs that have extreme tag density in the 3' of the terminal exons. Transcript pairs were grouped into tail-to-head (3' exon and downstream TU on same strand) or tail-to-tail (3' exon and downstream TU on opposite strand) configurations. Remaining TUs were used as control groups. For TUs with strong 3' transcriptional activity, the distance to the next TU is significantly smaller than expected when the gene pair is in a tail-to-tail configuration ( $P \leq 0.001107$ , Wilcoxon test), suggesting regulatory mechanisms based on natural antisense influencing the downstream gene (26).

**Table 1.** Data set resources.

	Total	Number of libraries	Safely mapped
RIKEN full-length cDNAs	102,801	237	100,313
Public (non-RIKEN) mRNAs	56,008		52,119
CAGE tags (mouse)	11,567,973	145	7,151,511
CAGE tags (human)	5,992,395	24	3,106,472
GIS ditags	385,797	4	118,594
GSC ditags	2,079,652		968,201
RIKEN 5'ESTs	722,642	266	607,462
RIKEN 3'ESTs	1,578,610	265	907,007
5'3'EST pairs of RIKEN cDNA	448,956	264	277,702

REPORTS

**Fig. 3.** Noncoding RNA promoters are highly conserved. (A) Human-mouse conservation of coding and noncoding RNAs compared with random genome sequence. (B and C) Promoters conservation of noncoding and coding mRNA evaluated (B) by identity and (C) by alignment. (D) Overlap of promoters of ncRNAs. (E and F) Promoters of coding mRNAs contain a larger fraction of low complexity and repeats than noncoding promoters. LINE, long interspersed nuclear elements; LTR, long terminal repeats; SINES, short interspersed nuclear elements.



(RPS) and a variant protein set (VPS) of phosphatases and kinases that have been comprehensively annotated (14) by looking at domain composition counts (table S4). These phosphoregulators could be functionally modulated through alteration in their intracellular location. Among the 21 receptor tyrosine phosphatase loci, we identified 25 variant transcripts from 14 loci with predicted changes to the subcellular localization and function of the encoded peptides. Of these, we identified two noncatalytic classes: secreted (10) and tethered (3). Furthermore, we identified two catalytic classes that lack the extracellular domains: catalytic only (5) and tethered catalytic (5). Similarly, among the 77 receptor kinase loci, we identified 41 variant transcripts from 33 loci which encode secreted (16), tethered (10), catalytic only (7), or other tethered catalytic (8) peptides. We then analyzed the membrane organization splicing

variants class within the full set of TUs (table S5), which revealed 1287 TUs that exhibit alternative initiation, splicing, and termination, likely to yield variant isoforms of membrane proteins that differ in their cellular location.

Of the 102,281 FANTOM3 cDNAs, 34,030 lack any protein-coding sequence (CDS) and are annotated as non-protein coding RNA (ncRNA) (6, 15) (table S1). Many putative ncRNAs were singletons in the full-length cDNA set. Among the FANTOM3 cDNA set there was additional support from ESTs, CAGE tags, or other cDNA clones overlapping both the starting and termination sites for 41,025 cDNAs, of which only 3652 were ncRNAs. This supported ncRNA set includes many known ncRNAs (SOM text 4), and many are dynamically expressed (SOM text 5). Following these same criteria, 3012 from 8961 cDNAs previously annotated as truncated

CDS were supported as genuine transcripts and are believed to be ncRNA variants of protein-coding cDNAs.

Many ncRNAs appear to start from initiation sites in 3' untranslated regions (3'UTRs) of protein-coding loci (16). The normalized distribution of CAGE tags along annotated exons of known transcripts with more than 300 mapped tags each is shown in Fig. 2A. As expected, the highest tag density on average occurs at the 5' end, but there is also a substantial increase of tags in the last one-fifth of the 3'UTR. Strong evidence of 3' end initiation was correlated with a short intergenic distance when in tail-to-tail orientation with a neighboring gene (Fig. 2B), suggesting a possible role in an intergenic regulatory interaction.

The function of ncRNAs is a matter of debate (17). Some ncRNAs are highly conserved even in distant species: 1117 out of 2886



**Table 2.** Transcript grouping and classification. The extent of splice variation was calculated by excluding T-cell receptor and immunoglobulin genes from the transcripts. The remaining 144,351 transcripts were grouped in 43,539 TUs, of which 18,627 (42.8%) consist of single-exon transcripts, 8110 (18.6%) contain a single multiexon transcript, and the remaining 16,802 TUs (38.6%) contain at least two spliced transcripts. Among these TUs, 5862 (34.9%) show no evidence of splice variation, whereas 10,940 (65.1%) contain multiple splice forms.

	Total	Average per TU cluster	Average per TK cluster
Total number of transcripts	158,807	7.59	7.30
RIKEN full-length	102,801		
Public (non-RIKEN) mRNAs	56,006		
Gfs	25,027	1.20	1.15
Framework clusters	31,992	1.53	1.47
TUs	44,147	2.11	2.03
With proteins	20,929	1.00	0.96
Without proteins	23,218	1.11	1.07
TK	45,142	2.16	2.07
With proteins	21,757	1.04	1.00
Without proteins	23,385	1.12	1.07
Splicing patterns	78,393	3.75	3.60

**Table 3.** Determination of transcripts start/end accuracy. Two pieces of evidence (cDNA, tag-dtags, EST, and 5'-3' EST pairs) are required when TSS/terminations lie inside larger transcripts, and one piece of evidence is required when they extend or identify new transcripts. Reliable indicates that both ends are associated with reliable tag clusters.

	Total	Reliable
Total 5'/3'-end pair sequence	1,507,122	1,336,397
5'/3'-end pair cluster	313,821	181,047

overlap chicken sequences, of which 780 do not overlap known CDS and 438 do not overlap known mRNAs on either strand, whereas 68 out of 2886 have BLAST-like alignment tool (BLAT) alignments to the Fugu genome, of which 40 do not overlap known CDS on either strand. These ncRNAs are at least as conserved as a reference set of known ncRNAs (Fig. 3A), contrary to a previous study (17). However, ncRNAs are slightly less conserved on average than 5' or 3'UTRs. In contrast, the promoter regions of ncRNAs are generally more conserved than the promoters of the protein-coding mRNA, not only between human and mouse but also down in the evolutionary scale to chicken (Fig. 3, B to F), and they contain binding sites for known transcription factors (18). We conclude that the large majority of ncRNAs that we analyzed display positional conservation across species. In considering function, one might conclude that the act of transcription from the particular location is either important or a consequence of genomic structure or sequence (for example, enhancers such as that of the globin locus can act as promoters), the transcript may function through some kind of sequence-specific interaction with the DNA sequence from which it is derived, or many noncoding

RNAs have other targets but are evolving rapidly (19, 20).

New databases have been created for cDNA annotation, expression, and promoter analysis (<http://fantom3.gsc.riken.jp/db/> and SOM text 6). The databases integrate common gene and tissue ontologies like eVOC mouse developmental ontologies (21), cross mapped to Edinburgh Mouse Atlas Project (EMAP) ontology terms (22). These eVOC terms allow analysis standardization of RNA samples used for cDNA and CAGE libraries in both mouse and human and were included into the DNA Database of Japan (DDBJ) data submission (23).

Analysis of the output of FANTOM2 suggested that there were many more transcripts still to be discovered (24). Here, we have confirmed that the majority of the mammalian genome is transcribed, commonly from both strands. Such transcriptional complexity implies caveats in interpretation of microarray experiments (25) and genome manipulation in mice, because these will commonly interrupt or interrogate more than one TK. Although the current overview gives us an indication of the complexity of the mammalian transcriptional landscape and a new set of tools to begin to understand transcriptional control (for example a very large set of promoters that can be ascribed to distinct classes) (16), we also gain insight into the scale of the task that remains. The dtag data indicate the existence of very long transcripts whose isolation and sequencing will require new cloning and sequencing strategies. Although we have isolated and sequenced many putative ncRNAs, the FANTOM3 collection only contains 40% of those already known. Finally, the focus has been on polyadenylated mRNAs that are processed and exported to the cytoplasm. Recently, Gingeras and colleagues (5) have

shown that the set of nonpolyadenylated nuclear RNAs may be very large, and that many such transcripts arise from so-called intergenic regions (7). The future can only reveal additional complexity in the mammalian transcriptome.

**References and Notes**

1. P. Carninci et al., *Genome Res.* 13, 1273 (2003).
2. T. Shiraki et al., *Proc. Natl. Acad. Sci. U.S.A.* 100, 15776 (2003).
3. P. Ng et al., *Nat. Methods* 2, 105 (2005).
4. R. H. Waterston et al., *Nature* 420, 520 (2002).
5. D. Kampa et al., *Genome Res.* 14, 331 (2004).
6. P. Bertone et al., *Science* 306, 2242 (2004).
7. J. Cheng et al., *Science* 308, 1149 (2005).
8. R. L. Strausberg et al., *Proc. Natl. Acad. Sci. U.S.A.* 99, 16899 (2002).
9. Y. Okazaki et al., *Nature* 420, 563 (2002).
10. A. Watahiki et al., *Nat. Methods* 1, 233 (2004).
11. V. Rajic, in preparation.
12. N. Maeda, R. Oyama, in preparation.
13. J. Cough, in preparation.
14. A. R. Forrest et al., *Genome Res.* 13, 1443 (2003).
15. Materials and methods are available as supporting material on Science Online.
16. P. Carninci et al., in preparation.
17. J. Wang et al., *Nature* 431, 1 p following 757; discussion following 757 (2004).
18. S. Cawley et al., *Cell* 116, 499 (2004).
19. T. Ravasi, D. A. Hume, in *Encyclopedia of Genetics, Genomics, Proteomics, and Bioinformatics*, L. B. Jorde, P. F. R. Little, M. J. Dunn, S. Subramaniam, Eds. (John Wiley & Sons, Chichester, UK, in press), part 2.3.
20. J. S. Mattick, I. V. Makunin, *Hum. Mol. Genet.*, in press.
21. J. Kelso et al., *Genome Res.* 13, 1222 (2003).
22. R. A. Baldoock et al., *Neuroinformatics* 1, 309 (2003).
23. All sequences (CAGE, and cDNA) are available through DDBJ to other public databases. The cDNA clones are available.
24. Y. Okazaki, D. A. Hume, *Genome Res.* 13, 1267 (2003).
25. E. Marshall, *Science* 306, 630 (2004).
26. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium, *Science* 309, 1564 (2005).
27. We thank H. Aitsuji, A. Hasegawa, Y. Hasegawa, K. Hayashida, H. Hirai, F. Hori, T. Iwahata, S. Kanagawa, C. Kawazu, M. Aoki, K. Murakami, M. Murata, H. Nishida, M. Nishikawa, K. Nomura, M. Ohno, Y. Onodera, N. Sakazume, H. Sato, Y. Shigemoto, N. Suzuki, Y. Takeda, Y. Tsujimura, K. Yoshida for discussion, encouragement, and technical assistance. We thank A. Wada, T. Ogawa, M. Muramatsu, and all the members of RIKEN Yokohama Research Promotion Division for support and encouragement. We also thank the Laboratory of Genome Exploration Research Group for secretarial and technical assistance, Yokohama City University for providing human samples, and computational resources of the RIKEN Super Combined Cluster (RSCC). This work was mainly supported by Research Grant for the Genome Network Project from MEXT, the RIKEN Genome Exploration Research Project from MEXT (Y.H.), Advanced and Innovative Research Program in Life Science (Y.H.), National Project on Protein Structural and Functional Analysis from MEXT (Y.H.), Presidential Research Grant for Intersystem Collaboration of RIKEN (P.C. and Y.H.) and a grant from the Six Framework Program from the European Commission (P.C.).

Supporting Online Material  
[www.sciencemag.org/cgi/content/full/309/5740/1559](http://www.sciencemag.org/cgi/content/full/309/5740/1559)  
 DOI: 10.1126/science.1112014  
 Materials and Methods  
 SOM Text  
 Figs. S1 to S4  
 Tables S1 to S10  
 References  
 DDBJ Accession Codes  
 9 March 2005; accepted 4 August 2005  
 10.1126/science.1112014



## Mice and Men: Their Promoter Properties

Vladimir B. Bajic<sup>1,2\*</sup>, Sin Lam Tan<sup>1,2</sup>, Alan Christoffels<sup>3,4</sup>, Christian Schönbach<sup>5</sup>, Leonard Lipovich<sup>6</sup>, Liang Yang<sup>7</sup>, Oliver Hofmann<sup>2</sup>, Adele Kruger<sup>2</sup>, Winston Hide<sup>2</sup>, Chikatoshi Kai<sup>8</sup>, Jun Kawai<sup>8,9</sup>, David A. Hume<sup>10</sup>, Piero Carninci<sup>8,9</sup>, Yoshihide Hayashizaki<sup>8,9</sup>

**1** Knowledge Discovery Laboratory, Institute for Infocomm Research, Singapore, **2** South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa, **3** Temasek Life Sciences Laboratory, National University of Singapore, Singapore, **4** School of Biological Sciences, Nanyang Technological University, Singapore, **5** Immunoinformatics Research Team, Advanced Genome Information Technology Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Yokohama, Japan, **6** Genome Institute of Singapore, Singapore, **7** Department of Obstetrics and Gynecology, National University Hospital, National University of Singapore, Singapore, **8** Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Yokohama, Japan, **9** Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, Wako, Japan, **10** Australian Research Council Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia

Using the two largest collections of *Mus musculus* and *Homo sapiens* transcription start sites (TSSs) determined based on CAGE tags, ditags, full-length cDNAs, and other transcript data, we describe the compositional landscape surrounding TSSs with the aim of gaining better insight into the properties of mammalian promoters. We classified TSSs into four types based on compositional properties of regions immediately surrounding them. These properties highlighted distinctive features in the extended core promoters that helped us delineate boundaries of the transcription initiation domain space for both species. The TSS types were analyzed for associations with initiating dinucleotides, CpG islands, TATA boxes, and an extensive collection of statistically significant *cis*-elements in mouse and human. We found that different TSS types show preferences for different sets of initiating dinucleotides and *cis*-elements. Through Gene Ontology and eVOC categories and tissue expression libraries we linked TSS characteristics to expression. Moreover, we show a link of TSS characteristics to very specific genomic organization in an example of immune-response-related genes (GO:0006955). Our results shed light on the global properties of the two transcriptomes not revealed before and therefore provide the framework for better understanding of the transcriptional mechanisms in the two species, as well as a framework for development of new and more efficient promoter- and gene-finding tools.

Citation: Bajic VB, Tan SL, Christoffels A, Schönbach C, Lipovich L, et al. (2006) Mice and men: Their promoter properties. *PLoS Genet* 2(4): e54. DOI: 10.1371/journal.pgen.0020054

The Genome Network  
Project – FANTOM3  
article collection  
**3**

### Introduction

The computational identification and functional analysis of mammalian promoters has, to date, been constrained by the relatively small datasets of experimentally confirmed transcription start sites (TSSs). For example, promoters within dbTSS were recently updated with the mapping of 195,116 FANTOM2 mouse full-length cDNA sequences to 6,875 RefSeq mouse genes [1,2]. Functional analyses of these mammalian promoters have been restricted to shared transcription factor binding sites (TFBSs) between human and mouse datasets [2]. Using the same collection of promoters contained in dbTSS, Aerts et al. embarked on a characterization of promoters by extending their study to *Drosophila melanogaster* and *Fugu rubripes* [3]. Further characterization of mammalian promoters is dependent on the availability of experimentally verified TSSs that would complement and extend existing datasets represented by the FANTOM collection, dbTSS, the H-Invitational database [4], and

RefSeq. The latest effort of the FANTOM3 consortium [5] has provided the scientific community with the largest collection of transcriptome data for *Mus musculus* (mouse), and has complemented this with CAGE tags of *Homo sapiens* (human). Based on these data we provide a comprehensive comparative analysis of mouse and human promoters that results in a number of new insights that help us to better understand the transcriptional scenario in these two species.

GC properties are well-known global factors that influence promoter characteristics and gene expression [3,6–9]. In addition, GC characteristics influence important DNA properties such as the “bendability” and curvature of the DNA helix and consequently influence the interplay of DNA and chromatin, which impacts transcription. We set out to

Editors: Judith Blake (The Jackson Laboratory, US), John Hancock (MRC Harwell, UK), Bill Pavlin (NIH/NIH, US), and Lisa Stubbs (Lawrence Livermore National Laboratory, US), together with *PLoS Genetics* EIC Wayne Frankel (The Jackson Laboratory, US)

Received August 15, 2005; Accepted February 27, 2006; Published April 28, 2006

DOI: 10.1371/journal.pgen.0020054

Copyright: © 2006 Bajic et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: CGI, CpG island; GO, Gene Ontology; Inr, initiator; ORI, over-representation index; PE, promoter element; TF, transcription factor; TFBS, transcription factor binding site; TSS, transcription start site

\* To whom correspondence should be addressed. E-mail: vlad@sanbi.ac.za

## Synopsis

Tens of thousands of mammalian genes are expressed in various cells at different times, controlled mainly at the promoter level through the interaction of transcription factors with *cis*-elements. The authors analyzed properties of a large collection of experimental mouse (*Mus musculus*) and human (*Homo sapiens*) transcription start sites (TSSs). They defined four types of TSSs based on the compositional properties of surrounding regions and showed that (a) the regions surrounding TSSs are much richer in properties than previously thought, (b) the four TSSs types are associated with distinct groups of *cis*-elements and initiating dinucleotides, (c) the regions upstream of TSSs are distinctly different from the downstream ones in terms of the associated *cis*-elements, and (d) mouse and human TSS properties relative to CpG islands (CGIs) and TATA box elements suggest species-specific adaptation. The authors linked TSS characteristics to gene expression through categories defined by the Gene Ontology and eVOC classifications and tissue expression libraries. They provided examples of the preference of immune response genes for TSS types and specific genomic organization. Their results shed light on the fine compositional properties of TSSs in mammals and could lead to better design of promoter- and gene-finding tools, better annotation of promoters by *cis*-elements, and better regulatory network reconstructions. These areas represent some of the focal topics of bioinformatics and genomics research that are of interest to a wide range of life scientists.

characterize the regions immediately surrounding TSSs based on such compositional properties. Our determination of tentative TSS locations has been based on the use of CAGE tags [10] and ditags [11] enriched with additional independent pieces of evidence of transcript existence including 5' expressed sequence tags, long 5'-SAGE, and the 5' ends of fully sequenced cDNAs from full-length libraries.

In this study, we report several distinctive features in the extended core promoters that helped us delineate the boundaries of the transcription initiation domain space for both mouse and human, as well as delineate species-specific characteristics within that space. We describe the association of TSS types with the initiating dinucleotide CGIs, TATA boxes, and an extensive collection of statistically significant

**Table 1.** Four TSS Types Defined Based on the GC Content Upstream and Downstream of the TSS

TSS Type	Upstream GC Content	Downstream GC Content
A	GC rich	GC-rich
B	GC rich	AT-rich
C	AT-rich	GC-rich
D	AT-rich	AT-rich

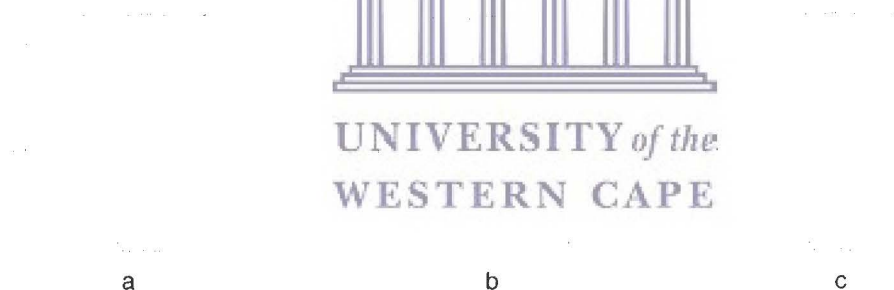
GC-rich means G + C > 50% in the considered region. AT-rich (i.e., GC-poor) means G + C < 50% in the considered region. In our case, the upstream region is [-100, -1], and the downstream region is [+1, +100] relative to the TSS.  
DOI: 10.1371/journal.pgen.0020054.t001

*cis*-elements in mouse and human, and correlate TSS properties with expression data through comparison with Gene Ontology (GO) [12] and eVOC [13] categories, tissue expression libraries, and specific genome organization.

## Results/Discussion

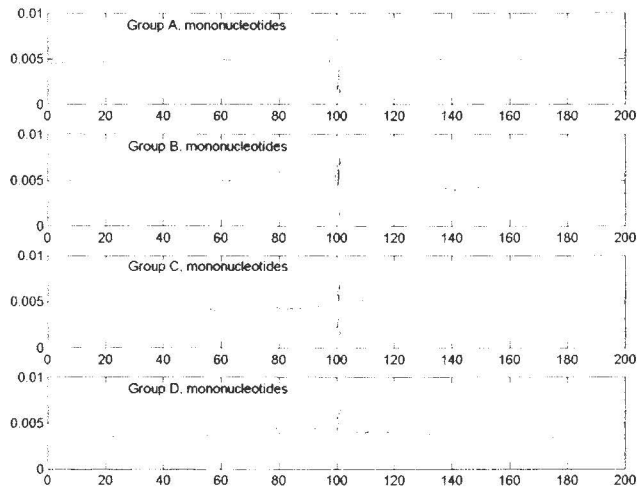
### GC Content and TSS Types

We considered TSS properties based on the GC characteristics of the segments immediately upstream and downstream of experimentally estimated TSSs. We split TSSs into four distinct classes based on the GC content upstream and downstream of the TSS, as shown in Table 1 (see Materials and Methods). These four tentative TSS types have been used as a tool to investigate different promoter features in mouse and human. Two TSS types do not differ in GC richness between the upstream and downstream regions. They are GC-rich (GC-GC, type A) or AT-rich (AT-AT, type D) both upstream and downstream. The other two are GC-rich upstream and AT-rich downstream (GC-AT, type B) and, vice versa, AT-rich upstream and GC-rich downstream (AT-GC, type C). The distributions of TSS positions in the case of mouse and human are depicted in Figure 1. A strong polarization of the TSS distribution exists, with TSS types A and D being most prevalent (Figure 1A). The number of TSSs in each of the TSS types remains almost unchanged if the length of the upstream and downstream regions changes



**Figure 1.** Transcription Initiation Domains for Mouse and Human

Distribution of mouse (red) TSSs overlapped by human (blue) TSSs based on (A) C + G content, (B) A + G content, and (C) T + G content. Nucleotide content is determined for upstream [-100, -1] and downstream [+1, +100] regions relative to the TSS. The distribution of TSS locations is more or less random when viewed in terms of A + G content (B) or T + G content (C). Strong polarization of distributions is evident only in the G + C case (A).  
DOI: 10.1371/journal.pgen.0020054.g001



**Figure 2.** Distribution of Mononucleotides in Mouse Promoters in the Region Surrounding the TSS

The nucleotides adenine, cytosine, guanine, and thymine are represented by blue, green, red, and light blue, respectively. The TSS types that are GC-poor upstream (C and D) show very characteristic enrichment in adenine and thymine nucleotides around [-35, -20], suggesting a potential dominant influence of TATA box and similar AT-rich elements in transcription initiation in these types. In type B and A TSSs, this influence does not seem to be dominant, but the presence of such elements is suggested by a significant reduction of the GC content in the [-35, -20] region. In principle, one could attempt to link the types of AT-rich upstream elements with initiating dinucleotides characteristic of different TSS types.

DOI: 10.1371/journal.pgen.0020054.g002

(Figure S1); it also only gradually changes with a change of threshold for GC richness (Figure S1). These findings suggest robustness of our TSS classification.

#### Are Two TSS Types (GC-Rich and AT-Rich) Sufficient to Consider?

Promoters are usually classified as either GC-rich or AT-rich, without separating such properties into upstream and downstream characteristics relative to the TSS [3]. In our study we observed that many of the TSSs that are not evidently GC-rich (both upstream and downstream of the TSS) have changing GC content when going from upstream to downstream regions (Figure 2). The types of patterns were AT→GC, AT→AT, and GC→AT, containing 1,911, 1,528, and 1,140 instances, respectively, in our mouse TSS dataset. We found it reasonable to assign the TSSs with a change of GC content around the TSS (AT→GC and GC→AT) to different classes because they represent about 2/3 of all non-GC-GC types. We use this profiling of TSS characteristics as a methodological convenience. However, the biological justification for this relies on the fact that many *cis*-elements have a preference for GC-rich or AT-rich domains, as found in studies of promoter groups [14,15]. Thus, considering separately the GC-rich (AT-rich) upstream and downstream segments around TSSs provides an opportunity to analyze different groups of binding sites that may confer different transcription initiation scenarios.

An essential support for the biological relevance of our introduced TSS classification relies on the fact that some eukaryotic genomes have dominant TSS characteristics of the classes we defined. For example, based on the work of Aerts et

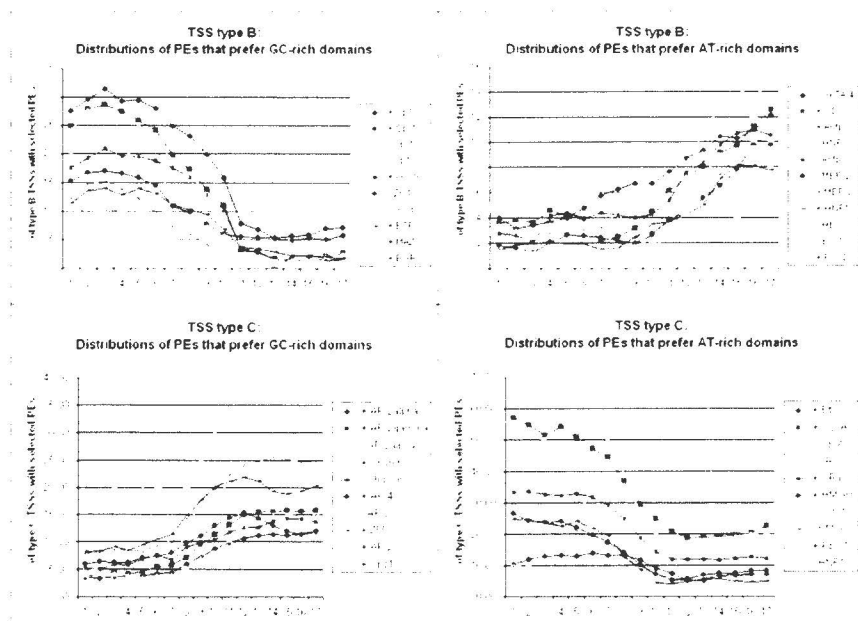
al. [3], TSS types B and C appear prevalent in *E. rubripes* and type D in *D. melanogaster*, while type A is characteristic of the human genome. There are other ways to classify promoters using certain functional rather than compositional properties. Kadonaga [16] used the presence of functional core promoter elements (PEs) such as TATA boxes, initiators, and downstream promoter elements (DPEs) to classify promoters into several types. A different approach was used by Kim et al. [17]: the properties of preinitiation complex binding to promoter and the observed transcript expression state were used to define four promoter groups.

We found through several sources of evidence that expanding a crude classification of GC-rich and AT-rich TSSs by two additional subclasses makes biological sense and presents certain fine details more explicitly than is possible if all TSSs are lumped into only two (GC-rich and AT-rich) classes. Very obvious examples of such details, in addition to largely different compositions of the putative *cis*-elements that reside in the upstream and downstream regions, are (a) specialized, but different, initiating dinucleotides overrepresented in a statistically significant manner in TSSs of different types, (b) great differences in the surrounding environment of the initiating dinucleotides between the four TSS types, and (c) different preferences of some functional gene groups for particular TSS types. These features cannot be observed if the groups are lumped.

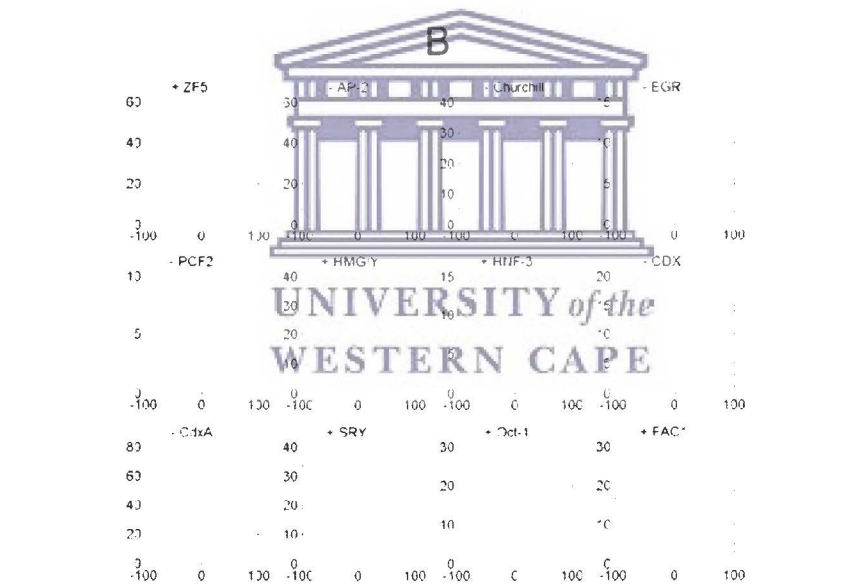
#### GC Content of TSS Surroundings Reflects Types of Putative *cis*-Elements

By considering the GC content upstream and downstream separately, we allowed for one more degree of freedom in

# A



# B





**Figure 3.** Distribution of Densities of Selected PEs in Promoters of the Four TSS Types in Mouse

The density of PEs is calculated from the region covering [-100, +100] relative to the TSS. Density is determined for bins of length 50 bp and shifted by 10 bp. In total, there are 17 bins. The vertical axis shows the percentage of TSSs of the considered type that contain the PE.  
 (A) Distribution of selected PEs that prefer GC-rich (left) and AT-rich (right) domains in type B (above) and type C (below) TSS groups. Bin number 9 is centered around the TSS. It can be seen that groups of PEs change significantly in their concentrations in transition from upstream to downstream regions and characterize two distinct TSS types (B and C).  
 (B) Distribution of selected PEs across all four TSS types. Blue, green, red, and light blue correspond to distributions characterized by type A, B, C, and D TSSs. The first five PEs are those that prefer GC-rich regions, and the last seven PEs prefer AT-rich regions (the plus or minus sign in front of the TFBS symbol denotes the strand where the TFBS is found).  
 DOI: 10.1371/journal.pgen.0020054.g003

observing global TSS properties. Here we denote a PE as a TFBS and the strand where it is found. Many PEs have preferences for either GC-rich or GC-poor regions [14,15]. For example, the well-known TATA box element, being AT-rich, will be found more frequently in AT-rich regions, while the Sp1-binding sites, being GC-rich, will be found more frequently in GC-rich regions. Thus, the four TSS types that we consider could be correlated in a global manner with the potential PEs that may control the respective genes. Support for the influence of potential PEs on specific TSS types is obtained from the distributions of PE densities (Figure 3). Density distributions of selected PEs that prefer GC-rich (AT-rich) domains in type B and type C TSSs are depicted in Figure 3A. We observe that PE groups change their concentrations significantly in transition from upstream to downstream regions. Moreover, in Figure 3B we present distributions for selected PEs across all four TSS types. The first five PEs in Figure 3B (+ZF5, -AP-2, Churchill, -EGR, and -PCF2) are those that prefer GC-rich regions (the plus and minus signs in front of the TFBS symbols denotes the strand where the TFBS is found). It is interesting to observe that these PEs occur in high concentrations in the type A group (GC-GC), occur in considerably lower concentrations in type D (AT-AT), and follow the change of GC content in types B and C. We observe the converse for the remaining seven PEs, which prefer AT-rich regions. These properties suggest that the four TSS types selectively associate with different groups of PEs.

#### Upstream and Downstream Regions Are Different: Enrichment by Specific PEs

We analyzed the preference of upstream and downstream regions in the four TSS types for significantly enriched (at least 3-fold) PEs in one region as opposed to the other region. The results are presented in Figure 4. To our surprise we found that for all TSS types the number of enriched PEs in the upstream region is much higher than in the downstream region. In three types (A, C, and D) the number of PEs in the downstream region is minimal compared to the upstream region. The only exception is type B, for which there are a significant number of enriched PEs in the downstream region. The data suggest for type A TSSs a high influence of PEs that reside upstream and prefer GC-rich domains, while for type C TSSs such influence is likely through PEs that are located upstream of the TSS but prefer AT-rich domains. Contrary to these patterns, promoters with type B TSSs seem to utilize a mix of both GC-rich-preferring and AT-rich-preferring PEs. A conclusion cannot be made for type D TSSs because of the very small number of highly enriched elements overall. Moreover, applying the Chi-square test for the equality of distributions in the upstream and downstream regions we get  $p = 1.34 \times 10^{-67}$ , which strongly rejects the null

hypothesis that these distributions are the same. All these findings suggest that upstream and downstream regions should be considered separately (as we do). The results emphasize enrichment of different PE groups associated with upstream and downstream regions in the promoters of the four TSS types.

#### Four TSS Types Associate with Different Sets of PEs

Different compositional properties of the four TSS types suggest that the TSSs may be controlled by specialized collections of transcription factors (TFs). Thus, we attempted to find the potential TFs that could play dominant roles in the four TSS types by identifying (a) the specificity of the top-ranked PEs (relative to overrepresentation index |ORI|; see Materials and Methods) in different TSS types, (b) unique and common motifs in the GC-rich/AT-rich upstream/downstream regions for different TSS types, and (c) the most significant PEs/TFs upstream/downstream of TSSs of types A, B, C, and D.

To carry out these analyses we initially compared the incidence of predicted DNA-binding sites of known TFs in the different promoter segments in mouse in the four TSS types against those in random mouse DNA. For the top 150 predicted motifs (representing approximately 10% of all elements found in these comparisons) determined based on ORI [15], we calculated Bonferroni corrected  $p$ -values for enrichment in the considered promoter segments. In the selection of these top 10% of motifs we required that they be present in at least 10% of the promoters in the target groups and that they have an ORI value not less than 1.5. In these comparisons we found that the corrected  $p$ -value was below the threshold of 0.05 for the great majority of cases. These comparisons indicate that most of the motifs for the considered TSS types are highly specific relative to random DNA (Table S1).

Next we aimed to see if promoter segments with the same GC richness share the same set of PEs. We compared the upstream regions of groups A versus B and C versus D, and the downstream regions of groups A versus C and groups B versus D. It is interesting to note that the upstream (GC-rich) regions of type A and B TSSs do share, as expected, a subset of predicted motifs, but each type is characterized also by a specialized collection of putative binding sites that do not appear in the top 150 ranked sites of the other type (for example, E1S appears in promoters of type B TSSs, but not in promoters of type A TSSs) (Table S2).

Even those TFs that are found to be common in the upstream parts of both type A and type B TSSs appear in significantly different proportions of promoters of these types, as summarized in Figure S2. For example, E1s (Table S1) appears in AT-rich downstream segments (types B and D). However, in type B TSSs it appears in 17.08% of promoters,

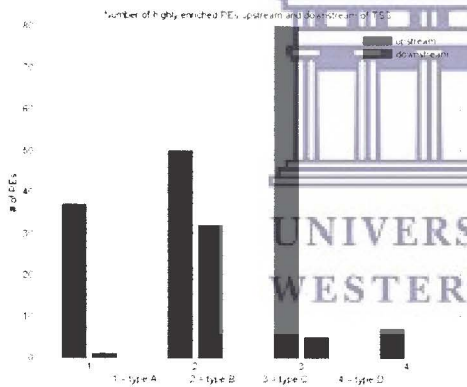
but only on the minus strand, while in type D it appears in 13.48% of promoters, but only on the plus strand.

Moreover, if we consider unique motifs that appear in different groups, they are commonly present in large proportions of promoters of those target groups. For example, in transcripts initiated from type D TSSs, we find only two unique PEs in the downstream region. One is DBP, a transcriptional activator in hepatic cells [18] and member of the C/EBP family, which appears in 26.77% of promoters with type D TSSs and only on the minus strand. The other element, Ncx, is enteric neuron homeobox and acts as an activator [19] that is required for proper positional specification, differentiative cell fate, and maintenance of proper function of enteric neurons [20,21]. It is present in 11.75% of promoters with type D TSSs and only on the plus strand.

Since any two of the four TSS types could differ in their GC content in the upstream, downstream, or both regions, and consequently harbor different sets of significant motifs, we conclude that, overall, TSS types contain sets of significant signature motifs (denoted by a plus sign next to the ORI value in Table S1 and a plus sign in Table S2) that potentially may contribute to orientation, and are likely to interact with distinct set of TFs. This concurs with the results of the preceding two subsections and suggests overall different transcriptional programs present in the transcripts of these TSS types. Lists of the most significant PEs that appear in the TSS groups are provided in Table S3.

#### The Initiating Dinucleotide and Its Environment

We analyzed in mouse and human datasets the initiating dinucleotide, that is, the one that occupies positions [-1, -1] relative to the TSS. We found that a number of different initiating dinucleotides are statistically significant across various TSS types and that they show certain regularities related to the GC content of upstream and downstream



**Figure 4.** Distribution of Selected Groups of PEs That Are Highly Enriched (at Least 3-Fold) Upstream or Downstream of the TSS

The upstream region considered covers [-100, -1], while the downstream region covers [+1, +100] relative to the TSS. In all TSS types, the upstream region contains significantly more enriched PEs than the downstream region.  
DOI: 10.1371/journal.pgen.0020054.g004

regions surrounding the TSS. Table 2 shows for mouse and human data all statistically significant cases based on the  $p$ -value obtained by the right-sided Fisher's exact test and corrected for multiplicity testing by the Bonferroni method. The association of initiating dinucleotide to TSS properties is very specific. It is interesting to note that the initiating dinucleotide TA is significantly enriched in TSS types that are AT-rich upstream, downstream, or both (B, C, and D), while dinucleotides that start with guanine (GA or GG) are significantly enriched in TSS types that are AT-rich specifically downstream (B and D). Type A TSSs are significantly enriched for dinucleotides that start with cytosine (CC, CG, and CT). However, the canonical initiating dinucleotide CA appears statistically significant only for TSS types that change GC richness (B and C). Finally, the TSS type C group contains AG and TG dinucleotides at a statistically significant level, while these do not appear significant in any other TSS type.

This compositional property of the initiating dinucleotide being linked in a statistically significant manner to the GC properties of the upstream and downstream regions would not be possible to discern if the TSS groups were lumped. We see that these properties characterize significant numbers of TSSs in our mouse dataset, namely, 10,547 (30.80%), 889 (61.74%), 1,372 (70.61%), and 534 (31.95%) of TSSs of type A, B, C, and D, respectively, and thus they do not appear to be artifacts of the proposed TSS classification that we have introduced. The conclusion is that the initiating dinucleotides show specific preferences at statistically significant levels to different TSS environments and that a significant portion of TSSs in our datasets are characterized by these initiating dinucleotides. Moreover, almost all of them are different from the canonical CA dinucleotide.

This last observation leads us to hypothesize that different TSS types may be controlled by different initiator (Inr) elements. Figure 2 depicts the quite different composition of the regions immediately surrounding tentative TSSs. The Inr elements—if they appear biologically relevant for these groups—should overlap TSSs and may be qualitatively different for different TSS types. Different initiating dinucleotides of highly statistically significant enrichment support such a hypothesis, and, at the same time, the variability of the observed initiating dinucleotides could explain the non-specific consensus of the octamer Inr element [22]. We have generated sequence logos of the TSS surroundings [-5, +5] in both mouse and human, and present them in Figure 5A. We observe that the nucleotide distributions for type A (GC-GC) TSSs are about the same in mouse and human. However, for TSS types B, C, and D, there is evident difference in these distributions in the region surrounding the TSS, which does not contradict our hypothesis of potentially different Inr elements for different TSS types. Figure 5B shows logos of regions [-33, +20] for the four TSS types in mouse and human. Again, we observe significant similarity between the species in the composition of the region for type A TSSs, while the other TSS types show significantly more variability. This may suggest species-specific organization of the core promoters for these minority TSS types (B, C, and D).

#### Relation of TSS Types to TATA Box Elements and CpG Islands

We analyzed the four TSS types in mouse and in human (Tables 3 and 4) for the presence of TATA box elements and

**Table 2.** Starting Dinucleotide [ 1, ·1] for Various TSS Types in Mouse and Human Datasets

Organism	Starting Dinucleotide	TSS Type	Number of Cases	Number of TSSs with Starting Dinucleotide	Total Number of TSSs in the Same TSS Group	Total Number of TSSs	Multiplicity Correction Factor	p-Value	Bonferroni Corrected p-Value	
Mouse	AG	C	172	1,943	2,524	39,156	16	$1.41 \cdot 10^{-5}$	$2.25 \cdot 10^{-1}$	
	CA	B	458	1,440	10,000	39,156	16	$3.25 \cdot 10^{-8}$	$5.20 \cdot 10^{-7}$	
	CA	C	558	1,943	10,000	39,156	16	$6.09 \cdot 10^{-1}$	$9.75 \cdot 10^{-1}$	
	CC	A	1,299	34,245	1,410	39,156	16	$7.17 \cdot 10^{-9}$	$1.15 \cdot 10^{-7}$	
	CG	A	8,669	34,245	9,076	39,156	16	$1.06 \cdot 10^{-18}$	$1.69 \cdot 10^{-18}$	
	CT	A	579	34,245	635	39,156	16	$1.80 \cdot 10^{-2}$	$2.88 \cdot 10^{-2}$	
	GA	B	16	1,440	171	39,156	16	$6.09 \cdot 10^{-1}$	$9.75 \cdot 10^{-1}$	
	GA	D	15	1,528	171	39,156	16	$2.99 \cdot 10^{-2}$	$4.79 \cdot 10^{-2}$	
	GG	B	264	1,440	2,952	39,156	16	$1.32 \cdot 10^{-12}$	$2.12 \cdot 10^{-11}$	
	GG	D	350	1,528	2,952	39,156	16	$8.28 \cdot 10^{-8}$	$1.33 \cdot 10^{-8}$	
	TA	B	151	1,440	2,703	39,156	16	$1.86 \cdot 10^{-7}$	$2.97 \cdot 10^{-6}$	
	TA	C	187	1,943	2,703	39,156	16	$2.30 \cdot 10^{-6}$	$3.68 \cdot 10^{-5}$	
	TA	D	169	1,528	2,703	39,156	16	$7.82 \cdot 10^{-10}$	$1.25 \cdot 10^{-8}$	
	TG	C	455	1,943	7,381	39,156	16	$1.55 \cdot 10^{-7}$	$2.48 \cdot 10^{-6}$	
	Human	AA	D	12	385	88	10,255	16	$1.03 \cdot 10^{-1}$	$1.65 \cdot 10^{-1}$
		CG	A	2,777	9,269	2,878	10,255	16	$2.37 \cdot 10^{-46}$	$3.79 \cdot 10^{-45}$
GG		D	85	385	578	10,255	16	$4.28 \cdot 10^{-29}$	$6.85 \cdot 10^{-28}$	
TA		B	25	244	575	10,255	16	$2.55 \cdot 10^{-1}$	$4.07 \cdot 10^{-2}$	
TA		C	35	357	575	10,255	16	$8.68 \cdot 10^{-1}$	$1.39 \cdot 10^{-2}$	

We show only statistically significant cases.  
DOI: 10.1371/journal.pgen.0020054.t002

association with CGIs. Globally, there are similarities in these properties of TSS types between these two species, but there are also significant differences. This mouse-human comparison must be treated with some caution, since the mouse and human datasets are based upon analysis of distinct tissues, and the human set is probably less comprehensive. In some measure, the distinctions may also relate to depth of coverage in the two species. However, since we considered a statistically large number of well-defined TSS locations in mouse (39,156) and in human (10,255), this makes comparison between the two species feasible.

Based on Bonferroni corrected *p*-values we find that the mouse and human datasets differ significantly in a number of promoter features (Tables 3 and 4). Mouse promoters are significantly enriched in (a) the number of promoters not associated with CGIs in TSS types A and B, and overall; (b) the number of TATA-less promoters in group A; (c) the overall number of promoters that have TATA boxes but are not associated with CGIs; and (d) the number of TATA-less promoters not associated with CGIs in TSS groups A and B, and overall. Conversely, human promoters are significantly enriched in (a) the number of promoters associated with CGIs in TSS types A and B, and overall; (b) the number of TATA-box-containing promoters in TSS type A; (c) the number of TATA-box-containing promoters associated with CGIs in TSS types A, B, and C, and overall; and (d) the number of TATA-less promoters associated with CGIs in TSS types A and B, and overall. These data suggest that there are species-specific solutions for transcriptional initiation in mouse and human for the analyzed TSS types.

There are a number of core PEs other than TATA boxes and Inr elements, such as the downstream promoter element (DPE) [23–26], the TFIIB response element (BRE) [27], the motif ten element (MTE) [28], and the downstream core

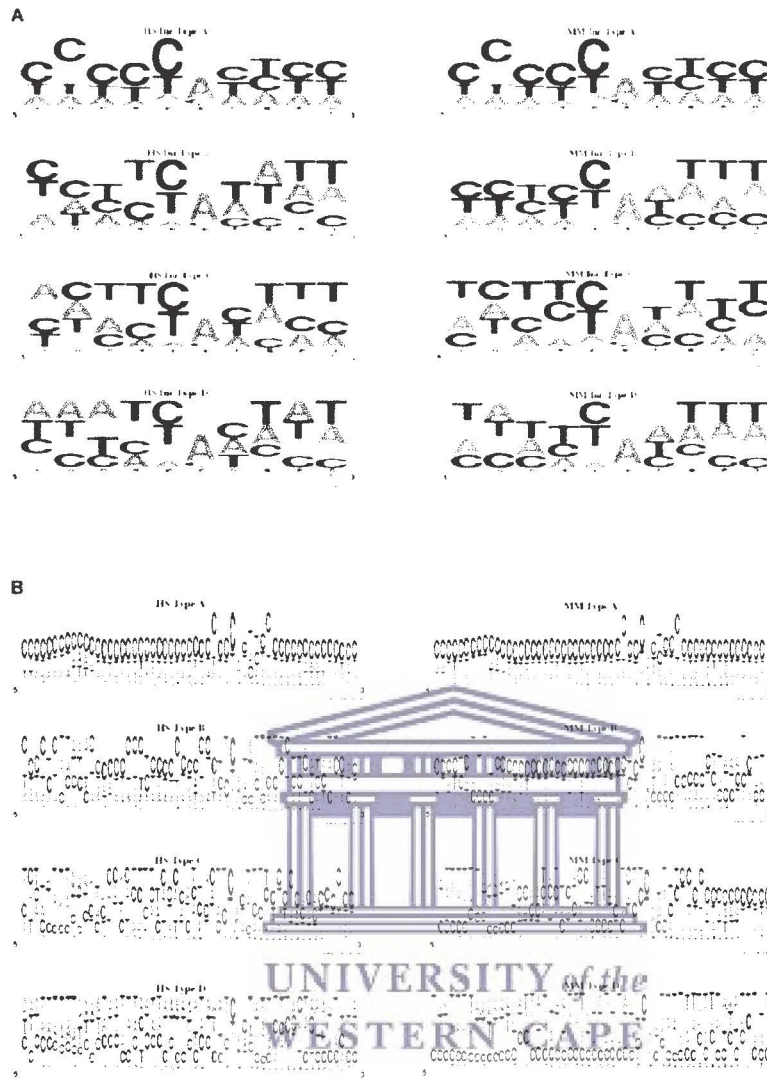
element (DCE) [29,30]. It would be of interest to investigate their presence around mammalian TSSs. Unfortunately, such an analysis represents a study on its own and requires reliable matrix models of these elements in mammals that are not yet available.

#### Linking TSS Properties and Gene Expression

We were interested to find out if the TSS types show any correlation with broad expression categories. We used association of transcripts with different GO [12] and eVOC [13] categories, as well as FANTOM3 tissue expression libraries, and analyzed their TSS distribution across the four types in mouse. While it is not possible to make definite conclusions because of incomplete GO, eVOC, tissue library, and transcript data, we were able to find a number of classes that associate with specific TSS types in a statistically significant manner (Tables 5, 6, S1, and S5). Moreover, we searched for ortholog transcript groups in mouse and human whose promoters preserve enrichment in specific TSS types in both species (Table S4). Under the conditions of our study we found that 100% of GO categories whose mapped transcripts emanate from type B TSSs preserve their enrichment; this is true for 61% of GO categories associated with type C TSSs and for 80% of GO categories associated with type D TSSs. These results suggest that between mouse and human the TSS character within the GO categories is largely conserved. Distributions of all mouse TSSs across the four TSS types for GO categories and FANTOM3 tissue libraries are provided in Table S5.

We further analyzed several specific cases. For many GO categories we found that transcripts associated with them prefer specific GC-rich/GC-poor transcription initiation frameworks (Table 5). For example, the immune response group (GO:0006955) (Figure 6) appears with 158-, 183-, and





**Figure 5. Sequence Logos**  
 (A) Sequence logos for Inr in human (left) and mouse (right) obtained using [-5, -5] segments relative to TSS locations. There is an evident bias in the nucleotide composition surrounding the TSS that effectively determines different Inr elements.  
 (B) Sequence logos for segments [-35, -20] relative to TSS locations. Strong similarity exists between human (left) and mouse (right) in TSS type A, while that similarity is considerably reduced for the other TSS types.  
 DOI: 10.1371/journal.pgen.0020054.g005



**Table 3.** Basic Statistics on Relation of TATA Box Motifs, CGIs, and Four TSS Types for MMS Transcripts

Category	TSS Type				Overall
	Type A	Type B	Type C	Type D	
Number of promoters	34,245	1,440	1,943	1,528	39,156
CGI	27,026 (78.92) [1]	253 (17.57) [1]	363 (18.68) [1]	9 (0.59) [1]	27,651 (70.62) [1]
No CGI	7,219 (21.08) [ $2.74 \times 10^{-11}$ ]	1,187 (82.43) [ $4.87 \times 10^{-7}$ ]	1,580 (81.32) [ $9.58 \times 10^{-2}$ ]	1,519 (99.41) [ $8.82 \times 10^{-2}$ ]	11,505 (29.38) [ $6.26 \times 10^{-16}$ ]
TATA	2,539 (7.41) [1]	188 (13.06) [1]	567 (29.18) [1]	434 (28.40) [1]	3,728 (9.52) [1]
TATA-less	31,706 (92.59) [ $1.63 \times 10^{-3}$ ]	1,252 (86.94) [ $1.43 \times 10^{-1}$ ]	1,376 (70.82) [1]	1,094 (71.60) [1]	35,428 (90.48) [ $2.02 \times 10^{-1}$ ]
CGI + TATA	1,613 (4.71) [1]	33 (2.29) [1]	58 (2.99) [1]	1 (0.07) [1]	1,705 (4.35) [1]
CGI + TATA-less	25,413 (74.21) [1]	220 (15.28) [1]	305 (15.70) [1]	8 (0.52) [1]	25,946 (66.26) [1]
No CGI + TATA	926 (2.70) [ $2.19 \times 10^{-1}$ ]	155 (10.76) [1]	509 (26.20) [1]	433 (28.34) [1]	2,023 (5.17) [ $12.09 \times 10^{-4}$ ]
No CGI + TATA-less	6,293 (18.38) [ $3.72 \times 10^{-11}$ ]	1,032 (71.67) [ $1.12 \times 10^{-1}$ ]	1,071 (55.12) [1]	1,086 (71.07) [1]	9,482 (24.22) [ $2.11 \times 10^{-12}$ ]

We present for each category (CGI, no CGI, etc.) the number of cases for each TSS type, the percent (in parentheses) of the total population in that TSS type, and the Bonferroni corrected *p*-value (in brackets) calculated from a right-sided Fisher's exact test based on the hypergeometric distribution.  
DOI: 10.1371/journal.pgen.0020054.t003

3.35-fold more transcripts having TSSs of type B, C, and D, respectively, than one would expect based on the proportion of transcripts in these groups in our reference mouse data. The enrichment in type C and D TSSs is statistically significant (Bonferroni-corrected right-sided Fisher's exact test,  $p = 1.33 \times 10^{-18}$  and  $p = 2.60 \times 10^{-1}$ , respectively). Based on this, we conclude that the transcript group GO:0006955 is characterized by increased participation of transcripts from TSS types that are AT-rich upstream or downstream. We analyzed in more detail the genomic organization of loci corresponding to genes from the most overrepresented TSS type (type C) for this GO. We found that TSSs of type C map to 36 nonredundant genes, of which two are in bidirectional promoters (2/36), which means these are underrepresented for type C TSSs relative to the genome average. There are 23 genes (61%) that are appearing in gene family clusters, that is, these genes are highly overrepresented for type C TSSs relative to the genome average. Finally, genes with type C TSSs have small genomic span: 34 out of 36 are less than 25 kb long, which is again more than one would expect based on the genome average. Most genes in the category GO:0006955 are short (the majority are actually less than 10 kb), are clustered with other members of the same families, and are not bidirectionally transcribed. This analysis illustrates a specific

genomic organization of genes with TSSs of type C in this GO group. Thus, TSS properties may influence genomic organization.

In Table 5, one can see that GC-rich TSSs relate to genes responsible for various binding and protein transport activities. These functions usually occur in different regions of the cell and are reflected in the diverse compartments that are enriched for type A TSSs. AT-rich TSSs (types C and D), on the other hand, are enriched in processes relating to defense responses to the environment, TSSs of the membrane attack complex (GO:0005579), defense response (GO:0006952), and immune response (GO:0006955) are enriched in type D TSSs, while the last two of these (defense and immune response) and cytochrome activity (GO:0005125) are enriched in type C TSSs. Globin group (GO:0001524) and hemoglobin complex (GO:0005833) are enriched in type B TSSs. These findings suggest a preference of different functional transcript groups for specific TSS types.

Similarly, for transcript groups based on eVOC terms, we find that they prefer GC-rich or GC-poor transcription initiation frameworks, depending on the eVOC category. For example, rhythm-expressed transcripts (EVM:2270063 and EVM:2280063) (Table 6) seem to prefer either type A or D TSSs. The same is the case for transcripts classified according

**Table 4.** Basic Statistics on Relation of TATA Box Motifs, CGIs, and Four TSS Types for H5T7 Transcripts

Category	TSS Type				Overall
	Type A	Type B	Type C	Type D	
Number of promoters	9,269	244	397	385	10,255
CGI	7,887 (85.09) [ $2.74 \times 10^{-11}$ ]	74 (30.33) [ $4.87 \times 10^{-7}$ ]	86 (24.09) [ $9.58 \times 10^{-2}$ ]	8 (2.08) [ $8.82 \times 10^{-2}$ ]	8,055 (78.55) [ $6.26 \times 10^{-16}$ ]
No CGI	1,382 (14.91) [1]	170 (69.67) [1]	271 (75.91) [1]	377 (97.92) [1]	2,200 (21.45) [1]
TATA	791 (8.53) [ $1.63 \times 10^{-3}$ ]	45 (18.44) [ $1.43 \times 10^{-1}$ ]	106 (29.69) [1]	101 (26.23) [1]	1,043 (10.17) [ $2.02 \times 10^{-1}$ ]
TATA-less	8,478 (91.47) [1]	199 (81.56) [1]	251 (70.31) [1]	284 (73.77) [1]	9,212 (89.83) [1]
CGI + TATA	574 (6.19) [ $7.00 \times 10^{-9}$ ]	16 (6.56) [ $7.01 \times 10^{-1}$ ]	22 (6.16) [ $2.99 \times 10^{-2}$ ]	0 (0.00) [1]	612 (5.97) [ $1.05 \times 10^{-10}$ ]
CGI + TATA-less	7,313 (78.90) [ $2.62 \times 10^{-29}$ ]	58 (23.77) [ $7.80 \times 10^{-1}$ ]	64 (17.93) [1]	8 (2.08) [ $5.64 \times 10^{-2}$ ]	7,443 (72.58) [ $4.31 \times 10^{-15}$ ]
No CGI + TATA	217 (2.34) [1]	29 (11.89) [1]	84 (23.53) [1]	101 (26.23) [1]	431 (4.20) [1]
No CGI + TATA-less	1,165 (12.57) [1]	141 (57.79) [1]	187 (52.38) [1]	276 (71.69) [1]	1,769 (17.25) [1]

We present for each category (CGI, no CGI, etc.) the number of cases for each TSS type, the percent (in parentheses) of the total population in that TSS type, and the Bonferroni corrected *p*-value (in brackets) calculated from a right-sided Fisher's exact test based on the hypergeometric distribution.  
DOI: 10.1371/journal.pgen.0020054.t004

**Table 5.** Enrichment of TSS Types in Selected GO Categories in Mouse

GO Category	GO ID	Term	Bonferroni Corrected p-Values for the TSS Types			
			A	B	C	D
Cellular component	GO:0005833	Hemoglobin complex	1	$1.74 \times 10^{-11}$	1	1
	GO:0005579	Membrane attack complex	1	1	1	$1.24 \times 10^{-5}$
	GO:0005576	Extracellular region	1	1	$4.79 \times 10^{-2}$	$2.09 \times 10^{-17}$
	GO:0005794	Golgi apparatus	$2.84 \times 10^{-12}$	1	1	1
	GO:0005634	Nucleus	$6.15 \times 10^{-12}$	1	1	1
	GO:0005737	Cytoplasm	$3.25 \times 10^{-4}$	1	1	1
	GO:0005739	Mitochondrion	$1.23 \times 10^{-9}$	1	1	1
	GO:0005829	Cytosol	$2.28 \times 10^{-2}$	1	1	1
	GO:0001524	Globin	1	$1.74 \times 10^{-11}$	1	1
	GO:0005125	Cytokine activity	1	1	$1.98 \times 10^{-7}$	1
Molecular function	GO:0003677	DNA binding	$1.63 \times 10^{-2}$	1	1	1
	GO:0003723	RNA binding	$3.38 \times 10^{-2}$	1	1	1
	GO:0003925	Small monomeric GTPase activity	$1.39 \times 10^{-4}$	1	1	1
	GO:0005524	ATP binding	$4.48 \times 10^{-7}$	1	1	1
	GO:0005525	GTP binding	$1.62 \times 10^{-4}$	1	1	1
	GO:0008565	Protein transporter activity	$2.11 \times 10^{-7}$	1	1	1
	GO:0016301	Kinase activity	$6.82 \times 10^{-5}$	1	1	1
	GO:0016740	Transferase activity	$3.19 \times 10^{-4}$	1	1	1
	GO:0006935	Chemotaxis	1	1	$1.32 \times 10^{-3}$	$1.36 \times 10^{-2}$
	GO:0006952	Defense response	1	1	$3.12 \times 10^{-6}$	$5.11 \times 10^{-2}$
	GO:0006955	Immune response	1	1	$1.33 \times 10^{-18}$	$2.60 \times 10^{-1}$
	GO:0006886	Intracellular protein transport	$1.77 \times 10^{-12}$	1	1	1
	GO:0007049	Cell cycle	$3.66 \times 10^{-3}$	1	1	1
GO:0007264	Small GTPase-mediated signal transduction	$2.76 \times 10^{-4}$	1	1	1	
GO:0015031	Protein transport	$3.36 \times 10^{-6}$	1	1	1	

The table shows some statistically significant examples of biased distribution of transcripts from different GO categories in specific TSS groups from all mouse data.  
DOI: 10.1371/journal.pgen.0020054.t005

to cardiovascular function (EVM:2280037 and EVM:2250015) (Table 6).

### Conclusions

We have introduced a different way to characterize TSSs, which connects TSS properties to the GC content of the immediately upstream and downstream regions. This implicitly links the TSS type with PEs that are residing in the TSS neighborhood. We were able to delineate transcription initiation active domains in the mouse and human genomes

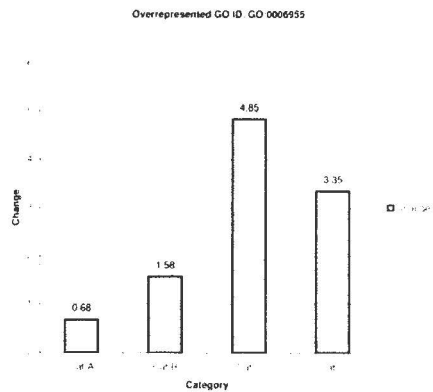
and observed fundamental similarities in the transcription initiation active domains in the two species. Looking separately at the GC content upstream and downstream of TSSs provides a useful paradigm to view certain phenomena in a clearer and more meaningful manner. We found that two of the TSS types, types C and D, possess positionally very well defined AT-rich regions [-35, +20] relative to the TSS, suggesting the significant role of AT-rich sequences such as TATA boxes in the control of TSSs of these types. Our analysis documents that various initiating dinucleotides show

**Table 6.** Enrichment of TSS Types in Selected eVOC Categories and Tissue Libraries in Mouse

EVOC ID or Tissue Library	Terms	Bonferroni Corrected p-Value for the TSS Types			
		A	B	C	D
EVM:2280168	Lung, male, adult	$2.22 \times 10^{-2}$	1	1	1
EVM:2120010	Whole body, mixture, embryo	$3.05 \times 10^{-2}$	1	1	1
EVM:2270063	Thymus, mixture, embryo	1	1	1	$7.51 \times 10^{-1}$
EVM:2280037	Aorta and vein, male, adult	1	1	1	$4.98 \times 10^{-11}$
EVM:2280087	Cortex, mixture, embryo	1	$4.56 \times 10^{-2}$	1	1
EVM:2280063	Thymus, mixture	$9.28 \times 10^{-4}$	1	1	1
EVM:2250045	Heart, mixture, embryo	$3.43 \times 10^{-3}$	1	1	1
I1	Blastocyst	$1.50 \times 10^{-4}$	1	1	1
I4	Osteoclast-like cell	$4.99 \times 10^{-2}$	1	1	1
I8	LPS treated bone marrow, macrophage	$4.85 \times 10^{-3}$	1	1	1
24	E5 cell	$8.20 \times 10^{-4}$	1	1	1
C7	Liver, tumor, adult	1	1	$3.13 \times 10^{-2}$	$2.51 \times 10^{-16}$

Some examples of statistically significant enrichment of different TSSs types in eVOC categories and tissue libraries from all mouse data.  
DOI: 10.1371/journal.pgen.0020054.t006

Type	Total	Number in category	Percent	Change
Cat A	254	152	60.08	0.68
Cat B	253	14	5.53	1.58
Cat C	253	57	22.53	4.85
Cat D	254	30	11.86	3.35



GO ID Description  
 GO:0006955  
 biological\_process immune response

**Figure 6.** Distribution of TSSs for Transcripts Related to Immune Response through GO:0006955

There are 1.58-, 4.85-, and 3.35-fold more transcripts having TSS types B, C, and D than one would expect based on the proportion of transcripts in these groups in our reference mouse data. Enrichment is statistically significant for types C and D based on Bonferroni corrected *p*-values obtained by the right-sided Fisher's exact test (Table 5).  
 DOI: 10.1371/journal.pgen.0020054.g006

very specific preferences for the TSS types we considered are present in statistically significant proportions of the TSSs in our datasets, and are almost all different from the consensus dinucleotide. Very specific sets of initiating dinucleotides are associated with different TSS types, and surrounding GC content is well correlated with the types of these dinucleotides. This suggests the potential presence of different Imr elements that may be characteristic for each of the TSS types and associated with different nucleotide characteristics of the surrounding domain.

We have shown that different TSS types associate with different PEs, that regions upstream and downstream of different TSS types are characterized by different collections of PEs, and that the putative PE content (in the top 10% of PEs) of the TSS surroundings generally differs for the TSS types. All these findings suggest likely control of the respective transcripts by different collections of significant PEs residing upstream or downstream of the TSS. Our results on TSS properties relative to CGIs, TATA boxes, and Imr elements in mouse and human suggest species-specific adaptation. Finally, we have shown a number of examples of transcript groups obtained on the basis of different ontologies or tissue libraries that have statistically significant enrichment in at least one of the TSS types. This has provided a link between TSS characteristics and expression data.

We believe that the results of this analysis will help in better understanding the general transcription regulation properties of mammalian promoters, and prove useful for further development and enhancement of promoter and gene prediction tools.

## Materials and Methods

**TSSs.** We constructed two highly accurate sets (one for mouse and one for human) of TSSs and of the promoter sequences covering the span  $[-100, +100]$  relative to these TSSs. These datasets are available at <http://www.sambic.ac.za> and were obtained as follows. If the first 5' nucleotide of the CAGE tag or 5' ditag ([http://antom31p.gs.riken.jp/cage\\_analysis/export](http://antom31p.gs.riken.jp/cage_analysis/export)) coincided with the first 5' nucleotide of the full-length cDNA (<http://antom.gs.riken.go.jp/download.html>), the TSS determined by this tag was selected. Also, in cases when this condition did not hold, we selected TSSs based on the following requirements: the TSS is a representative TSS location from a tag cluster that has at least ten tags, the representative TSS is supported by at least six tags, and there is at least one other piece of transcriptional evidence associated with this tag cluster (expressed sequence tag, full-length cDNA, or long SAGE; <http://antom.gs.riken.go.jp/download.html>). In this way, we compiled a mouse reference promoter set of 39,156 promoters and a human reference promoter set of 10,253 promoters. These two sets are used for all our analyses.

Randomly selected DNA sequences from mouse were used as the background set for analysis of TF binding sites in mouse promoters. These DNA sequences were 200 bp long and selected randomly from all mouse chromosomes, with the number of sequences from each chromosome proportional to the length of the chromosome. In total we selected 11,000 such random DNA sequences (Dataset S1).

**TSS types.** We determined the GC content of the  $[-100, +1]$  region and the  $[-1, +100]$  region relative to TSS location for each individual TSS. The TSS is considered to be between positions  $-1$  and  $+1$ . The upstream or downstream segment was defined as GC-rich if  $GC + C > 50\%$  in the region. Otherwise, the region was defined as AT-rich. Four types of TSSs were defined based on the GC richness in the upstream and downstream segments as follows (Table 1): type A, GC-rich upstream and downstream (GC-GC); type B, GC-rich upstream and AT-rich downstream (GC-AT); type C, AT-rich upstream and GC-rich downstream (AT-GC); and type D, AT-rich upstream and downstream (AT-AT). Each TSS can be represented as a point in the  $x$ - $y$  plane, where  $x$  corresponds to the GC content upstream and  $y$  corresponds to the GC content downstream of the considered TSS. For mouse and human these distributions are depicted in Figure 1A.

**TF binding sites in promoters.** We used all available matrix models of TFBS contained in the TRANSFAC Professional (version 8.4) database [31] and mapped them to the extracted sequences. We used minSUM profiles for the threshold of the matrix models since these contain the optimized threshold values for the core and matrix scores [32]. The thresholds in minSUM are based on optimization that provides the minimum sum of false positive and false negative TFBS predictions. To determine the overrepresentation of TFBSs found in the target set, we used the method of Bajic et al. [15]. All TFBSs mapped to target promoters were ranked based on their ORI as defined by Bajic et al. [15]. For ORI = 1 or close to this value, there is no overrepresentation of the motif in the target promoter group. We also estimated the likelihood of observing these TFBSs in the target set using the background random promoter set as a reference. The null hypothesis was that the proportion of sequences in the target set in which a particular PE was found was the same as that in the background set. The *p*-values were calculated using right-sided Fisher's exact tests based on hypergeometric distribution. The original *p*-value was subjected to Bonferroni correction for multiple testing. If the corrected *p*-value of the pattern was not greater than 0.05, we placed a plus sign after the ORI value in the provided tabular reports.

**Most significant PEs.** For each of the TSS types in mouse, we analyzed the 150 top-ranked PEs based on the values of ORI. This represents about 10% of all (1,128) PEs analyzed. We also required that the PEs have an ORI of at least 1.5 and that the PE be found in at least 10% of the target sequences. Details are explained in Tables S1–S3.

**TATA boxes.** The TATA box model used was based on that of Bucher [22]. The threshold used was 0.75, while score was normalized between zero and one (analogous to Bajic et al. [33]). A TATA box was considered detected if the maximum value of the score in the  $[-50, +1]$  region was higher than the threshold. Only one TATA box was assumed in the  $[-50, +1]$  region.



**eVOC, GO, and tissue expression libraries.** In order to assess the biological significance of our TSS classification system, we assigned TSSs according to different GO and eVOC categories, as well as tissue libraries in FANTOM3 collection (<http://fantom.gsc.riken.go.jp/download.html>). GO-FANTOM mapping data was downloaded from the RIKEN Web site (<http://fantom.gsc.riken.go.jp/FANTOM3/annotation/fantomdb-3.0/annotation.txt.gz>). The eVOC system consists of a set of orthogonal controlled vocabularies that unify gene expression data by mapping between the genome sequence and expression phenotype information. The eVOC human anatomy ontology [13] and the newly developed mouse adult and developmental ontologies (<http://www.evoontology.org>) have been mapped to the FANTOM3 library descriptions, providing a hierarchical representation of tissues, cell types, and developmental stage information. This allows for a standardized analysis of gene expression and promoter profiles independent of the original annotation vocabulary used in the original dataset.

For the generation of the results presented in Table S1, we used ortholog gene groups between mouse and human as defined at <http://ftp.ncbi.nih.gov/pub/HomoloGene>. Table S5 for mouse data contains statistics of all GO and tissue expression libraries from FANTOM3, complemented by the Bonferroni corrected *p*-value (right-sided Fisher's exact test based on hypergeometric distribution) for the null hypothesis that the proportion of TSSs of a specific type in the considered GO/tissue library is the same as what one can expect based on the distribution of these TSSs in mouse.

## Supporting Information

**Dataset S1.** Supplementary Nonpromoter Data

Found at DOI: 10.1371/journal.pgen.0020051.s001 (2.5 MB ZIP).

**Figure S1.** Number of TSSs of the Four Types in Human and Mouse Genomes under the Change of Parameters

Blue, green, red, and light blue correspond to TSSs of type A, B, C, and D, respectively. From graphs in the first row we observe that when the length of the region considered changes, the numbers of TSSs of the different types remain almost unchanged. We changed the length of upstream and downstream regions from  $\lfloor \alpha \cdot \lfloor \cdot \rfloor + 1 \rfloor$ ,  $\lceil \alpha \cdot \lceil \cdot \rceil - 1 \rceil$ , respectively, with values of  $\alpha$  from 50 to 150. From graphs in the second row we observe that the numbers of TSSs within the four types gradually change with the change of threshold for GC content. We changed this threshold from 10% to 60%.

Found at DOI: 10.1371/journal.pgen.0020051.s001 (21 KB PDF).

**Figure S2.** Distributions of TFS Found to Be Common among the Top 150 PEs in Comparisons of Different TSS Types

(A) Comparison of types A and B upstream regions.  
(B) Comparison of types B and D downstream regions.  
(C) Comparison of types A and C downstream regions.  
(D) Comparison of types C and D upstream regions.

Found at DOI: 10.1371/journal.pgen.0020051.s002 (10 KB PDF).

**Table S1.** List of Top 150 PEs That Appear with a Frequency of 10% or Greater in Upstream and Downstream Regions of Different TSS Categories

Comparison is carried out against a background of random mouse sequences. Ranking is based on ORI value. The higher the ORI, the higher the rank. We present results for the four TSS types (A, B, C, and D). For each PE we give the strand where it is found, C1 or  $\bar{C}1$ , name of TFS, ORI value, percentage of promoters in the target set that contain the PE, percentage of sequences in the background set that contain the PE, probability of finding the PE in the target set given as one prediction per nucleotide, probability of finding the PE in the background set given as one prediction per nucleotide, and Bonferroni corrected *p*-value. A plus sign added after the ORI value indicates that the PE is enriched in a statistically significant manner at the level 0.05. Almost all top-ranked elements appear to be statistically significantly enriched in the target sets.

Found at DOI: 10.1371/journal.pgen.0020051.s001 (329 KB PDF).

## References

- Suzuki Y, Yamashita R, Sugano S, Nakai K (2004) DBTSS: Database of transcriptional start sites. Progress report 2001. *Nucleic Acids Res* 32: D78–D81.
- Suzuki Y, Yamashita R, Shitota M, Sakakibara Y, Chiba J, et al. (2001) Large-

**Table S2.** Common and Specific TFSs in the Four TSS Types

PEs are compared relative to the same GC richness and same location (upstream or downstream) in different TSS types. The signs plus or minus indicate whether the PE was found to be significantly enriched in the considered region for the considered TSS type. For example, the first column (yellow), which shows comparison between the AT-rich downstream domains in TSS types B and D, contains a common element denoted as “+ TFS.” This means that TFS was found significantly enriched for the B type, but its enrichment was not significant for D type. When an element is unique for one or another group, then it is associated only with one plus or minus sign.

Found at DOI: 10.1371/journal.pgen.0020051.s002 (53 KB PDF).

**Table S3.** List of Significant PEs Unique and Common for Different TSS Types in the Upstream and Downstream Segments

The yellow highlighted TFS are unique for the considered groups when compared with the same upstream or downstream segment of another TSS type with the same GC richness.

Found at DOI: 10.1371/journal.pgen.0020051.s003 (47 KB PDF).

**Table S4.** GO Categories That Preserve Enrichment in Specific TSS Types between Human and Mouse

We considered only those TSSs whose generated transcripts belong to the same homology group as defined on the NCBI Web site (<http://ftp.ncbi.nih.gov/pub/HomoloGene>). We only considered GO categories that were supported by at least 60 TSSs and where the target TSS type was supported by at least three TSSs.

Found at DOI: 10.1371/journal.pgen.0020051.s004 (96 KB PDF).

**Table S5.** All GO and Tissue-Specific Libraries with Distribution of TSSs across the Four TSS Types in Mouse

The table presents the total number of TSSs associated with the category (GO or expression library), the number of TSSs of individual TSS type, the percentage of TSSs in that TSS type, enrichment of TSSs in the TSS type relative to what can be expected based on the distribution of all TSSs in mouse across all four TSS types, and Bonferroni corrected *p*-values calculated based on right-sided Fisher's exact tests for the null hypothesis that the proportion of TSS type found in the target group is the same as that of the general mouse distribution. For example, there are 253 transcripts associated with GO:0006955. Of these, 52 transcripts include a TSS of type C. For the number of transcripts in this GO category, one would expect only 11 transcripts with TSSs of type C. Thus, in this GO category, we have 4.7-fold enrichment of transcripts of this type (compared to what we would expect based on the distribution of all transcripts across the four TSS types). If in any of the GO/eVOC categories or tissue libraries, at least one of the TSS groups of transcripts has enrichment that is 1.5-fold or greater than the expected value, we consider such TSS type overrepresented.

Found at DOI: 10.1371/journal.pgen.0020051.s005 (2.9 MB PDF).

## Acknowledgments

**Author contributions.** VBB, JK, PC, and YH conceived and designed the experiments. VBB, SLE, and CK performed the experiments. VBB, SLE, AC, CS, JL, and OH analyzed the data. VBB, SLE, LY, OH, AK, WH, CK, JK, PC, and YH contributed reagents/materials/analysis tools. VBB, AC, CS, OH, and DAH wrote the paper.

**Funding.** This study was supported by a research grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan to YH, a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan to YH, and a grant for the Strategic Programs for R&D of RIKEN to YH.

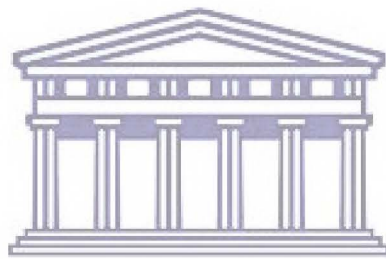
**Competing interests.** The authors have declared that no competing interests exist.

scale collection and characterization of promoters of human and mouse genes. *In Silico Biol* 1: 0036.

- Acetris, Huijs G, Dabrowski M, Morcau Y, De Moor B (2001) Comprehensive analysis of the base composition around the transcription start site in metazoan. *BMC Genomics* 5: 31.
- Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, et al. (2001)



- Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol* 2:e162. DOI: 10.1371/journal.pbio.0020162
5. Cantini P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. *Science* 309: 1539–1543.
  6. Vinogradov AE. (2005) Noncoding DNA, isochores and gene expression: Nucleosome formation potential. *Nucleic Acids Res* 33: 559–563.
  7. Vinogradov AE. (2005) Isochores and tissue-specificity. *Nucleic Acids Res* 33: 5212–5220.
  8. Vinogradov AE. (2003) DNA helix: The importance of being GC-rich. *Nucleic Acids Res* 31: 1838–1844.
  9. Levitsky AG, Podkolodnaya OA, Kolchanov NA, Podkolodny NI. (2001) Nucleosome formation potential of eukaryotic DNA: Calculation and promoters analysis. *Bioinformatics* 17: 998–1010.
  10. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* 100: 15770–15781.
  11. Ng P, Wei CL, Sung WK, Chin KP, Lipovich L, et al. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* 2: 105–111.
  12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29.
  13. KeBo J, Visagie J, Theiler G, Christoffels A, Barden S, et al. (2003) eYOC: A controlled vocabulary for unifying gene expression data. *Genome Res* 13: 1222–1230.
  14. Kel-Margolis OV, Tchekmenev D, Kel AE, Goessling E, Hornischer K, et al. (2003) Composition-sensitive analysis of the human genome for regulatory signals. *In Silico Biol* 3: 0013.
  15. Bajic VB, Choudhary V, Hock CK. (2001) Content analysis of the core promoter region of human genes. *In Silico Biol* 1: 109–125.
  16. Kadonaga JJ. (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116: 247–257.
  17. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. (2005) A high-resolution map of active promoters in the human genome. *Nature* 436: 876–880.
  18. Mueller CR, Maure P, Schibler U. (1990) DRP, a liver-enriched transcriptional activator, is expressed late in ontogeny and its tissue specificity is determined posttranscriptionally. *Cell* 61: 279–291.
  19. Shimizu H, Kang M, Iitsuka Y, Ichinose M, Tokuhisa T, et al. (2000) Identification of an optimal Ncs binding sequence required for transcriptional activation. *FEBS Lett* 475: 170–173.
  20. Shirasawa S, Yunker AM, Roth KA, Brown GA, Homing S, et al. (1997) Eux (BoxIII)-deficient mice develop intestinal neuronal hyperplasia and megaolon. *Nat Med* 3: 646–650.
  21. Hatano M, Aoki I, Dezawa M, Yusa S, Iitsuka Y, et al. (1997) A novel pathogenesis of megaolon in Nrx/BoxIII-1 deficient mice. *J Clin Invest* 100: 755–801.
  22. Bucher P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* 212: 563–578.
  23. Burke JW, Kadonaga JJ. (1996) *Drosophila* TFB2 binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* 10: 711–721.
  24. Burke JW, Kadonaga JJ. (1997) The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*. *Genes Dev* 11: 3029–3041.
  25. Kutach AK, Kadonaga JJ. (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol Cell Biol* 20: 4751–4764.
  26. Wills PJ, Kobayashi R, Kadonaga JJ. (2000) A basal transcription factor that activates or represses transcription. *Science* 290: 982–984.
  27. Lagrange J, Kapanidis AN, Lang H, Reinberg D, Ehrlich RH. (1998) New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IB. *Genes Dev* 12: 31–44.
  28. Lim CY, Santos B, Boulay T, Dong E, Odier C, et al. (2001) The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* 15: 1606–1617.
  29. Lewis BA, Kim TK, Orkin SH. (2000) A downstream element in the human beta-globin promoter: Evidence of extended sequence-specific transcription factor IID contacts. *Proc Natl Acad Sci U S A* 97: 7172–7177.
  30. Lee DH, Gershenzon S, Gupta M, Joshihles JP, Reinberg D, et al. (2005) Functional characterization of core promoter elements: The downstream core element is recognized by TAFII. *Mol Cell Biol* 25: 9674–9680.
  31. Mats V, Franke E, Gellers R, Gosling E, Handcock M, et al. (2003) TRANSAC: Transcriptional regulation from patterns to profiles. *Nucleic Acids Res* 31: 371–378.
  32. Kel AE, Gosling E, Ruter I, Cherenushkin E, Kel-Margolis OV, et al. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 31: 3576–3579.
  33. Bajic VB, Seth SJ, Chong A, Krishnan SP, Koh JL, et al. (2003) Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. *J Mol Graph Model* 21: 323–332.



UNIVERSITY of the  
WESTERN CAPE

### Appendix III Correlation coefficients of genes showing biased expression for the developmental brain in human and mouse

The correlation coefficients of the 90 genes showing bias for developmental expression in the human and mouse brain. The table lists the HomoloGene group identifier, Human Entrez Gene identifier, Human Entrez gene symbol, Mouse Entrez Gene identifier, Mouse Entrez gene symbol and the correlation coefficient between the expression profiles of the genes in each species.

Homolo-Gene ID	Human Gene	Human Symbol	Mouse Gene	Mouse Symbol	Correlation coefficient
7516	389075	RESP18	19711	Resp18	in mouse, only expressed in brain
78698	387876	LOC387876	380653	Gm872	in mouse, only expressed in brain
81871	56751	BARHL1	54422	Barhl1	in mouse, only expressed in brain
10774	57045	TWSG1	65960	Twsg1	in mouse, expressed in all tissues
27813	84865	FLJ14397	243510	A230058J24 Rik	0.646
16890	399664	RKHD1	237400	Rkhd1	0.548
2880	8835	SOCS2	216233	Soes2	0.531
1933	5050	PAFAH1B3	18476	Pafah1b3	0.531
55434	1289	COL5A1	12831	Col5a1	0.519
7291	10683	DLL3	13389	Dll3	0.471
84799	22835	ZFP30	22693	Zfp30	0.471
7667	1154	CISH	12700	Cish	0.458
32546	64410	KLHL25	207952	Klhl25	0.447
17078	387914	TMEM46	219134	Tmem46	0.447
32293	51018	CGI-115	67223	2810430M08 Rik	0.440
1871	4760	NEUROD1	18012	Neurod1	0.439
56774	54751	FBLIM1	74202	Fblim1	0.417
68973	1663	DDX11	320209	Ddx11	0.408

Homolo-Gene ID	Human Gene	Human Symbol	Mouse Gene	Mouse Symbol	Correlation coefficient
37917	1293	COL6A3	12835	Col6a3	0.408
55918	6882	TAF11	68776	Taf11	0.378
10695	57120	GOPC	94221	Gopc	0.316
14128	91107	TRIM47	217333	Trim47	0.300
68998	170302	ARX	11878	Arx	0.300
12418	124056	NOXO1	71893	Noxo1	0.289
55599	669	BPGM	12183	Bpgm	0.284
45198	65117	FLJ11021	208606	1500011J06 Rik	0.284
18123	140730	RIMS4	241770	Rims4	0.277
65328	7559	ZNF12	231866	Zfp12	0.273
68934	57016	AKR1B10	14187	Akr1b8	0.258
65280	286128	ZFP41	22701	Zfp41	0.258
22818	29850	TRPM5	56843	Trpm5	0.258
10663	57171	DOLPP1	57170	Dolpp1	0.251
45867	139189	DGKK	331374	Dgkk	0.240
17523	115290	FBXO17	50760	Fbxo17	0.207
4397	8971	H1FX	243529	H1fx	0.207
2212	6182	MRPL12	56282	Mrp12	0.194
11980	84262	MGC10911	66506	1810042K04 Rik	0.167
26702	93109	TMEM44	224090	Tmem44	0.149
56571	26503	SLC17A5	235504	Slc17a5	0.141
7717	24147	FJX1	14221	Fjx1	0.122
18903	440193	KIAA1509	68339	0610010D24 Rik	0.101
1028	1606	DGKA	13139	Dgka	0.101
4983	10991	SLC38A3	76257	Slc38a3	0.055
9813	55627	FLJ20297	77626	4122402O22 Rik	0.055
1368	1054	CEBPG	12611	Cebpg	0.055

Homolo-Gene ID	Human Gene	Human Symbol	Mouse Gene	Mouse Symbol	Correlation coefficient
64353	126374	WTIP	101543	Wtip	0.026
12993	84217	ZMYND12	332934	Zmynd12	0.000
7199	11054	OGFR	72075	Ogfr	0.000
46116	401399	LOC401399	101359	D330027H18 Rik	0.000
7500	5806	PTX3	19288	Ptx3	0.000
413	353	APRT	11821	Aprt	-0.026
49899	143282	C10orf13	72514	2610306H15 Rik	-0.026
12021	84557	MAP1LC3A	66734	Map1lc3a	-0.043
11920	84303	CHCHD6	66098	Chchd6	-0.050
32633	136647	C7orf11	66308	2810021B07 Rik	-0.050
7922	6150	MRPL23	19935	Mrpl23	-0.050
1290	9275	BCL7B	12054	Bcl7b	-0.050
9355	51637	C14orf166	68045	2700060E02 Rik	-0.077
40668	9646	SH2BP1	22083	Sh2bp1	-0.101
40859	27166	PX19	66494	2610524G07 Rik	-0.113
10494	58516	FAM60A	56306	Tera	-0.113
6535	11062	DUS4L	71916	Dus4l	-0.122
65318	23361	ZNF629	320683	Zfp629	-0.125
14180	115294	PCMTD1	319263	Pcmtd1	-0.145
32	435	ASL	109900	Asl	-0.145
68420	9559	VPS26A	30930	Vps26	-0.167
32331	51776	ZAK	65964	B230120H23 Rik	-0.175
11653	79730	FLJ14001	70918	4921525L17 Rik	-0.194
49970	83879	CDCA7	66953	Cdca7	-0.207
1330	857	CAV1	12389	Cav1	-0.213
14157	90416	CCDC32	269336	Ccdc32	-0.213



Homolo-Gene ID	Human Gene	Human Symbol	Mouse Gene	Mouse Symbol	Correlation coefficient
56005	6328	SCN3A	20269	Scn3a	-0.240
10026	55172	C14orf104	109065	1110034A24 Rik	-0.273
31656	27000	ZRF1	22791	Dnajc2	-0.273
41703	118881	COMTD1	69156	Comtd1	-0.289
14667	113510	HEL308	191578	Hel308	-0.300
268	5805	PTS	19286	Pts	-0.330
2593	7913	DEK	110052	Dek	-0.330
20549	4324	MMP15	17388	Mmp15	-0.354
18833	143678	LOC143678	75641	1700029I15 Rik	-0.354
9120	25851	DKFZP434B0335	70381	2210010N04 Rik	-0.372
15843	79591	C10orf76	71617	9130011E15 Rik	-0.372
3476	9197	SLC33A1	11416	Slc33a1	-0.389
21334	10912	GADD45G	23882	Gadd45g	-0.389
19028	146167	LOC146167	234788	Gm587	-0.408
10518	84273	C4orf14	56412	2610024G14 Rik	-0.411
35002	93082	LINCR	214854	Lincr	-0.411
12444	84902	FLJ14640	72140	2610507L03 Rik	-0.452
82250	150678	MYEOV2	66915	Myeov2	-0.646
24848	266629	SEC14L3	380683	RP23-81P12.8	-0.646

## Appendix IV Expression profile of genes showing biased expression for the developmental brain in human and mouse

The expression profiles of the 90 genes showing bias for developmental expression across major human and mouse tissues in the form of a binary pseudoarray. The tissues represented are female reproductive system, heart, kidney, liver, lung, male reproductive system and stem cell for both post-natal and developmental expression. The table lists the HomoloGene group identifier, Entrez Gene identifier and Entrez gene symbol for human and mouse, as well as the species each row represents. Values in the table are 1 if the genes (in rows) are expressed in the given tissues (in columns) and 0 if the genes are not found to be expressed in the tissues (PN – post-natal; D – development; FRS – female reproductive system; MRS – male reproductive system).

HomoloGene ID	GeneID	Gene Symbol	Species	PN: FRS	PN: heart	PN: kidney	PN: liver	PN: lung	PN: MRS	PN: stem cell	D: FRS	D: heart	D: kidney	D: liver	D: lung	D: MRS	D: stem cell
413	353	APRT	Human	1	0	1	0	1	1	1	1	1	0	1	1	1	1
32	435	ASL	Human	1	1	1	1	1	1	0	1	1	0	1	1	0	1
55599	669	BPGM	Human	1	0	1	1	1	1	0	1	1	1	1	1	1	1
1330	857	CAV1	Human	1	1	1	0	1	1	0	1	1	0	1	1	1	1
1368	1054	CEBPG	Human	1	0	0	1	1	1	0	1	1	0	1	1	1	1
7667	1154	CISH	Human	1	0	1	0	1	1	0	0	0	0	1	1	0	0
55434	1289	COL5A1	Human	1	0	1	0	1	1	0	1	1	0	1	1	1	1
37917	1293	COL6A3	Human	1	0	1	1	1	1	0	1	1	1	1	1	1	1
1028	1606	DGKA	Human	1	0	0	1	1	1	0	0	1	0	1	1	0	1
68973	1663	DDX11	Human	1	0	1	1	1	1	0	1	1	1	1	1	1	1
20549	4324	MMP15	Human	1	1	0	0	0	1	0	1	1	0	0	1	1	1
1871	4760	NEUROD1	Human	0	0	0	0	0	1	0	0	0	0	0	0	0	0
1933	5050	PAFAH1B3	Human	1	0	0	1	1	1	0	1	1	1	1	1	1	1

HomoloGene ID	GeneID	Gene Symbol	Species	PN: FRS	PN: heart	PN: kidney	PN: liver	PN: lung	PN: MRS	PN: stem cell	D: FRS	D: heart	D: kidney	D: liver	D: lung	D: MRS	D: stem cell
268	5805	PTS	Human	1	0	0	1	1	1	0	1	1	0	1	1	1	1
7500	5806	PTX3	Human	1	1	0	0	1	1	0	1	1	0	0	0	0	1
7922	6150	MRPL23	Human	1	0	0	1	1	1	0	1	1	0	1	1	1	1
2212	6182	MRPL12	Human	1	0	0	1	1	1	0	1	1	0	1	1	1	1
56005	6328	SCN3A	Human	1	0	1	0	0	1	0	0	0	0	1	1	1	1
55918	6882	TAF11	Human	1	0	1	0	1	1	0	0	1	0	1	1	1	1
65328	7559	ZNF12	Human	1	0	0	1	1	1	0	1	1	1	1	1	1	1
2593	7913	DEK	Human	1	1	0	1	1	1	0	1	1	1	0	1	1	1
2880	8835	SOCS2	Human	1	0	1	1	1	1	0	1	1	0	1	1	1	1
4397	8971	H1FX	Human	1	0	0	0	1	1	0	1	1	0	1	1	1	1
3476	9197	SLC33A1	Human	1	0	0	1	1	1	0	1	1	1	1	1	1	1
1290	9275	BCL7B	Human	1	0	1	1	1	1	0	0	1	0	1	1	1	1
68420	9559	VPS26A	Human	1	0	1	1	1	1	0	1	1	1	1	1	1	1
40668	9646	SH2BP1	Human	1	1	0	1	1	1	0	0	1	1	1	1	1	1
7291	10683	DLL3	Human	1	0	0	0	0	1	0	1	1	0	0	0	1	1
21334	10912	GADD45G	Human	1	0	0	1	1	1	0	1	1	0	0	1	1	1
4983	10991	SLC38A3	Human	1	0	0	1	0	0	0	0	0	0	1	1	0	0
7199	11054	OGFR	Human	1	0	0	1	1	1	0	0	0	0	0	1	1	1
6535	11062	DUS4L	Human	1	0	0	1	1	1	0	1	1	0	1	1	0	1
84799	22835	ZFP30	Human	1	0	0	0	1	1	0	1	1	1	1	1	1	1
65318	23361	ZNF629	Human	1	0	0	0	1	1	0	1	0	0	1	1	1	1
7717	24147	FJX1	Human	0	0	0	0	1	1	0	0	0	0	1	1	0	1
9120	25851	DKFZP434B0335	Human	1	0	0	0	1	1	0	0	0	0	0	0	1	1
56571	26503	SLC17A5	Human	1	0	0	1	1	1	0	1	1	0	1	1	1	1
31656	27000	ZRF1	Human	1	0	1	1	1	1	0	1	1	0	1	1	1	1



HomoloGene ID	GeneID	Gene Symbol	Species	PN: FRS	PN: heart	PN: kidney	PN: liver	PN: lung	PN: MRS	PN: stem cell	D: FRS	D: heart	D: kidney	D: liver	D: lung	D: MRS	D: stem cell
40859	27166	PX19	Human	1	0	1	1	1	1	0	1	1	1	1	1	1	1
22818	29850	TRPM5	Human	0	0	0	0	0	0	0	1	1	1	1	1	1	0
32293	51018	CGI-115	Human	1	0	0	1	1	0	0	1	1	1	1	1	1	1
9355	51637	C14orf166	Human	1	1	1	1	1	1	0	1	1	1	1	1	1	1
32331	51776	ZAK	Human	1	1	1	1	1	1	0	1	1	1	1	1	1	1
56774	54751	FBLIM1	Human	1	1	1	0	1	1	0	1	1	1	1	1	1	1
10026	55172	C14orf104	Human	1	0	1	1	1	1	0	1	1	0	1	1	1	1
9813	55627	FLJ20297	Human	1	0	1	0	1	1	0	1	1	1	1	1	1	1
81871	56751	BARHL1	Human	0	0	0	0	0	0	0	1	1	1	1	1	1	0
68934	57016	AKR1B10	Human	1	0	1	1	1	1	0	1	1	1	1	1	1	1
10774	57045	TWSG1	Human	1	0	0	0	1	1	0	1	1	0	1	1	1	1
10695	57120	GOPC	Human	1	0	1	0	1	1	0	1	1	0	1	1	1	1
10663	57171	DOLPP1	Human	1	0	0	0	1	0	0	1	1	0	1	1	1	1
10494	58516	FAM60A	Human	1	0	1	1	1	1	0	1	1	1	1	1	1	1
32546	64410	KLHL25	Human	1	0	0	0	0	1	0	1	1	0	0	1	1	1
45198	65117	FLJ11021	Human	1	0	1	1	1	1	0	1	1	0	1	1	1	1
15843	79591	C10orf76	Human	1	1	1	1	1	1	0	1	1	1	1	1	1	1
11653	79730	FLJ14001	Human	1	0	1	0	0	1	0	1	1	1	1	1	1	1
49970	83879	CDCA7	Human	1	0	0	0	1	1	0	1	1	0	1	1	1	1
12993	84217	ZMYND12	Human	0	0	0	0	0	1	0	1	1	1	1	1	1	0
11980	84262	MGC10911	Human	1	0	0	0	1	1	0	1	0	0	1	1	1	1
10518	84273	C4orf14	Human	1	0	0	1	0	1	0	1	1	1	1	1	1	1
11920	84303	CHCHD6	Human	1	0	1	0	1	1	0	1	1	0	1	1	1	1
12021	84557	MAP1LC3A	Human	1	0	1	0	1	1	0	1	1	0	0	1	0	1
27813	84865	FLJ14397	Human	1	0	0	1	0	1	0	0	0	0	1	0	0	1



HomoloGene ID	GeneID	Gene Symbol	Species	PN: FRS	PN: heart	PN: kidney	PN: liver	PN: lung	PN: MRS	PN: stem cell	D: FRS	D: heart	D: kidney	D: liver	D: lung	D: MRS	D: stem cell
12444	84902	FLJ14640	Human	1	0	1	0	1	1	0	1	1	1	1	1	1	1
14157	90416	CCDC32	Human	1	0	1	1	1	1	0	1	1	0	1	1	1	1
14128	91107	TRIM47	Human	1	0	1	1	1	1	0	1	1	0	1	1	1	0
35002	93082	LINCR	Human	0	0	0	0	1	1	0	1	1	1	1	1	1	0
26702	93109	TMEM44	Human	0	0	0	0	1	1	0	1	1	0	1	1	1	1
14667	113510	HEL308	Human	0	0	0	0	1	1	0	1	1	1	1	1	1	1
17523	115290	FBXO17	Human	0	1	1	1	0	1	0	0	1	0	1	1	1	1
14180	115294	PCMTD1	Human	0	1	1	1	1	1	0	1	1	0	1	1	1	1
41703	118881	COMTD1	Human	0	0	0	0	1	1	1	0	0	0	0	1	1	0
12418	124056	NOXO1	Human	0	0	0	0	0	1	0	1	1	0	0	1	1	1
64353	126374	WTIP	Human	0	0	1	1	1	1	0	1	1	1	1	1	1	1
32633	136647	C7orf11	Human	0	0	1	1	1	1	0	1	1	0	1	1	1	1
45867	139189	DGKK	Human	0	0	0	0	0	0	0	1	1	1	1	1	1	1
18123	140730	RIMS4	Human	0	0	0	0	1	1	0	0	1	0	0	1	1	1
49899	143282	C10orf13	Human	0	0	0	0	0	1	0	1	1	0	1	0	0	1
18833	143678	LOC143678	Human	0	0	0	0	0	0	0	1	1	1	1	1	1	0
19028	146167	LOC146167	Human	0	0	0	0	0	1	0	1	1	1	1	1	1	0
82250	150678	MYEOV2	Human	1	0	0	0	1	1	0	1	0	0	1	1	1	1
68998	170302	ARX	Human	1	0	0	0	1	1	0	1	0	0	0	0	0	0
24848	266629	SEC14L3	Human	0	0	0	0	0	0	0	1	1	1	1	1	1	0
65280	286128	ZFP41	Human	1	0	1	0	1	1	0	0	0	0	0	0	0	1
78698	387876	LOC387876	Human	0	0	0	0	0	0	0	1	1	1	1	1	1	0
17078	387914	TMEM46	Human	0	0	0	1	1	0	0	1	1	1	1	1	1	1
7516	389075	RESP18	Human	0	0	0	0	0	0	0	1	1	1	1	1	1	0
16890	399664	RKHD1	Human	1	0	0	0	1	1	0	1	1	1	1	1	1	1

HomoloGene ID	Gene ID	Gene Symbol	Species	PN: FRS	PN: heart	PN: kidney	PN: liver	PN: lung	PN: MRS	PN: stem cell	D: FRS	D: heart	D: kidney	D: liver	D: lung	D: MRS	D: stem cell
46116	401399	LOC401399	Human	0	0	0	0	0	0	0	1	1	1	1	1	1	1
18903	440193	KIAA1509	Human	1	0	1	0	1	1	0	1	1	1	1	1	1	1
3476	11416	Slc33a1	Mouse	1	1	1	1	1	1	1	0	0	0	0	0	1	1
413	11821	Aprt	Mouse	1	0	1	1	1	1	0	0	1	1	1	0	0	1
68998	11878	Arx	Mouse	0	0	0	0	0	1	0	1	1	0	0	0	1	0
1290	12054	Bcl7b	Mouse	1	1	1	1	1	1	0	1	1	1	1	0	0	0
55599	12183	Bpgm	Mouse	1	1	1	1	1	1	0	1	1	1	1	0	1	0
1330	12389	Cav1	Mouse	1	1	1	1	1	1	1	1	1	1	1	1	0	0
1368	12611	Cebpg	Mouse	1	1	1	1	1	1	0	0	1	1	1	0	1	1
7667	12700	Cish	Mouse	1	1	1	1	1	1	0	0	1	0	1	0	0	0
55434	12831	Col5a1	Mouse	1	1	1	0	1	1	0	0	1	0	1	0	1	1
37917	12835	Col6a3	Mouse	1	0	1	0	0	1	0	1	1	0	1	0	1	0
1028	13139	Dgka	Mouse	0	0	0	0	0	1	1	0	0	0	0	0	0	0
7291	13389	Dll3	Mouse	0	0	0	0	0	1	0	0	0	0	0	0	0	1
68934	14187	Akr1b8	Mouse	0	0	0	1	1	1	0	0	0	0	0	0	1	0
7717	14221	Fjx1	Mouse	1	0	0	0	1	0	0	0	0	0	0	0	0	0
20549	17388	Mmp15	Mouse	1	1	1	1	1	1	1	1	1	1	1	1	0	0
1871	18012	Neurod1	Mouse	0	1	1	1	0	1	0	0	0	0	0	0	0	0
1933	18476	Pafah1b3	Mouse	1	1	1	1	1	1	0	1	1	1	1	1	1	1
268	19286	Pts	Mouse	1	1	1	1	0	1	1	1	1	1	1	0	0	1
7500	19288	Ptx3	Mouse	0	1	0	0	0	0	0	0	1	0	1	0	1	0
7516	19711	Resp18	Mouse	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7922	19935	Mrpl23	Mouse	1	1	1	1	1	1	0	0	0	1	1	1	0	1
56005	20269	Scn3a	Mouse	0	0	0	0	1	0	0	0	0	0	0	0	0	0
40668	22083	Sh2bp1	Mouse	1	1	1	1	0	1	1	0	0	0	0	0	1	1

HomoloGene ID	GeneID	Gene Symbol	Species	PN: FRS	PN: heart	PN: kidney	PN: liver	PN: lung	PN: MRS	PN: stem cell	D: FRS	D: heart	D: kidney	D: liver	D: lung	D: MRS	D: stem cell
84799	22693	Zfp30	Mouse	0	0	0	0	0	0	0	1	1	1	1	1	0	0
65280	22701	Zfp41	Mouse	0	0	1	0	0	1	0	0	0	1	1	0	1	1
31656	22791	Dnajc2	Mouse	1	1	1	1	0	1	1	1	1	1	1	0	0	1
21334	23882	Gadd45g	Mouse	1	1	1	0	0	0	1	1	1	1	1	1	1	1
68420	30930	Vps26	Mouse	1	1	1	1	1	1	1	1	1	1	1	1	0	0
17523	50760	Fbxo17	Mouse	0	0	1	0	0	0	0	0	0	0	0	0	0	0
81871	54422	Barhl1	Mouse	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2212	56282	Mrpl12	Mouse	1	1	1	1	1	1	0	1	1	1	1	0	1	1
10494	56306	Tera	Mouse	1	1	1	1	1	1	1	1	1	1	0	1	1	1
10518	56412	2610024G14Rik	Mouse	0	1	1	1	0	1	1	0	0	0	0	0	1	0
22818	56843	Trpm5	Mouse	0	0	0	0	1	1	0	0	0	0	1	1	1	0
10663	57170	Dolpp1	Mouse	1	1	1	0	1	1	1	0	1	0	1	1	1	1
10774	65960	Twsg1	Mouse	1	1	1	1	1	1	1	1	1	1	1	1	1	1
32331	65964	B230120H23Rik	Mouse	1	1	1	0	1	1	1	1	0	0	1	0	1	1
11920	66098	Chchd6	Mouse	1	1	1	1	1	1	1	0	1	0	1	0	0	1
32633	66308	2810021B07Rik	Mouse	1	1	1	1	0	1	1	1	0	0	1	0	1	1
40859	66494	2610524G07Rik	Mouse	1	1	1	1	1	1	1	1	1	1	1	0	1	1
11980	66506	1810042K04Rik	Mouse	1	0	1	1	0	1	0	0	0	0	1	0	0	1
12021	66734	Map1lc3a	Mouse	1	1	1	1	0	1	1	0	1	0	0	0	1	1
82250	66915	Myeov2	Mouse	1	1	1	1	0	1	1	0	1	1	0	0	0	1
49970	66953	Cdca7	Mouse	1	1	1	1	0	1	1	1	1	1	1	1	1	1
32293	67223	2810430M08Rik	Mouse	1	0	0	0	1	1	1	1	1	1	1	1	1	1
9355	68045	2700060E02Rik	Mouse	1	1	1	1	1	1	1	1	1	0	1	1	1	1
18903	68339	0610010D24Rik	Mouse	1	0	1	0	0	1	1	0	1	0	1	0	0	0
55918	68776	Taf11	Mouse	1	1	1	1	1	1	0	0	1	0	0	0	1	1



HomoloGene ID	GeneID	Gene Symbol	Species	PN: FRS	PN: heart	PN: kidney	PN: liver	PN: lung	PN: MRS	PN: stem cell	D: FRS	D: heart	D: kidney	D: liver	D: lung	D: MRS	D: stem cell
41703	69156	Comtd1	Mouse	1	1	1	1	0	1	0	0	0	0	1	0	0	1
9120	70381	2210010N04Rik	Mouse	1	1	1	1	1	1	1	1	1	1	1	1	0	1
11653	70918	4921525L17Rik	Mouse	0	0	0	0	0	1	1	0	0	0	0	0	0	0
15843	71617	9130011E15Rik	Mouse	1	1	1	0	0	0	1	0	0	0	0	0	0	1
12418	71893	Noxol	Mouse	0	0	1	1	0	1	0	1	0	0	1	1	1	0
6535	71916	Dus4l	Mouse	0	0	0	0	0	0	1	0	0	0	0	0	0	1
7199	72075	Ogfr	Mouse	0	1	1	1	0	1	0	0	1	1	1	0	1	1
12444	72140	2610507L03Rik	Mouse	0	1	1	1	0	1	1	1	0	0	0	0	1	1
49899	72514	2610306H15Rik	Mouse	0	0	0	0	0	0	0	0	1	0	0	0	1	0
56774	74202	Fblim1	Mouse	1	1	1	1	1	1	0	1	1	1	1	1	0	1
18833	75641	1700029I15Rik	Mouse	0	0	0	1	0	1	0	0	0	0	0	0	0	0
4983	76257	Slc38a3	Mouse	0	0	1	1	0	0	1	0	0	0	0	0	0	0
9813	77626	4122402O22Rik	Mouse	1	0	0	1	0	1	1	1	1	1	0	1	1	1
10695	94221	Gopc	Mouse	0	0	0	0	0	1	1	1	1	0	0	0	1	1
46116	101359	D330027H18Rik	Mouse	0	0	0	0	0	0	0	0	1	0	0	0	0	0
64353	101543	Wtip	Mouse	1	1	0	0	1	0	0	0	1	0	0	1	0	0
10026	109065	1110034A24Rik	Mouse	0	1	1	1	0	1	1	1	1	1	1	0	1	1
32	109900	Asl	Mouse	1	1	1	1	1	1	1	1	1	1	1	0	1	1
2593	110052	Dek	Mouse	1	1	1	1	0	1	1	0	1	1	1	0	0	1
14667	191578	Hel308	Mouse	0	0	1	0	0	0	1	0	0	0	1	1	0	0
32546	207952	Klhl25	Mouse	1	0	1	0	0	1	1	1	1	1	0	1	0	1
45198	208606	1500011J06Rik	Mouse	1	0	1	1	1	1	1	1	1	1	1	0	1	1
35002	214854	Lincr	Mouse	1	0	1	0	1	0	1	0	0	0	0	0	0	0
2880	216233	Socs2	Mouse	1	1	1	1	1	1	0	1	1	1	1	1	1	1
14128	217333	Trim47	Mouse	1	1	1	1	1	1	1	0	1	0	1	0	1	0



## Appendix V The individual mouse developmental ontologies

### TS01

first polar body  
one-cell stage  
second polar body  
unclassifiable  
zona pellucida

### TS02

second polar body  
two-cell stage  
unclassifiable  
zona pellucida

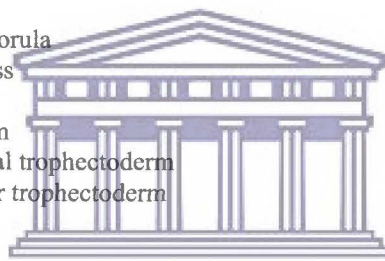
### TS03

4-8 cell stage  
compacted morula  
second polar body  
unclassifiable  
zona pellucida

### TS04

blastocoelic cavity  
embryo  
germ layers  
second polar body  
unclassifiable  
zona pellucida

compacted morula  
inner cell mass  
trophectoderm  
mural trophectoderm  
polar trophectoderm



UNIVERSITY of the  
WESTERN CAPE

### TS05

blastocoelic cavity  
embryo  
germ layers  
unclassifiable

inner cell mass  
trophectoderm  
mural trophectoderm  
polar trophectoderm

### TS06

blastocoelic cavity  
embryo  
germ layers  
unclassifiable

epiblast  
primitive endoderm  
trophectoderm  
mural trophectoderm  
polar trophectoderm

### TS07

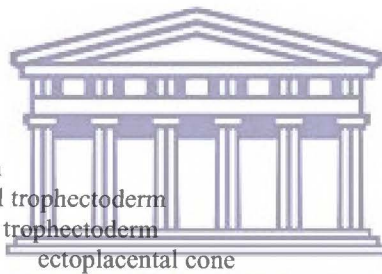
- embryo
  - epiblast
- germ layers
  - endoderm
  - trophectoderm
    - mural trophoctoderm
    - polar trophoctoderm
    - ectoplacental cone
- unclassifiable
- yolk sac cavity

**TS08**

- embryo
  - epiblast
- germ layers
  - ectoderm
  - endoderm
  - trophectoderm
    - mural trophoctoderm
    - polar trophoctoderm
    - ectoplacental cone
- unclassifiable
- yolk sac cavity

**TS09**

- embryo
- germ layers
  - ectoderm
  - endoderm
  - mesoderm
  - trophectoderm
    - mural trophoctoderm
    - polar trophoctoderm
    - ectoplacental cone
- primitive streak
- proamniotic cavity
- unclassifiable
- yolk sac cavity



UNIVERSITY of the  
WESTERN CAPE

**TS10**

- allantois
- embryo
- germ layers
  - ectoderm
  - endoderm
  - mesoderm
  - trophectoderm
    - mural trophoctoderm
    - polar trophoctoderm
    - ectoplacental cone
- primitive streak
- unclassifiable
- yolk sac

**TS11**

- allantois
- amnion
- anatomical site

- hematological system
  - blood island
- nervous system
  - central nervous system {CNS}
    - floor plate
    - future brain
      - future midbrain
      - future prosencephalon
      - future rhombencephalon
    - future spinal cord
      - neural tube
    - neural crest
    - notochord
  - peripheral nervous system {PNS}
    - auditory apparatus {ear}
      - internal ear
    - visual apparatus {eye}
- primitive streak
- unclassifiable
- yolk sac

**TS13**

- alimentary system
  - diverticulum
  - intestine {gut}
  - mesentery
- anatomical site
  - head
  - trunk
  - whole body
- branchial arch
- cardiovascular system
  - artery



- carotid artery
  - dorsal aorta
- heart
  - common atrial chamber
  - mesocardium
  - myocardium
  - primitive ventricle
  - sinus venosus
- vein
- endocrine system
  - thyroid primordium
- germ layers
  - ectoderm
  - endoderm
  - mesenchyme
- hematological system
  - blood
- nervous system
  - central nervous system {CNS}
    - floor plate
    - future brain
      - future midbrain
      - future prosencephalon
      - future rhombencephalon
    - future spinal cord

- neural tube
- neural crest
- notochord
- peripheral nervous system {PNS}
- auditory apparatus {ear}
- internal ear
- olfactory apparatus
- visual apparatus {eye}
- primitive streak
- unclassifiable
- urogenital system
  - nephric cord
  - presumptive nephric duct

**TS14**

- alimentary system
  - diverticulum
  - intestine {gut}
  - mesentery
- anatomical site
  - anterior limb bud
  - head
  - tail bud
  - trunk
  - whole body
- branchial arch
- cardiovascular system
  - artery
    - carotid artery
    - dorsal aorta
  - heart
    - common atrial chamber
    - mesocardium
    - myocardium
    - primitive ventricle
    - sinus venosus
  - vein
- endocrine system
  - pituitary gland
  - thyroid primordium
- germ layers
  - ectoderm
  - endoderm
  - mesenchyme
- hematological system
  - blood
- nervous system
  - central nervous system {CNS}
  - floor plate
  - future brain
    - future forebrain
    - future diencephalon
    - future midbrain
    - future rhombencephalon
    - prosencephalon
    - ventricular system
      - fourth ventricle
      - third ventricle

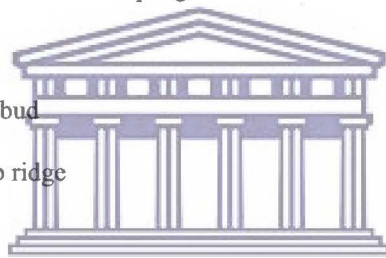




- future spinal cord
  - neural tube
- neural crest
- notochord
- peripheral nervous system {PNS}
  - auditory apparatus {ear}
    - internal ear
  - olfactory apparatus
  - visual apparatus {eye}
- primitive streak
- respiratory system
  - nose
- unclassifiable
- urogenital system
  - nephric cord
  - nephric duct
  - pronephros

**TS15**

- alimentary system
  - diverticulum
  - gall bladder primordium
  - intestine {gut}
  - mesentery
    - dorsal meso-oesophagus
  - oral cavity
  - pharynx
- anatomical site
  - anterior limb bud
  - head
  - posterior limb ridge
  - tail
  - trunk
  - whole body
- branchial arch
- cardiovascular system
  - artery
    - carotid artery
    - dorsal aorta
  - heart
    - atrium
      - common atrial chamber
    - mesocardium
    - myocardium
    - primitive ventricle
    - sinus venosus
  - vein
- endocrine system
  - pituitary gland
  - thyroid primordium
- germ layers
  - ectoderm
  - endoderm
  - mesenchyme
- hematological system
  - blood
- musculoskeletal system
  - pre-cartilage condensation

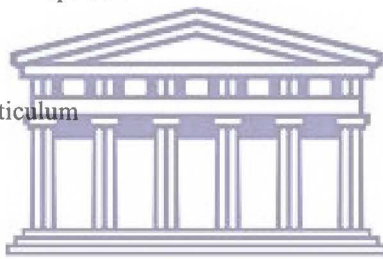


UNIVERSITY of the  
WESTERN CAPE

- nervous system
  - central nervous system {CNS}
    - floor plate
    - future brain
      - future forebrain
        - diencephalon
        - telencephalon
      - future midbrain
      - future rhombencephalon
    - ventricular system
      - fourth ventricle
      - third ventricle
  - future spinal cord
    - neural tube
  - neural crest
  - notochord
  - peripheral nervous system {PNS}
    - auditory apparatus {ear}
      - internal ear
        - otocyst
    - ganglion
    - olfactory apparatus
    - visual apparatus {eye}
      - intraretinal space
      - optic stalk

- respiratory system
  - lung
  - nose
  - tracheal diverticulum

- unclassifiable
- urogenital system
  - mesonephros
  - nephric cord
  - nephric duct



UNIVERSITY of the  
WESTERN CAPE

TS17

- alimentary system
  - diverticulum
  - intestine
    - large intestine
      - anal region
    - small intestine
  - liver and biliary system
    - cystic duct
    - gall bladder primordium
    - hepatic duct
    - liver
  - mesentery
  - oesophagus
  - oral cavity
  - pharynx
  - stomach
- anatomical site
  - anterior limb bud
  - head
  - posterior limb bud
  - tail
  - trunk

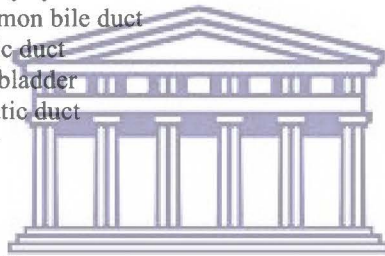
- whole body
- branchial arch
- cardiovascular system
  - artery
    - carotid artery
    - dorsal aorta
  - heart
    - atrium
      - common atrial chamber
    - mesocardium
    - myocardium
    - primitive ventricle
    - sinus venosus
    - valve
  - vein
- dermal system
  - dermis
  - epidermis
- endocrine system
  - pituitary gland
  - thyroid primordium
- germ layers
  - ectoderm
  - endoderm
  - mesenchyme
- hematological system
  - blood
- lymphoreticular system
- musculoskeletal system
  - cartilage condensation
  - pre-cartilage condensation
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
        - diencephalon
        - telencephalon
      - hindbrain
        - trigeminal V
      - midbrain
        - ventricular system
          - fourth ventricle
          - lateral ventricle
          - third ventricle
    - floor plate
    - future spinal cord
      - neural tube
    - notochord
  - peripheral nervous system {PNS}
    - auditory apparatus {ear}
      - external ear
      - internal ear
        - otocyst
      - middle ear
    - ganglion
    - olfactory apparatus
    - peripheral nerve
    - visual apparatus {eye}



- intraretinal space
- optic stalk
- respiratory system
  - bronchus
  - lung
  - nose
  - trachea
- unclassifiable
- urogenital system
  - reproductive system
    - gonadal component
  - urinary system
    - mesonephros
    - nephric cord
    - nephric duct

**TS18**

- alimentary system
  - diverticulum
  - intestine
    - large intestine
      - anal region
    - small intestine
      - duodenum
  - liver and biliary system
    - common bile duct
    - cystic duct
    - gall bladder
    - hepatic duct
    - liver
  - mesentery
  - oesophagus
  - oral cavity
    - tongue
  - pancreas primordium
  - pharynx
  - stomach
- anatomical site
  - anterior limb bud
  - head
  - posterior limb bud
  - tail
  - trunk
  - whole body
- branchial arch
- cardiovascular system
  - artery
    - carotid artery
    - dorsal aorta
  - heart
    - atrium
      - common atrial chamber
    - mesocardium
    - myocardium
    - pericardium
    - primitive ventricle
    - sinus venosus
    - valve



UNIVERSITY of the  
WESTERN CAPE



- vein
- dermal system
  - dermis
  - epidermis
- endocrine system
  - pituitary gland
  - thyroid
- germ layers
  - ectoderm
  - endoderm
  - mesenchyme
- hematological system
  - blood
- lymphoreticular system
- musculoskeletal system
  - cartilage condensation
  - pre-cartilage condensation
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
        - diencephalon
        - telencephalon
      - hindbrain
        - metencephalon
          - cerebellum primordium
          - facial VII
          - trigeminal V
        - myelencephalon
          - midbrain
          - ventricular system
            - fourth ventricle
            - lateral ventricle
            - third ventricle
  - floor plate
  - future spinal cord
    - neural tube
    - notochord
- peripheral nervous system {PNS}
  - auditory apparatus {ear}
    - external ear
    - internal ear
      - otocyst
    - middle ear
  - ganglion
    - sympathetic ganglion
  - olfactory apparatus
  - peripheral nerve
  - visual apparatus {eye}
    - cornea
    - lens vesicle
    - optic stalk
    - retina
- respiratory system
  - bronchus
  - lung
  - nose
  - trachea



- unclassifiable
- urogenital system
  - reproductive system
    - gonad primordium
  - urinary system
    - mesonephros
    - metanephros
    - nephric duct
    - ureteric bud

**TS19**

- alimentary system
  - diverticulum
  - intestine
    - large intestine
      - anal pit
    - small intestine
      - duodenum
  - liver and biliary system
    - common bile duct
    - cystic duct
    - gall bladder
    - hepatic duct
    - liver

- mesentery

- oesophagus

- oral cavity
  - mandibular process
    - mandible primordium
  - maxillary process
    - maxilla primordium
  - tongue

- pancreas primordium

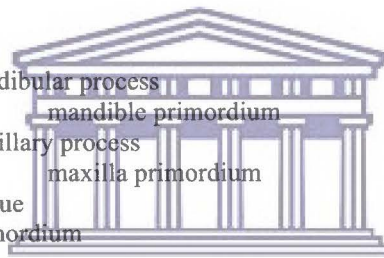
- pharynx

- stomach

- anatomical site
  - anterior limb bud
  - head
  - posterior limb bud
  - tail
  - trunk
  - whole body

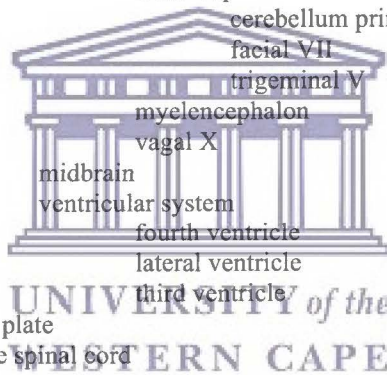
- branchial arch

- cardiovascular system
  - artery
    - carotid artery
    - dorsal aorta
  - heart
    - atrium
    - mesocardium
    - myocardium
    - pericardium
    - sinus venosus
    - valve
    - ventricle
  - vein
    - vena cava
      - inferior vena cava



UNIVERSITY of the  
WESTERN CAPE

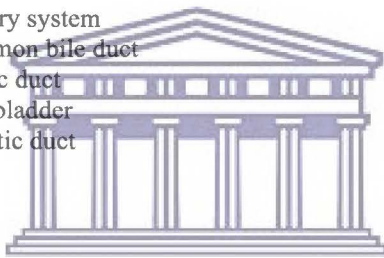
- dermal system
  - dermis
  - epidermis
- endocrine system
  - pituitary gland
  - thyroid
- germ layers
  - ectoderm
  - endoderm
  - mesenchyme
- hematological system
  - blood
- lymphoreticular system
- musculoskeletal system
  - cartilage condensation
  - pre-cartilage condensation
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
        - diencephalon
        - telencephalon
      - hindbrain
        - hypoglossal XII
        - metencephalon
          - cerebellum primordium
          - facial VII
          - trigeminal V
    - myelencephalon
      - vagal X
    - midbrain
      - ventricular system
        - fourth ventricle
          - lateral ventricle
          - third ventricle
  - floor plate
  - future spinal cord
  - neural tube
  - notochord
- peripheral nervous system {PNS}
  - auditory apparatus {ear}
    - external ear
    - future tympanum
    - internal ear
      - membranous labyrinth
        - saccule
        - utricle
      - osseous labyrinth
        - semicircular canal
    - middle ear
  - ganglion
    - sympathetic ganglion
  - olfactory apparatus
  - peripheral nerve
  - visual apparatus {eye}
    - cornea
    - lens vesicle
    - optic stalk



- retina
- respiratory system
  - bronchus
  - lung
  - nose
  - trachea
- unclassifiable
- urogenital system
  - reproductive system
    - genital tubercle
    - gonad primordium
  - urinary system
    - mesonephros
    - metanephros
    - nephric duct
    - ureteric bud

**TS20**

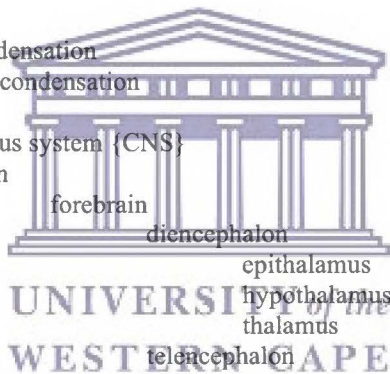
- alimentary system
  - diverticulum
  - intestine
    - large intestine
      - anal pit
    - small intestine
      - duodenum
  - liver and biliary system
    - common bile duct
    - cystic duct
    - gall bladder
    - hepatic duct
    - liver
  - mesentery
  - oesophagus
  - oral cavity
    - mandibular process
    - mandible primordium
    - maxillary process
    - maxilla
    - premaxilla
  - tongue
  - pancreas
  - pharynx
    - nasopharynx
  - stomach
- anatomical site
  - anterior limb
  - head
  - posterior limb
  - tail
  - trunk
  - whole body
- cardiovascular system
  - artery
    - carotid artery
    - dorsal aorta
  - heart
    - atrium
    - mesocardium



UNIVERSITY of the  
WESTERN CAPE



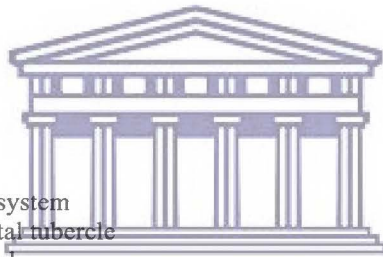
myocardium  
 pericardium  
 sinus venosus  
 valve  
 ventricle  
 vein  
     vena cava  
         inferior vena cava  
 dermal system  
     appendages  
         vibrissa  
     skin  
         dermis  
         epidermis  
 endocrine system  
     pituitary gland  
     thymus primordium  
     thyroid  
 germ layers  
     mesenchyme  
 hematological system  
     blood  
 lymphoreticular system  
 musculoskeletal system  
     bone  
     cartilage  
     cartilage condensation  
     pre-cartilage condensation  
 nervous system  
     central nervous system {CNS}  
         brain  
             forebrain  
                 diencephalon  
                     epithalamus  
                     hypothalamus  
                     thalamus  
                 telencephalon  
                     cerebral cortex  
                     corpus striatum  
             hindbrain  
                 medulla oblongata  
                     hypoglossal XII  
                     vagal X  
                 metencephalon  
                     cerebellum primordium  
                     pons  
                         facial VII  
                         trigeminal V  
                         vestibulocochlear VIII  
             midbrain  
                 oculomotor III  
             ventricular system  
                 fourth ventricle  
                 lateral ventricle  
                 third ventricle  
     floor plate  
     notochord  
     spinal cord



- peripheral nervous system {PNS}
- auditory apparatus {ear}
  - external ear
    - auricle
    - external acoustic meatus
  - middle ear
    - future tympanum
    - internal ear
      - membranous labyrinth
        - saccule
        - utricle
      - osseous labyrinth
        - cochlea
        - semicircular canal
  - ganglion
    - sympathetic ganglion
- olfactory apparatus
- peripheral nerve
- visual apparatus {eye}
  - cornea
  - lens vesicle
  - optic chiasma
  - optic stalk
  - retina

- respiratory system
  - bronchus
  - lung
  - nose
  - trachea
- unclassifiable
- urogenital system
  - reproductive system
    - genital tubercle
    - gonad
  - urinary system
    - mesonephros
    - metanephros
    - nephric duct
    - primitive ureter

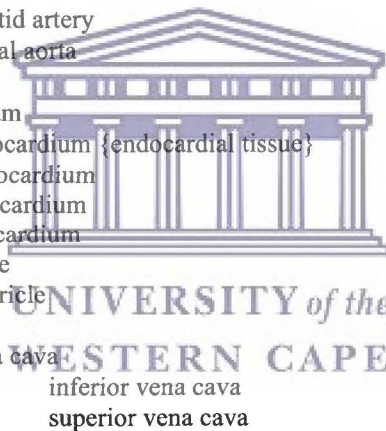


UNIVERSITY of the  
WESTERN CAPE

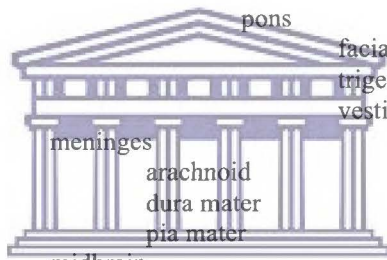
**TS21**

- alimentary system
  - intestine
    - large intestine
      - anal pit
      - colorectal
      - rectum
    - small intestine
      - duodenum
  - liver and biliary system
    - common bile duct
    - cystic duct
    - gall bladder
    - hepatic duct
    - liver
  - mesentery
  - oesophagus
  - omentum

- lesser omentum
- oral cavity
  - jaw
    - mandible
    - maxilla
    - premaxilla
    - tooth
      - molar
  - salivary gland
    - sublingual gland primordium
    - submandibular gland primordium
- tongue
- pancreas
- pharynx
  - nasopharynx
- stomach
- anatomical site
  - anterior limb
  - head
  - posterior limb
  - tail
  - trunk
  - whole body
- cardiovascular system
  - artery
    - carotid artery
    - dorsal aorta
  - heart
    - atrium
    - endocardium {endocardial tissue}
    - mesocardium
    - myocardium
    - pericardium
    - valve
    - ventricle
  - vein
    - vena cava
      - inferior vena cava
      - superior vena cava
- dermal system
  - appendages
    - vibrissa
  - skin
    - dermis
    - epidermis
- endocrine system
  - pituitary gland
  - thymus primordium
  - thyroid
- germ layers
  - mesenchyme
- hematological system
  - blood
- lymphoreticular system
- musculoskeletal system
  - bone
  - cartilage
  - cartilage condensation



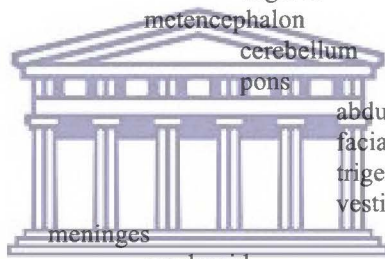
- joint
- ligament
- muscle
  - skeletal muscle {striated muscle}
- pre-cartilage condensation
- tendon
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
        - diencephalon
          - epithalamus
          - hypothalamus
          - thalamus
        - telencephalon
          - cerebral cortex
            - olfactory I
          - corpus striatum
          - olfactory lobe
        - hindbrain
          - medulla oblongata
            - hypoglossal XII
            - vagal X
          - metencephalon
            - cerebellum
            - pons
              - facial VII
              - trigeminal V
              - vestibulocochlear VIII
  - meninges
    - arachnoid
    - dura mater
    - pia mater
  - midbrain
    - oculomotor III
    - ventricular system
      - cerebral aqueduct
      - fourth ventricle
      - lateral ventricle
      - third ventricle
- peripheral nervous system {PNS}
  - auditory apparatus {ear}
    - auditory ossicle
    - external ear
      - auricle
      - external acoustic meatus
    - future tympanum
    - internal ear
      - membranous labyrinth
        - sacculle
        - utricle
      - osseous labyrinth
        - cochlea
        - semicircular canal
    - middle ear
  - ganglion



UNIVERSITY of the  
WESTERN CAPE



- joint
- ligament
- muscle
  - skeletal muscle {striated muscle}
- pre-cartilage condensation
- tendon
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
        - diencephalon
          - epithalamus
          - hypothalamus
          - thalamus
        - telencephalon
          - caudate nucleus
          - cerebral cortex
            - olfactory I
          - corpus striatum
          - lentiform nucleus
          - olfactory lobe
      - hindbrain
        - medulla oblongata
          - hypoglossal XII
          - vagal X
        - metencephalon
          - cerebellum
          - pons
            - abducent VI
            - facial VII
            - trigeminal V
            - vestibulocochlear VIII
- meninges
  - arachnoid
  - dura mater
  - pia mater
- midbrain
  - oculomotor III
  - tegmentum
  - trochlear IV
- ventricular system
  - cerebral aqueduct
  - fourth ventricle
  - lateral ventricle
  - third ventricle

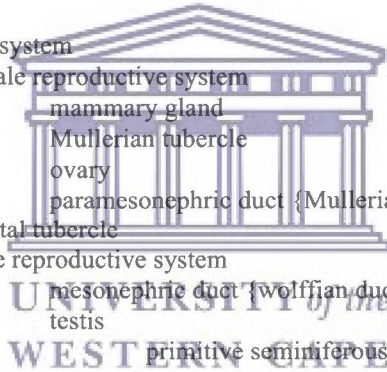
- floor plate
- spinal cord
- peripheral nervous system {PNS}
- auditory apparatus {ear}
  - auditory ossicle
  - external ear
    - auricle
    - external acoustic meatus
  - future tympanum
  - internal ear
    - membranous labyrinth
      - sacculle
      - utricle


UNIVERSITY of the  
WESTERN CAPE

osseous labyrinth  
     cochlea  
     semicircular canal  
 middle ear  
 ganglion  
     sympathetic ganglion  
 olfactory apparatus  
 peripheral nerve  
 visual apparatus {eye}  
     choroid  
     cornea  
     eyelid  
     lens  
     optic chiasma  
     optic stalk  
     retina  
     vitreous humor  
 respiratory system  
     bronchus  
     diaphragm  
     larynx  
     lung  
     nose  
     trachea  
 unclassifiable  
 urogenital system  
     reproductive system  
         female reproductive system  
             mammary gland  
             Mullerian tubercle  
             ovary  
             paramesonephric duct {Mullerian duct}  
         genital tubercle  
         male reproductive system  
             mesonephric duct {wolffian duct}  
             testis  
             primitive seminiferous tubule  
     urinary system  
         bladder  
         degenerating mesonephros  
         metanephros  
             nephron  
                 glomerulus  
         nephric duct  
         ureter

**TS23**

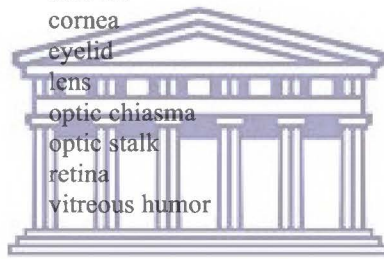
alimentary system  
     intestine  
         large intestine  
             anus  
             colorectal  
                 rectum  
         small intestine  
             duodenum  
             jejunum  
     liver and biliary system  
         common bile duct



- pineal primordium
- pituitary gland
- thymus primordium
- thyroid
- germ layers
  - mesenchyme
- hematological system
  - blood
- lymphoreticular system
  - lymph sac
  - spleen primordium
- musculoskeletal system
  - bone
  - cartilage
  - cartilage condensation
  - joint
    - ligament
  - muscle
    - skeletal muscle {striated muscle}
  - pre-cartilage condensation
  - tendon
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
        - diencephalon
          - epithalamus
          - hypothalamus
          - thalamus
        - telencephalon
          - caudate nucleus
          - cerebral cortex
          - olfactory I
      - hindbrain
        - medulla oblongata
          - floor plate
          - hypoglossal XII
          - vagal X
        - metencephalon
          - cerebellum
          - pons
            - abducent VI
            - facial VII
            - trigeminal V
            - vestibulocochlear VIII
    - meninges
      - arachnoid
      - dura mater
      - pia mater
    - midbrain
      - oculomotor III
      - tegmentum
      - trochlear IV
    - ventricular system
      - cerebral aqueduct



fourth ventricle  
 lateral ventricle  
 third ventricle  
 spinal cord  
 peripheral nervous system {PNS}  
 auditory apparatus {ear}  
 auditory ossicle  
 external ear  
 auricle  
 external acoustic meatus  
 future tympanum  
 internal ear  
 membranous labyrinth  
 saccule  
 utricle  
 osseous labyrinth  
 cochlea  
 semicircular canal  
 middle ear  
 ganglion  
 sympathetic ganglion  
 olfactory apparatus  
 peripheral nerve  
 visual apparatus {eye}  
 choroid  
 cornea  
 eyelid  
 lens  
 optic chiasma  
 optic stalk  
 retina  
 vitreous humor  
 respiratory system  
 bronchus  
 diaphragm  
 larynx  
 lung  
 nose  
 pleura {pleural cavity}  
 sinus {hindbrain}  
 trachea  
 unclassifiable  
 urogenital system  
 reproductive system  
 female reproductive system  
 mammary gland  
 Mullerian tubercle  
 ovary  
 paramesonephric duct {Mullerian duct}  
 genital tubercle  
 male reproductive system  
 penis  
 testis  
 primitive seminiferous tubule  
 vas deferens  
 urinary system  
 bladder  
 metanephros

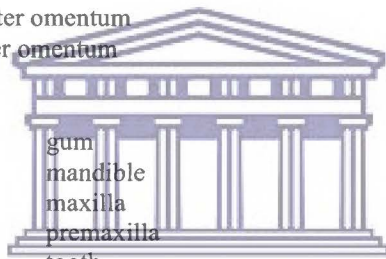


UNIVERSITY of the  
 WESTERN CAPE



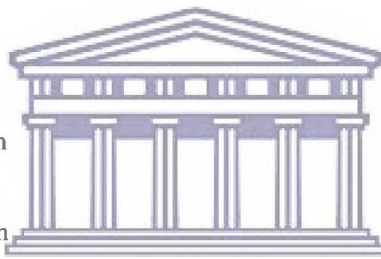
TS24

- nephron
  - glomerulus
- ureter
- urethra
- alimentary system
  - intestine
    - large intestine
      - anus
      - colorectal
        - colon
        - rectum
    - small intestine
      - duodenum
      - jejunum
  - liver and biliary system
    - common bile duct
    - cystic duct
    - gall bladder
    - hepatic duct
    - liver
  - mesentery
  - oesophagus
  - omentum
    - greater omentum
    - lesser omentum
  - oral cavity
    - jaw
      - gum
      - mandible
      - maxilla
      - premaxilla
      - tooth
        - molar
  - salivary gland
    - parotid gland
    - sublingual gland
    - submandibular gland
  - tongue
  - pancreas
  - pharynx
    - nasopharynx
  - stomach
- anatomical site
  - anterior limb
  - head
  - posterior limb
  - tail
  - trunk
  - whole body
- cardiovascular system
  - artery
    - aorta
    - carotid artery
  - heart
    - atrium
    - endocardium {endocardial tissue}



UNIVERSITY of the  
WESTERN CAPE

- mesocardium
- myocardium
- pericardium
- valve
- ventricle
- vein
  - vena cava
    - inferior vena cava
    - superior vena cava
- dermal system
  - appendages
    - hair
    - hair follicle
    - vibrissa
  - skin
    - dermis
    - epidermis
- endocrine system
  - adrenal gland
    - adrenal cortex
    - adrenal medulla
  - pineal gland
  - pituitary gland
  - thymus
  - thyroid
- germ layers
  - mesenchyme
- hematological system
  - blood
- lymphoreticular system
  - lymph sac
  - spleen
- musculoskeletal system
  - bone
  - cartilage
  - cartilage condensation
  - joint
    - ligament
  - muscle
    - skeletal muscle {striated muscle}
  - pre-cartilage condensation
  - tendon
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
        - diencephalon
          - epithalamus
          - hypothalamus
          - thalamus
        - telencephalon
          - caudate nucleus
          - cerebral cortex
            - olfactory I
          - corpus striatum
          - lentiform nucleus
          - olfactory lobe
          - temporal lobe



UNIVERSITY of the  
WESTERN CAPE

hindbrain  
     medulla oblongata  
         floor plate  
         hypoglossal XII  
         vagal X  
     metencephalon  
         cerebellum  
         pons  
             abducent VI  
             facial VII  
             trigeminal V  
             vestibulocochlear VIII

meninges  
     arachnoid  
     dura mater  
     pia mater

midbrain  
     oculomotor III  
     tegmentum  
     trochlear IV

ventricular system  
     cerebral aqueduct  
     fourth ventricle  
     lateral ventricle  
     third ventricle

spinal cord

peripheral nervous system {PNS}

auditory apparatus {ear}

auditory ossicle

external ear

    auricle

    external acoustic meatus

    future tympanum

internal ear

    membranous labyrinth

        sacculle

        utricle

    osseous labyrinth

        cochlea

        semicircular canal

middle ear

ganglion

    sympathetic ganglion

olfactory apparatus

peripheral nerve

visual apparatus {eye}

    choroid

    cornea

    eyelid

    lens

    optic chiasma

    optic stalk

    retina

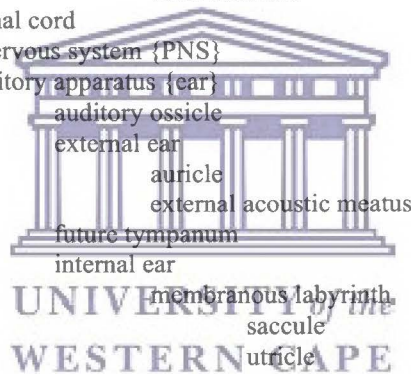
    sclera

    vitreous humor

respiratory system

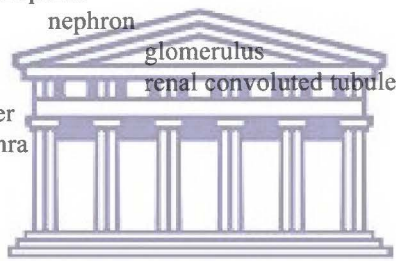
    bronchus

    diaphragm



- larynx
- lung
- nose
- pleura {pleural cavity}
- sinus {hindbrain,sinus}
- trachea
- unclassifiable
- urogenital system
  - reproductive system
    - female reproductive system
      - mammary gland
      - Mullerian tubercle
      - ovary
      - oviduct
      - vagina
    - genital tubercle
    - male reproductive system
      - penis
        - glans
      - testis
        - primitive seminiferous tubule
        - vas deferens
  - urinary system
    - bladder
    - metanephros
      - nephron
        - glomerulus
        - renal convoluted tubule
    - ureter
    - urethra
- alimentary system
  - intestine
    - large intestine
      - anus
      - colorecta
      - colon
      - rectum
    - small intestine
      - duodenum
      - jejunum
  - liver and biliary system
    - common bile duct
    - cystic duct
    - gall bladder
    - hepatic duct
    - liver
  - mesentery
  - oesophagus
  - omentum
    - greater omentum
    - lesser omentum
  - oral cavity
    - jaw
      - gum
      - mandible
      - maxilla

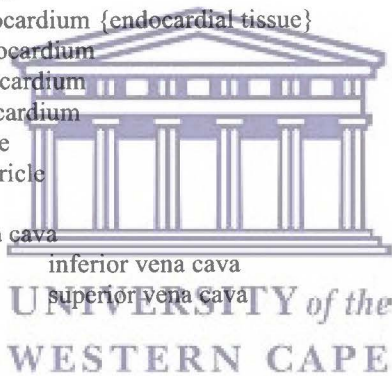
TS25



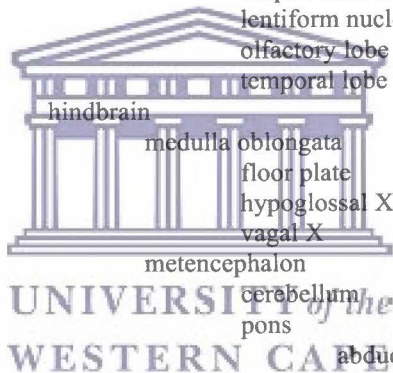
UNIVERSITY of the  
WESTERN CAPE



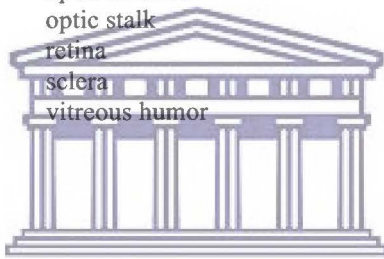
- premaxilla
- tooth
  - molar
- salivary gland
  - parotid gland
  - sublingual gland
  - submandibular gland
- tongue
- pancreas
- pharynx
  - nasopharynx
- stomach
- anatomical site
  - anterior limb
  - head
  - posterior limb
  - tail
  - trunk
  - whole body
- cardiovascular system
  - artery
    - aorta
    - carotid artery
  - heart
    - atrium
    - endocardium {endocardial tissue}
    - mesocardium
    - myocardium
    - pericardium
    - valve
    - ventricle
  - vein
    - vena cava
      - inferior vena cava
      - superior vena cava
- dermal system
  - appendages
    - hair
    - hair follicle
    - vibrissa
  - skin
    - dermis
    - epidermis
- endocrine system
  - adrenal gland
    - adrenal cortex
    - adrenal medulla
  - pineal gland
  - pituitary gland
  - thymus
  - thyroid
- germ layers
  - mesenchyme
- hematological system
  - blood
- lymphoreticular system
  - lymph sac
  - spleen



- musculoskeletal system
  - bone
  - cartilage
  - cartilage condensation
  - joint
    - ligament
  - muscle
    - skeletal muscle {striated muscle}
    - smooth muscle
  - pre-cartilage condensation
  - tendon
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
        - diencephalon
          - epithalamus
          - hypothalamus
          - thalamus
        - hippocampus
        - telencephalon
          - caudate nucleus
          - cerebral cortex
            - olfactory I
          - corpus striatum
          - lentiform nucleus
          - olfactory lobe
          - temporal lobe
      - hindbrain
        - medulla oblongata
          - floor plate
          - hypoglossal XII
          - vagal X
        - metencephalon
          - cerebellum
          - pons
            - abducent VI
            - facial VII
            - trigeminal V
            - vestibulocochlear VIII
  - meninges
    - arachnoid
    - dura mater
    - pia mater
  - midbrain
    - oculomotor III
    - tegmentum
    - trochlear IV
  - ventricular system
    - cerebral aqueduct
    - fourth ventricle
    - lateral ventricle
    - third ventricle
  - spinal cord
- peripheral nervous system {PNS}
  - auditory apparatus {ear}
    - auditory ossicle
    - external ear



auricle  
 external acoustic meatus  
 internal ear  
     membranous labyrinth  
         sacculle  
         utricle  
     osseous labyrinth  
         cochlea  
             spiral organ of Corti  
         semicircular canal  
  
 middle ear  
 tympanum primordium  
 ganglion  
     spinal ganglion  
     sympathetic ganglion  
 olfactory apparatus  
 peripheral nerve  
 visual apparatus {eye}  
     choroid  
     ciliary body  
     cornea  
     eyelid  
     iris  
     lens  
     optic chiasma  
     optic stalk  
     retina  
     sclera  
     vitreous humor  
  
 respiratory system  
     bronchus  
     diaphragm  
     larynx  
     lung  
         alveolus  
     nose  
     pleura {pleural cavity}  
     sinus {hindbrain, sinus}  
     trachea  
 unclassifiable  
 urogenital system  
     reproductive system  
         female reproductive system  
             mammary gland  
             Mullerian tubercle  
             ovary  
             oviduct  
             vagina  
         genital tubercle  
         male reproductive system  
             penis  
                 glans  
             testis  
                 primitive seminiferous tubule  
             vas deferens  
                 seminal vesicle  
  
 urinary system  
     bladder



metanephros  
 nephron  
 glomerulus  
 renal convoluted tubule  
 ureter  
 urethra

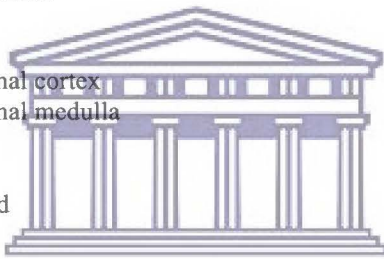
TS27

alimentary system  
 intestine  
 large intestine  
 anus  
 colorectal  
 cecum  
 colon  
 rectum  
 small intestine  
 duodenum  
 ileum  
 jejunum  
 liver and biliary system  
 bile duct  
 cystic duct  
 gall bladder  
 hepatic duct  
 liver  
 mesentery  
 oesophagus  
 omentum  
 greater omentum  
 lesser omentum  
 oral cavity  
 jaw  
 gum  
 mandible  
 maxilla  
 premaxilla  
 tooth  
 molar  
 salivary gland  
 parotid gland  
 sublingual gland  
 submandibular gland  
 tongue  
 pancreas  
 pharynx  
 hypopharynx  
 nasopharynx  
 oropharynx  
 stomach  
 anatomical site  
 anterior limb  
 head  
 posterior limb  
 tail  
 trunk  
 whole body  
 cardiovascular system



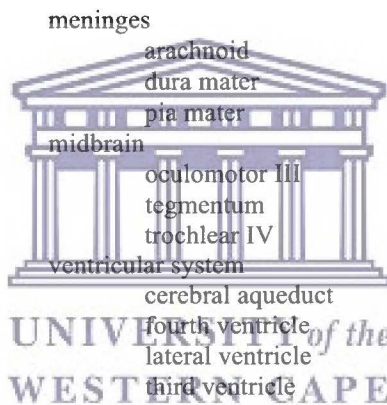


- artery
  - aorta
  - carotid artery
- capillary
- heart
  - atrium
  - cardiac valve
  - endocardium
  - myocardium
  - pericardium
  - ventricle
- vein
  - vena cava
    - inferior vena cava
    - superior vena cava
- dermal system
  - appendages
    - hair
    - hair follicle
    - sebaceous gland
    - sweat gland
    - vibrissa
  - skin
    - dermis
    - epidermis
- endocrine system
  - adrenal gland
    - adrenal cortex
    - adrenal medulla
  - parathyroid
  - pineal gland
  - pituitary gland
  - thymus
  - thyroid
- hematological system
  - blood
  - bone marrow
- lymphoreticular system
  - lymph node
  - spleen
  - tonsil
    - lingual tonsil
    - palatine tonsil
- musculoskeletal system
  - bone
  - cartilage
  - joint
    - ligament
    - synovium
  - muscle
    - skeletal muscle {striated muscle}
    - smooth muscle
  - tendon
- nervous system
  - central nervous system {CNS}
    - brain
      - forebrain
      - diencephalon



UNIVERSITY of the  
WESTERN CAPE

epithalamus  
 hypothalamus  
 thalamus  
 hippocampus  
 telencephalon  
     caudate nucleus  
     cerebral cortex  
         olfactory I  
     corpus striatum  
     lentiform nucleus  
     olfactory lobe  
     temporal lobe  
 hindbrain  
     medulla oblongata  
         hypoglossal XII  
         olivary nuclei  
         vagal X  
     metencephalon  
         cerebellum  
         pons  
             abducent VI  
             facial VII  
             trigeminal V  
             vestibulocochlear VIII



spinal cord  
 peripheral nervous system {PNS}  
     auditory apparatus {ear}  
         auditory ossicle  
         auditory tube  
         external ear  
             auricle  
             external acoustic meatus  
         internal ear  
             membranous labyrinth  
                 saccule  
                 utricle  
             osseous labyrinth  
                 cochlea  
                     spiral organ of Corti  
                 semicircular canal  
                 vestibule  
         middle ear  
         tympanum {tympanic membrane}  
 ganglion  
     spinal ganglion

sympathetic ganglion  
 olfactory apparatus  
 peripheral nerve  
 visual apparatus {eye}  
   choroid  
   ciliary body  
   conjunctiva  
   cornea  
   eyelid  
   iris  
   lacrimal gland  
   lens  
   optic chiasma  
   optic stalk  
   retina  
     fovea centralis  
     macula lutea  
   sclera  
   vitreous humor

respiratory system  
   bronchus  
   diaphragm  
   larynx  
   lung  
     alveolus

  nose  
   pleura {pleural cavity}  
   sinus {hindbrain, sinus}  
   trachea

unclassifiable

urogenital system

reproductive system

female reproductive system

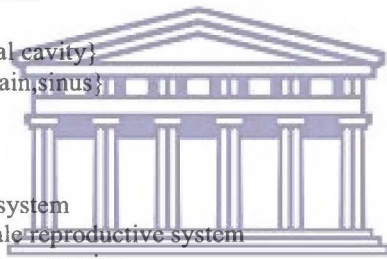
  amnion  
   breast  
   mammary gland  
   ovary  
   oviduct  
   placenta  
   uterus  
     cervix  
     endometrium  
     myometrium

  vagina  
   vulva

male reproductive system

  epididymis  
   penis  
     foreskin  
     glans  
   prostate  
   testis  
     seminiferous tubule  
   vas deferens  
     seminal vesicle

urinary system  
   bladder  
   kidney



UNIVERSITY of the  
 WESTERN CAPE

- nephron
  - renal corpuscle
    - glomerulus
  - renal tubule
    - loop of Henle
    - renal collecting duct
    - renal distal convoluted tubule
    - renal proximal convoluted tubule

- ureter
- urethra

**TS28**

alimentary system

intestine

large intestine

anus

colorectal

cecum

colon

rectum

small intestine

duodenum

ileum

jejunum

liver and biliary system

bile duct

cystic duct

gall bladder

hepatic duct

liver

mesentery

oesophagus

omentum

greater omentum

lesser omentum

oral cavity

jaw

gum

mandible

maxilla

premaxilla

tooth

molar

salivary gland

parotid gland

sublingual gland

submandibular gland

tongue

pancreas

pharynx

hypopharynx

nasopharynx

oropharynx

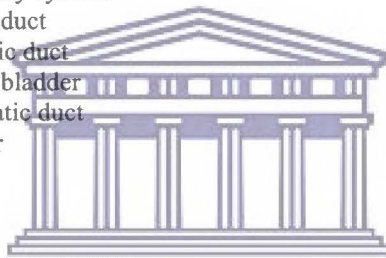
stomach

anatomical site

anterior limb

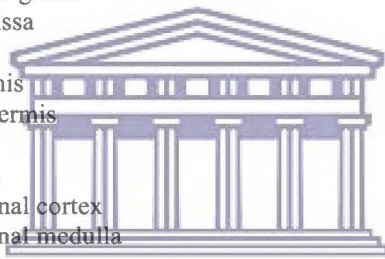
head

posterior limb



UNIVERSITY of the  
WESTERN CAPE

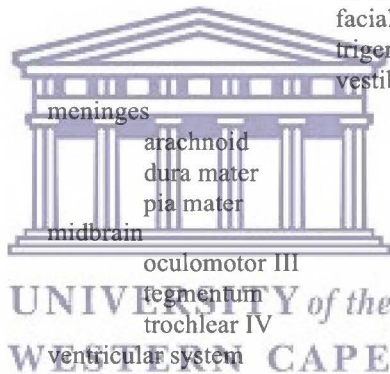
- tail
- trunk
- whole body
- cardiovascular system
  - artery
    - aorta
    - carotid artery
  - capillary
  - heart
    - atrium
    - cardiac valve
    - endocardium
    - myocardium
    - pericardium
    - ventricle
  - vein
    - vena cava
      - inferior vena cava
      - superior vena cava
- dermal system
  - appendages
    - hair
    - hair follicle
    - sebaceous gland
    - sweat gland
    - vibrissa
  - skin
    - dermis
    - epidermis
- endocrine system
  - adrenal gland
    - adrenal cortex
    - adrenal medulla
  - parathyroid
  - pineal gland
  - pituitary gland
  - thymus
  - thyroid
- hematological system
  - blood
  - bone marrow
- lymphoreticular system
  - lymph node
  - spleen
  - tonsil
    - lingual tonsil
    - palatine tonsil
- musculoskeletal system
  - bone
  - cartilage
  - joint
    - ligament
    - synovium
  - muscle
    - skeletal muscle {striated muscle}
    - smooth muscle
  - tendon
- nervous system



UNIVERSITY of the  
WESTERN CAPE



central nervous system {CNS}  
   brain  
     forebrain  
       diencephalon  
         epithalamus  
         hypothalamus  
         thalamus  
       hippocampus  
       telencephalon  
         caudate nucleus  
         cerebral cortex  
           olfactory I  
         corpus striatum  
         lentiform nucleus  
         olfactory lobe  
         temporal lobe  
     hindbrain  
       medulla oblongata  
         hypoglossal XII  
         olivary nuclei  
         vagal X  
       metencephalon  
         cerebellum  
         pons  
           abducent VI  
           facial VII  
           trigeminal V  
           vestibulocochlear VIII  
     meninges  
       arachnoid  
       dura mater  
       pia mater  
     midbrain  
       oculomotor III  
       trochlear IV  
       trigeminal V  
       trochlear IV  
       ventricular system  
       cerebral aqueduct  
       fourth ventricle  
       lateral ventricle  
       third ventricle  
   spinal cord  
 peripheral nervous system {PNS}  
   auditory apparatus {ear}  
     auditory ossicle  
     auditory tube  
     external ear  
       auricle  
       external acoustic meatus  
     internal ear  
       membranous labyrinth  
         sacculle  
         utricle  
       osseous labyrinth  
         cochlea  
           spiral organ of Corti  
         semicircular canal  
         vestibule



middle ear  
 tympanum {tympanic membrane}  
 ganglion  
     spinal ganglion  
     sympathetic ganglion  
 olfactory apparatus  
 peripheral nerve  
 visual apparatus {eye}  
     choroid  
     ciliary body  
     conjunctiva  
     cornea  
     eyelid  
     iris  
     lacrimal gland  
     lens  
     optic chiasma  
     optic stalk  
     retina  
         fovea centralis  
         macula lutea  
     sclera  
     vitreous humor  
 respiratory system  
     bronchus  
     diaphragm  
     larynx  
     lung  
         alveolus  
     nose  
     pleura {pleural cavity}  
     sinus  
     trachea  
 unclassifiable  
 urogenital system  
     reproductive system  
         female reproductive system  
             amnion  
             breast  
                 mammary gland  
             ovary  
             oviduct  
             placenta  
             uterus  
                 cervix  
                 endometrium  
                 myometrium  
             vagina  
             vulva  
         male reproductive system  
             epididymis  
             penis  
                 foreskin  
                 glans  
             prostate  
             testis  
                 semiferous tubule  
             vas deferens



UNIVERSITY of the  
 WESTERN CAPE

## Appendix VI The merged mouse developmental ontologies

### Mouse developmental ontology

4-8 cell stage

alimentary system

diverticulum

intestine

large intestine

anal pit

anal region

anus

colorectal

cecum

colon

rectum

small intestine

duodenum

ileum

jejunum

liver and biliary system

bile duct

common bile duct

cystic duct

gall bladder

gall bladder primordium

hepatic duct

liver

mesentery

dorsal meso-oesophagus

oesophagus

omentum

greater omentum

lesser omentum

oral cavity

jaw

UNIVERSITY of the  
WESTERN CAPE

gum  
mandible

maxilla

premaxilla

tooth

molar

mandibular process

mandible primordium

maxillary process

maxilla primordium

salivary gland

parotid gland

sublingual gland

sublingual gland primordium

submandibular gland

submandibular gland primordium

tongue

pancreas

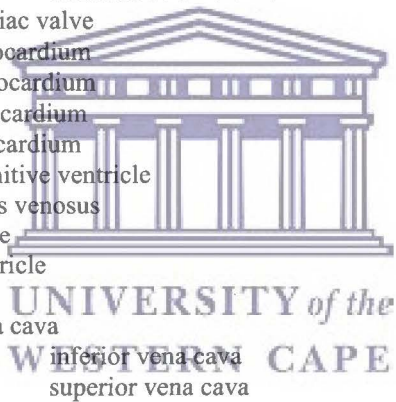
pancreas primordium

pharynx

hypopharynx

nasopharynx

- oropharynx
- stomach
- allantois
- anatomical site
  - anterior limb
  - anterior limb bud
  - head
  - posterior limb
  - posterior limb bud
  - posterior limb ridge
  - tail
  - tail bud
  - trunk
  - whole body
- blastocoelic cavity
- branchial arch
- cardiovascular system
  - artery
    - aorta
    - carotid artery
    - dorsal aorta
  - capillary
  - heart
    - atrium
      - common atrial chamber
      - cardiac valve
      - endocardium
      - mesocardium
      - myocardium
      - pericardium
      - primitive ventricle
      - sinus venosus
      - valve
      - ventricle
  - vein
    - vena cava
      - inferior vena cava
      - superior vena cava
- chorion
- dermal system
  - appendages
    - hair
    - hair follicle
    - sebaceous gland
    - sweat gland
    - vibrissa
  - skin
    - dermis
    - epidermis
- embryo
  - compacted morula
  - epiblast
  - inner cell mass
- endocrine system
  - adrenal gland
    - adrenal cortex
    - adrenal medulla
  - parathyroid

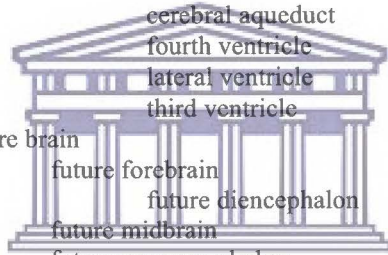


pineal gland  
 pineal primordium  
 pituitary gland  
 thymus  
 thymus primordium  
 thyroid  
 thyroid primordium  
 first polar body  
 germ layers  
   ectoderm  
   endoderm  
   mesenchyme  
   mesoderm  
   primitive endoderm  
   trophectoderm  
     mural trophoctoderm  
     polar trophoctoderm  
       ectoplacental cone  
  
 hematological system  
   blood  
   blood island  
   bone marrow  
 lymphoreticular system  
   lymph node  
   lymph sac  
   spleen  
   spleen primordium  
   tonsil  
     lingual tonsil  
     palatine tonsil  
 musculoskeletal system  
   bone  
   cartilage  
   cartilage condensation  
   joint  
     ligament  
     synovium  
   muscle  
     skeletal muscle  
     smooth muscle  
   pre-cartilage condensation  
   tendon  
 nervous system  
   central nervous system  
     brain  
       forebrain  
         diencephalon  
           epithalamus  
           hypothalamus  
           thalamus  
         hippocampus  
         telencephalon  
           caudate nucleus  
           cerebral cortex  
             olfactory I  
           corpus striatum  
           lentiform nucleus  
           olfactory lobe



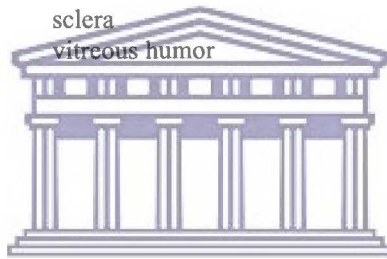


temporal lobe  
 hindbrain  
   medulla oblongata  
     floor plate  
     hypoglossal XII  
     olivary nuclei  
     vagal X  
   metencephalon  
     cerebellum  
     cerebellum primordium  
     pons  
       abducent VI  
       facial VII  
       trigeminal V  
       vestibulocochlear VIII  
   myelencephalon  
 meninges  
   arachnoid  
   dura mater  
   pia mater  
 midbrain  
   oculomotor III  
   tegmentum  
   trochlear IV  
 ventricular system  
   cerebral aqueduct  
   fourth ventricle  
   lateral ventricle  
   third ventricle  
 future brain  
   future forebrain  
   future diencephalon  
   future midbrain  
   future prosencephalon  
   future rhombencephalon  
   prosencephalon  
 future spinal cord  
   neural tube  
 neural crest  
 notochord  
 spinal cord  
 peripheral nervous system  
   auditory apparatus  
     auditory ossicle  
     auditory tube  
     external ear  
       auricle  
       external acoustic meatus  
     future tympanum  
     internal ear  
       membranous labyrinth  
         sacculle  
         utricle  
       osseous labyrinth  
         cochlea  
           spiral organ of Corti  
         semicircular canal  
         vestibule



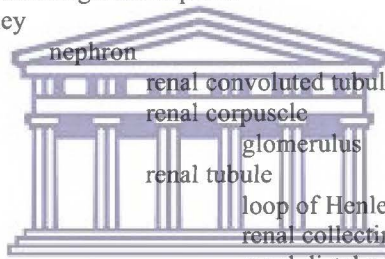
UNIVERSITY of the  
 WESTERN CAPE

otocyst  
 middle ear  
 tympanum  
 tympanum primordium  
 ganglion  
 spinal ganglion  
 sympathetic ganglion  
 olfactory apparatus  
 peripheral nerve  
 visual apparatus  
 choroid  
 ciliary body  
 conjunctiva  
 cornea  
 eyelid  
 intraretinal space  
 iris  
 lacrimal gland  
 lens  
 lens vesicle  
 optic chiasma  
 optic stalk  
 retina  
 fovea centralis  
 macula lutea  
 sclera  
 vitreous humor  
 notochordal plate  
 one-cell stage  
 primitive streak  
 proamniotic cavity  
 respiratory system  
 bronchus  
 diaphragm  
 larynx  
 lung  
 alveolus  
 nose  
 pleura  
 sinus  
 trachea  
 tracheal diverticulum  
 second polar body  
 two-cell stage  
 unclassifiable  
 urogenital system  
 presumptive nephric duct  
 pronephros  
 reproductive system  
 female reproductive system  
 amnion  
 breast  
 mammary gland  
 Mullerian tubercle  
 ovary  
 oviduct  
 paramesonephric duct  
 placenta



UNIVERSITY of the  
 WESTERN CAPE

uterus  
 cervix  
 endometrium  
 myometrium  
 vagina  
 vulva  
 genital tubercle  
 gonad  
 gonad primordium  
 gonadal component  
 male reproductive system  
 epididymis  
 mesonephric duct  
 penis  
 foreskin  
 glans  
 prostate  
 testis  
 primitive seminiferous tubule  
 seminiferous tubule  
 vas deferens  
 seminal vesicle  
 urinary system  
 bladder  
 degenerating mesonephros  
 kidney  
 nephron  
 renal convoluted tubule  
 renal corpuscle  
 glomerulus  
 renal tubule  
 loop of Henle  
 renal collecting duct  
 renal distal convoluted tubule  
 renal proximal convoluted tubule  
 mesonephros  
 metanephros  
 nephric cord  
 nephric duct  
 primitive ureter  
 ureter  
 ureteric bud  
 urethra  
 yolk sac  
 yolk sac cavity  
 zona pellucida



UNIVERSITY OF THE  
 WESTERN CAPE

**Theiler Stage**

adult

Theiler Stage 27 {TS 27; TS27}

Theiler Stage 28 {TS 28; TS28}

embryo

Theiler Stage 01 {TS 01; TS01}

Theiler Stage 02 {TS 02; TS02}

Theiler Stage 03 {TS 03; TS03}

Theiler Stage 04 {TS 04; TS04}

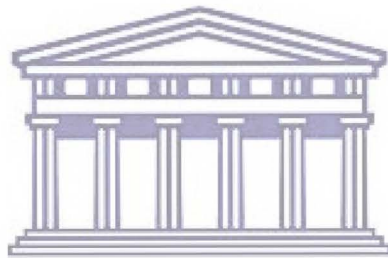
Theiler Stage 05 {TS 05; TS05}

Theiler Stage 06 {TS 06; TS06}

Theiler Stage 07 {TS 07; TS07}  
Theiler Stage 08 {TS 08; TS08}  
Theiler Stage 09 {TS 09; TS09}  
Theiler Stage 10 {TS 10; TS10}  
Theiler Stage 11 {TS 11; TS11}  
Theiler Stage 12 {TS 12; TS12}  
Theiler Stage 13 {TS 13; TS13}  
Theiler Stage 14 {TS 14; TS14}  
Theiler Stage 15 {TS 15; TS15}  
Theiler Stage 16 {TS 16; TS16}  
Theiler Stage 17 {TS 17; TS17}  
Theiler Stage 18 {TS 18; TS18}  
Theiler Stage 19 {TS 19; TS19}  
Theiler Stage 20 {TS 20; TS20}  
Theiler Stage 21 {TS 21; TS21}  
Theiler Stage 22 {TS 22; TS22}

fetus

Theiler Stage 23 {TS 23; TS23}  
Theiler Stage 24 {TS 24; TS24}  
Theiler Stage 25 {TS 25; TS25}  
Theiler Stage 26 {TS 26; TS26}  
Theiler Stage Unclassifiable {TS UN; TSUN}



UNIVERSITY *of the*  
WESTERN CAPE

# Appendix VIIa The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet.*

ARTICLES

nature  
genetics

## The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line

The FANTOM Consortium and the Riken Omics Science Center<sup>1</sup>

Using deep sequencing (deepCAGE), the FANTOM4 study measured the genome-wide dynamics of transcription-start-site usage in the human monocytic cell line THP-1 throughout a time course of growth arrest and differentiation. Modeling the expression dynamics in terms of predicted *cis*-regulatory sites, we identified the key transcription regulators, their time-dependent activities and target genes. Systematic siRNA knockdown of 52 transcription factors confirmed the roles of individual factors in the regulatory network. Our results indicate that cellular states are constrained by complex networks involving both positive and negative regulatory interactions among substantial numbers of transcription factors and that no single transcription factor is both necessary and sufficient to drive the differentiation process.

Development, organogenesis and homeostasis in multicellular systems involve the proliferation of precursor cells, followed by growth arrest and the acquisition of a differentiated cellular phenotype. Upon stimulation with phorbol myristate acetate (PMA), human THP-1 myelomonocytic leukemia cells cease proliferation, become adherent and differentiate into a mature monocyte- and macrophage-like phenotype<sup>1,2</sup>. This study aimed to understand the transcriptional network underlying growth arrest and differentiation in mammalian cells using THP-1 cells as a model system.

Most existing methods for regulatory network reconstruction collect genes into coexpressed clusters and associate these clusters with regulatory motifs or pathways (for example, see refs. 3–5). Alternatively, one can model the expression patterns of all genes explicitly in terms of predicted regulatory sites in promoters and the post-translational activities of their cognate transcription factors (TFs)<sup>6–8</sup>. Although this approach is challenging in complex eukaryotic genomes owing to large noncoding regions, ChIP-chip data<sup>9</sup> indicates that the highest density of regulatory sites is found near transcription start sites (TSSs) and regulatory regions originally thought to be distal may often be alternative promoters<sup>10,11</sup>. Precise identification of TSS locations is thus likely to be a crucial factor for accurate modeling of transcription regulatory dynamics in mammals.

In this study, we extend our previous observations of genome-wide TSS usage by Cap Analysis of Gene Expression (CAGE)<sup>12</sup> and using deep sequencing to identify promoters active during a time course of differentiation and quantify their expression dynamics. DeepCAGE data are used in combination with cDNA microarrays, other genome-scale approaches, novel computational methods and large-scale siRNA validation to provide a comprehensive analysis of growth arrest and differentiation in the THP-1 cell model.

### RESULTS

#### Outline of the analysis strategy

In most cell-line models, only a subset of cells undergoes growth arrest and differentiation. To maximize the sensitivity in this study, we identified a subclone of THP-1 cells in which the large majority of cells became adherent in response to PMA (Supplementary Fig. 1 online). Our strategy began with deepCAGE, which identified active TSSs at single-base-pair resolution, and simultaneously measured their time-dependent expression (using normalized tag frequency) as cells differentiated in response to PMA. The same RNA was subjected to cDNA microarray analysis on an Illumina platform. The differentiation of the cells was evident from the large increase in expression of macrophage-specific genes such as *CD14* and *CSF1R* detected by both deepCAGE and microarray in all replicates (Supplementary Fig. 2 online).

Figure 1 summarizes our Motif Activity Response Analysis (MARA) strategy. Promoters were defined as local clusters of coexpressed TSSs and promoter regions as their immediate flanking sequences (Fig. 1a,b). To reconstruct transcription regulatory dynamics we refined earlier computational methods<sup>6–8</sup> by incorporating comparative genomic information and each TF's positional preferences relative to the TSS in the prediction of regulatory sites. Binding sites for a comprehensive and unbiased collection of mammalian regulatory motifs were predicted in all proximal promoter regions (Fig. 1c) and the observed promoter expression profiles (Fig. 1d) were combined with the predicted site-counts (Fig. 1e) to infer time-dependent activity profiles of regulatory motifs (Fig. 1f). We inferred individual regulatory interactions (edges) between motifs and promoters by comparing the promoter expression and motif activity profiles (Fig. 1g). Rigorous Bayesian probabilistic methods were developed for all steps of the computational analysis. Finally, a core network was

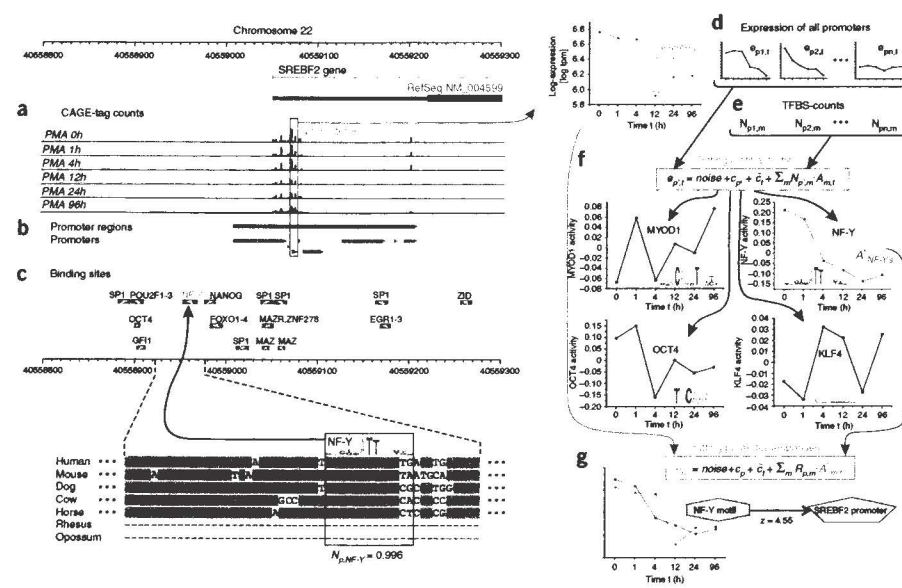
© 2009 Nature America, Inc. All rights reserved.



<sup>1</sup>A full list of authors and affiliations is provided at the end of this paper.

Received 16 July 2008; accepted 25 March 2009; published online 19 April 2009; doi:10.1038/ng.375





**Figure 1** Motif Activity Response Analysis (MARA). (a) CAGE tags are mapped to the human genome and their expression is normalized; vertical lines represent TSS positions, and their height is proportional to the normalized expression. (b) Mapped tags are clustered into promoters on the basis of their relative expression, and neighboring promoters are joined into promoter regions. (c) A window of -300 to +100 flanking each promoter region is extracted, multiply aligned and the MotEvo algorithm is used to predict binding sites for known motifs. (d-f) Observed expression of all promoters (d) and predicted site-counts (e) are used to infer motif activities (f). (g) The statistical significance of the regulatory edge from motif to promoter is calculated based on correlation of the promoter expression and motif activity profiles.

© 2009 Nature America, Inc. All rights reserved.

constructed by selecting the motifs that explained the greatest proportion of the expression variance, obtaining all predicted regulatory edges between TFs corresponding to these motifs, and selecting those regulatory edges that had independent experimental support. Using this approach, we reconstructed the transcriptional regulatory dynamics associated with cellular differentiation in human THP-1 cells, and validated a subset of predicted regulatory interactions.

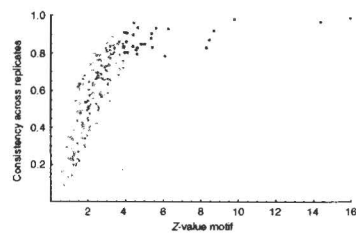
**DeepCAGE quantification of dynamic TSS usage**

CAGE tags generated from mRNA harvested at each time point were mapped to the human genome. Promoters were defined as clusters of nearby TSSs that showed identical expression profiles (within measurement noise) and were substantially expressed in at least one time point (Fig. 1a,b). Using these criteria we identified 29,857 promoters expressed in THP-1 cells containing 381,145 unique TSS positions (which is a subset of the nearly 2 million TSSs detected at least once in THP-1). These promoters were contained within 14,607 promoter regions (separated by at least 400 bp; Methods and Supplementary Fig. 3 online). The deepCAGE data was validated using genome tiling-array ChIP for markers of active transcription. Of the promoters identified, 79% and 78% were associated with H3K9Ac and RNA polymerase II, respectively (both markers of active transcription<sup>13,14</sup>), compared to 18% and 27% for inactive promoters (Supplementary Note online).

Among the identified promoters 84% (24,984) were within 1 kb of the starts of known transcripts and 81% (24,327) could be associated with 9,452 Entrez genes. Approximately half of the remaining promoters were more than 1 kb away from the loci of known genes (Supplementary Fig. 4 online). These newly identified promoters are conserved across mammals, suggesting that they are true transcription starts of currently unknown transcripts (Supplementary Fig. 5 online). The association of 24,327 promoters with 9,452 Entrez genes extends previous evidence of alternative promoter usage<sup>11</sup>—in this case even within a single cell type (Supplementary Table 1 online)—and demonstrates that promoter regions frequently contain multiple promoters with distinguishable expression profiles (Supplementary Table 2 online). In addition, for genes with known multiple promoters deepCAGE frequently identified only one promoter to be active in the THP-1 samples (Supplementary Fig. 6 online). Hence, deepCAGE samples a distinct aspect of transcriptional activity that can and does vary independently of mRNA abundances as measured by hybridization to representative microarray probes.

**Promoter expression**

Using the normalized tags per million (tpm) counts assigned to the promoters, we tested reproducibility among the three biological replicates and compared the outcome to the Illumina array from the same samples (Supplementary Fig. 7 online). DeepCAGE



**Figure 2** Statistical significance and consistency across replicates of the inferred motif activity profiles. Each dot corresponds to a motif. The significance of each motif in explaining the observed expression variation is quantified with the z value of its activity profile (horizontal axis, see Methods). The consistency of the inferred activity profile of each motif is quantified by the fraction of the variance (FOV) in the activity profile across all six replicates (three biological replicates for both CAGE and Illumina), which is reproduced in each replicate (vertical axis, see Methods).

expression measurements were comparatively noisy (Supplementary Fig. 7a). Nevertheless, the median Pearson correlation between the replicate-averaged expression profiles of CAGE and microarray was around 0.72 (Supplementary Fig. 7b), which is comparable to that observed with other deep transcriptome sequencing datasets<sup>15</sup>. As predicted, the correlation is lower for genes with multiple promoter regions (Supplementary Fig. 7b and discussed further in Supplementary Note).

#### Comprehensive regulatory site prediction

Known binding sites from the JASPAR and TRANSFAC databases<sup>16,17</sup> were used to construct a set of 201 regulatory motifs (position-specific weight matrices, WMs), which represent the DNA binding specificities of 342 human TFs. We predicted transcription factor binding sites (TFBSs) for all motifs within the proximal promoter regions (−300 to +100 bps) of all CAGE-defined promoters. Extending the proximal promoter regions beyond the −300 to +100 window decreased the quality of the fitted model described below (data not shown). In contrast to previous approaches that used simple WM scanning<sup>6</sup>, we incorporated information from orthologous sequences in six other mammals and used a Bayesian regulatory-site prediction algorithm that uses explicit models for the evolution of regulatory sites<sup>18,19</sup> (Fig. 1c and Methods). Notably, different motifs had distinct and highly specific positional preferences with respect to TSS (Supplementary Fig. 8 online), extending a previous genome-scale analysis<sup>20</sup>. Positional preferences were incorporated in the TFBS prediction by assigning each site a probability that it is under selection and correctly positioned. This analysis generated approximately 245,000 predicted TFBSs for the 201 motifs genome-wide. For each promoter–motif combination, the TFBS prediction was summarized by a count  $N_{pm}$ , which represents the estimated total number of functional TFBSs for motif  $m$  in promoter  $p$ . The TFBS predictions were compared with published high-throughput protein–DNA interaction datasets (ChIP-chip) and predicted target genes were significantly ( $P$  values ranged from 0.02 for *ETS1* to 6.60E−263 for *GABPA*) enriched among genes for which binding was observed (Supplementary Table 3 online).

#### Inferring key TFs and their time-dependent activities

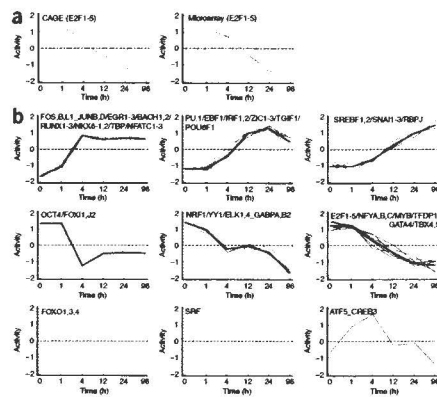
The details of our Motif Activity Response Analysis (MARA) are described in Methods. Briefly, for each motif  $m$  and each time point  $t$ ,

there is an (unknown) motif activity  $A_{mt}$ , which represents the time-dependent nuclear activity of positive and negative regulatory factors that bind to the sites of the motif (for example, the E2F activity will depend on nuclear E2F1-8, and DP1-2 levels, as well as RB1 phosphorylation status). As in previous work<sup>6–8,21</sup>, motif activities were inferred by assuming that the expression  $e_{pt}$  of promoter  $p$  at time  $t$  is a linear function of the activities  $A_{mt}$  of those motifs that have predicted sites in  $p$ . Additionally, the effect of motif  $m$  on the expression of promoter  $p$  is assumed to be proportional to the predicted number of functional sites  $N_{pm}$ . Assuming that the deviations of the predicted expression levels  $e_{pt}^{\text{obs}} = \text{constant} + \sum_m N_{pm} A_{mt}$  from the observed levels  $e_{pt}$  are Gaussian distributed, and using a Gaussian prior on the activities, we determine fitted activities  $A_{mt}^*$  that have maximal posterior probability (Methods).

The inferred motif activities were validated using a number of internal tests. First, our Bayesian procedure quantifies both the significance of each motif in explaining the observed expression variation as well as the reproducibility of its activity across replicates (Fig. 2 and Supplementary Table 4 online). The activity profiles of the top motifs are extremely reproducible across replicates and different measurement technologies (Figs. 2 and 3a and Supplementary Fig. 9 online). It should be stressed that, although motif activities are inferred by fitting the expression profiles of all promoters, the model cannot be expected to predict expression profiles of individual genes from the predicted TFBS in proximal promoters alone. The effects of chromatin structure, distal regulatory sites, nonlinear interactions between regulatory sites, and the contribution of the large numbers of human TFs for which no motif is known, are not considered. Furthermore, especially for genes that are dynamically regulated, mature mRNA abundance can be dynamically regulated independently of transcription initiation and promoter activity through selective mRNA elongation, processing and degradation. Our aim is not to predict expression profiles of individual genes but rather to predict the key regulators and their time-dependent activities, which can be inferred from integration of global expression information in a system undergoing dynamic change. We validated the significance of the inferred activity profiles by comparing the fraction of the 'expression signal' (expression variance minus replicate noise) that is explained by the model, compared to randomized versions, and under a tenfold cross-validation test (Supplementary Fig. 10 online). The explained expression signal is highly significant and this significance is maintained under tenfold cross-validation (Methods). In addition, the highly peaked positional profiles of TFBSs (Supplementary Fig. 8) suggest that knowing the exact TSS is important for accurate TFBS prediction. Indeed, the predicted TFBSs from CAGE promoters explain substantially more of the expression signal in microarrays than predicted TFBSs of the associated RefSeq promoters (Supplementary Fig. 10). We observe that the model better predicts the expression profiles of those promoters that are more strongly expressed, more reproducible across replicates, and have higher expression variance (Supplementary Fig. 11 online). Similarly, samples at the start and end of the differentiation time course are better predicted than those at intermediate time points (Supplementary Fig. 12 online), possibly because individual cells differentiate at different rates and leave the cell populations less homogeneous at intermediate time points.

Motif activities that were independently inferred from all 11,995 expressed microarray probes were combined with the inferred motif activities from all CAGE and microarray replicates into a final set of time-dependent motif activities (Methods). From these, we selected 30 'core' motifs that contribute most to explaining the expression





**Figure 3** Inferred time-dependent activities of the key regulatory motifs. (a) The time-dependent activity profile of the E2F1-5 regulatory motif as inferred from CAGE (left) and microarray (right) data. The three biological replicates are shown in red, blue and green. (b) The 30 most significant motifs with consistent activity profiles across all replicates (CAGE and microarray) were clustered into nine sets of motifs with similar dynamics. Each panel shows the activity of the members of the cluster (colored curves), the names of motifs contributing and the cluster average activity profile (black).

variation (red dots in Fig. 2) and segregated their activity profiles using a Bayesian procedure into nine clusters (Fig. 3b and Methods), including three clusters of upregulated motifs, three clusters of downregulated motifs and three clusters containing single motifs with profiles involving different transient dynamics. The genome-wide set of target promoters for each of the motifs was determined as described in Methods. The significance of each regulatory edge from a motif to a putative target promoter (containing a predicted TFBS) was quantified by the  $z$  value of the correlation between the motif's activity profile and the promoter's expression profile (Fig. 1e).

#### Core transcriptional regulatory network

The final aim in reconstructing transcriptional regulatory networks is to infer not only the key regulators and their target gene sets, but also the way in which the actions of these key regulators are coordinated. For this purpose, we collected all 199 predicted regulatory edges ( $z$  value  $\geq 1.5$ ) between the 30 core motifs. Recognizing that the prediction of individual regulatory edges is still prone to error, we constructed a core regulatory network (Fig. 4) of 55 highly-trusted edges by filtering the predicted edges according to experimental validation, either within our data or in existing literature (Supplementary Table 5 online). In addition, for each core motif we extracted the set of predicted target genes ( $z$  value  $\geq 1.5$ ) and checked for enrichment of gene ontology terms. A selection of significantly enriched terms is shown as oval nodes in Figure 4 (full set of GO enrichments are available as Supplementary Table 6 online).

Whereas our method infers the key regulators *ab initio*, the majority of factors within this core network are known to be important in the monocyte-macrophage lineage, thereby validating the method. In addition the predicted targets of these motifs

are enriched for biological processes known to be involved in differentiation of the monocytic lineage.

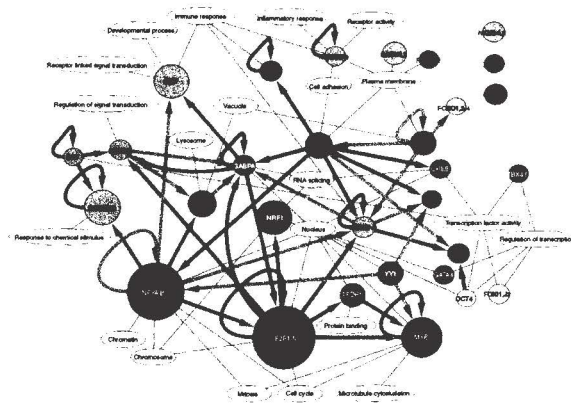
The gene ontology enrichments can broadly be divided into four groups. Downregulated motifs E2F1-5, NFYA,B,C and MYB are associated with cell cycle-related terms, consistent with the growth arrest observed during PMA-induced differentiation and the specific downregulation of numerous genes required for DNA synthesis and cell cycle progression within 24 h of PMA addition. Notably, MYB targets are also enriched specifically for microtubule-cytoskeleton-associated genes. Conversely, targets of upregulated motifs are associated with the terms immune response, cell adhesion, plasma membrane, vacuole and lysosome, all of which are consistent with differentiation into an adherent monocyte-like cell. The targeting of lysosomal genes by cholesterol-regulated SREBFs (sterol regulatory element-binding transcription factors) is of note, as lipid homeostasis is important in the macrophage in atherosclerosis and lysosomal storage diseases<sup>22</sup>. We also saw enrichment of signal transduction genes among targets of the early induced motifs EGR1-3 and TBP. Finally, there is a set of motifs whose targets are enriched in TFs. These motifs correspond to the transiently induced/repressed motifs, ATF5\_CREB3, FOXO1,3,4 and SRF, and the repressed pair of OCT4 and FOXH1,2 motifs.

#### Validation of edge predictions

THP-1 cells, even in an 'undifferentiated' state, are clearly a myeloid cell line. In seeking to validate the transcriptional network, we noted that there was a large set of TF genes expressed constitutively in the cells that were rapidly downregulated in response to PMA, of which *MYB* is an example, and another set that was only expressed later in the differentiation. To validate predicted edges empirically, we therefore chose to carry out siRNA knockdowns in undifferentiated THP-1 cells for genes encoding 28 TFs that are expressed in the undifferentiated state and for which we have associated motifs. To assess whether siRNA knockdown carried out in the undifferentiated state is appropriate to address factors that increase expression during the time course, we carried out the technically more difficult experiment of siRNA knockdown combined with PMA treatment for *SP11* (more commonly known in the literature as *PU.1*). All knockdowns were carried out in biological triplicate and qRT-PCR was used to confirm RNA-level knockdown, which in most cases was greater than 80% (Supplementary Table 7 online; in addition, protein-level knockdown was confirmed by protein blot for 14 siRNAs, see Supplementary Fig. 13 online). Changes in gene expression caused by TF knockdown were measured by Illumina microarrays. For each knocked-down TF gene, we obtained the list of predicted regulatory targets for the associated motif and divided the microarray probes into predicted targets and nontargets for a range of  $z$ -value thresholds. Higher-confidence targets in general show greater expression changes upon knockdown (Fig. 5a shows the example TF genes *MYB*, *SNAI3*, *EGR1* and *RUNX1*; additional examples are shown in Supplementary Fig. 14 online). For *SP11*, even in the absence of PMA treatment siRNA knockdown caused significant downregulation of predicted *SP11* targets, but the effects were much stronger when knockdown was combined with 1 h or 24 h of PMA treatment (Fig. 5b), confirming that PMA causes upregulation of *SP11* activity. A good correlation between target confidence ( $z$ -value cut-off) and average log expression ratio was observed for the large majority of experiments (Fig. 5c). For an intermediate cut-off of  $z = 1.5$  we quantified the difference in log expression ratio of predicted targets and nontargets (Fig. 5d) and







**Figure 4** Predicted core regulatory network of the 30 core motifs. An edge  $X \rightarrow Y$  is drawn whenever the promoter of at least one of the TFs associated with motif  $Y$  has a predicted regulatory edge for motif  $X$  ( $z$  value  $\geq 1.5$ ) and the edge has independent experimental support. The color of each node reflects its cluster membership and the size of the node reflects the significance of the motif. Edges confirmed in the literature, by ChIP or by siRNA are shown in red, blue and green, respectively. In cases where there are multiple lines of support only one evidence type is shown. **Supplementary Table 5** shows all predicted edges and their experimental support. GO terms significantly enriched among target genes are shown as white nodes with black edges. FOS/JUN (FOS,B,1,1\_JUNB,D), CREB (ATF5\_CREB3), GABPA (ELK1,4\_GABPA,B2).

found significant changes ( $z$ -value larger than 2) for 23 of 33 cases with *SP1* knockdown combined with 24 h of PMA treatment and *MYB* knockdown being the most significant (**Supplementary Fig. 15** online shows the entire distribution of log expression ratios of targets and nontargets for eight example TFs). Notably, for the TF genes *LMO2*, *MXI1* and *SP1*, the knockdown led to a significant upregulation of their targets, suggesting that the three encoded TFs act primarily as repressors in undifferentiated THP-1 cells (**Fig. 5d**, also see **Supplementary Fig. 14a**). Together these results provide compelling experimental validation of our predicted regulatory edges.

#### Single TF knockdowns affect multiple motif activities

Besides validating predicted targets, the siRNA knockdowns can also be used to assess the effects of the knockdown of one TF gene on the motif activities of other TFs. In addition to the 28 TFs perturbed above, we included a further 24 TFs that lacked motifs but were naturally repressed during PMA differentiation, or had been reported to have a role in myeloid differentiation or leukemia (**Supplementary Table 8** online).

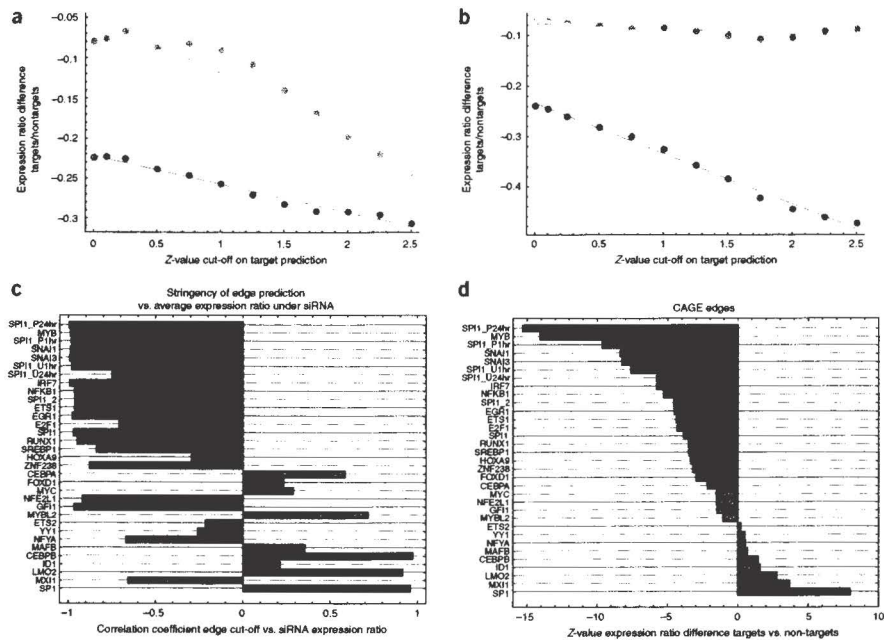
The motif activity inference method was used to determine the changes in activities of all motifs upon knockdown of each TF gene. To assess the role of each TF in differentiation, we defined the differentiative overlap between a TF gene knockdown and the PMA time course as the fraction of all motifs that significantly changed their activity in the same direction upon TF gene knockdown as in the PMA differentiation (Methods). By far the largest differentiative overlap (69%) was observed for the *MYB* knockdown, which not only affected *MYB* motif activity, but also the activity of most motifs in the core network, with the most significant activity changes all in the same direction as in the PMA time course (**Fig. 6e**). Knockdown of 13 other TF genes generated an overlap greater than the negative control (**Supplementary Table 9** online), and **Figure 6** shows three further examples (*E2F1*, *HOXA9* and *CEBPG*).

As for *MYB*, *E2F1* knockdown reproduced some of the downregulation of *MYB* and *E2F1* activity observed upon PMA stimulation, but it failed to reproduce the upregulation of *SREBF1,2*, *PU.1*, *NFATC1-3* and *FOS,B,1,1\_JUNB,D* activity (**Fig. 6b**). Similarly, the activity changes that *HOXA9* knockdown induced were mostly in the same direction as in the PMA differentiation; however, the *SNAI1-3*

and *IRF1,2* motif activities failed to be induced and the *GATA4* and *TBX4,5* motif activities failed to be downregulated (**Fig. 6c**).

Notably, knockdown of *CEBPG*, encoding one of the PMA-downregulated factors, for which we do not have a motif, also generated activity changes that significantly overlapped those observed in response to PMA (**Fig. 6d**). Finally, instead of comparing the motif activity changes that different knockdowns induced, we can also directly compare the expression changes of all genes with the expression changes observed in the PMA time course. We found that *MYB*, *HOXA9*, *CEBPG*, *GFI1*, *CEBPA*, *FLI1* and *MLL1,3* knockdowns all generated changes in gene expression that reiterated some of those observed with PMA treatment (**Supplementary Table 8**). *MYB* knockdown was exceptional, as it induced 35% (340/967) and repressed 19% (172/916) of the genes upregulated and downregulated with PMA, respectively. In addition the cells became adherent (**Supplementary Fig. 16** online) and began to express the monocytic markers *CD11B* (*ITGAM*), *CD54* (*ICAM1*), *CD14*, *APOE* and *CSF1R* (**Supplementary Fig. 2**), three of which we confirmed by flow cytometry (**Supplementary Table 10** online). This development of adherence could be linked to the GO enrichment for cytoskeleton-associated genes among *MYB* targets noted above. Given these observations one might wonder whether *MYB* is a master regulator of the differentiation process and whether stronger and longer knockdown would have reproduced the complete differentiation observed under PMA treatment. Several observations argue strongly against this. First, the gene sets perturbed by *MYB* and by the other pro-differentiative TFs overlap only partially (**Supplementary Table 11** online). Second, of the six other pro-differentiative TF genes only two (*CEBPG* and *GFI1*) are affected by *MYB* knockdown. Both these facts indicate that the other pro-differentiative TF genes are not simply downstream of *MYB*. Third, *MYB* downregulation does not occur until after the second hour of the PMA time course (**Fig. 3b**), which is at odds with the idea of *MYB* sitting at the top of the regulatory hierarchy. It is also worth noting that THP-1 cells harbor a leukemogenic fusion<sup>23</sup> between *MLL* (mixed-lineage leukemia) and *MLL1,3* (*MLL* translocation partner 3) and that the *MLL1,3* siRNA targets this leukemogenic fusion (note that full-length *MLL1,3* does not seem to be expressed in THP-1 as there is no CAGE 5' signal for this gene). Our data indicate that this fusion interferes with differentiation and that neither PMA treatment nor *MYB* knockdown affects *MLL-MLL1,3* levels, suggesting these stimuli can bypass the differentiative block. Conversely, *MLL1,3* knockdown had no effect on *MYB* levels. These





**Figure 5** Validation of predicted target promoter sets using siRNA knockdowns. (a) Difference in the average log expression ratio upon knockdown between predicted target promoters and predicted nontargets (vertical axis) as a function of the Z value cut-off on target prediction (horizontal axis), more stringent cut-offs are on the right) for knockdown of the TF genes *MYB* (red), *SNA13* (orange), *RUNX1* (green) and *EGR1* (light blue). (b) As in a but now for knockdown of *SPI1* followed by 1 h without treatment (light blue), 24 h without treatment (dark blue), 1 h of PMA treatment (orange) and 24 h of PMA treatment (red). All straight lines are linear regression fits. (c) Pearson correlation coefficients between the average log expression ratio difference of targets and nontargets and the cut-off on target predictions (horizontal axis). Red bars indicate correlation coefficients larger than 0.75 in absolute value; green bars, absolute values between 0.5 and 0.75; and blue bars, less than 0.5. (d) Significance (Z value) of the difference in log expression ratio between predicted targets and nontargets (cut-off  $z = 1.5$ ) for all 28 TFs associated with a motif, measured as a Z value (number of standard errors). Red bars correspond to significant changes, that is, greater than two standard errors; green bars, changes between 1 and 2 standard errors; and blue bars, changes less than 1 standard error. siRNA knockdowns were carried out in biological triplicate and knockdown was assessed by qRT-PCR (Supplementary Table 7).

results agree with previous RNAi studies that conclude that downregulation of *MLL* leukemogenic fusion proteins can promote growth arrest but is not required for terminal differentiation<sup>24,25</sup>. Thus, individual TF gene knockdowns affect the activities of multiple motifs and elicit different, but overlapping, subsets of the regulatory changes observed in the PMA time course. Taken together, the data indicate that the independent perturbation of expression of multiple TFs in response to PMA is both necessary and sufficient to initiate partial differentiation.

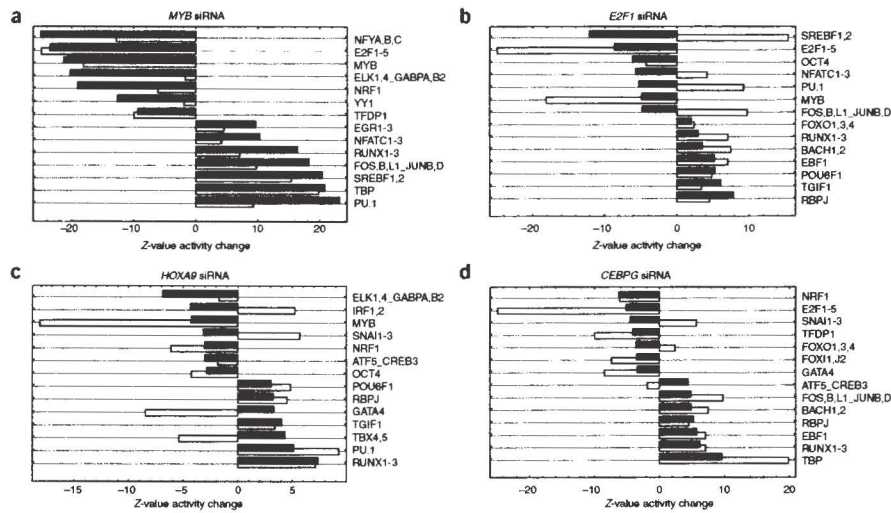
#### Many TFs are involved in the differentiation process

The network predictions and the siRNA results above suggest that upregulation and downregulation of the activities of multiple co-operating TFs is required for differentiation. Of a curated list<sup>26</sup> of 1,322 human TFs, 610 were detected by both CAGE and microarray in at least one time point (Supplementary Table 12 online); however, only 155 of these are covered by weight matrices, suggesting that other

factors may well be important in these cells. Of the 610 expressed TFs 64 were most highly expressed in the undifferentiated and 34 in the differentiated state. In addition, 101 TFs were transiently induced or repressed during differentiation. To elucidate the connection of these TFs to the inferred network, we compared the predicted regulatory inputs of co-regulated subsets of TFs with the predicted regulatory inputs of the set of all 610 expressed TFs.

Whereas no motifs are overrepresented among inputs of statically expressed TFs, inputs of dynamically expressed TFs showed enrichment for a subset of motifs. TFs downregulated from 0 to 96 h PMA were most enriched for three downregulated motifs of the core network: *OCT4* (3.4 $\times$ ), *GATA4* (3.3 $\times$ ) and *NFYA,B,C* (2.2 $\times$ ) (Supplementary Table 13a online). Similarly, TFs upregulated from 0 to 96 h were most enriched for core network motifs that increase activity during differentiation: *SNA1-3* (4.6 $\times$ ) and *TBP* (5.2 $\times$ ) (Supplementary Table 13b). Finally, transiently regulated TFs were enriched for the *SRF* (3.5 $\times$ ) and *NHLH1,2* (3 $\times$ ) motifs (Supplementary Table 13c).





**Figure 6** Most significant motif activity changes (as measured by z value, red bars) for four TF gene knockdowns that induce motif activity changes that have a differentiative overlap with the PMA time course of more than 50%. The corresponding motif activity changes observed in the PMA time course are shown as gray bars.

Notably, TFs that are predicted targets of SRF are mostly induced in the first hour of PMA-induced differentiation. During this first hour 55 of the 57 genes whose expression was perturbed are induced and 30% encode TFs (Supplementary Fig. 17a online). The regulatory inputs of these early-induced TFs are enriched for the motifs SRF, TBP and FOSL2 (Supplementary Table 13d), which all correspond to known PMA-responsive TFs<sup>27–30</sup>. Among the early-induced TFs, five correspond to upregulated core network motifs themselves (FOXB, EGR1-3 and SNAI1) and two (MAFB and EGR1) are known to induce pro-differentiative changes<sup>31,32</sup>. It is also worth noting that significant downregulation did not occur until the second hour, and this may require both early induction of transcriptional repressors and the RNA degradation proteins BTG2 and ZFP36 (tristetraprolin)<sup>33,34</sup> (Supplementary Fig. 17b). Together, these results suggest that induction of SRF target genes in the first hour is critical to establishing the differentiative program and is required before factors maintaining the undifferentiated state are downregulated (Supplementary Fig. 17b,c).

#### Web interface to data and analysis results

To facilitate the use of the data and analysis of results amassed here, we provide an online tool, EdgeExpressDB, as part of the PANTOM4 web resource, which allows users to explore our annotations of the structure, expression and regulation of promoters genome-wide. It also integrates published TF-promoter interactions, the siRNA perturbations and genome-wide chromatin immunoprecipitation experiments. Our complete set of regulatory-interaction predictions provides a large collection of hypotheses that can be targeted for validation, for example, through chromatin immunoprecipitation, gel shift assays or reporter assays. The value of this resource is illustrated

by detailed examination of individual loci. For example, the osteopontin gene (*SPPI*) is massively induced from 12 h of differentiation (Supplementary Note). Our predictions confirm RUNX and PU.1 as regulators and support a previous analysis in mouse implicating the TGF $\beta$  factor. In addition our analysis identifies NEAT, STAT, NKX6.2 and LIM domain and homeobox proteins as candidates for further testing.

Finally, our set of human promoters, TF motifs, genome-wide annotation of TF binding sites and their predicted effects on the expression of the target promoters are available through the SwissRegulon website. A web interface, allowing researchers to automatically perform Motif Activity Response Analysis (MARA) of their own expression data in terms of our genome-wide predictions of TFBSs, is also available at SwissRegulon.

#### DISCUSSION

We have devised a new integrated approach that combines genome-wide identification of TSSs and their time-dependent expression with computational modeling to reconstruct the transcriptional regulatory dynamics of a differentiating human cell line. The CAGE tag sequencing used here is tenfold deeper than in previous studies<sup>11</sup>, and this is the first study to our knowledge to quantitatively monitor dynamic expression changes of individual TSSs genome-wide. Using this data we developed a new computational method in which promoter expression profiles were modeled directly in terms of the TFBSs occurring in their proximal promoter regions. This method allowed us to infer which regulatory motifs are most predictive of expression changes and the time-dependent activities of the corresponding TFs *ab initio*. We identified more than two dozen different regulatory motifs that significantly change their activity during PMA-induced



- A.L., A.R.R.F., C.A.W., C. Kai, C. Kawazu, C.O., C.P., C. Simon, C.W., D.A.H., E.B., E.M.-S., F.B., G.S.L., H. Koga, H. Miura, H.N., H.O.-Y., H.S., H.Y., J.B., J.C., J.K., J.O., J.S.M., J.Y., K.F., K. Imamura, K.M., K.M.J., K.N., K. Schroder, K. Shirahige, L.W., M.A., M.C.K., M.F., M. Hashimoto, M. Hatakeyama, M.J.S., M.K.-K., M. Kojima, M. Murata, M.N., M.R., M. Suzuki, M.T., N.A.M., N.I., N.N., N.P., R.K., R.D.T., S.M.G., S.H., S.L., S. Miyamoto, S. Noma, S. Nygaard, S. Takeeda, T.A., T. Kawashima, T. Kojima, T. Sano, T. Suzuki, Y.O., Y.A., Y. Hasegawa, Y.L., Y. Kitazume, Y.N., Y.O., Y. Takahashi and Y. Tomaru were involved in biological aspects of the project. A.M.C., A.R.R.F., A.S., B.L., C.O.D., D.F., E.A., E.v.N., G.J.F., H.A., H.S., J.D., J.M., J.Q., J.S.M., K.W., M. Lindow, M.Z., N.C., N.M., O.H., P.J.B., P.C., R.J.T., R.S., S.M.G., S. Kondo, T.L., T.R. and V.O. were involved in the genome-wide and RNA analyses. E.v.N. and P.J.B. designed and carried out the motif activity response analysis. A.R.R.F., E.v.N., Y. Tomaru and M.K.-K. carried out the siRNA analysis. A.R.R.F., C.O.D., D.A.H., E.v.N., H.S., J.K., P.C. and Y. Hayashizaki oversaw the project. H.S., A.R.R.F., E.v.N. and D.A.H. wrote the manuscript with assistance from T.R., T.L., M.J.S., Y. Hasegawa, M.d.H., K.M.J., K.Schloder, P.J.C., P.J.B., E.A., N.P., M.R., S.M.G., C.A.W., J.Q., W.H., A. Kubosaki, Y. Tomaru, V.B.B., M. Suzuki and Y. Hayashizaki.
- Published online at <http://www.nature.com/naturegenetics/>  
Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>
1. Tsuchiya, S. *et al.* Induction of maturation in cultured human monocytic leukemia cells by a phorbol diester. *Cancer Res.* **42**, 1530–1536 (1982).
  2. Abirink, M., Gobi, A.E., Huang, R., Nilsson, K. & Hellman, L. Human cell lines U-937, THP-1 and Mono Mac 6 represent relatively immature cells of the monocyte-macrophage cell lineage. *Leukemia* **8**, 1579–1584 (1994).
  3. Beer, M.A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
  4. Ramsey, S.A. *et al.* Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. *PLoS Comput. Biol.* **4**, e1000021 (2008).
  5. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
  6. Das, D., Nahle, Z. & Zhang, M.Q. Adaptively inferring human transcriptional subnetworks. *Mol. Syst. Biol.* **2**, 2006.0029 (2006).
  7. Gao, F., Foat, B.C. & Bussemaker, H.J. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5**, 31 (2004).
  8. Nguyen, D.H. & D'Haeseleer, P. Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Syst. Biol.* **2**, 2006.0012 (2006).
  9. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
  10. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
  11. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
  12. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**, 15776–15781 (2003).
  13. Roh, T.Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552 (2005).
  14. Sandoval, J. *et al.* RNA Pol-CHIP: a novel application of chromatin immunoprecipitation to the analysis of real-time gene transcription. *Nucleic Acids Res.* **32**, e88 (2004).
  15. Clouan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
  16. Wiegand, D. *et al.* A new generation of JASPAR: the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34**, D95–D97 (2006).
  17. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24**, 238–241 (1996).
  18. Moses, A.M., Chiang, D.Y., Pellard, D.A., Iyer, V.N. & Eisen, M.B. MONKEY: identifying conserved transcription factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* **5**, R98 (2004).
  19. van Nimwegen, E. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics* **8** (Suppl. 6), S4 (2007).
  20. Frith, M.C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Res.* **18**, 1–12 (2008).
  21. Bussemaker, H.J., Foat, B.C. & Ward, L.D. Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 329–347 (2007).
  22. Schmitz, G. & Grandl, M. Lipid homeostasis in macrophages—implications for atherosclerosis. *Rev. Physiol. Biochem. Pharmacol.* **160**, 93–126 (2008).
  23. Odero, M.D., Zeleznik-Le, N.J., Chinwalla, V. & Rowley, J.D. Cytogenetic and molecular analysis of the acute monocytic leukemia cell line THP-1 with an MLL-AF9 translocation. *Genes Chromosomes Cancer* **29**, 333–338 (2000).
  24. Martino, V. *et al.* Down-regulation of MLL-AF9, MLL and MYC expression is not obligatory for monocyte-macrophage maturation in AML-M5 cell lines carrying t(9;11)(p22;q23). *Oncol. Rep.* **15**, 207–211 (2006).
  25. Pession, A. *et al.* MLL-AF9 oncogene expression affects cell growth but not terminal differentiation and is downregulated during monocyte-macrophage maturation in AML-M5 THP-1 cells. *Oncogene* **22**, 8671–8676 (2003).
  26. Roach, J.C. *et al.* Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. *Proc. Natl. Acad. Sci. USA* **104**, 16245–16250 (2007).
  27. Biggs, J.R., Ahn, N.G. & Kraft, A.S. Activation of the mitogen-activated protein kinase pathway in U937 leukemic cells induces phosphorylation of the amino terminus of the TATA-binding protein. *Cell Growth Differ.* **9**, 667–676 (1998).
  28. Iyer, D. *et al.* Serum response factor MADS box serine-162 phosphorylation switches proliferation and myogenic gene programs. *Proc. Natl. Acad. Sci. USA* **103**, 4516–4521 (2006).
  29. Morton, S., Davis, R.J. & Cohen, P. Signaling pathways involved in multisite phosphorylation of the transcription factor ATF-2. *FEBS Lett.* **572**, 177–183 (2004).
  30. Trejo, J. *et al.* A direct role for protein kinase C and the transcription factor Jun/AP-1 in the regulation of the Alzheimer's beta-amyloid precursor protein gene. *J. Biol. Chem.* **269**, 21682–21690 (1994).
  31. Kelly, L.M., Engstler, U., Lafon, I., Siewek, M.H. & Graf, T. MafB is an inducer of monocyte differentiation. *EMBO J.* **19**, 1987–1997 (2000).
  32. Krishnaraju, K., Hoffman, B. & Liebermann, D.A. The zinc finger transcription factor Egr-1 activates macrophage differentiation in M1 myeloblastic leukemia cells. *Blood* **92**, 1957–1966 (1998).
  33. Mauviel, F., Faux, C. & Seraphin, B. The BTG2 protein is a general activator of mRNA deadenylation. *EMBO J.* **27**, 1039–1048 (2008).
  34. Blackshear, P.J. Tristetraprolin and other COOH tandem zinc-finger proteins in the regulation of mRNA turnover. *Biochem. Soc. Trans.* **30**, 945–952 (2002).
  35. Carey, J.O., Posekany, K.J., deVente, J.E., Pettit, G.R. & Ways, D.K. Phorbol-ester-stimulated phosphorylation of PU.1: association with leukemic cell growth inhibition. *Blood* **87**, 4316–4324 (1996).
  36. Foster, N., Lea, S.R., Preshaw, P.M. & Taylor, J.J. Pivotal advance: vasoactive intestinal peptide inhibits up-regulation of human monocyte TLR2 and TLR4 by LPS and differentiation of monocytes to macrophages. *J. Leukoc. Biol.* **81**, 893–903 (2007).
  37. Xu, X. *et al.* A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res.* **17**, 1550–1561 (2007).
  38. Anfosso, G., Gewirtz, A.M. & Calabretta, B. An oligomer complementary to c-myc-encoded mRNA inhibits proliferation of human myeloid leukemia cell lines. *Proc. Natl. Acad. Sci. USA* **86**, 3379–3383 (1989).
  39. Reddy, M.A. *et al.* Opposing actions of c-ets/PU.1 and c-myc protooncogene products in regulating the macrophage-specific promoters of the human and mouse colony-stimulating factor-1 receptor (c-fms) genes. *J. Exp. Med.* **180**, 2309–2319 (1994).
  40. Feng, R. *et al.* PU.1 and DEB/alpha/beta convert fibroblasts into macrophage-like cells. *Proc. Natl. Acad. Sci. USA* **105**, 6057–6062 (2008).
  41. Carter, J.R. & Tournebise, W.G. Early growth response transcriptional regulators are dispensable for macrophage differentiation. *J. Immunol.* **178**, 3038–3047 (2007).
  42. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
  43. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
  44. Altschuler, S.J., Weinhold, B., Oelgeschlaeger, M., Ruther, U. & Nordheim, A. Serum response factor is essential for mesoderm formation during mouse embryogenesis. *EMBO J.* **17**, 6289–6299 (1998).
  45. Cooper, S.J., Trinkl, N.D., Nguyen, L. & Myers, R.M. Serum response factor binding sites differ in three human cell types. *Genome Res.* **17**, 136–144 (2007).
  46. Flieger, A. *et al.* Serum response factor contributes selectively to lymphocyte development. *J. Biol. Chem.* **282**, 24320–24328 (2007).
  47. Poser, S., Impey, S.J., Trifiro, K., Xia, Z. & Storm, D.R. SRF-dependent gene expression is required for P13-kinase-regulated cell proliferation. *EMBO J.* **19**, 4955–4966 (2000).
  48. Huang, S. & Ingber, D.E. Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell Res.* **261**, 91–103 (2000).
  49. Kauffman, S. *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, New York, 1993).

The full list of authors and affiliations is as follows:  
The FANTOM Consortium:

Harukazu Suzuki<sup>2,50–52</sup>, Alistair R R Forrest<sup>2,3,50–52</sup>, Erik van Nimwegen<sup>4,50–52</sup>, Carsten O Daub<sup>2,51,52</sup>, Piotr J Balwiercz<sup>4,51</sup>, Katharine M Irvine<sup>5,51,52</sup>, Timo Lassmann<sup>2,51,52</sup>, Timothy Ravas<sup>5,51,52</sup>, Yuki Hasegawa<sup>2,51</sup>, Michiel J L de Hoon<sup>2,51</sup>, Shintaro Katayama<sup>2,51</sup>, Kate Schroder<sup>2,51,52</sup>, Piero Carninci<sup>2,51</sup>, Yasuhiro Tomaru<sup>2,51</sup>, Mutsumi Kanamori-Katayama<sup>2,51</sup>, Atsutaka Kubosaki<sup>2,51</sup>, Altuna Akalin<sup>7</sup>, Yoshinari Ando<sup>7</sup>, Erik Amer<sup>7</sup>, Maki Asada<sup>7</sup>, Hiroshi Asahara<sup>7</sup>, Timothy Bailey<sup>7</sup>,

ARTICLES

Vladimir B Bajic<sup>2,51</sup>, Denis Bauer<sup>2</sup>, Anthony G Beckhouse<sup>1</sup>, Nicolas Bertin<sup>2</sup>, Johan Björkregren<sup>10</sup>, Frank Brombacher<sup>11</sup>, Erika Bulger<sup>2</sup>, Alistair M Chalk<sup>3</sup>, Joe Chiba<sup>12</sup>, Nicole Cloonan<sup>13</sup>, Adam Dawe<sup>14</sup>, Josec Dostie<sup>14</sup>, Pär G Engström<sup>15</sup>, Maghubah Fassack<sup>2</sup>, Geoffrey J Faulkner<sup>14</sup>, J Lynn Fink<sup>14</sup>, David Fredman<sup>16</sup>, Ko Fujimori<sup>16</sup>, Masaaki Furuno<sup>17</sup>, Takashi Gōjōbōri<sup>17,51</sup>, Julian Gough<sup>18</sup>, Sean M Grimmond<sup>17,51</sup>, Mika Gustafsson<sup>19</sup>, Megumi Hashimoto<sup>2</sup>, Takehiro Hashimoto<sup>2</sup>, Mariko Hatakeyama<sup>20</sup>, Susanne Heinzel<sup>21</sup>, Winston Hide<sup>22,23</sup>, Oliver Hofmann<sup>22</sup>, Michael Hörnquist<sup>24</sup>, Lukasz Hummnicki<sup>25</sup>, Kazuhiko Ikey<sup>17</sup>, Naoko Inamoto<sup>24</sup>, Satoshi Inoue<sup>25</sup>, Yusuke Inoue<sup>26</sup>, Ryoko Ishihara<sup>27</sup>, Takao Iwayanagi<sup>27</sup>, Anders Jacobsen<sup>28</sup>, Mandeep Kaur<sup>19</sup>, Hideya Kawai<sup>2</sup>, Markus C Kerr<sup>15</sup>, Ryoichiro Kimura<sup>12</sup>, Syuhei Kimura<sup>29</sup>, Yasumasa Kimura<sup>30</sup>, Hiroaki Kirano<sup>30</sup>, Hisashi Koga<sup>31</sup>, Toshio Kojima<sup>30</sup>, Shinji Kondo<sup>2</sup>, Takeshi Konno<sup>17</sup>, Anders Krogh<sup>28</sup>, Adele Kruger<sup>2</sup>, Aiji Kumari<sup>12</sup>, Boris Lenhard<sup>25,1</sup>, Andreas Lennartsson<sup>2</sup>, Morten Lindow<sup>32</sup>, Marina Lizio<sup>2</sup>, Cameron MacPherson<sup>2</sup>, Norihiro Maeda<sup>2</sup>, Christopher A Maher<sup>2</sup>, Monique Maquungu<sup>2</sup>, Jessica Mar<sup>33</sup>, Nicholas A Matigian<sup>34</sup>, Hideo Matsuda<sup>35</sup>, John S Mattick<sup>3</sup>, Stuart Meier<sup>3</sup>, Sei Miyamoto<sup>17</sup>, Etsuko Miyamoto-Sato<sup>35</sup>, Kazuhiko Nakabayashi<sup>17</sup>, Yutaka Nakachi<sup>36</sup>, Mika Nakano<sup>37</sup>, Sanne Nygaard<sup>38</sup>, Toshitsugu Okayama<sup>17</sup>, Yasushi Okazaki<sup>39</sup>, Haruka Okuda-Yabukami<sup>2</sup>, Valerio Orlando<sup>37</sup>, Jun Otomo<sup>38</sup>, Mikhail Pachkov<sup>4</sup>, Nikolai Petrovsky<sup>21</sup>, Charles Plessy<sup>40</sup>, John Quackenbush<sup>43,51</sup>, Aleksandar Radovanovic<sup>2</sup>, Michael Rehl<sup>39</sup>, Rintaro Saito<sup>40</sup>, Albin Sandelin<sup>28</sup>, Sebastian Schmeier<sup>2</sup>, Christian Schönbach<sup>41</sup>, Ariel S Schwartz<sup>42</sup>, Colin A Semple<sup>42</sup>, Miho Sera<sup>17</sup>, Jessica Severin<sup>2</sup>, Katsuhiko Shirahige<sup>43</sup>, Cas Simons<sup>13</sup>, George St Laurent<sup>32</sup>, Masanori Suzuki<sup>44</sup>, Takahiro Suzuki<sup>45</sup>, Matthew J Sweet<sup>46</sup>, Ryan J Taft<sup>13</sup>, Shizu Takeda<sup>28</sup>, Yoichi Takenaka<sup>34</sup>, Kai Tan<sup>6</sup>, Martin S Taylor<sup>45</sup>, Rohan D Teasdale<sup>4</sup>, Jesper Tegner<sup>10,46,51</sup>, Sarah Teichmann<sup>47</sup>, Eivind Valen<sup>48</sup>, Claes Wahlestedt<sup>49</sup>, Kazunori Waki<sup>2</sup>, Andrew Waterhouse<sup>2</sup>, Christine A Wells<sup>51</sup>, Ole Winther<sup>28</sup>, Linda Wu<sup>21</sup>, Kazumi Yamaguchi<sup>2</sup>, Hiroshi Yanagawa<sup>35</sup>, Jun Yasuda<sup>2</sup>, Mihaela Zavolan<sup>4</sup> & David A Hume<sup>40,51,52</sup>

Riken Omics Science Center.  
 Takahiro Arakawa<sup>2</sup>, Shiro Fukuda<sup>2</sup>, Kengo Imamura<sup>2</sup>, Chikatoshi Kai<sup>2</sup>, Ai Kaiho<sup>2</sup>, Tsugumi Kawashima<sup>2</sup>, Chika Kawazu<sup>2</sup>, Yayoi Kitazume<sup>2</sup>, Miki Kojima<sup>2</sup>, Hisashi Mura<sup>2</sup>, Kayoko Murakami<sup>2</sup>, Mitsuyoshi Murata<sup>2</sup>, Noriko Ninomiya<sup>2</sup>, Hiromi Nishiyori<sup>2</sup>, Shohei Noma<sup>2</sup>, Chihiro Ogawa<sup>2</sup>, Takuma Sano<sup>2</sup>, Christophe Simon<sup>2</sup>, Michihira Tagami<sup>2</sup>, Yukari Takahashi<sup>2</sup> & Jun Kawai<sup>2,51</sup>

General Organizer:  
 Yoshihide Hayashizaki<sup>2,51,52</sup>

<sup>2</sup>RIKEN Omics Science Center, RIKEN Yokohama Institute, Kanagawa, Japan. <sup>3</sup>The Ekitis Institute for Cell and Molecular Therapies, Griffith University, Australia. <sup>4</sup>Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland. <sup>5</sup>Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia. <sup>6</sup>Department of Bioengineering, Jacobs School of Engineering, University of California, San Diego, La Jolla, California, USA. <sup>7</sup>Bergen Center for Computational Science, Bergen, Norway. <sup>8</sup>National Research Institute for Child Health and Development, Tokyo, Japan. <sup>9</sup>South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa. <sup>10</sup>Computational Medicine Group, Atherosclerosis Research Unit, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital Solna, Stockholm, Sweden. <sup>11</sup>Institute of Infectious Disease and Molecular Medicine (IDMM), Wolfson Pavilion Level 2, Faculty of Health Sciences, University of Cape Town, Observatory, South Africa. <sup>12</sup>Department of Biological Science and Technology, Tokyo University of Science, Japan. <sup>13</sup>Australian Research Council (ARC) Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Australia. <sup>14</sup>Department of Biochemistry, McGill University, Montreal, Quebec, Canada. <sup>15</sup>Australian Research Council (ARC) Centre of Excellence in Bioinformatics, Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Australia. <sup>16</sup>Laboratory of Biodefense and Regulation, Osaka University of Pharmaceutical Sciences, Osaka, Japan. <sup>17</sup>Research Organization of Information and Systems, Center for Information Biology and DNA Data Bank of Japan (DDBJ), National Institute of Genetics, Shizuoka, Japan. <sup>18</sup>Department of Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol, UK. <sup>19</sup>Department of Science and Technology, Linköping University, Norrköping, Sweden. <sup>20</sup>Computational and Experimental Systems Biology Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, Japan. <sup>21</sup>Department of Diabetes and Endocrinology, Flinders University and Medical Centre, Bedford Park, Adelaide, Australia. <sup>22</sup>Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. <sup>23</sup>Department of Cell and Molecular Biology (CMB), Karolinska Institutet, Stockholm, Sweden. <sup>24</sup>Cellular Dynamics Laboratory, Discovery and Research Institute, RIKEN Wako Institute, Saitama, Japan. <sup>25</sup>Graduate School of Medicine and Faculty of Medicine, the University of Tokyo, Tokyo, Japan. <sup>26</sup>Department of Biological and Chemical Engineering, Gunma University Faculty of Engineering, Gunma, Japan. <sup>27</sup>R&D Solution Center, Research & Development Group, Hitachi Ltd., Tokyo, Japan. <sup>28</sup>The Bioinformatics Centre, Department of Biology and Biotech Research & Innovation Centre, University of Copenhagen, Copenhagen, Denmark. <sup>29</sup>Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University, Tottori, Japan. <sup>30</sup>The Systems Biology Institute, Shibuya, Tokyo, Japan. <sup>31</sup>Department of Human Gene Research, Kazusa DNA Research Institute, Chiba, Japan. <sup>32</sup>Department of Biochemistry and Molecular Biology, the George Washington University Medical Center, Washington, D.C., USA. <sup>33</sup>Department of Biostatistics, Harvard School of Public Health, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. <sup>34</sup>Department of Bioinformatics Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan. <sup>35</sup>Department of Biosciences and Informatics, Faculty of Science and Technology, Keio University, Yokohama, Japan. <sup>36</sup>Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Saitama Medical School, Saitama, Japan. <sup>37</sup>Dulbecco Telethon Institute, IRCCS Fondazione Santa Lucia at EBRI, Rome and IGB CNR, Naples, Italy. <sup>38</sup>Central Research Laboratory, Hitachi Ltd., Tokyo, Japan. <sup>39</sup>Department of Hematology and Oncology, University of Regensburg, Hospital, Regensburg, Germany. <sup>40</sup>Faculty of Environment and Information Studies, Keio University, Fujisawa, Kanagawa, Japan. <sup>41</sup>School of Biological Science, University of Genomics and Genetics, Nanyang Technological University, Singapore. <sup>42</sup>MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, UK. <sup>43</sup>Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, Japan. <sup>44</sup>Institute for Molecular Bioscience, School of Molecular and Microbial Sciences, CRC for Chronic Inflammatory Diseases, The University of Queensland, St. Lucia, Australia. <sup>45</sup>EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. <sup>46</sup>Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden. <sup>47</sup>Structural Studies Division MRC Laboratory of Molecular Biology, Hills Rd., Cambridge, UK. <sup>48</sup>Biochemistry/Neuroscience, the Scripps Research Institute, Jupiter, Florida, USA. <sup>49</sup>The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin, UK. <sup>50</sup>These authors contributed equally to this work. <sup>51</sup>These authors are the core writing group. <sup>52</sup>These authors are affiliated with the FANTOM 4 headquarters. Correspondence should be addressed to D.H. (david.hume@roslin.ed.ac.uk) or Y.H. (yoshihide@gsr.riken.jp).

© 2009 Nature America, Inc. All rights reserved.

UNIVERSITY of the  
 WESTERN CAPE



## **Appendix VIIb: Clusters of genes from Illumina microarray expression experiment with early, mid and late response characteristics**

### **Data selection**

For each time-point, the Rank Invariant normalization values, as well as the Flag Detection scores for each probe, were extracted from the files supplied by the Consortium. The Flag Detection scores are determined as follows:

- for each probe, the bead standard deviation (defined as the ‘average standard deviation associated with bead-to-bead variability for the sample in the group’ – Illumina BeadStudio User Guide) was divided by the intensity value to determine the variance of the measurements, yielding the flag detection score
- for flag detection scores equal to 1, the probe is flagged as ‘present’ (P)
- for flag detection scores between 0.99 and 1.00, the probe is flagged as ‘marginal’ (M)
- for flag detection scores less than 0.99, the probe is flagged as ‘absent’ (A)

We excluded from consideration all probes that were flagged as ‘absent’ at any time-point. This resulted in a total of 9 187 probes. The probe identifiers were converted to EntrezGene identifiers. Many of the probe identifiers did not have a corresponding gene identifier and were excluded from further analysis. This filtering step finally yielded 7 932 genes associated with the probes.

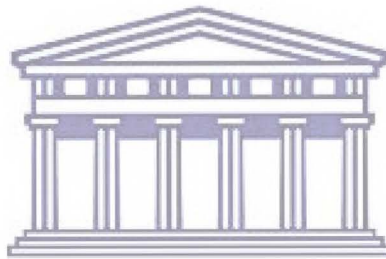
### **Data transformation**

The 7 932 genes selected were subjected to the following transformation steps:

- add a value of 50 to all data-points to eliminate negative values
- perform a log<sub>2</sub> transformation on the dataset
- normalize the data of the 0hr by making zero mean and standard deviation of 1

- transform all other time point values using the mean and standard deviation determined for 0 hr.
- to determine the change  $x$  in the expression over time for each probe relative to the expression level at point 0 hr, subtract the 0 hr value from all the other time-point values for each probe
- to calculate the fold-change in expression for each time-point relative to 0 hr, calculate  $2^x$  for each time-point value  $x$ .

The result of the data transformation is a fold-change value varying from 0 to infinity. A fold-change value between 0 and 0.5 indicates that the expression of the probe is half or less of what it was originally (at 0 hr), and therefore the respective gene is considered significantly down-regulated. A fold-change value of 2 or more indicates that the expression of the probe is 2 or more fold greater than it was originally (at 0 hr) and we considered it to represent a significant up-regulation of the gene.



## Clustering

The transformed data was binned into the following categories for clustering:

- Down-regulated: all values in the range  $0 \leq X \leq 0.5$ 
  - clustering value = -1
- No regulation: all values in the range  $0.6 < X < 2$ 
  - clustering value = 0
- Up-regulated: all values  $\geq 2$ 
  - clustering value = +1

The tool used to perform clustering was TIGR MultiExperiment Viewer (version 3.1), which is freely available from <http://www.tm4.org>. For clustering we applied a Hierarchical Clustering algorithm using the Euclidean distance metric and average linkage clustering.



## Selection of clusters


Of the transformed 7 932 genes, 1 807 genes were not regulated throughout the time-points, 710 genes were down-regulated at the 24h time-point only, and 5 220 genes were up-regulated at the 24h time-point only. These three clusters of genes were not selected.

The remaining clusters were visually inspected and divided into 10 categories based on their regulation over time as presented in Table 1 (see Figure 2 for graphical representation). In Table 1 we used the following classification of the time intervals in the gene response:

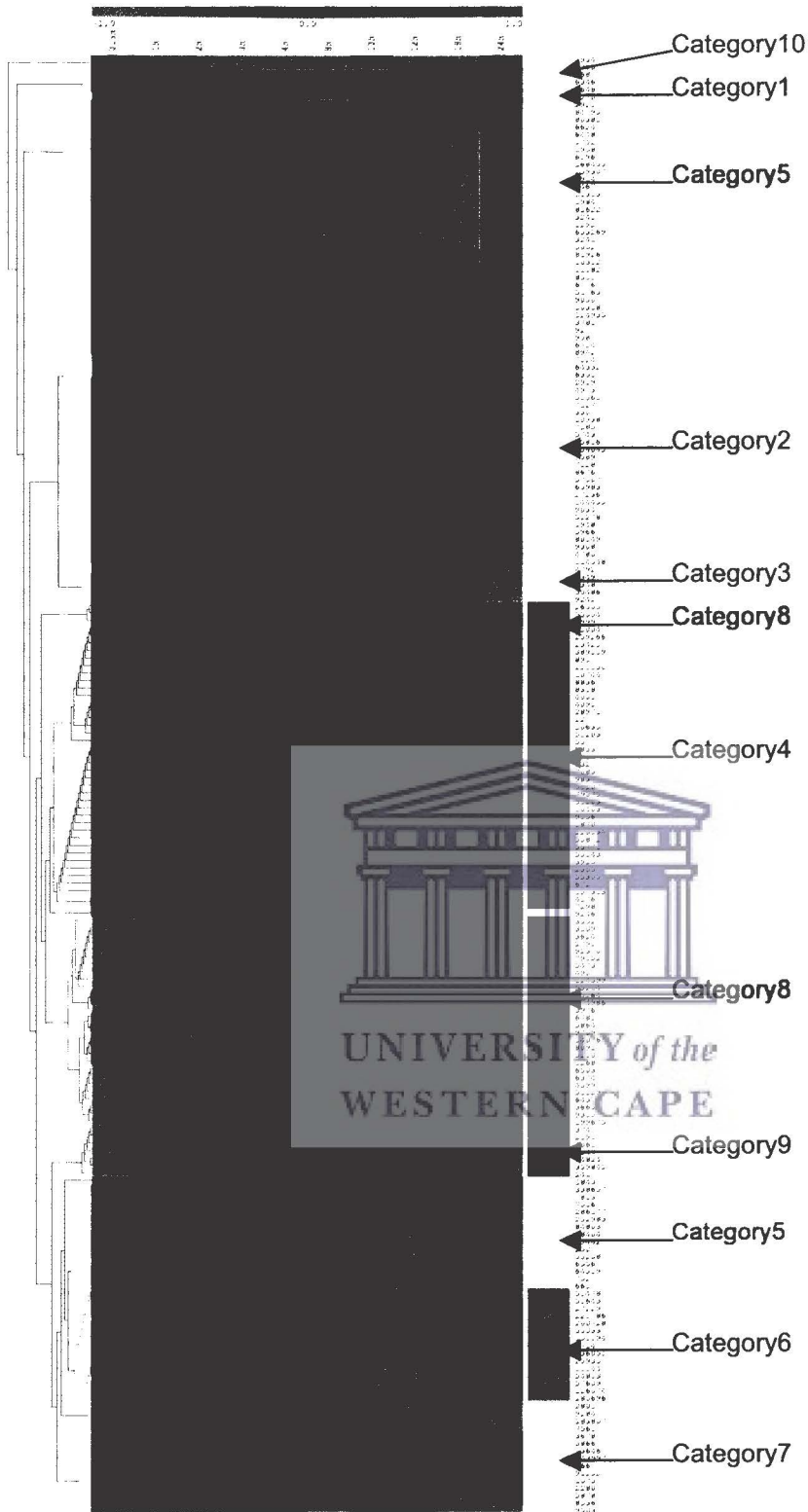
- early regulation refers to the first four time-points (0.5h, 1h, 2h, 3h)
- middle regulation refers to the next three time-points (4h, 8h, 10h)
- late regulation refers to the last three time-points (12h, 18h, 24h)

The heat-map of the selected clusters is depicted in Figure 1.

**Table 1:** Clustering categories for Illumina data based on the time of the response of genes to LPS stimulation.

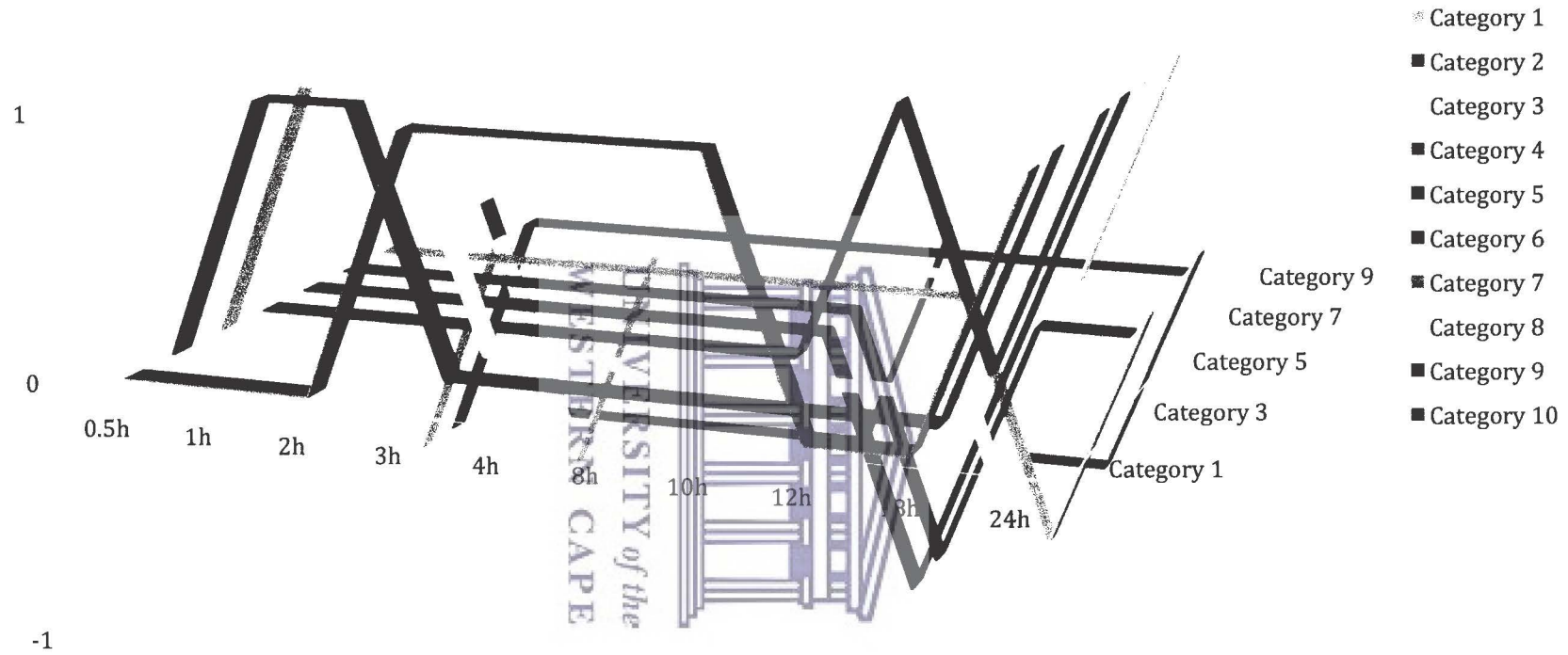


	Early Regulation	Middle regulation	Late regulation	GeneCount
Category 1	Up	Up	Up	4
Category 2	Up	None	Up	40
Category 3	Up	None	Down	5
Category 4	None	None	Up	38
Category 5	None	Down	Up	36
Category 6	None	Down	None	15
Category 7	None	Down	Down	15
Category 8	Down	None	Up	31
Category 9	Down	None	Down	7
Category 10	Down	Down	Down	2



**Figure 1**

**Clustering image from TMeV. Clusters were selected based on the visual inspection of expression profiles. Each cluster was classified into an expression category based on their expression over time.**



**Figure 2**

**Average expression profiles for the expression categories. The average expression profile for each category was plotted along the time-points. Values in the graph range from -1 (down-regulated) through 0 (no regulation) to 1 (up-regulation).**

### Appendix VIII Expression profile of transcription factors showing tissue restriction

The expression profile of the 145 transcription factors expressed in 25% of tissues. (FRS – female reproductive system; MRS – male reproductive system).

GeneID	Gene Symbol	adrenal gland	blood	blood vessel	bone	bone marrow	brain	heart	kidney	liver	lung	lymph	MRS	mucosa	pancreas	pineal gland	pituitary gland	spleen	stem cell	thymus	thyroid	other	
326	AIRE	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
430	ASCL2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
579	BAPX1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
668	FOXL2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1032	CDKN2D	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1
1053	CEBPE	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1745	DLX1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
1746	DLX2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1748	DLX4	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
1761	DMRT1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
1961	EGR4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
1993	ELAVL2	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1
2016	EMX1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2020	EN2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2103	ESRRB	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
2118	ETV4	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1



GeneID	Gene Symbol	adrenal gland	blood	blood vessel	bone	bone marrow	FRS	heart	kidney	liver	lung	lymph	MRS	mucosa	pancreas	pineal gland	pituitary gland	spleen	stem cell	thymus	thyroid	other
2294	FOXF1	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
2295	FOXF2	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
2297	FOXD1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1
2302	FOXJ1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
2304	FOXE1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
2306	FOXD2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2623	GATA1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
2672	GFI1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1
3007	HIST1H1D	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
3008	HIST1H1E	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
3009	HIST1H1B	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1
3110	HLXB9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3198	HOXA1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
3205	HOXA9	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
3207	HOXA11	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
3209	HOXA13	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3231	HOXD1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1
3234	HOXD8	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1
3642	INSM1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0
3975	LHX1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1
4210	MEFV	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4656	MYOG	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4796	NFKBIL2	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1



GeneID	Gene Symbol	adrenal gland	blood	blood vessel	bone	bone marrow	FRS	heart	kidney	liver	lung	lymph	MRS	mucosa	pancreas	pineal gland	pituitary gland	spleen	stem cell	thymus	thyroid	other
4821	NKX2-2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4861	NPAS1	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1
4901	NRL	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
5013	OTX1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
5076	PAX2	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
5077	PAX3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1
5079	PAX5	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1
5081	PAX7	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5453	POU3F1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1
5454	POU3F2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
5455	POU3F3	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
5462	POU5F1P1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5992	RFX4	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	1
6474	SHOX2	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1
6493	SIM2	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
6496	SIX3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6664	SOX11	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1
6689	SPIB	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1
6877	TAF5	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0	1
6899	TBX1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1
6913	TBX15	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1
7023	TFAP4	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1
7161	TP73	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1

GeneID	Gene Symbol	adrenal gland	blood	blood vessel	bone	bone marrow	FRS	heart	kidney	liver	lung	lymph	MRS	mucosa	pancreas	pineal gland	pituitary gland	spleen	stem cell	thymus	thyroid	other
7291	TWIST1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1
7310	U2AF1L1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1
7546	ZIC2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1
7621	ZNF70	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7673	ZNF222	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
7675	ZNF121	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
7710	ZNF154	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1
7768	ZNF225	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	1	0	0	1
8092	CART1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
8193	DPF1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8320	EOMES	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	1
8345	HIST1H2BH	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0
8820	HESX1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8970	HIST1H2BJ	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1
9970	NR1I3	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1
10215	OLIG2	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
10655	DMRT2	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
10794	ZNF272	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0
11077	HSF2BP	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	1
11281	POU6F2	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	1	1
25806	VAX2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
26038	CHD5	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1
26108	PYGO1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1

GeneID	Gene Symbol	adrenal gland	blood	blood vessel	bone	bone marrow	FRS	heart	kidney	liver	lung	lymph	MRS	mucosa	pancreas	pineal gland	pituitary gland	spleen	stem cell	thymus	thyroid	other
26468	LHX6	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1
27023	FOXB1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1
27164	SALL3	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1
27288	HNRNPG-T	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
27439	CECR6	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1
30009	TBX21	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
30012	TLX3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
50805	IRX4	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1
51022	GLRX2	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1
51402	LW-1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
51450	PRRX2	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
54626	HES2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1
55552	HSZFP36	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
55659	ZNF416	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	1
56938	ARNTL2	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1
56978	PRDM8	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1
57116	ZNF695	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
57332	CBX8	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1
57343	ZNF304	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
57801	HES4	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
58495	OVOL2	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1
60529	ALX4	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1
63978	PRDM14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1



GeneID	Gene Symbol	adrenal gland	blood	blood vessel	bone	bone marrow	FRS	heart	kidney	liver	lung	lymph	MRS	mucosa	pancreas	pineal gland	pituitary gland	spleen	stem cell	thymus	thyroid	other
79192	IRX1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	1
79722	FLJ11795	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0	0	0	1
79816	TLE6	1	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
79862	ZNF669	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
80032	ZNF556	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84127	RUNDC2A	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
84911	ZNF382	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0	1
85409	NKD2	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	1
85446	ZFHX2	1	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
89870	TRIM15	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
90649	ZNF486	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1
94039	ZNF101	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	1
94234	FOXQ1	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1
116448	OLIG1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
126295	LOC126295	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0
129025	SUHW1	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1
136051	DKFZp7621137	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
138474	TAF1L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
140883	SUHW2	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1
142689	ASB12	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
146434	ZNF597	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1
148268	ZNF570	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
148979	GLIS1	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1

GeneID	Gene Symbol	adrenal gland	blood	blood vessel	bone	bone marrow	FRS	heart	kidney	liver	lung	lymph	MRS	mucosa	pancreas	pineal gland	pituitary gland	spleen	stem cell	thymus	thyroid	other	
161253	FLJ38964	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	
162979	ZNF342	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1
163059	ZNF433	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0
163071	ZNF114	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
170302	ARX	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
171392	ZNF675	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1
221527	ZBTB12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
245806	VGLL2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
253738	EBF3	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
283078	MKX	0	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1
285676	ZNF454	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
339416	ANKRD45	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
339488	TFAP2E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
341405	ANKRD33	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1



# Appendix IX Genome-wide analysis of cancer/testis gene expression. *Proc Natl Acad Sci U S A.*

PNAS PNAS PNAS

## Genome-wide analysis of cancer/testis gene expression

Oliver Hofmann<sup>a,b,1</sup>, Otavia L. Caballero<sup>c</sup>, Brian J. Stevenson<sup>d,e</sup>, Yao-Tseng Chen<sup>f</sup>, Tzeela Cohen<sup>g</sup>, Ramon Chua<sup>h</sup>, Christopher A. Maher<sup>b</sup>, Sumir Panji<sup>b</sup>, Ulf Schaefer<sup>b</sup>, Adele Kruger<sup>b</sup>, Minna Lehtvaslaih<sup>b</sup>, Piero Carninci<sup>i,j</sup>, Yoshihide Hayashizaki<sup>k,l</sup>, C. Victor Jongeneel<sup>d,e</sup>, Andrew J. G. Simpson<sup>f</sup>, Lloyd J. Old<sup>c,1</sup>, and Winston Hide<sup>a,b</sup>

<sup>a</sup>Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, SPH2, 4th Floor, Boston, MA 02115; <sup>b</sup>South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa; <sup>c</sup>Ludwig Institute for Cancer Research, New York Branch at Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021; <sup>d</sup>Ludwig Institute for Cancer Research, Lausanne Branch, 1015 Lausanne, Switzerland; <sup>e</sup>Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; <sup>f</sup>Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021; <sup>g</sup>Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; and <sup>h</sup>Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, 2-1 Hirasawa, Wako, Saitama, 3510198, Japan

Contributed by Lloyd J. Old, October 28, 2008 (sent for review June 6, 2008)

**Cancer/Testis (CT) genes, normally expressed in germ line cells but also activated in a wide range of cancer types, often encode antigens that are immunogenic in cancer patients, and present potential for use as biomarkers and targets for immunotherapy. Using multiple in silico gene expression analysis technologies, including twice the number of expressed sequence tags used in previous studies, we have performed a comprehensive genome-wide survey of expression for a set of 153 previously described CT genes in normal and cancer expression libraries. We find that although they are generally highly expressed in testis, these genes exhibit heterogeneous gene expression profiles, allowing their classification into testis-restricted (39), testis/brain-restricted (14), and a testis-selective (85) group of genes that show additional expression in somatic tissues. The chromosomal distribution of these genes confirmed the previously observed dominance of X chromosome location, with CT-X genes being significantly more testis-restricted than non-X CT. Applying this core classification in a genome-wide survey we identified >30 CT candidate genes; 3 of them, PEPP-2, OTOA, and AKAP4, were confirmed as testis-restricted or testis-selective using RT-PCR, with variable expression frequencies observed in a panel of cancer cell lines. Our classification provides an objective ranking for potential CT genes, which is useful in guiding further identification and characterization of these potentially important diagnostic and therapeutic targets.**

gene index | prediction

**C**ancer/Testis (CT) genes are a heterogeneous group that are normally expressed predominantly in germ cells and in trophoblasts, and yet are aberrantly activated in up to 40% of various types of cancer types (1). A subset of the CT genes has been shown to encode antigens that are immunogenic and elicit humoral and cellular immune responses in cancer patients (2). Because of their restricted expression profile in normal tissues and because the testis is an immunoprivileged site, the CT antigens are emerging as strong candidates for therapeutic cancer vaccines, as revealed by early-phase clinical trials (3–10). Biologically, the CT genes provide a model to better understand complex gene regulation and aberrant gene activation during cancer.

Any gene that exhibits an mRNA expression profile restricted to the testis and neoplastic cells can be termed a CT gene. Existing definitions of CT genes vary in the literature, from genes expressed exclusively in adult testis germ cells and malignant tumors (1, 11) to dominant testicular expression (12), possible additional presence in placenta and ovary and epigenetic regulation (13), or membership of a gene family and localization on the X chromosome (14). Reflecting this lack of a consensus definition, an increasing number of heterogeneous CT candidates have appeared in the literature, with available

expression profile information frequently limited to the original defining articles. In some cases, e.g., ACRBP, the original CT-restricted expression in normal tissues could not be confirmed by subsequent experiments (1). Partially due to this lack of a clear and broadly applicable definition, or “type specimen,” for a CT gene, it has become increasingly challenging to identify the CT genes that are most suitable for cancer vaccine development. Moreover, this incoherent classification increases the risk of pursuing unsuitable clinical targets. However, with more expression data becoming available, CT gene transcripts of genes originally thought to have the CT expression profile are being detected in additional tissues (1), resulting in the more stringent “testis-restricted” description being altered to one of “testis-preference.” Based on a compilation from the published literature, the CT database now lists >130 RefSeq nucleotide identifiers as CT genes that belong to 83 gene families ([www.cta.lncb.br](http://www.cta.lncb.br)). An analysis of the human X chromosome has also suggested that as many as 10% of the genes on this chromosome may be CT genes (15). Given this increasing number of CT and CT-like genes, their comprehensive classification based on expression profiles is essential for our understanding of their biological role and regulation of expression.

In an attempt to resolve this and to identify new CT antigens, we have taken an in silico approach to produce a comprehensive survey of CT gene expression profiles by combining expression information from an existing corpus of >8,000 cDNA libraries (16) together with the depth and resolution provided by massively parallel signature sequencing (MPSS) expression libraries (17), cap-analysis of Gene Expression (CAGE) libraries (18), and a survey using semiquantitative reverse-transcription PCR (RT-PCR) on a panel of 22 normal tissues. As a result, we have created a coherent classification of CT genes, and new CT genes have been identified using well-informed, structured prediction and confirmation criteria.

### Results and Discussion

CT classification. CT genes were classified into 3 groups, testis-restricted, testis/brain-restricted and testis-selective, based on

Author contributions: O.H., O.L.C., C.A.M., U.S., A.K., A.J.S., L.J.O., and W.H. designed research; O.H., O.L.C., T.C., and R.C. performed research; B.J.S., Y.-T.C., T.C., C.A.M., S.P., U.S., M.L., A.K., P.C., Y.H., and C.V.J. contributed new reagents/analytic tools; O.H., O.L.C., and B.J.S. analyzed data; and O.H. wrote the paper.

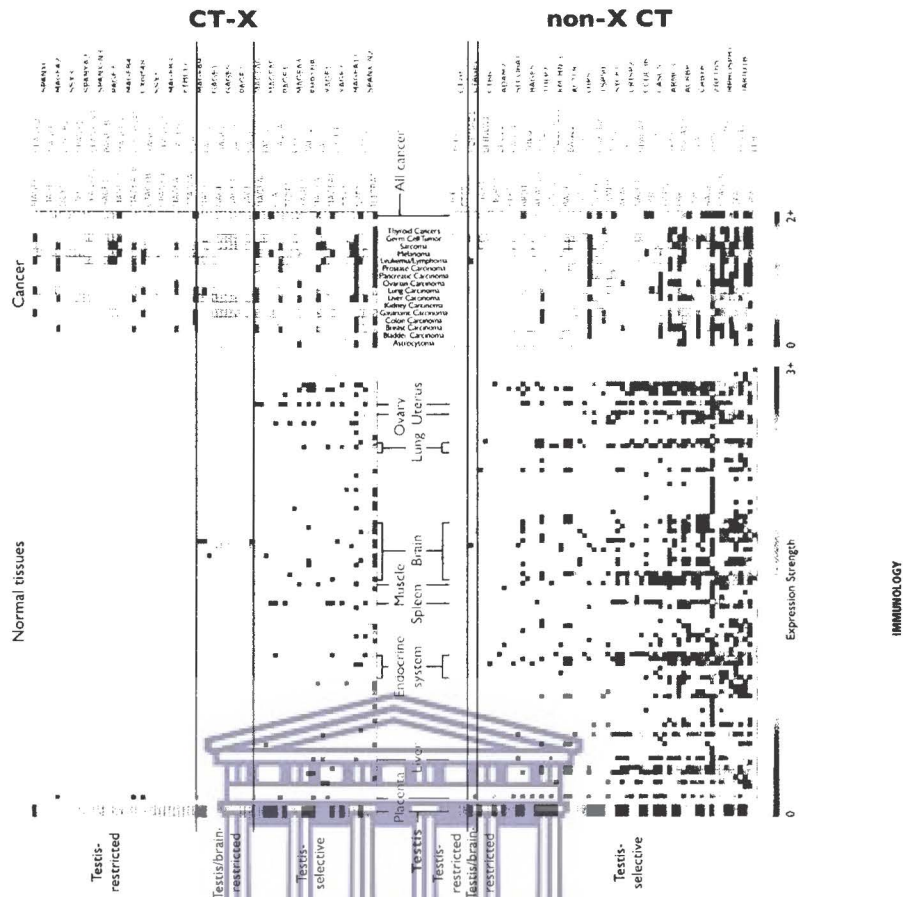
The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

<sup>1</sup>To whom correspondence may be addressed. E-mail: [ohofmann@sph.harvard.edu](mailto:ohofmann@sph.harvard.edu) or [loid@icr.org](mailto:loid@icr.org).

This article contains supporting information online at [www.pnas.org/cgi/content/full/081077105-DC1](http://www.pnas.org/cgi/content/full/081077105-DC1) Supplemental.

© 2008 by The National Academy of Sciences of the USA.



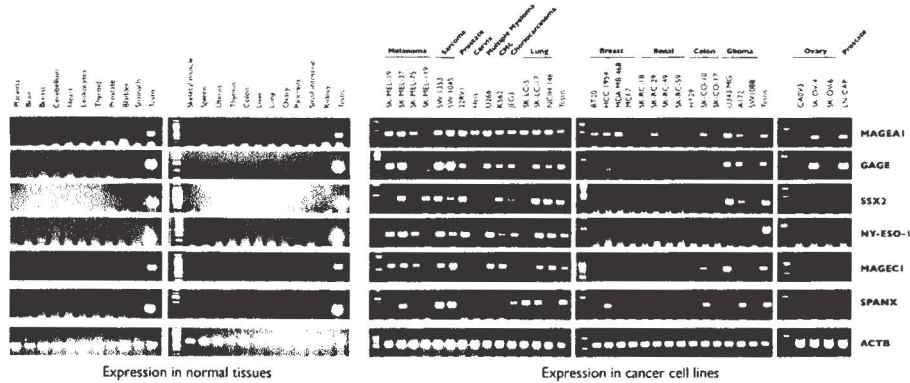
**Fig. 1.** Merged expression profiles of CT-X (left array) and non-X CT genes (right) based on expression data from RT-PCR and cDNA, MPSS and CAGE libraries from tissues sources annotated as normal and “adult” (lower) or “cancer.” Expression in normal testis, placenta, and selected tissues is marked. Color reflects the support for the expression of a CT genes in a given anatomical site (blue for low combined expression evidence  $\geq 1$ , red for strong support from at least 3 sources (for the normal tissue panel) with a total score  $\geq 3$ ) or 2 sources (the cancer panel lacking RT-PCR data), respectively. The most abundant expression (red) is seen in testis for most genes, particularly in the non-X CT group. Expression values were normalized on a per-gene basis relative to the combined normal testis/placenta expression confidence (Lower) or the source of the highest cancer expression confidence (Upper). The 3 CT annotation groups (testis-restricted, testis/brain-restricted and testis-selective) are highlighted. See Dataset S3 for the full list of CT classifications.

WESTERN CAPE

their expression profiles obtained from a manually curated corpus of cDNA, MPSS, CAGE expression libraries and RT-PCR (see Dataset S1 for MPSS and CAGE library annotation and <http://evoontology.org> for the cDNA annotation). By merging expression information using different technology platforms, we were able to leverage their individual strengths—the breadth of tissue coverage associated with the cDNA/EST expression libraries, the high sensitivity of CAGE/MPSS and the ability to

custom-tailor PCR primers. Of 153 genes, 39 with transcripts present only in adult testis and no other normal adult tissue except for placenta were classified as testis-restricted; 14 CT genes with additional expression in other adult immunorestricted sites (all regions of the brain) were classified as testis/brain-restricted, and 85 genes, designated as testis-selective, were ranked by the ratio of testis/placenta expression relative to other expression in normal adult tissues (see Fig. 1 for

Hofmann et al.



**Fig. 2.** RT-PCR analysis of selected CT genes in the testis-restricted category (MAGEA1, GAGE, SSX2, NY-ESO-1, MAGEC1, and SPANX). Expression profiles are shown for a range of 22 normal tissues (Left) and 31 cancer cell lines (Right).

the expression array, Fig. 2 for the PCR panel of selected testis-restricted CT genes, and Fig. S1 and Dataset S2 for arrays from individual expression sources).

An uneven chromosomal distribution of the CT genes was observed, with 83 of 153 genes (54%) being on the X chromosome, and 70 on non-X chromosomes (Fig. S3). Furthermore, 35 CT-X genes were classified as testis-restricted, whereas only 4 non-X CT genes belong to this group. An additional 12 CT-X genes were found to be testis/brain-restricted, compared with 2 non-X testis/brain-restricted CT genes. CT-X gene family members thus appear to be under more stringent transcriptional restriction in somatic tissues, whereas non-X CT genes are more broadly expressed. This validates the CT gene classification into CT-X and CT non-X groups, with the CT-X group being of particular interest for therapeutic approaches.

Twenty-six CT-X and 59 non-X CT genes belong to the testis-selective category, and 36 of these genes (5 CT-X and 31 non-X CT) had >50% of the expression evidence derived from non-testis or placental libraries, indicating that these might not qualify as CT genes.

Seven CT genes were not identified in any library at all (2 CT-X and 5 non-X CT). An additional 8 CT-X genes (SPANX, PAGE1, CSAG1, SSX5/6/7/9, and CT45-2) were not present in any testis-annotated library. Of these, SSX5 and SSX7 have been shown to be expressed in testis by RT-PCR (19), suggesting a likely discrepancy in mapping short sequence tags to their genomic counterparts, an expected phenomenon for large and highly homologous gene families like SSX. In contrast, the absence of testicular expression of SSX6 and SSX9 was confirmed in that study, indicating that some of the currently recognized CT genes could either be silent or expressed at extremely low levels in testis. The full list with classification and raw expression scores across the merged expression array can be found in Dataset S3.

Associations between different CT gene properties and their assigned classification were analyzed using the APRIORI algorithm. Besides being more likely testis-restricted, CT-X genes were found to be more often members of multigene families than non-X CTs. In addition, Gene Ontology terms showed CT-X genes to be more often in the “molecular function unknown” and “biological process unknown” categories, whereas the non-X CTs are associated with known functions such as meiosis, sexual

reproduction, and gametogenesis (see Dataset S4 for all attributes and annotations).

While the description of CT-X genes such as NY-ESO-1 (20), SSX2 (21), and MAGE-A1 (22) match our classification—all are in the testis-restricted category—not all CT genes were found to be as testis-restricted as described in the literature. BAGE, SPO11, LIPI, LDHC, and BRDT, considered to be testis-restricted based on a tissue panel of 13 non-gametogenic normal tissues (1), fall into the testis-selective category in our screen, most likely due to a larger amount of expression sources sampled. Despite the broader coverage we could not confirm an expression of MAGE-A1, MAGE-C1, and NY-ESO-1 at low levels in the pancreas reported in the same study. In agreement with the study in ref. 1, we found IL13RA1, ACRBP, and SPA17 to be expressed in a wide variety of tissues, falling into the lower end of the testis-selective category.

In the present study, we have ranked the testis-selective genes based upon the ratios of their expression evidence in testis and placenta relative to other somatic tissues, rather than using fixed thresholds and the number of somatic tissues in which a CT candidate is allowed as the distinguishing criteria for CT versus non-CT genes (2). Genes without any somatic expression have unique potential for cancer vaccines and other therapeutic approaches to cancer. From past work involving screening of larger sets of genes (23), a cutoff was introduced that defined CT candidate genes as genes with 2-fold higher expression evidence in testis and placenta relative to all other somatic normal tissues. This approach was complementary to our current one and will not require updated thresholds as the number of sampled tissue sources increases.

Intriguingly, a number of CT genes were found to be expressed in no somatic tissues except for brain, suggesting the presence of a distinctive transcriptional control mechanism that functions with tissue specificity in germ cells and in brain. There have been relatively few studies of CT gene expression in different anatomical regions of normal brain and similarly not many in brain tumors (24, 25), except for NXF2, which was shown to be expressed in normal brain (26). Our in silico study has discovered a broader subset of CT genes with brain expression, among them members of the otherwise fully testis-restricted GAGE and MAGE families, found to be expressed in the hippocampus and cerebral cortex. A previous study has similarly identified a group



of cancer/testis/brain (CTB) antigens (27). However, despite the bioinformatic evidence, we have not been able to confirm the expression of selected CT genes (MAGEA9, MAGEC2, PASD1, and GAGE) in tissue samples from total brain, cerebellum, caudate nucleus, thalamus, frontal cortex, occipital cortex, pons, or amygdala by RT-PCR (data not shown), and whether these genes are expressed in brain remains to be proven.

**Distribution of CT Genes in Cancer Tissues.** Our ranking by the number of different cancer types and anatomical sites of CT genes expressed in cancer-annotated libraries distinguishes CT-“rich” and CT-“poor” tumors based on the *in silico* analysis obtained from cDNA, CAGE, and MPSS libraries (Fig. 1 and Dataset S5). The broadest distribution of CT genes was found in germ cell tumors, melanomas and lung carcinomas, adenocarcinomas and chondrosarcomas. Breadth of cancer expression was uncorrelated with tissue restriction in normal tissues ( $r = 0.18$  for CT-X genes,  $r = 0.02$  for non-X CT genes using Spearman rank correlation); for instance, the fully testis-restricted CT genes, such as MAGEA2/A2B and CTAG2, were found to be present in a variety of different tumor tissues.

Melanoma, non-small-cell lung cancer, hepatocellular carcinoma and bladder cancer have been identified as high CT gene expressors, with breast and prostate cancer being moderate and leukemia/lymphoma, renal and colon cancer low expressors (1). Our *in silico* analysis confirms this distinction, in particular for tumor tissues well represented by the available libraries, showing a broad distribution of CT genes expressed in cancers of skin including melanoma (43% of CT genes with cancer expression were found in at least one melanoma library), lung (37%), and liver (34%). Strong presence of CT expression found in the present study but not by previous RT-PCR studies includes tumors from germ cells (39%), stomach (28%), and cartilage (chondrosarcomas, 26%). One reason for this discrepancy could be the lack of RT-PCR data for certain tumors, e.g., gastric cancer is much rarer than other carcinomas in the Western world, and mesenchymal tumors are also not well represented in many of the RT-PCR studies to date. Our *in silico* information may thus serve as a guide for future experimental investigations, especially useful for recently described CT genes not yet analyzed in great detail. Discrepancies are also likely to occur due to the potential inclusion of cancer cell line samples in the survey that, unlike normal tissue samples explicitly labeled as normal, are often not distinguished from primary tumor samples. A third reason for this observed discrepancy could be the bias that resulted from differences in library numbers studied for each tumor type: for instance, ovarian cancer is CT-rich by RT-PCR but not evident from our *in silico* study, possibly due to the low number of available ovarian cDNA libraries. However, colon cancer, a CT-poor tumor, was correctly shown to have low frequency of CT genes despite the large number of colon libraries in the databases, and this would argue that the difference in library numbers may not have been a significant factor. Last, the *in silico* finding of high CT expression in germ cell tumor represents a special situation that can be explained by two reasons. One is that a subset of CT genes, particularly the non-X CTs, encode proteins with known specific functions in germ cells, and their expression in germ cell tumors represents the preserved expression of lineage-specific markers—rather than aberrant gene activation, conceptually similar to the expression of thyroglobulin by thyroid cancer or prostate specific antigen by prostate cancer. The other reason would be that the germ cell tumors from which the mRNA expression profiles were derived could have been contaminated by the adjacent or entrapped testicular tissue, which provides the source for CT gene transcripts when the germ cell tumor was actually negative for the CT gene in question.

**CT Candidate Prediction.** Prediction of CT candidates based on their expression profiles in cDNA, MPSS, and CAGE libraries resulted in 28 genes supported by 2 expression platforms in the testis- or testis/brain-restricted category, including 10 known CT genes and 18 novel CT candidates (Fig. S3 and Dataset S6). An additional, less stringent screen for CT-X genes identified 47 genes in the same categories, including 34 known CT genes and 13 novel candidates. After manual curation, the list of novel candidates was extended to include the highest scoring testis-selective CT-X candidates, TKTL1 and NXF3, the latter being a known CT gene, a member of the NXF2 CT family (28).

Of 33 novel CT candidate genes, 12 most promising genes were manually selected for experimental validation by RT-PCR based on an evaluation of available gene expression data in human cancer. Of the 5 X- and 7 non-X-chromosomal candidates, 11 transcripts could be amplified, whereas transcripts from VCX2 were not detected in any of the 23 normal tissue RNA samples. Three of the amplified gene transcripts exhibited testis-restricted (AKAP4) or testis-selective (PEPP-2, OTOA) expression (data not shown). RT-PCR products of these genes were also detected in samples from a panel of 30 cancer cell lines.

PEPP-2, an X-linked human homeobox gene, encodes a transcriptional factor with similar cancer/testis restricted expression patterns in both human and mouse (29); it is also a member of a top 50 list of genes under strong positive selection between human and chimpanzee (30). Otoacrorin (OTOA) was reported to be specific to sensory epithelia of the inner ear (31), but has also been associated with ovarian and pancreatic cancer due to its homology with mesothelin, a cancer immunotherapy target (32). AKAP4 (CT-X), identified in the 2-platform screen, exhibits weak expression in different cancer cell lines and encodes a kinase anchor protein (33) involved in the cAMP-regulation of motility (34) and was recently suggested as a CT gene in an independent study (35).

All 3 confirmed genes are candidates for immunotherapy based on their restricted expression, and further investigation of their mRNA and protein expression in various tumors is warranted and ongoing. Given the comprehensive nature of our study and the limited number of confirmed novel CT candidates, it seems that the number of true CT genes matching the criterion of stringent testis-restricted expression profile has reached a plateau.

Although it is clear that the CT designation has been inappropriately given to a large number of genes with wide normal tissue expression, it is less evident how precisely the term CT should be applied. There is no difficulty with CT genes whose expression profile have a classic CT pattern; we estimate  $\approx 39$  genes presently in this category and  $\approx 90\%$  of them reside on the X chromosome. The challenge for the remaining CT genes, most of which are non-X coded, is that they are expressed in testis and cancer, but are also expressed in a limited number of normal tissues. Should these be designated CT? Perhaps the best solution at this point would be to assemble further information about CT genes and their products, including function, binding partners, evolutionary selection (36), control of gene expression, identification of expressing normal somatic cells, aberrant non-lineage expression in cancer, and immunogenicity, before establishing a uniform classification of CT genes.

#### Methods

**Selection of CT Genes.** A total of 153 CT genes (200 unique RefSeq transcript identifiers) were selected from the CT Antigen DB (<http://www.cta.lncx.br>) and by manual curation of the literature. Genes were annotated with their most current gene identifiers and merged based on shared National Center for Biotechnology Information RefSeq nucleotide identifiers (Dataset S7). Additional gene identifiers were obtained from RefSeq release 11 (37), IPI version 3.29 (38); genomic coordinates were taken from the University of California, Santa Cruz Genome Browser hg18 human genome build (39). Of these 153



genes, 83 that encode 107 RefSeq transcripts were mapped to the X chromosome (CT-X genes) whereas 70 genes were on autosomes (non-X CT genes). Subcellular localization was based on predictions in the human version of the LOCATE system (40). SEREX information was obtained from the Cancer Immunome Database website (<http://ludwig-sun5.unil.ch/CancerImmunoDB>). Ambiguities were resolved by manual curation.

**Source of Expression Information.** Gene expression profiles were determined based on 4 different sources: 99 CAGE libraries from the RIKEN FANTOM3 project (18), 47 MPSS libraries (17, 23, 41), a collection of 8401 cDNA expression libraries from the eVOC system (16), and semiquantitative RT-PCR across 22 normal tissue samples. Source materials were annotated with regards to the anatomical site and pathological status of their source tissues. In cases where the anatomical source was unclassified, cell type information was used. Bone marrow/blood libraries were designated bone marrow, and all combinations with mucosa (colon, stomach) were merged into "mucosa." Libraries not explicitly annotated as "normal" were considered as unclassified. Libraries from pooled tissue sources were ignored, and pooled samples were kept as long as the pathological and anatomical status was identical for all donors (see Dataset S1 for annotated libraries).

**Pseudoarrays.** Expression information was organized into "pseudoarrays" based on expression information obtained from CAGE-, MPSS-, and cDNA-libraries in the case of cancer expression and merged with RT-PCR results in the case of normal tissue expression. Columns reflect the class of library in which a CT transcript was identified and rows represent individual RefSeq transcripts. Annotation was based on the general library class description (normal, cancer or unclassified) combined with pathological state and anatomical site. To evaluate the relative levels of CT expression we converted expression signals from the 4 sources into "expression evidence": For CAGE- and MPSS-based expression data, expression evidence was based on detected tags per million (TPM), with matches  $<3$  TPM ( $\approx 1$  transcript per cell) filtered out. Normalized and subtracted EST libraries prevent quantitation of expression strength based on EST counts, therefore expression evidence is represented by the number of cDNA libraries in which a given transcript was identified. RT-PCR results were manually binned into 5 groups of expression, ranging from 0 (not expressed) to 4 (strongly expressed). For each expression source, evidence values were normalized on a per-transcript basis by setting the highest expression evidence in normal tissues to a value of 1, reflecting relative changes in expression levels across tissues and pathological states. Pseudoarrays from the 4 expression sources were merged by summing the individual expression evidence scores for a given transcript from each platform. Expression profiles for multiple transcripts associated with the same gene were merged into a single representation, keeping the highest expression score for overlapping annotations. In arrays where annotation was "merged" into single columns based on their class (e.g., all cancer expression information), the highest expression score across all annotated libraries was kept for each gene.

**Visualization and Ranking.** Genes were divided into CT-X and non-X CT panels, then individually ranked by their expression properties in normal tissues and classified into the following 3 categories: (i) expression in testis and placenta only (testis-restricted); (ii) expression in testis, placenta and brain-regions only (testis/brain-restricted); and (iii) all other genes (testis-selective). Final ranking within each category was obtained by sorting based on decreasing level of normal tissue specificity as measured by the combined testis and placenta expression evidence divided by all normal expression evidence. All arrays were visualized using MeV 4.0 ([www.tm4.org](http://www.tm4.org)).

**Clustering Methods.** Associations between CT annotation and their classification were investigated by recording their assigned class; presence or absence

in placenta, brain, testis, and developing ovary; their testis/placenta tissue specificity; their X vs. non-X chromosomal status; membership in a gene family; subcellular localization; and evolutionary status (36) followed by an analysis with the APRIORI algorithm (42), which identifies association rules matching a predefined threshold of support (30%) and confidence ( $\geq 0.8$ )

**Search Criteria for CT Candidates.** CT candidates were identified using the same in silico expression sources, but with no filters for minimum TPM value and satisfying the following criteria: (i) exhibit expression in testis and at least one cancer-associated tissue at 10 TPM (CAGE, MPSS) or presence in at least one EST/cDNA library with testis and cancer annotation; (ii) not be present above those levels in any other tissue except for placenta, ovary, and brain; and (iii) be supported independently by 2 platforms. Identified candidates were ranked using the same approach used to classify known CT genes. To increase coverage of CT-X genes, a second genome-wide search was conducted requiring support from only a single platform. Candidates were selected for RT-PCR validation by manual curation, removing hypothetical proteins, predicted genes and candidates with multiple publications indicating expression in somatic tissues.

**RT-PCR.** RNA preparations were purchased from the normal tissue panels of Clontech and Ambion or prepared from cancer cell lines using the RNeasy kit (Qiagen) and were used to prepare cDNA for RT-PCR. A total of 1.0  $\mu$ g of RNA was reverse transcribed into cDNA in a total volume of 20  $\mu$ l using the Omniscript RT kit (Qiagen) according to the manufacturer's protocol using oligo(dT)<sub>18</sub> primers (Invitrogen). The cDNA was diluted 5 times and 3  $\mu$ l was used in the PCR with primers specific to each analyzed gene in a final volume of 25  $\mu$ l. Primers used for PCR amplification were designed to have an annealing temperature  $\approx 60$  °C using Primer3 software ([www.genome.wi.mit.edu/cgi-bin/primer/primer3www.cgi](http://www.genome.wi.mit.edu/cgi-bin/primer/primer3www.cgi)) and were chosen to encompass introns between exon sequences to avoid amplification of genomic DNA. DNase treatment was undertaken before cDNA synthesis to analyze intronless genes. Primers were designed to target all known variants of a gene in RefSeq and their specificity was confirmed by aligning with the National Center for Biotechnology Information sequence databases using BLAST ([www.ncbi.nlm.nih.gov/blast/blast.cgi](http://www.ncbi.nlm.nih.gov/blast/blast.cgi)). Primer sequences and amplicon sizes are provided in Dataset S2.

JumpStart REDTaq ReadyMix (Sigma Aldrich) was used for amplification according to the manufacturer's instructions. Samples were amplified with a pre-cycling hold at 95 °C for 3 min, followed by 35 specific cycles of denaturation at 95 °C for 15 seconds, annealing for 30 seconds (10 cycles at 60 °C, 10 cycles at 58 °C and 15 cycles at 56 °C) and extension at 72 °C for 30 seconds followed by a final extension step at 72 °C for 7 min.  $\beta$ -actin was amplified as control. PCR products were separated on 1.5% agarose gels stained with ethidium bromide. For semiquantitative PCR analysis, RT-PCR products were classified into 0 (negative) to 4 (strongest signal) based on the intensity of the product on ethidium bromide-stained gels.

**ACKNOWLEDGMENTS.** We thank Dmitry Kuznetsov for providing access to the SEREX information on CT genes and Erika Ritter (Ludwig Institute for Cancer Research, New York Branch at Memorial Sloan-Kettering Cancer Center, New York) for providing cell lines. This project was supported by the South African National Bioinformatics Network; National Institutes of Health Stanford-South African Informatics Training for Global Health Grant TW-03-008; Atlantic Philanthropies; The Oppenheimer Memorial Trust; a Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government (to Y. H.); and a grant from the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan. This work was conducted as part of the Hilton-Ludwig Cancer Metastasis Initiative, funded by the Conrad N. Hilton Foundation and the Ludwig Institute for Cancer Research.

- Scanlan MJ, Simpson AJG, Old LJ (2004) The cancer/testis genes: Review, standardization, and commentary. *Cancer Immunol* 4:1.
- Simpson AJG, Caballero OL, Jungbluth A, Chen YT, Old LJ (2005) Cancer/testis antigens, gametogenesis and cancer. *Nat Rev Cancer* 5:615–625.
- Murchand M, et al. (1999) Tumor regressions observed in patients with metastatic melanoma treated with an antigenic peptide encoded by gene MAGE-3 and presented by HLA-A1. *Int J Cancer* 80:219–230.
- Davis ID, et al. (2004) Recombinant NY-ESO-1 protein with iscomatrix adjuvant induces broad integrated antibody and CD4(+) and CD8(+) t cell responses in humans. *Proc Natl Acad Sci USA* 101:10697–10702.
- Jäger E, et al. (2006) Recombinant vaccinia/fowlpox NY-ESO-1 vaccines induce both humoral and cellular NY-ESO-1-specific immune responses in cancer patients. *Proc Natl Acad Sci USA* 103:14453–14458.
- Valmori D, et al. (2007) Vaccination with NY-ESO-1 protein and CPG in montanide induces integrated antibody/Th1 responses and CD8 T cells through cross-priming. *Proc Natl Acad Sci USA* 104:8947–8952.
- Uenaka A, et al. (2007) T cell immunomonitoring and tumor responses in patients immunized with a complex of cholesterol-bearing hydrophobized pullulan (chp) and NY-ESO-1 protein. *Cancer Immunol* 7:9.
- Odunsi K, et al. (2007) Vaccination with an NY-ESO-1 peptide of HLA class II specificities induces integrated humoral and t cell responses in ovarian cancer. *Proc Natl Acad Sci USA* 104:12837–12842.
- Atanackovic D, et al. (2006) Expression of cancer-testis antigens as possible targets for antigen-specific immunotherapy in head and neck squamous cell carcinoma. *Cancer Biol Ther* 5:1218–1225.
- Gnjatic S, et al. (2006) NY-ESO-1: Review of an immunogenic tumor antigen. *Adv Cancer Res* 95:1–30.
- Scanlan MJ, et al. (2002) Identification of cancer/testis genes by database mining and mRNA expression analysis. *Int J Cancer* 98:485–492.
- Zendman AJW, Ruiter DJ, Muijen GNPV (2003) Cancer/testis-associated genes: Identification, expression profile, and putative function. *J Cell Physiol* 194:272–288.

13. Costa FF, Blanc KL, Brodin B (2007) Concise review: Cancer/testis antigens, stem cells, and cancer. *Stem Cells* 25:707–711.
14. Kalejs M, Erenpreisa J (2005) Cancer/testis antigens and gametogenesis: A review and "brain-storming" session. *Cancer Cell Int* 5:4.
15. Ross MT, et al. (2005) The DNA sequence of the human X chromosome. *Nature* 434:325–337.
16. Kelso J, et al. (2003) eVOC: A controlled vocabulary for unifying gene expression data. *Genome Res* 13:1222–1230.
17. Jongeneel CV, et al. (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res* 15:1007–1014.
18. Carninci P, et al. (2005) The transcriptional landscape of the mammalian genome. *Science* 309:1559–1563.
19. Güre AO, Wei U, Old LJ, Chen YT (2002) The S5X gene family: Characterization of 9 complete genes. *Int J Cancer* 101:448–453.
20. Chen YT, et al. (1997) A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. *Proc Natl Acad Sci USA* 94:1914–1918.
21. Tórceli O, et al. (1996) The S5X-2 gene, which is involved in the t(1;18) translocation of synovial sarcomas, codes for the human tumor antigen HOM-MEL-40. *Cancer Res* 56:4766–4772.
22. van der Bruggen P, et al. (1991) A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* 254:1643–1647.
23. Chen YT, et al. (2005) Identification of cancer/testis-antigen genes by massively parallel signature sequencing. *Proc Natl Acad Sci USA* 102:7940–7945.
24. Sahin U, et al. (2000) Expression of cancer testis genes in human brain tumors. *Clin Cancer Res* 6:3916–3922.
25. Scarcella DL, et al. (1999) Expression of MAGE and GAGE in high-grade brain tumors: A potential target for specific immunotherapy and diagnostic markers. *Clin Cancer Res* 5:335–341.
26. Zhang M, Wang Q, Huang Y (2007) Fragile X mental retardation protein FMRP and the RNA export factor NXF2 associate with and destabilize NXF1 mRNA in neuronal cells. *Proc Natl Acad Sci USA* 104:10057–10062.
27. Scanlan MJ, Güre AO, Jungbluth AA, Old LJ, Chen YT (2002) Cancer/testis antigens: An expanding family of targets for cancer immunotherapy. *Immunol Rev* 188:22–32.
28. Loriot A, Boon T, Siret CD (2003) Five new human cancer-germline genes identified among 12 genes expressed in spermatogonia. *Int J Cancer* 105:371–376.
29. Wayne CM, MacLean JA, Cornwall G, Wilkinson MF (2002) Two novel human x-linked homeobox genes, hPEPP1 and hPEPP2, selectively expressed in the testis. *Gene* 301:1–11.
30. Nielsen R, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* 3:e170.
31. Zwaenepoel I, et al. (2002) Otoancorin, an inner ear protein restricted to the interface between the apical surface of sensory epithelia and their overlying scalular cells, is defective in autosomal recessive deafness DFNB22. *Proc Natl Acad Sci USA* 99:6240–6245.
32. Muminova ZE, Strong TV, Shaw DR (2004) Characterization of human mesothelin transcripts in ovarian and pancreatic cancer. *BMC Cancer* 4:19.
33. Turner RM, Johnson LR, Haig-Ladewig L, Gerton GL, Moss SB (1998) An X-linked gene encodes a major human sperm fibrous sheath protein, HAKAP2, genomic organization, protein kinase a-rii binding, and distribution of the precursor in the sperm tail. *J Biol Chem* 273:32135–32141.
34. Michel JJC, Scott JD (2002) AKAP mediated signal transduction. *Annu Rev Pharmacol Toxicol* 42:235–257.
35. Chiriva-Internati M, et al. (2008) AKAP-4: A novel cancer testis antigen for multiple myeloma. *Br J Haematol* 140:465–468.
36. Stevenson BJ, et al. (2007) Rapid evolution of cancer/testis genes on the X chromosome. *BMC Genomics* 8:129.
37. Wheeler DL, et al. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res* 31:28–33.
38. Kersey PJ, et al. (2004) The International Protein Index: An integrated database for proteomics experiments. *Proteomics* 4:1985–1988.
39. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12:996–1006.
40. Fink JL, et al. (2006) LOCATE: A mouse protein subcellular localization database. *Nucleic Acids Res* 34:D213–7.
41. Grigoriadis A, et al. (2006) Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data. *Breast Cancer Res* 8:R56.
42. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. *Proc ACM SIGMOD Management Data* 22:207–216.



UNIVERSITY of the  
WESTERN CAPE

## **Appendix X: Manual curation steps applied in filtering the expression array generated for the investigation of 63 potential mouse cancer/testis genes**

Remove column if annotation is:

- Unclassifiable pathology
- Pooled from different tissues
- Non-cancer pathology
- Whole body, head, neck, trunk, anatomical site, maxillary process, anterior limb or diaphragm

Remove developmental stage information from annotation

Remove cell type information from annotation unless there is no anatomical system information

- Exception: keep cell type and discard anatomical system for 'fibroblast|synovium'

Remove 'unclassifiable\_AS' from annotation (unclassifiable anatomical system)

Remove column if annotation is now only normal

Merge:

- Carcinoma = adenocarcinoma, teratocarcinoma
- Bone = bone marrow
- Brain = cerebellum, cerebral cortex, corpus striatum, diencephalon, hippocampus, hypothalamus, lateral ventricle, medulla oblongata, midbrain, olfactory lobe
- Intestine = cecum, colon, small intestine
- Visual apparatus = choroid, retina
- Auditory apparatus = internal ear, spiral organ of Corti
- Blood = B-lymphocyte, erythroblast
- Lymphoreticular system = lymph node

For all annotations that are identical, merge them into one column and sum the values in each column for every gene.



UNIVERSITY *of the*  
WESTERN CAPE



