DOCTOR OF PHILOSOPHY

Development of bioinformatics analytical approach to identify novel human cancer testis gene candidates

Feichtinger, Julia

Award date:
2012

Awarding institution:
Bangor University

Link to publication
Development of a Bioinformatic Analytical Approach to Identify Novel Human Cancer Testis Gene Candidates

A thesis submitted to Bangor University in candidature for the degree of Doctor of Philosophy in Cancer Studies

Julia Feichtinger

North West Cancer Research Fund Institute, Bangor University, Bangor, Gwynedd LL57 2UW, UK
December 2012
Summary

The identification of tumour antigens (TAs) represents an ongoing challenge to the development of novel cancer diagnostic, prognostic and therapeutic strategies. A group of proteins, the cancer testis (CT) antigens are promising targets for such clinical applications. Their encoding genes show expression restricted to the immunologically privileged testes but their expression is also found in cells with a cancerous phenotype. To facilitate and automate the identification of novel CT genes, bioinformatic analytical pipelines based on publicly available microarray and expressed sequence tag (EST) data were developed and implemented as web tools to support wider application. Human germline-associated datasets were generated and the developed screening pipelines were subsequently used to analyse these datasets, leading to the identification of a novel cohort of meiosis-specific genes, the meiCT genes that exhibit the characteristics of CT genes and may have oncogenic features. In general, frequent germline gene expression found in cancer could reflect a soma-to-germline transformation occurring in human cells in the course of the development of cancer. The expression of germline-specific genes, in particular of meiotic genes, could lead to the production of proteins that cause oncogenic events and thus contribute to tumorigenesis and to the acquisition of tumour characteristics.
Contents

Please note that chapters with * are presented as papers/manuscripts and thus do not follow regular numbering. The numbering for these sections is omitted.

Acknowledgements xiii
Declaration xv
Abbreviations xix

1 Introduction 1
   1.1 Cancer and Cancer Testis Antigens ................................. 1
   1.1.1 Cancer .......................................................... 1
   1.1.2 Carcinogenesis .................................................. 1
   1.1.3 Tumour Antigens ............................................... 3
   1.1.4 Cancer Testis Antigens ....................................... 5
   1.2 Meiosis ............................................................. 11
   1.2.1 The Principle of Meiosis ........................................ 11
   1.2.2 Mammalian Gametogenesis ...................................... 12
   1.2.3 Meiotic Entry .................................................. 15
   1.2.4 Meiotic Recombination ......................................... 16
   1.2.5 The Synaptonemal Complex .................................... 18
   1.2.6 Dependency of Pairing, Meiotic Recombination and Synapsis .. 20
   1.3 Cancer and Germ Cells ........................................... 21
1.4 Microarray Meta-analysis* .............................................. 23
   Abstract ........................................................................ 23
   Introduction ............................................................ 23
   Microarray Technology and Data Analysis .............................. 23
   Meta-Analysis: The Analysis of Analyses ............................... 23
   Visualization of Complex Data ........................................ 23
   References ............................................................. 23
2 Aims and Objectives

3 Meta-analysis of Clinical Data Identifies a Novel Cohort of Highly
Cancer-specific Marker Genes*

4 CancerMA: a Web-based Tool for Automatic Meta-analysis of
Public Cancer Microarray Data*

5 CancerEST: a Web-based Tool for Automatic Meta-analysis of
Public EST Data*
## Contents

6 Meta-analysis of Germline Gene Expression* 123  
Abstract  
Introduction  
Materials and Methods  
Results  
Discussion  
References  
Supplemental Material

7 Overall Discussion 165

8 Conclusions 181  
8.1 Project Aim I 181  
8.2 Project Aim II 182

9 Further Works 183

Appendices 185

References 233
List of Tables

Please note that chapters with * are presented as papers/manuscripts and thus do not follow regular numbering. The numbering for these tables is omitted.

Chapter 1*

A selection of internet repositories and search interfaces for microarray data .
A selection of useful annotation tools for microarrays . . . . . . . . . . . . . . .
A selection of available meta-analysis tools . . . . . . . . . . . . . . . . . . . . . . .
List of Figures

Please note that chapters with * are presented as papers/manuscripts and thus do not follow regular numbering. The numbering for these figures is omitted.

Chapter 1*

1.1 The hallmarks of cancer proposed by Hanahan and Weinberg ................. 3
1.2 Scheme of spermatogenesis and expression of cancer testis (CT) genes in human germ cells ................................................................. 7
1.3 Epigenetic events regulating cancer testis (CT) gene expression ............ 9
1.4 Overview of the meiotic process during mammalian gametogenesis .... 13
1.5 Timing of gametogenesis/meiosis in mammalian males and females .... 14
1.6 Schematic composition of the seminiferous epithelium in the mammalian testes ................................................................. 15
1.7 Distinct pathways of meiotic recombination .................................... 17
1.8 Schematic composition of the synaptonemal complex (SC) ............. 19
1.9 Assembly of the synaptonemal complex (SC) .............................. 19
1.10 Temporal course of pairing, meiotic recombination and synaptonemal complex (SC) formation .................................................. 20
Example Forest plot ........................................................................ 20

Chapter 3*

Examples of gene expression and protein production profiles for meiCT genes.
Schematic flow diagram for the selection of candidate meiCT genes ........
Grid representation of gene expression profiles for the 33 meiCT genes identified
Circos plot showing the meta-change in gene expression in relation to corresponding cancer types for the 25 meiCT genes ..............
An example of a Forest plot for a meiCT gene, PRDM9 .......................
Circos plot showing the meta-change in gene expression in relation to corresponding cancer types for the 21 genes with a testis-only expression profile .................................................................
List of Figures

Chapter 4*
CancerMA workflow .................................................................
Circos plots showing the meta-change in gene expression in relation to
corresponding cancer types ..................................................
An example of a Forest plot showing the expression of gene CAV1
downregulated in lung cancer .................................................
Circos plot showing the meta-change in gene expression in relation to
corresponding cancer types ..................................................

Chapter 5*
CancerEST workflow ...............................................................
Circos plot showing the gene expression in relation to the corresponding
cancer types for the 39 testis-restricted genes .........................
An example of a bar chart showing the integrated expression profile of the
MAGEA1 gene ........................................................................

Chapter 6*
Circos plot showing the gene expression in relation to the corresponding
cancer types for the 43 human orthologues based on the EST meta-analysis
Circos plot showing the meta-change in gene expression in relation to
corresponding cancer types for the 41 human homologues based on the
microarray meta-analysis ..........................................................

Chapter 7
7.1 High-throughput expression profiling techniques in comparison with
high-sensitivity methods ......................................................... 167
7.2 Examples of applications and workflows for both pipelines .......... 169
7.3 Data selection process for CancerEST ................................. 170
7.4 Data selection process for CancerMA .................................. 171
7.5 Visualisations of the results generated by CancerMA ............... 176
## Appendices

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix A</td>
<td>185</td>
</tr>
<tr>
<td>Appendix B</td>
<td>189</td>
</tr>
<tr>
<td>Appendix C</td>
<td>193</td>
</tr>
<tr>
<td>Appendix D</td>
<td>197</td>
</tr>
<tr>
<td>Appendix E</td>
<td>209</td>
</tr>
<tr>
<td>Appendix F</td>
<td>227</td>
</tr>
<tr>
<td>Electronic Appendices</td>
<td>231</td>
</tr>
</tbody>
</table>
I dedicate this thesis to my mother, who died during the course of the thesis. She raised me, supported me, taught me and loved me. She has always encouraged and believed in me. Without her I would not have been able to reach the present position in life.

My thesis would not have been possible without the help and support of many people. I would like to express my gratitude for their assistance, support and encouragement.

First and foremost, I thank my supervisors Ramsay J. McFarlane and Lee D. Larcombe for their excellent supervision and their continuous advice throughout the course of this thesis. Their support, academic experience, encouragement and guidance were essential and invaluable for my work.

I am also indebted to my fellow Ph.D. students at Bangor University for the lab work validating my analyses. I would also like to acknowledge my colleagues at the Technical University of Graz for stimulating discussions, assistance and friendship. I would like to give my special thanks to Gerhard G. Thallinger for allowing me to work at the Technical University of Graz as a visiting scientist.

I thank the Welsh National Institute for Social Care and Health Research (NISCHR) and North West Cancer Research Fund (NWCRF) Institute for financially making this thesis possible.

I am grateful for the understanding, love and support of my family and friends. Most especially, I thank my partner Helge Lütgendorf for his love and support. His motivation and guidance have been invaluable to me.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>Axial element</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>BTB</td>
<td>Blood-testis-barrier</td>
</tr>
<tr>
<td>CE</td>
<td>Central element</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>CIBEX</td>
<td>Center for Information Biology Gene Expression Database</td>
</tr>
<tr>
<td>CNS</td>
<td>Central nervous system</td>
</tr>
<tr>
<td>CO</td>
<td>Crossover</td>
</tr>
<tr>
<td>CR</td>
<td>Chromatin remodelling co-repressor complex</td>
</tr>
<tr>
<td>CT antigen (or CTA)</td>
<td>Cancer testis antigen</td>
</tr>
<tr>
<td>CT gene</td>
<td>Cancer testis gene</td>
</tr>
<tr>
<td>CT-X (or X-CT) gene</td>
<td>Cancer testis gene encoded on the X chromosome</td>
</tr>
<tr>
<td>D loop</td>
<td>Displacement loop</td>
</tr>
<tr>
<td>DDD</td>
<td>Digital Differential Display</td>
</tr>
<tr>
<td>dHJ</td>
<td>Double Holliday junction</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNMT</td>
<td>DNA methyltransferase</td>
</tr>
<tr>
<td>DSB</td>
<td>Double-strand break</td>
</tr>
<tr>
<td>DSBR</td>
<td>Double-strand break repair</td>
</tr>
<tr>
<td>EBI</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed sequence tag</td>
</tr>
<tr>
<td>FC</td>
<td>Fold change</td>
</tr>
<tr>
<td>FDA</td>
<td>US Food and Drug Administration</td>
</tr>
<tr>
<td>G0 phase</td>
<td>Growth 0 phase</td>
</tr>
<tr>
<td>G1 phase</td>
<td>Growth 1 phase</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>GO</td>
<td>Gene ontology</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Lysine 4 of histone H3</td>
</tr>
<tr>
<td>HAT</td>
<td>Histone acetyltransferase</td>
</tr>
</tbody>
</table>
**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDAC</td>
<td>Histone deacetylase</td>
</tr>
<tr>
<td>HGNC</td>
<td>HUGO Gene Nomenclature Committee</td>
</tr>
<tr>
<td>HMT</td>
<td>Histone methyltransferase</td>
</tr>
<tr>
<td>HPrSM cell</td>
<td>Non-cancerous primary human prostate smooth muscle cell</td>
</tr>
<tr>
<td>iPSC</td>
<td>Induced pluripotent stem cell</td>
</tr>
<tr>
<td>LE</td>
<td>Lateral element</td>
</tr>
<tr>
<td>log2FC</td>
<td>Log 2-fold change</td>
</tr>
<tr>
<td>MAQC project</td>
<td>Microarray Quality Control project</td>
</tr>
<tr>
<td>MAS 5.0</td>
<td>Microarray Suite 5.0 from Affymetrix</td>
</tr>
<tr>
<td>MBP</td>
<td>Methyl-binding protein</td>
</tr>
<tr>
<td>mDEDS</td>
<td>Meta Differential Expression via Distance Synthesis</td>
</tr>
<tr>
<td>meiCT gene</td>
<td>Meiosis-specific cancer testis gene</td>
</tr>
<tr>
<td>MIAME</td>
<td>Minimum Information About a Microarray Experiment</td>
</tr>
<tr>
<td>MM sequence</td>
<td>Mismatching sequence</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MSCI</td>
<td>Meiotic sex chromosome inactivation</td>
</tr>
<tr>
<td>NCO</td>
<td>Non-crossover</td>
</tr>
<tr>
<td>NISCHR</td>
<td>National Institute of Social Care and Health Research</td>
</tr>
<tr>
<td>NLS</td>
<td>Nuclear transport signal</td>
</tr>
<tr>
<td>non-CT-X (or non-X-CT) gene</td>
<td>Cancer testis gene not encoded on the X chromosome</td>
</tr>
<tr>
<td>NuRD complex</td>
<td>Nucleosome remodelling and histone deacetylase complex</td>
</tr>
<tr>
<td>NWCRF</td>
<td>North West Cancer Research Fund</td>
</tr>
<tr>
<td>PAR</td>
<td>Pseudoautosomal region</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PGC</td>
<td>Primordial germ cell</td>
</tr>
<tr>
<td>PlantGDB</td>
<td>Plant Genome Database</td>
</tr>
<tr>
<td>PM sequence</td>
<td>Perfectly matching sequence</td>
</tr>
<tr>
<td>Poly(A) tail</td>
<td>Polyadenylation tail</td>
</tr>
<tr>
<td>PTC</td>
<td>Papillary thyroid carcinoma</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>Quantitative reverse transcription PCR</td>
</tr>
<tr>
<td>RA</td>
<td>Retinoic acid</td>
</tr>
<tr>
<td>Rb</td>
<td>Retinoblastoma</td>
</tr>
<tr>
<td>RMA</td>
<td>Robust multi-array average</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>Reverse transcription PCR</td>
</tr>
<tr>
<td>S phase</td>
<td>Synthesis phase</td>
</tr>
<tr>
<td>SAGE</td>
<td>Analysis of gene expression</td>
</tr>
<tr>
<td>SAM</td>
<td>Significance Analysis of Microarrays</td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>Synaptonemal complex</td>
</tr>
<tr>
<td>SDSA</td>
<td>Synthesis-dependent strand annealing</td>
</tr>
<tr>
<td>SEREX</td>
<td>Serological analysis of cDNA expression libraries</td>
</tr>
<tr>
<td>SMD</td>
<td>Stanford Microarray Database</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>ssDNA</td>
<td>Single-strand DNA</td>
</tr>
<tr>
<td>TA</td>
<td>Tumour antigen</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>tpm</td>
<td>Transcripts per million</td>
</tr>
<tr>
<td>TVF</td>
<td>Transverse filament</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>VSN</td>
<td>Variance Stabilization and Normalization</td>
</tr>
<tr>
<td>WCE</td>
<td>Whole cell extract</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Cancer and Cancer Testis Antigens

1.1.1 Cancer

Cancers are a group of genetic diseases, which result in aberrant cell proliferation, decreased cell death, tissue invasion and other malignant characteristics, and may affect any tissue of the body. It is the leading cause of death worldwide, accounting for 7.6 million deaths in 2008 [1]. More than 100 distinct cancer types are known, each being classified, in general, according to the type of cell that is presumed to be its origin. The main categories include (i) carcinoma (cancers arising from epithelia, responsible for 80% of the cancer deaths); (ii) sarcoma (cancers deriving from mesenchymal cells); (iii) leukaemia (cancers arising from haematopoietic tissue); (iv) lymphoma or myeloma (cancers originating from cells of the immune system); and (v) neuroectodermal tumours (cancers deriving from the central or peripheral nervous system) [2]. 90–95% of the cancer cases are associated with environment and lifestyle, whilst only the remaining 5-10% are due to inherited genetic defects. Lifestyle or environmental risk factors include smoking, sun exposure, alcohol consumption, diet, obesity, physical inactivity, environmental exposures (physical carcinogens such as radiation, or chemicals such as alkylating agents) as well as certain viral and bacterial infections [3].

1.1.2 Carcinogenesis

Carcinogenesis or tumorigenesis is a multistep process, whereby genetic and epigenetic alterations are accumulated, each providing a growth advantage and thus transforming normal cells into cancer cells in a stepwise fashion [4, 5]. Multiple alterations in three types of genes are responsible for this transformation: (i) oncogenes; (ii) tumour suppressor genes; and (iii) genome stability genes [6]. Mutational activation of oncogenes and mutational inactivation of tumour suppressor genes enhance cell growth by stimulating the cell cycle, by impairing apoptosis, or by enhancing nutrient supply via angiogenesis [7, 8]; for example, the products of the RAS oncogenes regulate cell proliferation in response to growth factors and often fail to be inactivated in various cancer types due
to mutations [9, 10], whereas aberrant inactivation of the TP53 tumour suppressor gene leads to failure of growth inhibition and apoptosis [11, 12]. In contrast, genome stability genes such as genes involved in DNA repair, recombination or chromosomal segregation keep the rate of genetic change to a minimum and thus when mutated can increase the chance of alterations in other genes [13].

Mutations in a vast number of genes have been reported to be associated with cancer, yet disruption of only a few key pathways gives rise to the characteristics of cancer [6]. Genetic alterations in different genes often entail a similar or common phenotype, where these genes are related as part of the same pathway. The TP53 pathway, for example, is often disrupted in cancer either due to point mutations in the TP53 gene or due to one of the numerous alternative gene mutations that may lead to disruption of this pathway at key points [11, 12]. Hanahan and Weinberg go one step further by reducing the complexity of cancer and carcinogenesis to a number of underlying principles, the so-called hallmarks of cancer [14, 15]. They suggest that cells have to acquire certain physiological characteristics during tumour development to become malignant: (i) sustaining proliferative signalling; (ii) evading growth suppressors; (iii) sustained angiogenesis; (iv) limitless replicative potential; (v) tissue invasion and metastasis; (vi) evading cell death; (vii) reprogramming of energy metabolism (emerging hallmark); and (viii) evading immune destruction (emerging hallmark). Additionally, two enabling characteristics drive tumour progression: (i) genome instability; and (ii) inflammation (Figure 1.1).

In addition to the current dogma that the development of human tumours occurs in a stepwise fashion, the concept of chromothripsis has recently been postulated, whereby a single catastrophic event leads to hundreds of genomic rearrangements promoting cancer development [16].

Over the last decade, researches have increasingly demonstrated the complexity of tumorous tissue. Tumours are composed of numerous distinct cell types within a tumour microenvironment [17]. Furthermore, a model has been proposed, where a small subpopulation of cells, the cancer stem cells are driving tumour growth. These cells are capable of self-renewal and differentiation as well as to seed new tumours due to broken regulatory mechanisms of normal stem cell developmental pathways [18–21]. Cancer stem cells may arise due to genetic and/or epigenetic defects in normal stem cells or in their progenitors. Another possibility is that cancer cells may acquire a stem cell-like character [22].
1 Introduction

1.1 Cancer and Cancer Testis Antigens

Figure 1.1: The hallmarks of cancer proposed by Hanahan and Weinberg [14, 15]. (A) The acquired capabilities of cancer cells: (i) sustaining proliferative signalling; (ii) evading growth suppressors; (iii) sustained angiogenesis; (iv) limitless replicative potential; (v) tissue invasion and metastasis; and (vi) evading cell death.

(B) The two hallmarks emerging from recent cancer research (reprogramming of energy metabolism and evading immune destruction) as well as the two characteristics (genome instability and inflammation) enabling and fostering multiple hallmark characteristics.

1.1.3 Tumour Antigens

As cancer cells may display tumour-associated proteins, they can evoke an immune response and thus be selectively removed by the immune system. Nevertheless, the immune system of cancer patients mostly fails to fight tumours effectively and therefore the development of immunotherapies focuses on enhancing and directing its response [23–27]. The potential of tumour antigens (TAs) in cancer immunotherapy has led to the identification of a number of molecules, which are predominantly produced in cancer cells, with MAGEA1 being the first one reported to evoke a T cell response [28]. TAs have been classified into a shared and a unique group. Unique TAs arise from random mutations and are unique to the tumour of an individual patient, whereas shared TAs are expressed in many independent tumours [29, 30]. Three main subclasses can be iden-
1.1 Cancer and Cancer Testis Antigens

Introduction

Cancer testis (CT) antigens are defined among shared TAs: (i) cancer testis (CT) antigens; (ii) differentiation antigens; and (iii) overexpressed antigens [30]. CT genes are solely expressed in normal testes and aberrantly expressed in a range of cancer types (e.g., MAGEA1) [31, 32]. Genes encoding differentiation antigens are expressed in tumours and in the corresponding normal tissue of origin, mainly in melanomas/melanocytes (e.g., MART1) [30, 33]. In contrast, genes encoding overexpressed TAs are weakly expressed in normal tissues, but are overexpressed in various tumours (e.g., ERBB2) [30, 34].

Conventional cancer therapies such as surgery, radiotherapy and chemotherapy have limitations as they fail to completely target all cancer cells and are often associated with severe side effects. Immunotherapy offers an alternative, as it specifically targets TAs in cancer cells and spares normal cells [23-27]. There are two distinct immunotherapeutic strategies: (i) active immunotherapy with vaccines; and (ii) passive immunotherapy using monoclonal antibodies or adoptive cell therapy.

Passive immunotherapy based on the administration of therapeutic antibodies has recently been considerably successful. The therapeutic antibodies bind to tumour cells and thus stimulate their destruction either through activating cytotoxic effects and phagocytosis (immune-mediated destruction), through antagonising oncogenic pathways to cause reduced proliferation and/or apoptosis, or through delivering conjugated drugs [34, 35]. Twelve antibodies have been developed to date, targeting cancer-associated proteins to treat various solid tumours or haematological malignancies, including Trastuzumab/Herceptin (targeting ERBB2-positive breast cancer), Rituximab/Mabthera (specific to CD20 in Non-Hodgkin lymphoma and chronic lymphocytic leukaemia) and Alemtuzumab/Campath (targeting CD52 in B and T cell lymphomas) [34–37].

Adoptive cell therapy relies on the extraction of the patient’s lymphocytes, which are cultured and enriched *ex vivo* to be specific to either multiple undefined TAs or to a single defined TA, and are subsequently transfused back to the patient [35, 38]. Recent clinical trials for adoptive cell therapy have shown encouraging results, in particular in melanoma patients [39].

Active immunotherapy approaches, in contrast, prime the immune system to attack the cancer cells and use either whole tumour cell vaccines, which expose the immune system to many putative TAs but could cause autoimmunity, or defined TA vaccines, which provoke a very specific immune response to a single epitope limiting the danger of autoimmunity. Various strategies are employed in developing defined TA vaccines such as dendritic cell-based approaches, particle-based vaccines and vaccines on the basis of
1 Introduction

1.1 Cancer and Cancer Testis Antigens

DNA, RNA or viral-vectors [23, 24, 29, 30]. So far only one cancer vaccine, Sipuleucel-T [40] has been approved by the US Food and Drug Administration (FDA), but several vaccines are currently under investigation in clinical trials.

Until recently, cancer immunotherapy has had limited success, as cancer cells acquire various distinct strategies to escape being targeted by the immune system, including suppression of dendritic cells, inhibiting T cell responses by suppressing their activation or by impairing their penetration into the tumour as well as recruitment of regulatory T cells, which in turn inactivate cytotoxic T lymphocytes [23, 26, 35, 41]. However, understanding how immune tolerance and suppression affect anti-tumour immune responses as well as the development of novel immunotherapeutics and combined chemotherapies have led to advances in this field [41].

Many tumour antigens are also used as cancer markers to detect, diagnose or classify a malignancy as well as to predict or control responses to treatment, to monitor patients with diagnosed malignancies, or to assess their prognosis [42]. More than 20 cancer markers are commonly in use to date; for example, assessing the expression of ERBB2 is used to predict the responses to treatment with Trastuzumab/Herceptin or Tamoxifen in patients with breast cancer [43, 44].

1.1.4 Cancer Testis Antigens

Cancer testis (CT) antigens are a class of proteins whose genes show expression restricted to testicular cells, but are also aberrantly expressed in a wide range of cancer types [31, 32, 45–48]. The immunological privilege of the testis [49, 50] and the fact that they can induce an immunological response in the body [31] makes the CT antigens to promising candidates for immune targeting. A number of CT antigens are currently under investigation for their potential as cancer immunotherapeutics such as MAGEA1 and CTAG1A/CTAG1B/NY-ESO-1 [39, 51].

The first CT antigen, MAGEA1 was identified through autologous typing by van der Bruggen et al. [28], followed by the discovery of GAGE [52] and BAGE [53]. Since then, various approaches have been employed to identify novel CT antigens including serological analysis of cDNA expression libraries (SEREX) [54, 55] as well as gene expression techniques comparing normal and cancerous tissues [56–59]. Research to date has led to the identification of over 200 genes belonging to over 70 gene families, which have been reported to exhibit CT gene characteristics. Information on these genes has been gathered in CTdatabase, a publicly accessible database [60].
In addition to expression in the testis, some CT genes have recently been found to be expressed in tissues of the central nervous system (CNS), which also represent immunologically privileged areas. These are referred to as CT/CNS antigens [57–59]. However, a number of CT genes were subsequently determined to show even a broader expression in somatic tissues than first assumed. This resulted in the classification of the CT genes in testis- or testis/CNS-restricted and testis- or testis/CNS-selective [58]. This re-assessment of their classification clarifies and significantly reduces the number of currently known, bona fide restricted CT genes (expression tightly restricted to cancer and testicular cells).

CT antigens have been further divided into those encoded on the X-chromosome (CT-X antigens) and those that are not (non-CT-X antigens) [31]. CT-X antigens tend to belong to multigene families and reside within genomic repeats [31, 46, 61]. It has been estimated that up to 10% of the X chromosome consists of CT-X genes, with the Xq24-q28 region harbouring the highest density of them [62]. Non-CT-X antigens, in contrast, are usually encoded by single copy genes and show a random distribution throughout the human genome [31]. The two classes of CT genes show distinct expression in spermatogenesis. CT-X genes such as MAGEA1 are mainly expressed in spermatogonia, whereas non-CT-X genes such as SYCP1/SCP1 are commonly expressed in meiotic cells such as spermatocytes and spermatids and are often active or involved in meiotic functions (Figure 1.2) [31, 47].

In general, tumours appear to be either CT-rich or poor [58] and particularly CT-X antigens often show co-expression in cancer [63, 64]. High expression of CT genes is observed in melanoma, non-small-cell lung cancer, hepatocellular carcinoma, bladder cancer and ovarian cancer. Breast cancer and prostate cancers, in contrast, show moderate CT gene expression, whereas renal cancer, colon cancer, leukaemia and lymphoma exhibit low frequency of CT gene expression [45, 58]. High grade and late stage cancers have been correlated with higher CT gene expression [46].

The function of most CT antigens in both germline and tumours still remains poorly understood, but clues have emerged that at least some CT antigens may contribute to tumorigenesis [32, 47]. A number of CT antigens are associated with transcriptional regulation and/or could affect cellular processes such as signalling, translation and chromosome recombination. CT antigens can be grouped according to their putative function [47, 65]: (i) transcriptional regulators such as MAGEA1 [28], SSX2 [66], BRDT [67] and BORIS [68]; (ii) signal transduction involvement such as MAGEA1 [69];
1 Introduction

1.1 Cancer and Cancer Testis Antigens

Figure 1.2: Scheme of spermatogenesis and expression of cancer testis (CT) genes in human germ cells [31]. Male primordial germ cells (PGCs) migrate to the genital ridge, proliferate and arrest until puberty. At puberty, spermatogenesis is initiated, which continuously maintains spermatogonia through mitosis as well as produces mature sperm cells through meiosis. Spermatogonia, the male germline stem cells are diploid cells and proliferate throughout the adult life. Some of these cells, however, undergo meiosis. After DNA replication, primary spermatocytes produce secondary spermatocytes after the first meiotic division, which in turn divide to spermatids in the second meiotic division. Spermatids subsequently evolve to spermatozoa. CT genes encoded on the X chromosome (CT-X genes) are mainly expressed in mitotic spermatogonia, whereas those that are autosomal encoded (non-CT-X genes) are commonly expressed in meiotic spermatocytes and spermatids [31, 47].

(iii) structural components of spermatozoa such as TSGA10 [70] and AKAP4 [71]; (iv) role in cell-to-cell adhesion and/or cell migration such as SP17 [72] and ADAM2 [73]; (v) apoptosis inhibitors such PIWIL2 [74] and GAGE7C [75]; (vi) helicase-like or other enzymatic functions such as DDX53/CAGE [76] and TSP50 [77]; (vii) involvement in spermatogonial mitotic self-renewal such as PIWIL2 [74]; and (viii) role in spermatogonial meiosis such as SYCP1 [78] and SPO11 [79]. The functional roles of non-CT-X
1.1 Cancer and Cancer Testis Antigens 1 Introduction

Antigens are generally better known, as most of them have been conserved during evolution [46]. They have known functions in spermatogenesis, meiosis and fertilisation [31, 46, 47] such as SYCP1 [78], SPO11 [79], BORIS [68], HORMAD1 [80] and ACRBP [81].

Mainly epigenetic events, in particular methylation processes, appear to regulate CT gene expression in both normal and cancer cells (Figure 1.3). CT genes are silenced in somatic tissues, whereas their expression is activated in testicular cells and malignancies. This correlates with the findings that the DNA in their promoter regions is methylated in somatic tissues and unmethylated in germ cells [82]. It also has been shown that CT gene expression can be induced by demethylating agents in cancer cell lines [82–84]. The expression of various CT genes in tumours or cell lines was associated with hypomethylation of their promoters, and various CT gene promoter experiments using reporter genes further showed that such hypomethylation is the primary mechanism of CT gene regulation [85–87]. Histone modifications also influence CT gene expression, suggesting an accessory role to DNA methylation [32].

Simpson et al. suggested that a dysfunctional control of germline genes (e.g., the aberrant activation of a genetic master switch) could initiate a silenced gametogenic programme in cancer, activating CT genes in cancer cells that are usually expressed at various stages of gametogenesis [31]. The recently discovered CT antigen and CTCF paralogue, BORIS possibly functions as an epigenetic regulator. Its upregulation correlates with the depletion of methylation during germ cell development as well as with the downregulation of CTCF, which has antagonistic features [68, 88]. BORIS might mediate activation of at least some CT genes [89, 90]. However, BORIS was shown to be insufficient for DNA hypomethylation and CT gene activation in ovarian cell lines [91]. Furthermore, heterogeneity in CT gene expression as well as rare CT gene expression in colorectal cancer, where hypomethylation is common [92], point to an additional transcriptional regulation [46].
1 Introduction

1.1 Cancer and Cancer Testis Antigens

Figure 1.3: Epigenetic events regulating cancer testis (CT) gene expression [32]. Hypermethylation of CT gene promoters causes gene silencing in normal somatic cells by means of two mechanisms. First, methylation represents a physical barrier to transcription factors (TFs). Methylation predominately occurs on cytosines at CpG sites (red circles), catalysed by DNA methyltransferases (DNMTs). Second, complexes containing methyl-binding proteins (MBPs) bind to CpG sites. They prevent transcription either by impairing access to TFs or by recruiting chromatin remodelling co-repressor complexes (CRs) responsible for histone modifications. These CRs in turn contain histone methyltransferases (HMTs) and histone deacetylases (HDACs), which cause chromatin condensation, making the DNA inaccessible to TFs. Transcriptionally active CT genes, in contrast, are demethylated (green circles), preventing binding of MBPs and CRs. These demethylated promoter regions can be occupied by TFs and histone acetyltransferases (HATs), leading to transcription [32, 93].
1.2 Meiosis – Halving the Chromosome Content

1.2.1 The Principle of Meiosis and Its Temporal Course

A diploid cell contains two homologous copies of each autosome and a pair of sex chromosomes originating from the parental haploid gametes. During the process of mitosis a somatic diploid cell divides to generate two identical diploid daughter cells. Meiosis, in contrast, yields four genetically distinct haploid cells (gametes – egg and sperm cells in higher eukaryotes) and involves two rounds of cell division (Figure 1.4) [94–96]. The first division is a reductional division that requires the establishment of sister centromere monopolarity. Before a cell undergoes meiotic chromosome segregation, however, meiotic DNA replication occurs generating sister chromatid pairs, which are bound together via cohesin complexes. This is followed by a number of meiosis-specific processes [94–97]. First, homologue alignment occurs, whereby homologous chromosomes associate to homologous pairs (pairing), and a protein structure, the synaptonemal complex (SC) is formed, running between paired homologues (synapsis) [98–101]. Second, homologue alignment is accompanied by meiotic recombination, which leads to the formation of bivalent structures by establishing so-called chiasmata or crossovers (COs) [102–104]. Third, segregation of homologues into two daughter cells takes place [105, 106]. In the second meiotic division, the chromosomes segregate in a mitotic-like fashion without a DNA replication step but with a reversion to bipolarity of sister centromeres, resulting in four haploid cells [94, 96, 97]. The resulting gametes are genetically distinct from one another due to chromosome shuffling and recombination events between homologues during the first meiotic division [94, 95, 97, 107].

The distinct meiotic phases of the first and second meiotic division are listed and described in detail below [96, 106]:

**Mammalian meiosis I: The first meiotic division**

- Interphase I
  - Growth 0 (G0) phase: Resting phase
  - Growth 1 (G1) phase: Synthesis of enzymes and structural proteins needed for growth.
  - Meiotic synthesis (S) phase: DNA replication occurs forming sister chromatids.

*Julia Feichtinger*
1.2 Meiosis

- Prophases I: Pairing, synapsis and meiotic recombination takes place.
  - Leptotene: Homologues condense and pairing is initiated. Recombination is also initiated.
  - Zygotene: Pairing continues. SC assembly begins.
  - Pachytene: Pairing and SC assembly are completed.
  - Diplotene/Diakinesis: SC is disassembled and recombination is completed. Chromosomes are further condensed.

- Metaphase I: Bivalents align along an equatorial plate with monopolar sister centromere configurations.

- Anaphase I: Bivalents separate, resulting in a haploid set. They are linked by spindle fibers to opposite poles, moving towards them. The cell elongates and the cleavage furrow forms.

- Telophase I: Nuclei form in the daughter cells and the chromosomes unwind.

Mammalian meiosis II: The second meiotic division

- Interphase II: No DNA replication occurs.

- Prophase II: Disappearance of the nuclear membrane and chromosome condensation. The spindle fibers are arranged for the second division.

- Metaphase II: Sister chromatids align along an equatorial plate.

- Anaphase II: Sister chromatids separate. They are linked by spindle fibers to opposite poles, moving towards them. The cell elongates and the cleavage furrow forms.

- Telophase II: Nuclei form in the daughter cells and the chromosomes unwind.

1.2.2 Mammalian Gametogenesis

Meiotic divisions are an essential aspect of gametogenesis (Figure 1.4), which describes the complete process of producing gametes (either mature eggs or sperm cells), whereby cells undergo mitotic and meiotic cell divisions as well as different stages of differentiation.

There are crucial differences in regulation and timing of gametogenesis as well as in characteristics of its associated cells in distinct species and in distinct sexes of the same
1 Introduction

1.2 Meiosis

Figure 1.4: Overview of the meiotic process during mammalian gametogenesis [96]. The distinct meiotic phases of the first and second meiotic division are schematically illustrated (Interphase, Prophase, Metaphase, Anaphase and Telophase). Meiosis includes two rounds of cell division resulting in four genetically distinct haploid cells (egg or sperm cells). The first meiotic division, meiosis I employs a number of meiosis-specific processes: (i) homologue pairing and synapsis; (ii) meiotic recombination; and (iii) chromosome segregation. This reductional division halves the chromosome content. In the second meiotic division, meiosis II, the cells divide in a mitotic-like fashion without a DNA replication step [94, 96, 97].

species. The male form of gametogenesis is called spermatogenesis, whereas the female one is described as oogenesis (Figure 1.5) [108–110].

In vertebrates, primordial germ cells (PGCs) migrate to the developing ovaries or testes, where they become either oogonia or spermatogonia [108, 111]. In females, oogonial cells undergo several mitotic divisions and some of these cells, now called primary oocytes enter meiosis I but arrest after diplotene (Figure 1.5). During this arrest phase the cells grow and become surrounded by somatic cells, forming follicles. The completion of meiosis I occurs at puberty, whereby small groups of primary oocytes periodically resume meiosis I and mature to secondary oocytes. At ovulation, the follicle is ruptured and the egg is released. In most mammalian females, the secondary oocytes arrest at
meiosis II, which is only completed upon fertilisation. Meiosis I and II can only result in one mature egg, as the rest of the cells are produced as polar bodies, which eventually degenerate [109, 112–115].

In males, spermatogenesis occurs in the epithelium of the seminiferous tubules of the testes, where the spermatogonial cells transform into spermatozoa, the mature sperm cells [116–119]. In contrast to oogonial cells in oogenesis, meiosis in spermatogonial cells is not initiated until puberty (Figure 1.5). At puberty, however, spermatogonial cells proliferate to maintain themselves as well as continuously enter meiosis. After DNA replication, now so-called primary spermatocytes produce secondary spermatocytes after the first meiotic division, which in turn divide to spermatids in the second meiotic division, resulting in four gametes. The spherical spermatids subsequently evolve to condensed, elongate mature spermatozoa, which are released to the lumen of the seminiferous tubules [117, 120, 121].

![Figure 1.5: Timing of gametogenesis/meiosis in mammalian males and females [109]. There are crucial differences in regulation and timing of gametogenesis in distinct sexes. In males, meiosis is initiated at puberty and occurs continuously, whereas in females, meiosis begins already in the foetal ovaries but then arrests until puberty. At puberty, some oocytes periodically complete meiosis to produce mature egg cells.](image)

The seminiferous epithelium consists of germ cells at various stages of differentiation as well as of Sertoli cells (Figure 1.6). The Sertoli cells are large cells reaching from the basal lamina to the lumen of the seminiferous tubules and support the germ cells throughout their development, similar to the follicle cells during oogenesis [116, 118, 119]. Tight junctions between Sertoli cells form the blood-testis-barrier (BTB) and separate the seminiferous epithelium into a basal compartment for spermatogonia as well as into an adluminal compartment for the differentiating spermatocytes (Figure 1.6) [122]. This seals off the adluminal compartment from the rest of the body, resulting in an immunologically privileged site for meiotic and postmeiotic cells [49, 50, 122–124].
Successful meiosis is dependent on pairing and synapsis of homologous chromosomes. In male meiosis, however, pairing and synapsis of the X and Y chromosome can only occur at a small pseudoautosomal region (PAR), which ensures proper segregation of the sex chromosomes into distinct daughter cells [125]. The unsynapsed regions undergo chromatin remodelling, leading to transcriptional silencing of the X and Y chromosome, which is a process called meiotic sex chromosome inactivation (MSCI) [126, 127].

1.2.3 Meiotic Entry

Cellular differentiation crucially depends on the establishment of the specific expression needed at a given moment in a given tissue. However, very little is known about the regulation of meiotic expression, although thousands of genes are differentially expressed between testicular Sertoli and germ cells [128, 129]. A few epigenetic factors have been associated with regulating the meiotic expression programme such as PRDM9 [130, 131], BRDT [132], BORIS [68, 88] and TET1 [133]. Furthermore, recent findings show that meiotic entry is initiated by retinoic acid (RA) coupled with STRA8 gene expression. In foetal ovaries, RA and STRA8 induce meiotic entry, whereas the initiation of meiosis in males is thought to be regulated by stage-specific expression of the CYP26B1 gene, which encodes a RA-degrading enzyme and is present in the Sertoli cells of the testes, possibly impairing meiotic entry until puberty [96, 134]. In general, meiosis-specific genes are tightly regulated, which is most likely due to a deleterious impact of their associated proteins for mitotic growth of somatic cells [129].
1.2 Meiosis

1.2.4 Meiotic Recombination

Meiotic recombination involves the creation of double-strand breaks (DSBs) catalysed by SPO11 [79] and their subsequent repair, creating crossovers (COs) or non-crossovers (NCOs) [102–104]. Meiotic COs are responsible for the exchange of alleles to ensure genetic diversity and facilitate proper chromosome segregation, as they support homologue alignment and accurate attachment to the spindle [102–104, 107].

Meiotic recombination is initiated by the formation of DSBs upon the action of SPO11, followed by the creation of 3′ single-strand DNA (ssDNA) overhangs [79, 135]. These overhangs provide the substrates for strand exchange factors (RAD51/DMC1) to search for complementary sequences in homologous chromosomes, leading to single-end strand invasions [136] to create displacement (D) loop recombination intermediates (Figure 1.7). DSBs catalysed by SPO11 (or an SPO11 orthologue) appear to be conserved in all sexually reproducing species, whereas the subsequent DSB repair may vary in distinct organisms and research in this area is largely based on model organisms. The canonical model is the double-strand break repair (DSBR) pathway, whereby the second DSB end is captured, leading to the formation of a double Holliday junction (dHJ), which can be resolved, yielding either COs or NCOs depending on the orientation of the enzymatic cleavage of the dHJ (Figure 1.7) [137, 138]. A few endonucleases capable of cleaving these dHJs have been identified so far (MUS81-EME1, SIX1-BTBD12 and GEN1) [139, 140]. However, recent findings suggest that the decision to repair the DSBs to COs or NCOs is made before dHJ resolution and that the DSBR pathway mainly gives rise to COs [104, 141]. Studies in Saccharomyces cerevisiae showed that the formation of dHJs and their resolution to COs through the DSBR pathway is dependent on ZMM proteins [142]. This ZMM-dependent pathway is interference-dependent, reducing the probability of further COs nearby to ensure that COs are evenly distributed [141, 142]. A second pathway yielding COs has been identified in Schizosaccharomyces pombe. Here, D loop intermediates may be resolved to COs upon the action of the MUS81-EME1 complex, ensuring the formation of COs (Figure 1.7) [143]. The creation of such COs, in contrast, is interference-independent [142]. NCOs are formed in an alternative pathway, the so-called synthesis-dependent strand annealing (SDSA) pathway, whereby after strand invasion and DNA synthesis, the D loop dissociates, the strand pulls out and the synthesised stretch anneals with the DSB end of the original strand to yield NCOs (Figure 1.7). This unwinding of the D loop intermediate may occur through the activity of the helicases such as BLM [144] or RTEL [145]. Alternatively, these helicases can dissolve dHJs through the double-junction dissolution pathway, also leading to NCOs (Figure 1.7) [146].
Figure 1.7: Distinct pathways of meiotic recombination [102]. All pathways are initiated by double-strand breaks (DSBs), which are resected creating 3’ overhangs. Subsequent strand invasion leads to displacement (D) loop recombination intermediates. Here the pathways become distinct. The double-strand break repair (DSBR) pathway yields crossovers (COs) and non-crossovers (NCOs) through the formation and subsequent cleavage of double Holiday junctions (dHJs). ZMM proteins may be involved in this pathway, directing the resolution of dHJs to produce COs. dHJs may also be dissolved upon the action of helicases (double-junction dissolution pathway), leading to NCOs. The main pathway producing NCOs, however, is the synthesis-dependent strand annealing (SDSA) pathway, whereby the D loop intermediate dissociates and the invading strand anneals with the DSB end of the original strand. Alternatively, the D loop intermediate may be converted to COs upon action of the MUS81-EME1 complex [102–104, 142].
1.2 Meiosis

In general, DSB initiation sites are non-randomly distributed but concentrated at specific sites in the genome, so-called recombination hotspots [147, 148], which tend to be close to genes but are localised outside of transcribed regions. DSB initiation sites appear to be both genetically and epigenetically marked. In particular, trimethylation of lysine 4 of histone H3 (H3K4me3) as well as degenerate 13-mer sequence motifs have been associated with DSB initiation sites. The gene PRDM9 is specifically expressed in meiotic germ cells and encodes a protein, which binds to these degenerate motifs as well as possesses H3K4me3 activity. Thus, it is thought to play an important role in marking recombination hotspots and regulating meiotic recombination [130, 131, 149].

1.2.5 The Synaptonemal Complex

The synaptonemal complex (SC) is a protein structure [150], which is formed during meiosis, connecting the two homologous chromosomes [98–101]. The function of the SC remains poorly understood. It has been proposed that it functions in some aspect of CO control, although it is dispensable in some organisms [99, 100]. The mature SC is a zipper-like structure and consists mainly of meiosis-specific proteins [101]. In order for the SC to assemble, homologous chromosomes align and axial elements (AEs) form along each of the homologues, which mature into lateral elements (LEs) when the homologues are fully synapsed. Homologous LEs are linked together by the formation of a central region, which consists of transverse filaments (TVFs) and a central element (CE), resulting in the mature SC (Figure 1.8 and 1.9) [98, 99, 101].

Various functions have been associated with AEs/LEs such as chromosome condensation, pairing, regulating DSB repair, CO formation and CO interference as well as assembly of the mature SC by organising the formation of TVFs [99]. The main structural components of AEs/LEs are SYCP2 [151] and SYCP3 [152] (Figure 1.9). Cohesin proteins including STAG3, SMC1α, SMC1β, SMC3 and REC8 are thought to mediate the localisation of proteins for correct assembly and formation of the AEs/LEs [99, 153]. SYCP2 and SYCP3 also appear to support cohesin core integrity and chromosome condensation [154]. Further components of AEs are the HORMA-domain proteins, HORMAD1 and HORMAD2 [80], which are localised to the unsynapsed chromosome axis. They are thought to direct DSB repairs preferentially from homologues and not from sister chromatids, which is required for homology search and the formation of COs [80]. TVFs serve as a bridge between the LEs and the CE. They consist of SYCP1 [78] molecules assembled to parallel coiled-coil homodimers, forming the zipper-like structure of the SC (Figure 1.9). The N-termini of SYCP1 molecules overlap head-to-head in
Figure 1.8: Schematic composition of the synaptonemal complex (SC) [99]. The SC is a zipper-like structure consisting of two lateral elements (LEs), which are linked by transverse elements (TVFs) and a central element (CE). It forms and runs between paired homologues and is thought to mediate and maintain pairing as well as to facilitate recombination events at least in some organisms [99, 100].

Figure 1.9: Assembly of the synaptonemal complex (SC) [98]. Axial elements (AEs)/lateral elements (LEs) consist of SYCP2, SYCP3, HORMA-domain proteins as well as of cohesin proteins including STAG3, SMC1α, SMC1β, SMC3 and REC8. Transverse elements (TVFs) represent a bridge between LEs and the central region, and consist of SYCP1 homodimers. The central element (CE) is composed of SYCE1, SYCE2, SYCE3 and TEX12 [98].

the CE, whereas their C-termini are associated with the LEs [155]. Moreover, SYCP1 may also support the maturation of COs before acting as a building block for the SC [156], as well as recruits proteins forming the CE [157]. The CE is composed of SYCE1 [157], SYCE2 [157], SYCE3 [158] and TEX12 [159] (Figure 1.9). CE proteins have a function in the formation and stabilisation of the SC and might also be involved in CO events [158, 160, 161].
1.2.6 Dependency of Pairing, Meiotic Recombination and Synapsis

Most organisms show a dependency between the processes of pairing, recombination and synapsis. These processes exhibit a tight temporal correlation (Figure 1.10) and perturbation of one process can lead to failure of the others [98]. In particular, recombination initiation is essential for stable pairing [162] and for synapsis [163]. Complete SC assembly, however, is necessary for completion of recombination and CO maturation but not for recombination initiation [98, 99, 156, 158]. CO formation in turn is required for proper chromosome segregation [98, 99, 142]. However, the level of dependency between pairing, recombination and synapsis appears to differ in distinct species, and some organisms even pose extremes such *Schizosaccharomyces pombe*, which does not establish a SC [164].

Figure 1.10: Temporal course of pairing, meiotic recombination and synaptonemal complex (SC) formation [165]. These processes exhibit a tight temporal correlation and occur during Prophase I. Prophase I consists of Leptotene, Zygotene, Pachytene, Diplotene and Diakinesis. During Leptotene chromosomes condense and pairing is initiated, axial elements (AEs) form and double-strand breaks (DSBs) appear. In the next phase, Zygotene, assembly of the synaptonemal complex (SC) begins, pairing continues and DSB repair initiates. In Pachytene, pairing and SC assembly are completed and double Holiday junctions (dHJs) appear. During Diplotene the SC is disassembled and meiotic recombination is completed. In the last phase of Prophase I, Diakinesis, chromosomes are further condensed [94–96, 106].
1.3 Cancer and Germ Cells

Cancer and germ cells exhibit profound commonalities such as rapid proliferation, undifferentiated phenotype, migratory behaviour and immortality/lack of senescence [166, 167]. The frequent expression of CT genes in cancer cells could reflect the aberrant induction of a silenced gametogenic programme [31, 166], which in turn could lead to a soma-to-germline transformation, driving tumorigenesis. According to this hypothesis, the distinct CT expression profiles observed in cancer may represent the normal CT expression profiles during the different stages of gametogenesis [166]. In support of this view, ectopic expression of germline genes has been reported in brain tumours of *Drosophila melanogaster* animals caused by mutation of the *l(3)mbt* gene [168] as well as in *Caenorhabditis elegans* strains with similar mutations [169, 170]. The aberrant activation of CT genes could be induced through a dysfunctional control of germline-specific genes, for example, caused by a defective master switch [31]. As both cancer and germ cells share genome-wide demethylation, it is plausible that genes controlling demethylation are defective [31, 166].

The ectopic activation of germline genes could have a fundamental role in tumorigenesis, as their associated proteins could contribute to the acquisition of tumorous characteristics [47]. Most germline genes, in particular meiotic genes, are tightly regulated and thus mostly restricted to germ cells. Derepression of these genes may have severe consequences. CT antigens function in a number of processes including transcriptional regulation, meiotic spermatogenesis, spermatogonial mitotic self-renewal, signal transduction, germ cell apoptosis, and cell adhesion/migration [47, 65]. Acquiring self-renewal, evading apoptosis and sustaining proliferative signalling are all hallmarks of cancer [14]. Dysfunction of cell adhesion/migration contributes to activating invasion and metastasis [171], which in turn is also a hallmark of cancer [14]. These malignant characteristics could be hijacked through the expression of CT genes encoding proteins that function in these pathways. The expression of meiotic genes, in particular those producing proteins with chromosome modulating potential, in somatic cells could lead to perturbation of the mitotic process and thus could result in inappropriate recombination events, leading to oncogenic changes such as translocations, aberrant chromosome segregation and aneuploidy [31, 129, 172], which in turn drive genomic instability, again a hallmark of cancer [14]. Kalejs et al., for example, reported the upregulation of meiosis-specific genes in tumour cells, which appears to be associated with arrested mitosis and polyploidy [173]. Moreover, the ectopic expression of a few testis-specific factors, whose associated proteins act as epigenetic and transcriptional regulators, could further drive soma-to-germline transformations and tumorigenesis [172].
1.4 Microarray Meta-analysis: From Data to Expression to Biological Relationships

The purpose of this section is to describe the process of meta-analysing microarrays and to discuss objectives, data collection/resources, annotation, analysis methods including a section of tools, as well as suggested data visualisations.

Please note that this chapter is presented as author-created version of a book chapter in Computational Medicine, Tools and Challenges edited by Zlatko Trajanoski and published by Springer Vienna (available at: \url{http://link.springer.com/chapter/10.1007/978-3-7091-0947-2_4} (chapter), \url{http://www.springer.com/biomed/molecular/book/978-3-7091-0946-5} (complete book)) [174]. The content structure, layout, language and reference style follow the specifications of Springer Vienna.
Microarray Meta-Analysis: From Data to Expression to Biological Relationships

Julia Feichtinger*,†, Gerhard G. Thallinger‡, Ramsay J. McFarlane§, and Lee D. Larcombe¶

1North West Cancer Research Fund Institute, Bangor University, Bangor, Gwynedd LL57 2UW, UK
2Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
3Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology (ACIB GmbH), Petersgasse 14, 8010 Graz, Austria
4Cranfield Health, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK

Since the introduction of microarray technology, it has become the workhorse for mRNA expression profiling. Its application ranges from investigating gene function, regulation, and co-expression, to clinical use in diagnosis and prognosis. Over the last decade, a large number of microarray experiments have become available in public repositories often addressing similar or related hypotheses. The large compendia of gene expression data provide the opportunity to conduct meta-analyses by combining data from various independent but related studies. Such data integration has the potential to enhance the reliability and generalizability of the results of individual microarray studies. This chapter describes the meta-analysis process including objectives, data collection, annotation, analysis methods, and visualizations. For each step we present a selection of tools available and discuss associated problems and difficulties.

*Author to whom correspondence should be addressed
†bspa33@bangor.ac.uk
‡gerhard.thallinger@tugraz.at
§r.macfarlane@bangor.ac.uk
¶leelarcombe@gmail.com
1 Introduction

Microarray technology enables the investigation of tens of thousands of genes simultaneously in a single experiment. It is possible to capture the transcriptional state of a cell under different environmental, genetic, physiological, and pathologic conditions or at different stages of development, resulting in expression signatures that characterize such a state (Bullinger et al. 2004; Dhanasekaran et al. 2001; Furlong et al. 2001; Gasch et al. 2000; Ivanova et al. 2002; Lu et al. 2004; McDonald and Rosbash 2001; Ramalho-Santos et al. 2002; Ramaswamy et al. 2003; White et al. 1999). Expression signatures have high medical and clinical value, with the utility of microarrays in medical research being evidenced by the ability to classify subtypes of diseases and to predict targets for prognosis, diagnosis, and therapy (Alizadeh et al. 2000; Bullinger et al. 2004; Dhanasekaran et al. 2001; Golub et al. 1999; Perou et al. 1999; van’t Veer et al. 2002) as well as to measure drug responses (Dan et al. 2002; Shimizu et al. 2004; Staunton et al. 2001; Zembutsu et al. 2002). The clinical potential of the technology as a diagnostic or prognostic tool (Li et al. 2008) can be highlighted by the US Food and Drug Administration (FDA) approval of the AmpliChip CYP450 from Roche (de Leon et al. 2006) and the MammaPrint from Agenda (Slodkowska and Ross 2009).

However, the identification of such sets of differentially expressed genes distinguishing one condition from another (e.g., healthy and diseased) continually proves challenging. Variation from differences in experimental settings, lack of validation, and, in particular, the small sample size of many microarray studies complicates the interpretation of the results, and calls the reliability and reproducibility of individual studies into question (Marshall 2004; Michiels et al. 2005; Ntzani and Ioannidis 2003). An integrative data analysis, a so-called meta-analysis, can serve as a remedy by combining information from independent but related studies in order to enhance the statistical power, reliability, and generalizability of results (Normand 1999; Ramasamy et al. 2008). In addition to refining and validating hypotheses between analogous studies (Arasappan et al. 2011; Griffith et al. 2006; Grutzmann et al. 2005; LaCroix-Fralish et al. 2011; Parmigiani et al. 2004; Rhodes et al. 2002; Shen et al. 2004; Smith et al. 2008; Vierlinger et al. 2011; Wang et al. 2004), meta-analyses can be used to identify a meta-signature across related studies (Anders et al. 2011; Daves et al. 2011; Pihur et al. 2008; Rhodes et al. 2004); to address novel questions (Chang et al. 2011; Cheng et al. 2011; Wennmalm et al. 2005); and/or to infer co-expression patterns and gene function (Lee et al. 2004; Stuart et al. 2003; Wren 2009; Zhou et al. 2005). Ultimately, meta-analyses can provide the opportunity to maximize the use of available data to help to uncover underlying biological mechanisms.
2 Microarray Technology and Data Analysis

This section serves as a brief introduction into microarray technology and data analysis to provide a basic understanding of this methodology.

2.1 Introduction to Microarray Technology

Despite the completion of the human genome sequencing project, questions remain addressing expression, function, and regulation of genes, which can be studied among others by mRNA expression profiling. Techniques such as serial analysis of gene expression (SAGE, Velculescu et al. 1995), expressed sequence tags (ESTs, Adams et al. 1991), and microarrays (Schena et al. 1995) enable evaluation of the expression of tens of thousands of genes in parallel. However, since the introduction of microarray technology by Schena et al. (1995), it has developed to become the most widely used method for profiling mRNA expression. In addition, microarrays can characterize the genome by investigating single nucleotide polymorphisms (SNPs, Kennedy et al. 2003; Teh et al. 2005), alternative RNA splicing (Pan et al. 2004), or DNA copy number changes (Pollack et al. 1999).

The underlying principle of microarray methodology relies on hybridization between nucleic acids (Southern et al. 1999). For expression studies, an RNA sample under investigation is reverse transcribed into complementary DNA (cDNA), labeled, and hybridized on an array. The array represents a defined matrix of tens of thousands of cDNA or oligonucleotide probes each corresponding to a gene of interest and arrayed onto a solid surface at distinct sites. After hybridization and washing, a scanner is used to detect fluorescence intensities at each probe site (Burgess 2001; Gershon 2002; Schena et al. 1995). In contrast to such a single channel experiment, a two channel experiment uses cDNA samples from two diverse populations labeled with different fluorophores. These are hybridized to the same array, which results in relative amounts of transcripts between the two populations, detectable as relative fluorescence intensities (Shalon et al. 1996).

A number of microarray platforms have been developed with the most popular being Affymetrix (http://www.affymetrix.com/), NimbleGen (http://www.nimblegen.com/), and Agilent (http://www.home.agilent.com). In general, there are two methods of microarray fabrication: to manufacture cDNA arrays, probes are spotted onto a surface such as glass or silicon (Cheung et al. 1999), whereas the production of oligonucleotide arrays is based on direct synthesis of the probes onto the array surface using
photolithographic methods (Lipshutz et al. 1999; Singh-Gasson et al. 1999) or ink-jet printing (Blanchard et al. 1996).

Oligonucleotide arrays are widely used with Affymetrix having the highest market share. Affymetrix arrays typically consist of 11–20 probe pairs per gene, where each probe pair represents a perfectly matching (PM) and a mismatching sequence (MM) of 25 bp in length to distinguish between specific and nonspecific hybridization events (Lipshutz et al. 1999). Newer arrays, however, do not provide MM sequences anymore, as studies showed that MM sequences could not reliably be used to detect nonspecific hybridization events (Irizarry et al. 2003; Wang et al. 2007).

2.2 Introduction to Standard Microarray Data Analysis

After quantifying the raw images into fluorescence intensity values for each probe, the data must undergo various preprocessing steps to account for variation caused during the experimental procedure (Nadon and Shoemaker 2002) followed by a statistical analysis to compute differentially expressed genes. The data analysis process described here focuses on profiling mRNA expression for single channel experiments using Affymetrix arrays, but is generally applicable with slight adaptations to other applications.

Preprocessing steps include (1) background correction, (2) normalization, (3) summarization of probe intensities, and (4) filtering (reviewed by Gentleman 2005; Suarez et al. 2009). Background correction is essential to eliminate the noise originating from nonspecific hybridization and the laser scanning process (Gentleman 2005). Various methods have been developed with the most popular being the robust multi-array average (RMA, Irizarry et al. 2003) and the MAS 5.0 background (MicroArray Suite from Affymetrix, Hubbell et al. 2002). Normalization is used to detect and correct for systematic differences in the overall distribution of probe intensity values, and allows the comparison of data from different chips (Owzar et al. 2008). Bolstad et al. (2003) compare and review a number of normalization methods including cyclic loess (Dudoit et al. 2002), quantile (Bolstad et al. 2003), scaling (Affymetrix), and nonlinear methods (Li and Hung Wong 2001; Schadt et al. 2001). Another widely used normalization method, Variance Stabilization and Normalization (VSN), was introduced by Huber et al. (2002). After normalization, summarization of probe intensities is necessary to establish a single expression value for each gene. Among the more common methods are Li–Wong (Li and Wong 2001), median polish (Tukey 1977), and summarization methods from Affymetrix. Finally, filtering may be applied to eliminate genes, which for example exhibit relatively low variability across the samples. Filtering increases the
3 Meta-Analysis

statistical power, as it reduces the number of hypotheses to be tested (Gentleman 2005; Owzar et al. 2008).

Following preprocessing, a statistical analysis serves to identify significant genes that are differentially expressed under certain conditions. Various approaches have been developed (Cui and Churchill 2003; Suarez et al. 2009) ranging from simple fold-change (FC) criteria (DeRisi et al. 1997) and ordinary $t$ tests (Callow et al. 2000), to more sophisticated methods including moderated $t$ test (Limma, Smyth 2004), Bayesian methods (Lo and Gottardo 2007), rank product statistics (Breitling et al. 2004), Analysis of Variance (ANOVA, Sahai and Ageel 2000), or permutation methods such as Significance Analysis of Microarrays (SAM, Tusher et al. 2001).

Microarray data analysis is often now performed using R (R Core Team 2012) and the Bioconductor libraries (http://www.bioconductor.org/). Numerous R packages have been developed to facilitate microarray data analysis such as the popular “affy” R package (Gautier et al. 2004a). Alternatives are Matlab (http://www.mathworks.co.uk/) or other platforms developed such as Genesis (Sturn et al. 2002) and TM4 (Saeed et al. 2006), with commercial data analysis software such as GeneSpring GX from Agilent Technologies (http://www.genespring.com/) and GeneMaths XT (http://www.appliedmaths.com/) also available.


Large amounts of microarray data are now available in public repositories and provide researchers the opportunity to retrieve, integrate, and reanalyze the data (Moreau et al. 2003). So-called meta-analysis techniques aim to combine the data available and integrate information from multiple independent but related microarray experiments to identify significant genes (Normand 1999; Ramasamy et al. 2008).

A meta-analysis consists of (1) objective definition, (2) data collection; (3) data preprocessing and selection of differentially expressed genes, (4) annotation, (5) analysis of differentially expressed genes across studies, and (6) data interpretation and presentation (Ramasamy et al. 2008). Points (1)–(5) are discussed below including the associated advantages, problems, and difficulties. The last point is discussed in the final section by presenting a couple of examples for visualization of complex data.
3.1 Advantages of Meta-Analysis and Its Objectives

Combining studies can enhance reliability and generalizability of the results (Ramasamy et al. 2008) and is generally used to obtain a more precise estimate of gene expression. In particular, the benefit of increasing the statistical power can help to overcome probably the most profound limitation of microarray studies: testing tens of thousands of hypotheses, relying only on relatively few samples (Campion and Yang 2010; Normand 1999).

Combining microarray datasets is only sensible if the individual microarray experiments address similar or related questions. It may be used (1) to reveal a more valid set of differentially expressed genes in analogous studies (Arasappan et al. 2011; Griffith et al. 2006; Grutzmann et al. 2005; LaCroix-Fralish et al. 2011; Parmigiani et al. 2004; Rhodes et al. 2002; Shen et al. 2004; Smith et al. 2008; Vierlinger et al. 2011; Wang et al. 2004), (2) to identify an overlap of genes in related studies – a meta-signature (Anders et al. 2011; Daves et al. 2011; Pihur et al. 2008; Rhodes et al. 2004), (3) to test new hypotheses (Chang et al. 2011; Cheng et al. 2011; Wennmalm et al. 2005), or (4) to gain insights into co-expression patterns and gene function (Lee et al. 2004; Stuart et al. 2003; Wren 2009; Zhou et al. 2005). Meta-analyses can aid to determine subtypes of diseases, targets for prognosis, treatment, diagnosis, and monitoring as well as treatment effects or signatures for biological mechanisms and conditions, and thus can lead to a more accurate understanding of underlying biological mechanisms.

Meta-analyses can eliminate artifacts of individual but analogous studies (e.g., a given cancer type) or resolve conflicting results between analogous studies to refine and validate primary hypotheses (Normand 1999; Rhodes et al. 2002). Biological, experimental, and technological variations including differences in experimental conditions, tissues, cell lines, species, platforms, sample treatment, and processing can lead to inconsistencies in gene expression, which reflect the differences in the experimental setting in addition to the objective studied (Cahan et al. 2007). Combining studies can eliminate these variations and identify a more valid set of differentially expressed genes; for example, LaCroix-Fralish et al. (2011) analyzed the results of existing tonic/chronic pain microarray studies and could identify a more accurate set of differentially expressed genes. Similarly, Arasappan et al. (2011) found a refined expression signature for systemic lupus erythematosus, and Vierlinger et al. (2011) reported the identification of a potential biomarker for papillary thyroid carcinoma (PTC) by merging of microarray datasets comparing PTC nodules to benign nodules.
Another closely related objective is to identify a common transcriptional profile – a meta-signature. A meta-signature is an overlap of genes, which is shared within a given group across related studies (e.g., across cancer studies). In one microarray experiment hundreds of genes can be declared as significant, of which numerous might be spurious or system-specific and thus are expected to show no change across related studies. In contrast, core features are expected to be overrepresented (Pihur et al. 2008; Rhodes et al. 2004). Daves et al. (2011), for example, reported a common meta-signature for metastasis by means of comparing primary to metastatic tumors in various types of cancer, while Anders et al. (2011) detected angiogenesis-related meta-signatures in cancer.

Gene function can also be inferred through meta-analysis. Since genes are conditionally expressed, groups of co-expressed genes can aid functional annotation by assuming associated functions (Troyanskaya 2005). Conducting a global meta-analysis can predict gene function based on the recurrent expression pattern of co-regulated genes across various conditions (Wren 2009).

Novel questions may be addressed by conducting a meta-analysis. According to Wennmalm et al. (2005) the observed expression pattern of senescence in cell cultures resembles that of aging in mouse but not in humans. Chang et al. (2011) identified housekeeping and tissue-selective genes across 43 tissues by means of meta-analysis, whereas Cheng et al. (2011) reported potential reference genes for 13 tissue types across 4 physiological states, which may be used for normalization of quantitative real-time polymerase chain reaction (qRT-PCR).

3.2 Data Collection and Quality Control

There are two types of data suitable for meta-analyses: the raw data (probe intensities) and the published results (gene lists, Cahan et al. 2007; Ramasamy et al. 2008). Processed data is more frequently available than raw data; Larsson and Sandberg (2006) stated that only 48% of published microarray experiments in public repositories such as ArrayExpress are available in the form of raw data. Nevertheless, it is recommended to use raw data, as results of microarray analyses depend on the genes covered in the study, the preprocessing steps, the annotation methods, and the data analysis techniques used. In addition, information about all other genes not in the list is lost (Suarez-Farinas et al. 2005).
3.2 Data Collection and Quality Control

Public repositories (Table 1) such as ArrayExpress (Parkinson et al. 2009), Gene Expression Omnibus (GEO, Barrett et al. 2011), Center for Information Biology Gene Expression database (CIBEX, Ikeo et al. 2003), and Stanford Microarray Database (SMD, Hubble et al. 2009) collect raw and processed microarray data from diverse platforms and provide it to the public. Recently, a number of more specialized databases (Table 1) have become available. Databases such as M²DB (Cheng et al. 2010) and M³D (Faith et al. 2008) collected microarray data and uniformly preprocessed it. Databases such as L2L (Newman and Weiner 2005) provide published gene lists, which can be compared to the user’s own microarray data. To facilitate querying of popular databases (Table 1), interfaces have been implemented such as Geometadb (Zhu et al. 2008), the “GEOquery” R package (Davis and Meltzer 2007), the “ArrayExpress” R package (Kauffmann et al. 2009), and MaRe (Ivliev et al. 2008).

Varying results observed between studies raised concerns about the comparability of microarray experiments and led to the questioning of reproducibility, repeatability, and validation of microarrays in the research community (Marshall 2004; Michiels et al. 2005; Ntzani and Ioannidis 2003). Intensive studies were carried out to assess the reproducibility across platforms and laboratories, in particular driven by the MicroArray Quality Control (MAQC) project (Shi et al. 2006). In general, agreement in cross-platform and cross-laboratory experiments was achieved if the preparation and the consumables were appropriately handled (Irizarry et al. 2005; Larkin et al. 2005; Shi et al. 2006; Shippy et al. 2006). In particular, Affymetrix platforms provided the most consistent results across multiple laboratories (Irizarry et al. 2005). Recent studies suggest that poor quality data is responsible for the differences between experiments, which may be due to lack of standards as well as inadequate experimental procedures, statistical analysis, validation, and/or reporting of the studies (Dupuy and Simon 2007; Jafari and Azuaje 2006; Shi et al. 2005). According to a study by Larsson and Sandberg (2006), only 23% of the raw data in GEO and ArrayExpress meet the quality requirements for RNA integrity and hybridization sensitivity to be considered as reliable datasets (Larsson and Sandberg 2006). Hence one should access the data quality, and poor quality data should then be excluded to assure comparability. The introduction of the Minimum Information About a Microarray Experiment (MIAME) standard led to an improvement by requiring comprehensive reporting of sample, experimental, and array design to allow proper interpretation of microarray experiments (Brazma et al. 2001). A number of journals now require the submission of microarray data to a public repository preferable in a format that agrees with MIAME (Ball et al. 2004). Ultimately, good laboratory practice and well-controlled experiments assure reproducibility, as the quality of the meta-analysis can only be as good as the quality of the underlying data (Shi et al. 2006).
**Table 1:** A selection of internet repositories and search interfaces for microarray data. * indicates MIAME-supportive databases.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArrayExpress R</td>
<td>R/Bioconductor package to access ArrayExpress</td>
<td><a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a></td>
</tr>
<tr>
<td>ArrayExpress*</td>
<td>Repository for raw and/or processed data</td>
<td><a href="http://www.ebi.ac.uk/arrayexpress/">http://www.ebi.ac.uk/arrayexpress/</a></td>
</tr>
<tr>
<td>CIBEX*</td>
<td>Repository for raw and/or processed data</td>
<td><a href="http://cibex.nig.ac.jp/index.jsp">http://cibex.nig.ac.jp/index.jsp</a></td>
</tr>
<tr>
<td>GEO*</td>
<td>Repository for raw and/or processed data</td>
<td><a href="http://www.ncbi.nlm.nih.gov/geo/">http://www.ncbi.nlm.nih.gov/geo/</a></td>
</tr>
<tr>
<td>GEOmetadb</td>
<td>Web-based search interface to GEO and R/Bioconductor package</td>
<td><a href="http://gbnci.abcc.ncifcrf.gov/geo/">http://gbnci.abcc.ncifcrf.gov/geo/</a></td>
</tr>
<tr>
<td>GEOquery</td>
<td>R/Bioconductor package to access GEO</td>
<td><a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a></td>
</tr>
<tr>
<td>L2L</td>
<td>Repository for published gene lists</td>
<td><a href="http://depts.washington.edu/121/">http://depts.washington.edu/121/</a></td>
</tr>
<tr>
<td>M^2DB</td>
<td>Repository for raw and preprocessed Affymetrix data</td>
<td><a href="http://metadb.bmes.nthu.edu.tw/m2db/">http://metadb.bmes.nthu.edu.tw/m2db/</a></td>
</tr>
<tr>
<td>M^3D</td>
<td>Repository for raw and preprocessed Affymetrix data for three microbial species</td>
<td><a href="http://m3d.bu.edu/">http://m3d.bu.edu/</a></td>
</tr>
<tr>
<td>MaRe</td>
<td>Web-based search interface to GEO and ArrayExpress</td>
<td><a href="http://www.lgtc.nl/MaRe/">http://www.lgtc.nl/MaRe/</a></td>
</tr>
<tr>
<td>SMD*</td>
<td>Repository for raw, preprocessed and processed data and analysis tools for microarray data</td>
<td><a href="http://smd.stanford.edu/">http://smd.stanford.edu/</a></td>
</tr>
</tbody>
</table>
3.3 Data Preprocessing and Selection Criteria

Numerous preprocessing and data analysis methods for microarrays have been proposed over the years to include background correction, normalization, measure summarization, and filtering (cf. Sect. 2).

The impact of preprocessing on reproducibility of microarrays has been intensively studied (Irizarry et al. 2005; Owzar et al. 2008; Patterson et al. 2006; Shippy et al. 2006). Inconsistencies were found when comparing differently preprocessed datasets, even when different methods were applied to the same dataset (Gagarin et al. 2005; Owzar et al. 2008). Irizarry et al. (2005) claim that alternative preprocessing methods such as RMA (Irizarry et al. 2003) can improve cross-study and cross-platform agreement. In contrast, other studies showed that preprocessing methods had very little impact on the resulting gene lists when following the manufacturer’s recommendations (Patterson et al. 2006; Shi et al. 2006; Shippy et al. 2006). Nevertheless, it is advisable to uniformly preprocess the raw data to account for any systematic differences. Unfortunately, this may be difficult for meta-analyses combining cross-platform data, as few preprocessing methods can be applied to all platforms (Ramasamy et al. 2008).

The selection criteria for differentially expressed genes can also affect the reproducibility. The use of FC criteria proved to generate more reproducible results than solely relying on p value criteria. Also more sophisticated methods such as SAM (Tusher et al. 2001) did not improve the reproducibility (Shi et al. 2005, 2006). However, using a non-stringent p value cutoff in addition to FC criteria generated the highest overlap in differentially expressed gene lists (Guo et al. 2006).

3.4 Annotation

In order to interpret the results of a microarray study, probe-level identifiers (e.g., Affymetrix IDs or I.M.A.G.E. cloneIDs) need to be linked to the corresponding gene identifiers (e.g., Entrez Gene IDs or Ensembl IDs). To do so, one can use the annotation files provided by Affymetrix (http://www.affymetrix.com/support/technical/annotationfilesmain.affx), the annotation packages from Bioconductor (http://www.bioconductor.org/packages/release/data/annotation/), or web tools such as IDconverter (Alibes et al. 2007), SOURCE (Diehn et al. 2003), RESOURCERER (Tsai et al. 2001), DAVID (Dennis et al. 2003), or MADGene (Baron et al. 2011) (Table 2). Several tools not only allow mapping of identifiers but also provide annotation with additional biological information and/or comparison to other species. Alternatively, alignment algorithms such as BLAST (Altschul et al. 1990) may be used to map probes
based on sequence similarity (Shi et al. 2006) by means of databases such as RefSeq (Pruitt et al. 2009) or the TIGR Gene Index databases (now the DFCI Gene Index, Lee et al. 2005).

Annotation poses a challenge for various reasons. It has been suggested that the annotation method used could have an effect on the resulting gene list and therefore might be responsible for inconsistencies between platforms (Irizarry et al. 2005). Thus, the annotation method should be consistent across all microarray experiments if possible (Ramasamy et al. 2008). First, diverse platforms do not use a unique nomenclature or common identifiers, which impairs gene annotation and thus comparability of results (Cahan et al. 2007). Second, as gene annotation is not yet complete, genome databases are incomplete, which in turn affects microarray annotation (Brors 2005; Shi et al. 2006). Third, probe disparities can cause inconsistencies, as probes used to measure gene expression may differ depending on the platform. These disparities may be due to different sensitivity and/or specificity, in particular if splice variants are involved (Cahan et al. 2007; Shi et al. 2006). Fourth, the available platforms differ not only in hybridization technique but also in coverage. A lot of arrays do not cover the complete genome and thus the transcript coverage could cause differences in the resulting gene list (Cahan et al. 2007). Finally, not all probes map to one gene and vice versa, as probes might not be specific enough due to cross-hybridization from splice variants or closely related genes (Ramasamy et al. 2008; Shi et al. 2006). This led to the proposal of alternative mappings of probes to genes for Affymetrix chips (Gautier et al. 2004b; Harbig et al. 2005).

3.5 Analysis Methods

Numerous meta-analysis approaches have been developed over the last century that have more recently been adapted for application to microarray experiments. These can generally be divided into two categories: relative and absolute methods, where the former analyzes each study (microarray experiment) and combines the results, and the latter combines the data first and subsequently analyzes it with traditional techniques (reviewed by Campain and Yang 2010; Hong and Breitling 2008; Larsson et al. 2006; Ramasamy et al. 2008).

A brief summary of developed strategies and a selection of available tools are given in Table 3. The strategies discussed here are based on a two class comparison (e.g., cancer vs. normal) for single channel experiments with the focus mainly on relative approaches.
### Table 2: A selection of useful annotation tools for microarrays

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrix</td>
<td>Annotation files of the manufacturer</td>
<td><a href="http://www.affymetrix.com/support/technical/annotationfilesmain.affx">http://www.affymetrix.com/support/technical/annotationfilesmain.affx</a></td>
</tr>
<tr>
<td>MADGene</td>
<td>ID converter</td>
<td><a href="http://cardioserve.nantes.inserm.fr/madtools/">http://cardioserve.nantes.inserm.fr/madtools/</a></td>
</tr>
<tr>
<td>Resourcerer</td>
<td>Annotation for common platforms including comparisons within and across species</td>
<td><a href="http://compgen.bio.unipd.it/bioinfo/amadman/">http://compgen.bio.unipd.it/bioinfo/amadman/</a></td>
</tr>
<tr>
<td>Source</td>
<td>Mapping of feature identifiers and annotation with additional information from various databases</td>
<td><a href="http://smd.stanford.edu/cgi-bin/source/sourceSearch">http://smd.stanford.edu/cgi-bin/source/sourceSearch</a></td>
</tr>
</tbody>
</table>

### Table 3: A selection of available meta-analysis tools

<table>
<thead>
<tr>
<th>Name</th>
<th>Strategy</th>
<th>Implementation</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Expression Atlas</td>
<td>Vote counting</td>
<td>Web platform</td>
<td><a href="http://www.ebi.ac.uk/gxa/">http://www.ebi.ac.uk/gxa/</a></td>
</tr>
<tr>
<td>Genevestigator</td>
<td>Absolute method</td>
<td>Web platform</td>
<td><a href="https://www.genevestigator.com/">https://www.genevestigator.com/</a></td>
</tr>
<tr>
<td>Package</td>
<td>Description</td>
<td>Type</td>
<td>Website</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------------------------------------------</td>
<td>---------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MAMA</td>
<td>9 in 1 package includes e.g., RankProd and metaMA</td>
<td>R package</td>
<td><a href="http://cran.r-project.org/src/contrib/Archive/MAMA/">http://cran.r-project.org/src/contrib/Archive/MAMA/</a></td>
</tr>
<tr>
<td>metaArray</td>
<td>Integrative correlation strategy</td>
<td>R package</td>
<td><a href="http://www.bioconductor.org/packages/2.10/bioc/html/metaArray.html">http://www.bioconductor.org/packages/2.10/bioc/html/metaArray.html</a></td>
</tr>
<tr>
<td>metaMA</td>
<td>Effect size and p-value combination</td>
<td>R package</td>
<td><a href="http://cran.r-project.org/web/packages/metaMA/">http://cran.r-project.org/web/packages/metaMA/</a></td>
</tr>
<tr>
<td>METRADISC</td>
<td>Rank aggregation</td>
<td>Compaq Visual Fortan90 software</td>
<td><a href="http://biomath.med.uth.gr">http://biomath.med.uth.gr</a></td>
</tr>
<tr>
<td>Oncomine</td>
<td>Vote counting</td>
<td>Web platform</td>
<td><a href="https://www.oncomine.org/">https://www.oncomine.org/</a></td>
</tr>
<tr>
<td>RankAggreg</td>
<td>Rank aggregation</td>
<td>R package</td>
<td><a href="http://cran.r-project.org/web/packages/RankAggreg/">http://cran.r-project.org/web/packages/RankAggreg/</a></td>
</tr>
</tbody>
</table>
3.5 Analysis Methods

3.5.1 Vote Counting Strategies

Vote counting strategies are based on the number of studies reporting gene $i$ to be differentially expressed (Bushman 1994). Rhodes et al. (2004) applied such a vote counting approach to microarray experiments and assessed the significance by random permutation testing. Additionally, they collected and analyzed cancer microarray data that is publicly accessible via the data mining platform Oncomine (Rhodes et al. 2007).

The Gene Expression Atlas is another data mining platform that is provided by the European Bioinformatics Institute (EBI) and relies on curated microarrays derived from the ArrayExpress repository. The expression profile of a given gene $i$ across numerous conditions, developmental stages, and tissues can be viewed (http://www.ebi.ac.uk/gxa/).

Other related approaches have been developed. Griffith et al. (2006), for example, applied a vote counting strategy to processed data and calculated the significance by means of a Monte Carlo simulation.

3.5.2 Rank Aggregation Strategies

Rank combination strategies consider the individual rank orders of each gene $i$ across $k$ lists (individual results of microarray experiments) to merge them to an aggregated rank order. One possibility is to aggregate relative preferences of paired items across $k$ lists. Fagin et al. (2003) described possible distance measures between top $x$ lists and Dwork et al. (2001) proposed aggregating these relative preferences by means of Markov algorithms. DeConde et al. (2006) adopted this technique for microarray experiments by first computing pairwise comparisons of the ranks of gene $i$ and $i'$ relative to each other based on an extension of Kendall’s tau for two nonidentical but overlapping top $x$ lists $\tau_1$ and $\tau_2$ (Fagin et al. 2003; Kendall 1938), described as:

$$K^{(p)}(\tau_1, \tau_2) = \sum_{i,i' \in P(\tau_1, \tau_2)} K^{(p)}_{i,i'}(\tau_1, \tau_2) \text{ where } i \neq i'$$

(1)

$P(\tau_1, \tau_2)$ is the set of all paired items $i$ and $i'$ and $K^{(p)}_{i,i'}(\tau_1, \tau_2)$ represents the penalty value for the paired items $i$ and $i'$. The values range from 0 to 1 depending on whether the pair is concordantly (0) or discordantly ranked (1) across the two lists. If the ordering cannot be inferred, a penalty parameter $p$ is defined ($0 < p < 1$). Second, DeConde et al. (2006) converted the pairwise comparisons into aggregate rankings by means of three different algorithms: a Thurstone’s order-statistics model (Thurstone 1931) and two Markov chain algorithms Dwork et al. (2001). The two Markov chain algorithms...
algorithms use the pairwise comparisons to define a transition matrix $M$. For a set of genes (states) $G$, $M$ represents the relative preference for gene $i$ over gene $i'$ across $k$ microarray experiments, and the aggregation of the rankings is given by the stationary distribution $\pi$ of the Markov chain. The stationary distribution $\pi$ is the principal left eigenvector of the transition matrix $M$ associated with an eigenvalue of 1. $\pi$ reflects a natural order for $G$, where the highest value in $\pi$ corresponds to the gene with the highest rank order (DeConde et al. 2006; Dwork et al. 2001) as shown below:

$$M = g \times g$$

$$G = \{1, 2, \ldots, g\}$$

$$\pi = \pi \cdot M$$

Pihur et al. (2008) developed a closely related approach, which is also based on distance measures between top $x$ lists. The method is publicly available as the “RankAggreg” R package (Pihur et al. 2009). For measuring the distance between two top $x$ lists, one can choose between Spearman’s foot rule (Spearman 1904) and Kendall’s tau (Kendall 1938), and both distance measures can be additionally weighted. They provide two rank aggregation algorithms: the Cross-Entropy Monte Carlo and the Genetic algorithm (Pihur et al. 2008, 2009).

A further rank aggregation strategy, “RankProd”, was developed by Breitling et al. (2004) and is also available as an R package (Hong et al. 2006). The method is based on FC criteria, where for two experimental conditions $A$ and $B$, $M = n_A \times n_B$ represents the pairwise FC ratios for each gene $i$ in a dataset $j$ with $n = n_A + n_B$ samples. A rank product $RP_i$ is computed based on the ranks $r_{ijc}$ for gene $i$ across all $k$ datasets and all $s$ pairwise comparisons, and significance is determined by permutation testing (Breitling et al. 2004; Hong et al. 2006):

$$\frac{A_{n_1}}{B_{n_1}}, \frac{A_{n_1}}{B_{n_2}}, \ldots, \frac{A_{n_A}}{B_{n_B}} \Rightarrow n_A \times n_B$$

$$RP_i = \left(\prod_{j=1}^{k} \prod_{c=1}^{s} r_{ijc}\right)^{\frac{1}{s}}$$

Another rank combination method, Meta-Analysis of Ranked Discovery Datasets (METRADISC), was proposed by Zintzaras and Ioannidis (2008), and allows the incorporation of heterogeneity between studies. An average rank $r^*$ and a heterogeneity metric $q^*$ are computed for each gene $i$ across $k$ datasets as:
The significance of $r^*$ and $q^*$ is calculated via Monte Carlo permutation testing. The METRADISC software is publicly available for download (http://biomath.med.uth.gr).

### 3.5.3 p Value Combination Strategies

p Value combination strategies pool p values from independent studies to determine if a variable (gene) $i$ is significant (reviewed by Loughin 2004). A popular method is the sum of logs strategy proposed by Fisher (1932), whereby the p values of each study $j$ are used to generate a summary statistic $S_i$: 

$$S_i = -2 \sum_{j=1}^{k} \log p_{ij}$$

To determine the p value for $S_i$, $S_i$ can be assumed to follow a $\chi^2$ distribution with $2k$ degrees of freedom. This method was applied to microarray data by Rhodes et al. (2002), whereby they computed p values for gene $i$ in a dataset $j$ by random permutation $t$ tests, combined them by means of Fisher’s method (Fisher 1932), and assessed $S_i$ by permutation testing (Rhodes et al. 2002).

Alternatively, $z$ scores may be used instead of p values. This so-called inverse normal method was introduced by Stouffer (1949) and allows assigning a weight $w_i$ to each individual study $j$ (Hedges et al. 1992; Stouffer 1949; Whitlock 2005). This can be described as follows, where $p_{ij}$ is the one-tailed p value corresponding to the $t$ test of study $j$ for gene $i$ and $\Phi$ represents the normal distribution function:

$$z_{ij}(p_{ij}) = -\Phi^{-1}(p_{ij})$$

$$z_i = \frac{\sum_{j=1}^{k} z_{ij}}{\sqrt{k}}$$

$$z_i = \frac{\sum_{j=1}^{k} w_j z_{ij}}{\sqrt{\sum_{j=1}^{k} w_j^2}}$$

The “metaMA” R package provides p value combination strategies based on Fisher’s and Stouffer’s methods (Marot et al. 2009).
3.5.4 Effect Size Combination Strategies

The effect size is a standardized, scale-free measure of the magnitude of a difference between two groups (Cohen 1988). For a meta-analysis, the effect size estimates from each individual study \( j \) can be combined to an overall estimate of the size of the effect. Choi et al. (2003) proposed an effect size-based meta-analysis method for microarrays, whereby they calculate the effect size using Cohen’s \( d \) (Cohen 1988) modified to Hedges’ \( g^* \) (Hedges and Olkin 1985) using a correction factor, which accounts for the sample size bias:

\[
d_{ij} = \frac{\bar{x}_{Aij} - \bar{x}_{Bij}}{s_{ij}}
\]

\[
g_{ij} = \frac{\bar{x}_{Aij} - \bar{x}_{Bij}}{s^*_j}
\]

\[
g^*_{ij} = g_{ij} \cdot (1 - \frac{3}{4n - 9})
\]

Here \( x_A \) and \( x_B \) are the means of the two groups for \( n \) samples with the standard deviation \( s \) and the pooled standard deviation \( s^* \). The overall mean of differential expression \( \mu_i \) for each gene \( i \) across all \( k \) datasets can be extracted from the following model, where \( g^*_{ij} \) is the effect size and \( \theta_{ij} \) the study-specific mean of study \( j \). \( \epsilon_{ij} \) represents the within-study effect with the corresponding variance \( s^2_{ij} \), whereas \( \delta_{ij} \) is the between-study effect with the corresponding variance \( \tau^2_i \):

\[
g^*_{ij} = \theta_{ij} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, s^2_{ij})
\]

\[
\theta_{ij} = \mu_i + \delta_{ij}, \delta_{ij} \sim N(0, \tau^2_i)
\]

For the random-effects model, \( \tau^2 \) can be estimated using a method developed by DerSimonian and Laird (1986). In contrast to the random-effects model, the fixed-effects model assumes that the between-study variances are 0 and may be used when the studies show biological and technical uniformity (Ramasamy et al. 2008). “GeneMeta” (Lusa et al. 2006) is an available R package for microarray meta-analysis based on Choi et al. (2003) as described above.

Marot et al. (2009) suggested an adaptation of such approaches, whereby they use moderated effect sizes by relating to moderated \( t \) tests. The effect size resembles the \( t \) test apart from the factor \( \sqrt{n} \), which accounts for the sample size:

\[
t_{ij} = d_{ij} \cdot \sqrt{n_{ij}}
\]
\begin{equation}
\tilde{n}_{ij} = \frac{n_{A_{ij}} \cdot n_{B_{ij}}}{n_{A_{ij}} + n_{B_{ij}}}
\end{equation}

\( t \) can be calculated using Limma (Smyth 2004) or other variance shrinkage approaches. The “metaMA” R package (Marot et al. 2009) provides various meta-analysis strategies including the methods based on Choi et al. (2003) and Marot et al. (2009).

### 3.5.5 Other Strategies

An integrative correlation (IC) strategy to define reproducible genes was proposed by Parmigiani et al. (2004), which was, in combination with the generation of a probability of expression matrix \( E_j \) (Parmigiani et al. 2002), implemented into the “metaArray” R package (Choi et al. 2007). First, the expression values for \( g \) genes across \( n \) samples are transformed into \( E_j \):

\[
E_j = g \times n
\]

\[
e_{in} \begin{cases} 
-1 & \text{underexpressed} \\
0 & \text{not differentially expressed} \\
1 & \text{overexpressed}
\end{cases}
\]

Second, correlations for all pairs of genes \( i \) and \( i' \) in a study \( j \) are computed by means of the Pearson correlation coefficient \( \rho_{ii'j} \) and summarized as a mean of the correlations per study \( \bar{\rho}_j \). Third, the integrative correlation \( I_{ijj'} \) for gene \( i \) for two datasets \( j \) and \( j' \) is given by:

\[
I_{ijj'} = \sum_{i'=1}^{g} (\rho_{ii'j} - \bar{\rho}_j) \cdot (\rho_{ii'j'} - \bar{\rho}_{j'}) \text{ where } i \neq i' \text{ and } j \neq j'
\]

For more than two studies, the average of all integrative correlations for a certain gene \( i \) represents a reproducibility score. All genes with a score above a certain threshold are deemed to be reproducible (Parmigiani et al. 2004).

Campain and Yang (2010) proposed a method described as meta Differential Expression via Distance Synthesis (mDEDS), which relies on combining multiple statistical measures such as FC, SAM, and \( t \) values from standard or moderated statistics to identify true differently expressed genes.

Various absolute approaches have been reported including the web-based platforms GeneSapiens (Kilpinen et al. 2008) and Genevestigator (Hruz et al. 2008). Such platforms enable to compare the expression values of samples that have been pooled and
uniformly preprocessed, and provide insight in gene expression across numerous conditions and tissues. In addition, the web application A-MADMAN (Bisognin et al. 2009) allows retrieving, annotating, and pooled preprocessing of microarray datasets. It outputs expression values, which can be fed into a custom R analysis.

Furthermore, approaches incorporating additional information were developed, such as the literature-aided meta-analysis reported by Jelier et al. (2008) and the pathway-based approach proposed by Arasappan et al. (2011). The introduction of additional quality weights can further enhance the statistical power of meta-analyses (Hu et al. 2006).

### 3.5.6 Strategy Comparison

Opinions about the performance of meta-analysis strategies differ. Hong and Breitling (2008) conducted a comparison of three meta-analysis strategies: rank aggregation by Breitling et al. (2004), Fisher’s method (Fisher 1932), and effect size combination by Choi et al. (2003). According to Hong and Breitling (2008) the rank aggregation strategy demonstrates greater sensitivity and reproducibility, in particular concerning small sample sizes and high between-study variations. A comparison presented by Campain and Yang (2010) evaluated eight different methods including six methods described above and two absolute strategies. Most methods performed reasonably well for similar-platform meta-analyses, but struggled with cross-platform analyses. Fisher’s method (Fisher 1932), the integrative correlation strategy (Parmigiani et al. 2002; Parmigiani et al. 2004), and mDEDS (Campion and Yang 2010) outperformed the other methods under such conditions. In contrast, Ramasamy et al. (2008) favored the effect size combination strategies, primarily due to the value of weighting each study. The disadvantage of vote counting strategies is that only significant genes of the individual studies are considered for meta-analysis. Combining p values can increase the significance of the results, but does not provide the magnitude of the effect or a direction of significance if two-sided p values are used (Ramasamy et al. 2008).

### 4 Visualization of Complex Data

Results of single microarray studies are usually presented in the form of heatmaps or clustered heatmaps (Eisen et al. 1998) to illustrate the similarities of expression patterns across groups of genes or samples. Heatmaps are grids, in which color ranges are used to reflect the expression value. Coherent color patterns derive from hierarchical clustering
Visualization of Complex Data

and are indicated through tree-like structures (Wilkinson and Friendly 2009). For short gene lists or small numbers of combined studies, clustered heatmaps or Venn diagrams (Venn 1880) can be used to visualize meta-analysis results quite successfully. With increasing complexity of the data, however, other visualization techniques must be used to simplify the interpretation of large quantities of data and to highlight the relationships within the data. The most popular way to visualize meta-analysis results (Lalkhen and McCluskey 2008) is a Forest plot (Lewis and Clarke 2001), where each study is illustrated by a square; the position on the x-axis representing the measure estimate (e.g., FC ratio), the size proportional to the weight of the study, and the horizontal line through it reflecting the confidence interval of the estimate (Fig. 1). Alternatively, new approaches to visualize complex data are developing, such as the circular layout visualizations produced by Krona (Ondov et al. 2011) or Circos plots (Krzywinski et al. 2009). Such tools are becoming increasingly popular in comparative genomics and metagenomics, and could, for example, be used to illustrate the weighted relationships between gene expression and different study datasets. The challenge for successful meta-analysis visualization methods is to demonstrate the variations between studies and facilitate biological interpretation of the overall result.

**Figure 1: Example Forest plot.** The Forest plots illustrate the variations in FC values for the genes *ALG3* (Hs.478481, most significantly upregulated) and *FGD4* (Hs.117835, most significantly down-regulated) between various cancer studies (Ramasamy et al. 2008).
4 Visualization of Complex Data

References


Book chapter in Computational Medicine published by Springer Vienna


4 Visualization of Complex Data


Book chapter in Computational Medicine published by Springer Vienna
4 Visualization of Complex Data


4 Visualization of Complex Data


4 Visualization of Complex Data

S, Kallioniemi O (2008) Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. Genome Biol 9(9):R139


4 Visualization of Complex Data


4 Visualization of Complex Data


Book chapter in Computational Medicine published by Springer Vienna


4 Visualization of Complex Data


1.5 Expressed Sequence Tag Meta-analysis

There are several genome-scale technologies available to date to investigate gene expression that have been applied towards the understanding of fundamental biology and disease as well as towards the discovery of new biomarkers and therapeutic targets. However, there is always a gap between technique development, data availability and widespread use. Over the last decade an overwhelming amount of EST data has become available, providing the opportunity to conduct meta-analyses by combining the large compendia of public EST data. This chapter describes the process of integrating EST data including objectives, data collection/resources and analysis methods with a selection of tools.

1.5.1 Introduction to Expressed Sequence Tag Technology

Expressed sequence tags (ESTs) are short DNA sequences (200-500 nucleotides) generated by sequencing the 5’ and/or 3’ ends of cDNAs that are subsequently clustered to gene transcripts based on sequence similarity followed by EST counting for quantification and/or functional annotation [175]. An EST library represents an expression signature reflecting the transcriptional state of the cell or tissue type analysed under different physiological and/or pathological conditions, or at different stages of development. EST analysis does not only enable quantitative evaluation of expression levels [176, 177], but also allows the discovery of novel genes [178–180], the investigation of alternative splicing [181, 182], the detection of single nucleotide polymorphisms (SNPs) [183, 184], the characterisation of tissue- and/or disease-specific expression [57, 58, 185–187], the identification of co-expressed genes [188, 189], the prediction of gene structures [190, 191] as well as the improvement and support of genome annotation [192].

1.5.2 Expressed Sequence Tag Data Analysis

The mRNA sequences under investigation reflect the parts of the genome that are transcriptionally active. For expression studies, mRNA is reverse transcribed into cDNA, which can be cloned to construct cDNA libraries. The 5’ and/or 3’ ends of these cDNA clones are subsequently sequenced [193, 194]. Base calling software reads the generated sequencer traces and converts them to inferred base sequences. Phred is the most popular and widely used base calling software and also ascribes each base a quality score, which can be used to excise low quality bases from the sequences [195, 196]. These EST sequences need to be further preprocessed in order to remove vector, linker or adapter fragments as well as poly(A) tails and contaminant DNA, which would impair
1.5 Expressed Sequence Tag Meta-analysis

subsequent clustering and assembly. BLAST searches [197], tools such as Cross_match [198] and Lucy2 [199], or the EST preprocessing EMBoss packages [200] can identify such sequence fragments for removal. Also low complexity and repetitive regions should be masked using tools such as DUST [201] or RepeatMasker [202], as they may lead to erroneous sequence assembly. After preprocessing, EST clustering and assembly are necessary to group overlapping ESTs deriving from transcripts of a single gene based on sequence similarity and to compute consensus sequences. Among the more common methods are CAP3 [203], MIRA [204] and Phrap [198]. Pipelined analysis tools such as EGassembler [205], ESTAP [206] and the commercial software suite SeqMan from DNASTAR, Inc. incorporate preprocessing steps, sequence clustering and assembly.

Finally, the assembled consensus sequences may be searched against databases of annotated gene or protein sequences to annotate and to assign functionality if possible [194]. The sequences can either be searched directly against a nucleotide sequence database using BLAST [197] or against a protein sequence database using BLASTX, which first translates the sequences in all six reading frames. Moreover, quantitative expression levels may be inferred from EST data, as the number of tags is proportional to the abundance of gene transcripts of the cell or tissue type analysed [207]. To compare EST libraries with certain library sizes $n$, the EST counts $m_g$ for a given gene $g$ need to be converted to transcripts per million $tpm_g$:

$$tpm_g = \frac{m_g}{n} \cdot 10^6$$

To evaluate the significance of differential expression between two libraries either the Audic and Claverie’s test [207] or the Fisher’s exact test [208] may be used. For comparing more than two libraries, significance calculations such as the R-statistic may be employed [209].

1.5.3 Expressed Sequence Tag Meta-analysis

The large amount of EST data available in public repositories provides researchers the opportunity to retrieve, integrate and investigate the data. So-called meta-analysis techniques aim to combine the data available and integrate information from multiple independent EST libraries [210].

Advantages of Meta-analysis and Its Objective

The advantages and objectives of an EST meta-analysis may be severalfold. Combining information over different independent but related studies (e.g., a given tissue type) in-
creases the statistical power, enhances the reliability and generalizability of the results and compensates for artefacts of individual studies such as biological, experimental, and technological variations [210]. Integrated EST data analysis across independent but unrelated studies (e.g., across numerous tissue types) can reveal comprehensive gene expression profiles, which may be used to uncover information about tissue- and disease-specific expression [211]. Without data integration, the establishment of such comprehensive expression maps for the complete human body would pose an immense challenge due to the difficulty of obtaining such data empirically. As tissue-specific gene expression plays a fundamental role in human biology and disease, the identification of genes with restricted/specific expression patterns helps to understand development, function and homeostasis of the distinct cell/tissue types as well as aetiology, gene-tissue relationships and gene functions [212–214]. For example, Stanton and Green [215] identified stage- and developmental-specific genes expressed during preimplantation embryo development by means of EST data integration, whereas Yu et al. [216] analysed EST data to determine tissue-specific genes for numerous tissue types as well as tissue-specific combinatorial gene regulation. Identification of genes with certain tissue- and disease-specific expression patterns such as the cancer testis (CT) genes can also aid the discovery of novel marker, prognostic or therapeutic target genes [57, 58, 185, 186]; for example, Kim et al. [217], Hofmann et al. [58], and Campagne and Skrabanek [186] identified potential cancer biomarkers using EST data integration.

**EST Data Resources**

In the last decade, large amounts of EST data have been deposited in public repositories with dbEST [218] being the largest one, which currently holds records of 8,692,773 human ESTs [219]. Unigene [220–222] and DCFI Gene Indices (previously TIGR Gene Indices) [223, 224] have grouped these expression data into clusters and assigned them to genes, facilitating the indexing of the EST data. Other specialist EST repositories such as the Plant Genome database (PlantGDB) [225] have been developed and mainly focus on one or a group of organisms [194].

Establishing indexed EST data comprises various steps including data preprocessing, clustering, cluster joining, assembly and consensus sequence computation [194, 226]. Clustering and assembling the large amounts of EST data available in dbEST poses an immense challenge in contrast to individual EST projects. Clustering in general is the process of finding sets of sequence fragments that belong together and arise from transcripts of a single gene [194]. It is usually based on pairwise sequence similarity, which is converted to binary values depending on whether the match is significant or not [194,
In general, two main approaches are employed to cluster EST data: (i) stringent clustering; and (ii) loose clustering. Stringent clustering employs single pass clustering, which results in higher quality clusters but lower coverage. Loose clustering is less conservative and repeats the clustering process many times, producing larger clusters that may contain paralogous sequences. The Unigene database adopts a loose clustering approach, whereas the DCFI Gene Indices database is based on stringent clustering [194, 226, 227].

Unigene is by far the most popular and most frequently updated database holding clustered and indexed EST data and covers numerous organisms [222, 227]. EST sequences derive from dbEST and are preprocessed prior to clustering. Unigene’s clustering algorithm initially builds clusters by comparing gene sequences (mRNA or genomic sequences from Genbank). By comparing EST sequences with the initial clusters using megaBLAST [228], significant similar sequences are added to the clusters. EST to EST sequence comparisons further extend the initial clusters. Overlapping clusters may be merged, and generated clusters without at least one sequence containing a poly(A) tail are discarded, resulting in so-called anchored clusters. ESTs not assigned to an anchored cluster undergo a further round of sequence comparisons at a lower level of stringency [222, 227].

The DCFI Gene Indices database provides additionally to clustered and indexed EST data, consensus sequences and also covers multiple organisms [223, 224]. The clusters are generated using their in-house analysis pipeline, TGICL [229]. The underlying EST data is also extracted from dbEST and subsequently preprocessed. The clusters are seeded using gene sequences (mRNA or coding sequences from Genbank). Pairwise similarity searches by means of megaBLAST [228] compare EST and gene sequences, adding sequences to clusters which are sufficiently similar. Finally, the clusters are assembled using CAP3 [203], producing tentative consensus sequences [223, 224, 229].

**Data Collection, Quality and Preprocessing**

There are two possibilities for EST data collection: (i) using the previously clustered and indexed EST data from repositories such as Unigene; or (ii) selecting raw EST libraries of interest. The disadvantage of the latter is that the data still needs to be preprocessed, clustered and indexed, but allows employing preprocessing parameters and clustering algorithms of choice (cf. section 1.5.2). In general, normalised and subtracted cDNA libraries need to be excluded, as they are not suitable for integration and impair quantitative gene expression analysis based on EST counts [211].
Analysis Approaches and Tools

Several approaches and tools exist that exploit EST data to construct integrated expression profiles. However, compared with microarray meta-analysis approaches, these are underrepresented. The underlying principle of EST meta-analysis mostly relies on generating meta-libraries by merging all ESTs belonging to a given tissue. These meta-libraries should contain more than 10,000 ESTs to ensure reliable quantification and comparisons. Furthermore, normalisation of the expression levels of each meta-library for a given tissue/cell type is required in order to make the different meta-libraries comparable with each other [211].

The significance of differential expression between two conditions (e.g., cancerous vs. normal tissue) may be evaluated using the Fisher’s exact test [208] or other significance calculations [207, 209].

The tool TissueInfo [187] allows determining the tissue-specificity for a given gene $g$ or tissue-specific genes for a given tissue $t$. The tool relies on the principle of meta-libraries as described above. A given gene $g$ is considered as expressed if $m_{t,g} > 0$, where $m_{t,g}$ is the number of ESTs for gene $g$ corresponding to the tissue $t$. The tissue-specificity is computed by testing if this gene $g$ is predominately expressed in a given tissue $t$ with a stringency parameter $\alpha$, where $a_g$ is the total number of ESTs belonging to gene $g$:

$$\frac{m_{t,g}}{a_g} > \alpha$$

Similarly, the web tool TiGER [216, 230] allows the evaluation of the tissue-specificity for a given gene $g$ or provides tissue-specific genes for a given tissue $t$, but also contains information about combinatorial regulation and cis-regulatory modules. They calculate the expected number of ESTs $f_{t,g}$ for a given gene $g$ in all $k$ tissues, where $a_g$ is the total number of ESTs belonging to gene $g$ and $n_t$ is the library size of tissue $t$:

$$f_{t,g} = a_g \cdot p_t$$

$$p_t = \frac{n_t}{\sum_{i=0}^{k} n_t}$$
The expression enrichment $e_{t,g}$ is subsequently computed as follows, where $m_{t,g}$ is the number of ESTs for gene $g$ corresponding to the tissue $t$:

$$ e_{t,g} = \frac{m_{t,g}}{f_{t,g}} $$

The expression enrichment $e_{t,g}$ is the ratio between observed to expected number of tags for a given gene $g$ in a given tissue $t$. Genes are considered as tissue-specific, if the significance of enrichment is below and the expression enrichment above a certain threshold [216].

Unigene itself offers two tools, Digital Differential Display (DDD) and the EST Profile Viewer. In contrast to TiGER and TissueInfo, DDD [222] compares EST profiles of user-defined EST libraries to identify genes with significantly different expression levels, and the EST Profile Viewer [222] shows the approximate expression profile for a given gene. Both tools rely on the generation of meta-libraries. DDD ranks differentially expressed genes of two user-defined meta-libraries according to the significance of differential expression computed by the Fisher’s exact test [208, 222]. The EST Profile Viewer establishes a comprehensive expression profile for a given gene $g$ by means of EST counting. The EST counts are normalised by calculating the transcripts per million $tpm_{t,g}$, where $m_{t,g}$ is the number of ESTs for a given gene $g$ and for a given tissue type $t$, and $n_t$ is the total number of ESTs for that given tissue type $t$ compiled in a meta-library [222]:

$$ tpm_{t,g} = \frac{m_{t,g}}{n_t} \cdot 10^6 $$

Several other tools were published but appear to be currently unavailable (DigiNorthern [231], ZooDDD [232], GBA server [233]).

Comparison of Analysis Approaches and Tools

Most tools evaluate the tissue-specificity of one gene at a time such as the EST Profile Viewer, TiGER and TissueInfo, and do not allow the analysis of a set of related genes. TiGER provides tissue-specific genes of one user-defined tissue type, whereas TissueInfo allows the selection of several tissues for computation of tissue-specific genes. In general, TiGER and TissueInfo do not compute tight tissue-specific genes, as their calculations produce highly enriched or predominately expressed genes for a given tissue type. The EST Profile Viewer does not provide a tissue-specific list as it only focuses on the EST profile of a certain gene of interest. Furthermore, TiGER and TissueInfo do not evaluate disease-specific expression, whereas the EST Profile Viewer breaks the expression profile
down according to the tissue type, health state and developmental stage. DDD allows user-defined EST library selection and focuses on differential expression between these two EST pools, but does not provide information about tissue-specificity.
2 Aims and Objectives

Tumour antigens (TAs) provide the basis for immunodiagnostics and immunotherapeutic tools. Their identification is a key element in this developing field and can be facilitated by the use of in silico screening pipelines prior to experimental validation, decreasing the risk of pursuing unsuitable clinical targets. Among the most promising TAs are a group of proteins encoded by testis-specific genes, the cancer testis (CT) genes, whose frequent expression in cancer could reflect the aberrant induction of a silenced gametogenic programme in cancer cells. Some of these CT antigens could function in the testes to mediate the meiotic programme. Their aberrant activation in somatic cells could lead to oncogenic changes, which in turn drive tumorigenesis. Thus, the primary purpose of this project is to first develop an integrative bioinformatic analytical approach to automate and optimise the identification of novel CT candidate genes. Once this has been established, the study can move on to the second aim, testing the hypothesis that a group of meiosis-specific genes is aberrantly expressed in cancer, forming a novel subset of the CT genes.

Specific Objectives

1. Establish in silico screening pipelines for automated microarray and expressed sequence tag (EST) meta-analysis across publicly available data for a gene set of interest with intuitive, user-friendly web interfaces.

2. Validate the pipelines with published experimentally derived data.


4. Use the pipelines to analyse and interpret the expression of germline-associated gene sets to identify new CT candidate genes.
3 Meta-analysis of Clinical Data Using Human Meiotic Genes Identifies a Novel Cohort of Highly Restricted Cancer-specific Marker Genes

This chapter describes the development and employment of a bioinformatic analytical strategy analysing high throughput expression data to identify new cancer testis (CT) genes, leading to the exposure of a novel group of meiosis-specific genes, the meiCT genes, which are aberrantly expressed in a wide range of cancers and represent potential clinically relevant cancer biomarkers. Their associated proteins might have oncogenic features and could serve as targets for novel cancer diagnostic, prognostic and therapeutic strategies. The work presented in this chapter contributes to project objectives 1, 3 and 4.

Please note that all practical work performed for this paper was carried out by fellow Ph.D. students (Ibrahim Aldeajel, Rebecca Anderson, Mikhlid Almutairi, Ahmed Almatrari and Naif Alsiwiehri). This chapter is presented as paper published in the open-access journal Oncotarget (available at: [http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=article&op=view&path[]=580](http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=article&op=view&path[]=580)) [57]. The content structure, layout, language and reference style follow the specifications of Oncotarget.
Meta-analysis of clinical data using human meiotic genes identifies a novel cohort of highly restricted cancer-specific marker genes

Julia Feichtinger1, Ibrahim Aldeailej1,*, Rebecca Anderson1,*, Mikhlid Almutairi1, Ahmed Almatrafi1, Naif Alsawihehri1, Keith Griffiths2, Nicholas Stuart1,3,5, Jane A. Wakeman1, Lee Larcombe4 and Ramsay J. McFarlane1,5

1 North West Cancer Research Fund Institute, Bangor University, Bangor, LL57 2UW, UK
2 Therapies and Health Sciences, Betsi Cadwaladr University Health Board, Bangor, LL57 2UW, UK
3 Medical Sciences, Bangor University, Bangor, LL57 2UW, UK
4 Bioinformatics Group, Cranfield Health, Cranfield University, Beds, MK43 0AL, UK
5 NISCHR Cancer Genetics Biomedical Research Unit

* These authors made an equal contribution

Correspondence to: Jane A. Wakeman/Ramsay J. McFarlane. e-mails: j.a.wakeman@bangor.ac.uk/r.macfarlane@bangor.ac.uk

Keywords: cancer biomarkers, cancer testes antigens, oncogenes, meiosis, PRDM9, cohesins drug targets

Received: July 26, 2012, Accepted: August 2, 2012, Published: August 13, 2012

ABSTRACT:

Identifying cancer-specific biomarkers represents an ongoing challenge to the development of novel cancer diagnostic, prognostic and therapeutic strategies. Cancer/testis (CT) genes are an important gene family with expression tightly restricted to the testis in normal individuals but which can also be activated in cancers. Here we develop a pipeline to identify new CT genes. We analysed and validated expression profiles of human meiotic genes in normal and cancerous tissue followed by meta-analyses of clinical data sets from a range of tumour types resulting in the identification of a large cohort of highly specific cancer biomarker genes, including the recombination hot spot activator PRDM9 and the meiotic cohesin genes SMC1beta and RAD21L. These genes not only provide excellent cancer biomarkers for diagnostics and prognostics, but may serve as oncogenes and have excellent drug targeting potential.

INTRODUCTION

The demarcation of neoplastic cells from healthy tissue represents an important goal in clinical oncology; this is of particular interest given the need for diagnostic markers to enable early intervention strategies, such as surgical resection, and the re-emergence of immunotherapeutics, cancer vaccines and targeted drug delivery via antibody-drug conjugates [1-10]. To achieve this goal, the identification of tumour-associated antigens is of central importance [for example, see 11-13]. Whilst almost all cancer cells have an altered gene expression profile, including many up regulated genes, most of the associated antigens are recognised as ‘self’ by the immune system, limiting their use in immune therapeutic, prognostic and diagnostic technologies. One family of proteins, the so-called cancer/testis (CT) antigens, represents an excellent group of cancer-specific biomarkers [14-21]. These are produced in the testes of healthy male adults and can also be found in cells with a cancerous phenotype. The immunological privilege of the testes [22,23] makes the CT antigens excellent immunological targets and a number of CT antigens have been employed successfully in a range of clinical applications, including adoptive therapeutics for late stage cancer treatment [for example, see 24]. Some CT antigens are also present in other immunologically privileged tissues of the central nervous system.
CNS (CT/CNS) antigens [25]. Many genes have been purported to encode CT antigens [21], however, not all of these have endured continued scrutiny and many of the genes have subsequently been found to have some degree of expression in normal somatic tissues [25]. This has lead to the redefining of CT genes into testis (and CNS)-restricted and testis (and CNS)-selective, where there is some evidence to indicate the latter class are expressed in at least one non-immune privileged, normal tissue type [25,26].

CT antigens have been further sub-classified into those which are encoded by genes on the X chromosome (X-CT genes) and those which are encoded by genes on autosomes (non-X-CT genes) [14-21]. The majority of characterised testis-restricted CT genes are X-CT genes and many of these reside within large families of orthologous genes, such as the MAGE family [14-21,27]. In addition, some CT genes are co-expressed in the same cancerous tissue, suggesting a dysfunction in one or more, as yet uncharacterised, testis-specific transcriptional regulatory pathway(s) [for example, see 28].

It has been demonstrated that some CT antigens have the potential for oncogenic activity or contribute to maintaining or enhancing the neoplastic state [19]. For example, MAGE-A2 has been demonstrated to induce the down regulation of one of the primary tumour suppressor genes, p53 [29]. Furthermore, MAGE-A2 and another MAGE family member, MAGE-A6, have been demonstrated to have the potential to induce resistance to chemotherapeutic agents [30]. However, the function, the oncogenic activity and the drug resistance-inducing potential of CT antigens remains poorly studied considering the potential importance of these proteins.

There has been speculation that some CT antigens could function in the testes to mediate the meiotic programme [31,32]. During meiosis the chromosomes of diploid progenitor cells (spermatogonia in testis) become reductionally segregated to produce haploid gametes (sperm cells in testis) [33,34]. This meiotic chromosome segregation involves a complex and poorly understood series of events, which include the pairing of homologous chromosomes followed by a covalent conjoining to generate a bivalent which is required for chromatid alignment at the first meiotic division. It has been postulated that the aberrant production of CT antigens with chromosome modulating potential in mitotically dividing somatic cells could result in inappropriate nonallelic inter-/intra-chromosomal recombination and inter-homologue recombination events which could generate oncogenic genetic changes such as translocations and losses of heterozygocity [21,31,32]. In addition, the aberrant expression of meiotic chromosome regulators in matched induced pluripotent stem cells (iPSCs) has been demonstrated to illicit an immune response to iPSC-induced teratomas in mice indicating a broader importance to understanding the consequences of aberrant expression of meiotic genes [35].

In male mammals there exists a unique mechanism for the meiosis-specific transcriptional silencing of the X chromosome during the meiotic zygotene to pachytenene transition, which is dependent upon meiotic double-strand break formation in unpaired chromatin [36]. This meiotic X inactivation suggests that most of the genes encoding known testis-restricted CT antigens are silenced during meiosis, as most of these are X-encoded and so may have largely non-meiotic roles in the testis.

These findings lead us to postulate that there is a family of human meiosis-specific genes, which are autosomally encoded and therefore not subjected to meiotic X inactivation. If these genes are aberrantly expressed in cancers and iPSCs they might represent a clinically important, novel sub-class of the testis-restricted CT gene family. Moreover, we speculated that such genes might have oncogenic activity by encoding proteins which interfere with chromosome dynamics and cell division when aberrantly expressed in mitotically dividing somatic cells. Here we identify human meiosis-specific genes showing the characteristics of CT genes, which we designate meiCT genes. This work defines a novel, meiosis-specific sub-class of clinically-relevant CT genes which includes previously uncharacterised human testis-specific genes, the human meiotic hotspot regulator gene PRDM9, the meiotic regulator gene STRA8 and meiosis-specific sister chromatid cohesion regulator genes.

RESULTS

Analysis of selected meiotic chromosome regulatory genes for CT gene candidate

Some important meiosis-specific genes which encode chromosome modulators have previously been reported to be CT genes, including SPO11 which encodes a meiosis-specific nuclease required for the initiation of meiotic recombination [15]; however, many of these previously identified meiotic regulators (including, SPO11, HORMADI, SYCE1, SYCP1) have subsequently been found to be selective in their expression profile, suggesting they are not strictly testis-restricted [25]. As a first step to address the possibility that additional meiosis-specific genes might encode highly restricted CT antigens, we selected from the literature a sample cohort of human genes predicted to have meiosis-specific expression (Supplementary Table S1). These included genes encoding subunits of the meiosis-specific cohesin complex (REC8, STAG3, SMC1beta, RAD21L), which is responsible for modulating meiotic
recombination, meiotic centromere monopolarity and meiotic sister chromatid cohesion [37,38]. To assess the meiotic specificity of the selected genes, we obtained RNA extracted from a panel of normal human tissues, including testis, ovary and CNS tissue. We designed intron-spanning primer sets for the genes of interest and carried out reverse transcriptase polymerase chain reaction (RT-PCR) analysis as described in the materials and methods. Surprisingly, a number of these genes were expressed in a wide range of normal tissues, suggesting that their expression is not exclusively meiosis-specific (Fig. 1B; Supplementary Fig. S1A). These included two genes encoding cohesin components, \textit{REC8} and \textit{STAG3}, the protein products of which are widely accepted as being meiosis-specific from studies in other organisms, including the mouse [for example, 39]. Taking this further, we detected some expression of both \textit{REC8} and \textit{STAG3} in mouse non-meiotic somatic tissue using our RT-PCR conditions (Supplementary Fig. S1B), suggesting that expression of these genes is not fully restricted to meiosis, rather, they are subjected to a meiotic up regulation, consistent with previous analyses in the mouse [for example, 40]. DNA sequencing was used to confirm the identity of all RT-PCR products from both mouse and human. RT-PCR followed by DNA sequencing confirmed that a number of other genes we had postulated would be testis / meiosis-specific were expressed more extensively in non-meiotic tissues (Supplementary Fig. S1A; Supplementary Table S1). Despite the detection of \textit{REC8} and \textit{STAG3} expression in normal tissue, the other two meiosis cohesin genes tested, \textit{RAD21L} and \textit{SMC1-beta}, exhibited a tight testis-specific expression profile (Fig. 1A). To explore whether these testis-specific cohesin genes could potentially encode CT antigens we analysed their expression using RT-PCR on RNA samples taken from cancer cell lines and tumour samples from a range of cancer types. For both genes, expression was detected in a number of the cancer cells indicating they are CT genes (Fig. 1A). Of the other

Figure 1: Examples of gene expression and protein production profiles for predicted meiCT genes. A. Agarose gels showing RT-PCR profiles generated from a range of normal human tissues obtained post mortem (left two panels) and cDNA generated from a range of cancer cell lines or solid tumours (right hand two panels). The expression profile for \textit{betaACT} is a positive control (top row). The profile for \textit{SSX2} provides an example of a previously characterised X-CT gene. Five examples of testis-restricted meiCT genes are shown (\textit{RAD21L}/\textit{SMC1beta}/\textit{PRDM9}/\textit{C1orf65}/\textit{STRA8}) along with the expression profile of one testis selective meiCT gene, \textit{TEX19}. The \textit{C5orf47} profile provides an example of genes which were testis-restricted with no evidence of expression in any of the cancer cells tested. B. Agarose gels showing RT-PCR profiles for normal human tissues for the mitotic cohesion gene \textit{RAD21} and the two cohesin genes \textit{REC8} and \textit{STAG3}. C. Western blots showing the presence of the PRDM9 protein in the cancer cell line NTERA-2, but not in primary cultures of human prostate smooth muscle cells (HPrSMC).
manually selected genes a further five exhibited a CT gene expression profile (a mixture of cancer/testis-restricted and cancer/testis/CNS-restricted); these were the meiotic recombination hotspot activator gene PRDM9, the nuclear protein in testis (NUT) gene, the testis-specific serine/threonine kinase 1 (TSSK1) gene, the synaptonemal complex component gene SYCP1 (which has previously been reported as a CT gene [18]) and the meiotic regulatory gene STRA8 (Fig. 1A). A sixth gene, TEX19, exhibited a testis-selective expression profile as it was expressed in the testis and the thymus, the latter being a tissue known to undergo atrophy in older individuals such as those from which this tissue was derived, which may account for the expression of this gene in the thymus; TEX19 also exhibited extensive expression in many cancer types (Fig. 1A).

To address whether an expressed gene might be translated into a protein product, which might therefore provide an antigenic target in clinical applications, we carried out western blot analysis to detect the protein product of one of the CT genes identified above, PRDM9, for which commercial antibodies were available. The intracellular nature of these antigens does not preclude them from serving as targets for monoclonal antibody therapies or other immunotherapeutic approaches [for example, 24,41,42]. We generated whole cell extracts (WCEs) from one of the cancer cell lines in which PRDM9 gene expression had been observed, NTERA-2 (Fig. 1A), and a culture of non-cancerous primary human prostate smooth muscle (HPrSM) cells, in which no PRDM9 gene expression could be detected by RT-PCR. PRDM9 was readily detectable in the cancer cells, but not in the HPrSM cells (Fig. 1C) demonstrating that the expression of the PRDM9 CT gene results in protein production, leading to the possibility that the de-repression of the PRDM9 gene generates a protein which could be antigenic and thus be of clinical and oncogenic importance.

**Identification and validation of novel meiosis-associated CT genes using computational analysis of EST data**

Whilst the above approach has identified a number of new CT genes, it is limited by the fact that a manual curation of the literature is not only time-consuming, but also exposes relatively few mammalian meiosis-specific genes. We took advantage of a previous large scale microarray study which identified an extensive cohort of genes with expression associated specifically with meiosis and spermatocyte development in mammals [43]. We used this to conduct a systematic approach to identify new meiosis-associated human CT genes. The mouse study provided a starting point of 744 mammalian meiosis-specific genes. After human orthologue assignment and filtering to eliminate non-testis-specific genes, 375 human genes remained and these were fed into an expressed sequence tag (EST) analysis pipeline based on the complete Unigene database (Fig. 2; support text to Fig. 2 is given in Supplementary Information).

**Figure 2:** Schematic flow diagram of the approach used for the selection of candidate meiCT genes. Based on a large scale microarray study [43], 744 mouse meiosis-specific genes were selected as a starting point: 408 human orthologues could be identified and 375 human meiosis-specific genes remained after filtering to eliminate non-testis-specific genes. All 375 candidates as well as 3 controls (MAGE-A1, GAGE1 and SSX2) were fed into the EST analysis pipeline, which returned 105 candidate genes which were subjected to RT-PCR validation/microarray meta-analysis. Support text to Fig. 2 is given in Supplementary Information.
Supplementary Information). Briefly, if a candidate gene is represented in a non-testis / non-CNS normal tissue EST library, it was dismissed. The remaining genes were further assessed for representation in cancer EST libraries. This screen identified 177 candidate genes, of which 9 were cancer/testis-restricted (class 1), 75 were testis-restricted (class 2), 21 were cancer/testis/CNS-restricted (class 3) and 72 were testis/CNS-restricted (class 4). We favoured an EST screen, since microarray technology is limited by the number of published cancer arrays available as well as by the number of genes which can be analysed due to lack of gene coverage on arrays. Moreover, an EST screen can confirm the testis-restricted expression pattern and therefore functions as an additional filter to eliminate non-testis-restricted genes. However, microarray data sets were not ignored and following experimental validation of the candidates (see below) we carried out meta-analysis of clinically-relevant cancer microarray data sets (see below).

Having identified candidates in the four classes outlined above we validated those in classes 1-3 using RT-PCR. We included the genes in class 2, which are predicted to be testis-specific, but have not been identified in the EST data sets of cancer tissue. We initially carried out RT-PCR on RNA isolated from a range of normal human tissues, including testis-derived RNA, as described above. Of the 105 genes in classes 1-3 we could not obtain RT-PCR products in our control tissue (testis) for 12 genes, resulting in 93 genes which were subjected to RT-PCR validation. Of these, 39 genes were expressed in more than two non-testis/ CNS normal tissues and were therefore dismissed at this stage. Of the remaining 54 genes, 41 had expression restricted to testis in normal tissues, 3 had expression restricted to testis and CNS tissue and 10 were testis-specific, or testis/CNS-specific and yet exhibited expression in one or two normal tissues. RT-PCR analysis was carried out to assess the expression profiles of these 54 genes in cancer cells, as above. From these analyses it was determined that 29 genes exhibited no expression in any of the cancerous material and appeared to be tightly testis-specific (Supplementary Table S2; the example of C5orf47 is shown in Fig. 1A); 12 were CT-restricted genes and they were expressed in the testis and at least one cancer type (the example of C1orf65 is shown in Fig. 1A; Fig. 3); 3 were cancer/testis/CNS-restricted genes as they were expressed in testis, CNS and at least one cancer type (Fig. 3); 6 were cancer/testis-selective as they were expressed in testis, one or two other non-testis/CNS normal tissues and at least one cancer type (Fig. 3); 4 were cancer/testis/CNS-selective as they were expressed in testis, CNS tissue, one or two other normal tissue types and at least one cancer type (Fig. 3). This resulted in the identification of a total of 25 genes distributed in the various CT classes. This, in combination with the 8 CT genes identified in the preliminary study (see above) resulted in the identification of 33 restricted / selective CT or CT/CNS genes, most of which have not been previously characterised as CT genes and are largely autosomally encoded; we will refer to these as meiCT genes (Fig. 3).

Meta-analysis of validated candidate genes

To explore the clinical relevance of the 33 meiCT genes we have identified, we developed a meta-analysis pipeline for patient-derived cancer microarray data including 13 cancer types (Fig. 2; support text to Fig. 2 is given in Supplementary Information; Supplementary Table S3). We analysed the meta-change in gene expression of patient-derived, untreated cancerous tissues compared to normal tissues for the candidates as well as for 3 known X-CT control genes (MAGE-A1, GAGE, SSX2) in a total of 80 microarray data sets (Supplementary Table S3). Of the 33 candidates, 25 were covered by the array sets and could be evaluated (Supplementary Table S4). This revealed that 15 of the meiCT genes exhibited statistically significant, cancer-specific mean up regulation in at least one cancer type for combined data sets for specific cancer types where enough clinically-derived data sets were available (Fig. 4; Supplementary Table S4). The Circos plot (Fig. 4) shows the meta-change in gene expression in relation to the corresponding cancer type. This provides evidence that the meiCT genes are expressed in clinically-relevant material and shows examples of more extensive tumour expression patterns. Some notable patterns emerge from these analyses; firstly, many of the meiCT genes show a mean up regulation in ovarian, brain and lung cancers; secondly, a number of cancer types exhibit no mean up regulation of any of the analysed meiCT genes, these include breast and colorectal, for which 11 and 13 microarray data sets were available. However, a limited number of microarray data sets were available for many cancer types and thus designating cancer specificity from these data has limitations.

Whilst a significant mean up regulation is observed for a number of genes in combined data sets for distinct cancer types (Fig. 4), this does not reflect a uniform up regulation of a specific gene in all samples for a given cancer type. For example, PRDM9 exhibits a significant mean up regulation in the ovarian cancer microarray sets used (Fig. 4); however, it is not significantly up regulated in all the individual cancer samples tested, despite the significant mean elevation (Fig. 5). This indicates that these markers may not be universally up regulated in specific cancer sub-types or cancer samples. Extending this, we determined that a further number of clinically-derived cancer samples exhibited up regulation of a wider range of
the 25 meiCT genes represented on the arrays in individual (not combined) cancer data sets (cancer vs. normal) (Supplementary Fig. S2A) indicating the meiCT genes are expressed in clinically-relevant samples covering a broad range of cancer types.

The meta-analysis approach was extended to address whether any of the 29 genes ascribed as testis-specific (no evidence for expression in any of the cancer cells we tested) by RT-PCR analysis, were upregulated in the clinically-derived microarray data sets. Of the 29 genes, 21 were represented on the arrays (Supplementary Table S2). Meta-analysis of combined cancer data sets revealed that 9 of these genes showed a significant mean up regulation in leukaemias and lung and ovarian cancers (Fig. 6). These findings indicate that these further 9 genes qualify as meiCT genes, bringing the total number of meiCT genes identified in this study to 42 (Supplementary Table S5), many of which are novel genes which have not been classified as cancer biomarkers. Additionally, analysis of individual (not combined) cancer data sets revealed up regulation of 19 of the 21 genes in a broader range of cancer types (Supplementary Fig. S2B), indicating that a further 10 genes could be considered as meiCT genes.
DISCUSSION

The restricted regulation of CT genes has resulted in the emergence of their associated antigens as important oncological biomarkers. However, the classification of CT genes remains fraught with difficulties and it has been proposed that a uniform classification is premature until a greater insight into the biology and clinical importance of CT genes is revealed [25]. Here we add a new group of genes to this expanding family, the meiCT genes, which not only have expression restricted to the testis, but are likely to be further restricted to the highly immunologically privileged meiotic spermatocytes. These are more widely represented on the autosomes than previously characterised CT genes (45 of the 52 meiCT genes are autosomally encoded; Supplementary Table S5) and their identification opens up new possibilities in terms of both tumour distribution and oncogenic activities. Analysis of the meiCT genes demonstrates that these are expressed in a wide range of cancer types. For example, our RT-PCR validation demonstrates expression of a number of meiCT genes, including PRDM9, in lymphoma and leukaemia lines (Fig. 1A; Fig. 3). Of the 46 genes subjected to meta-analysis (combined cancer data sets), 20 were expressed in ovarian cancers. The use of CT antigens for immunotherapies to treat ovarian cancers
has yielded positive results (for example, see 44), and so the finding that the meiCT genes are extensively expressed in ovarian cancers could provide an additional suite of markers for a tumour type which is amenable to immunotherapeutic approaches. In addition to the meta-analysis, study of individual cancer data sets (cancer vs. normal) suggested expression of additional meiCT genes (42 out of the 44 represented on microarrays) in a broader range of cancer sub-types (Fig. 6; Supplementary Fig. S2; Supplementary Table S5); however, the extent to which these single data set analyses reflect extensive expression in a given cancer type will rely on the generation of further clinically-derived data sets and their subsequent analyses, whereas meta-analyses are generally accepted to indicate a more precise and reliable estimate of gene expression for a given cancer type.

In our original computational analysis we applied relatively stringent conditions to the classification of the meiCT genes. We only selected candidates for validation which were not represented in EST libraries of any non-testis / non-CNS normal tissue types. Chen and co-workers [26] challenged EST data sets with testis-specific genes identified by massive parallel sequencing and retained those genes with expression in one or two other non-testis normal tissues. Indeed, when we re-ran our analysis setting the criteria with this same lower stringency we identified a significant number of additional candidates. We took the more stringent approach so as to target meiCT genes which were tightly restricted CT genes. However, on validation we revealed a number of meiCT genes which fall into the CT (and CT/CNS) selective class. This might be due to the nature of the so-called normal tissue; as in our case, RNA from many normal tissues are extracted from tissue obtained post mortem and are often pooled from tissues from a number of individuals, many of whom were aged at time of death. It remains a possibility that some of these tissues had undergone undiagnosed neoplastic change and might have been aberrantly expressing one or more of the candidate genes. In support of this, Chen and co-workers [26] observed expression of some genes in tissues from one panel of normal tissues, but detected no measurable expression in similar tissue types from a distinct second source. Thus, genes which exhibit meiCT selective profiles, such as TEX19, might indeed be CT restricted genes and be of clinical use.

**Meiotic chromosome regulators as CT genes**

Here we have identified a number of previously uncharacterised genes as meiCT genes; for example, C12orf12. However, we also find that a number of relatively well characterised meiotic genes are meiCT genes. It has been previously proposed that the aberrant expression of CT genes may have an oncogenic effect [18,32] and indeed, aberrant expression of germ line genes in Drosophila contributes to malignant growth [45]; when this idea is applied to the genes identified here it opens up some interesting possibilities, which might indicate that the meiCT genes might not only be oncogenic, but might also provide drug targeting opportunities. For example, the meiotic cohesin genes RAD21L and SMC1beta may produce proteins which are incorporated into functional cohesin complexes within mitotically dividing tissues; this may not only result in aberrant modulation of chromosome segregation resulting in genome instability, but might also provide a cancer cell-specific drug target to inhibit chromosome segregation.

The expression of the meiotic recombination hotspot activator gene PRDM9 is intriguing as the gene product is a sequence-specific zinc finger histone methyltransferase known to regulate the epigenetic programme for hotspot chromatin activation [46,47] and in mice the orthologue, Meisetz, has a function in transcriptional regulation where it activates expression of the testis-specific RIK gene, amongst others [48]. We could find no evidence that any of the human orthologues of RIK were differentially activated in cancer cells expressing PRDM9, but the possibility remains that active PRDM9 protein in somatic cells might trigger unscheduled transcriptional activity and/or generate regions with altered chromatin structure which could form unstable chromatin lesions, both of which could be oncogenic in nature.

In addition to identifying the meiCT genes, we found expression in non-testis tissues of a number of genes which are reported as meiosis-specific, including REC8 and STAG3. This is not inconsistent with previous studies, where these genes are reported to be up regulated in the testis and are not testis-restricted. Why might some genes, which encode meiosis-specific functions, be less tightly regulated than others? The answer to this could come from studies in the fission yeast were the production of Rec8 protein in mitotic cells is inhibited by specific post-transcriptional mRNA degradation [49,50]. If an analogous system were operating in mammals then many CT antigen genes might be missed using transcriptional profiling alone; REC8 and STAG3 might prove to be good genes on which to test this idea. This raises the possibility that CT antigens can be generated not only by transcriptional dysfunction, but also by the de-regulation of translational repression programmes which ensure spermatocyte/testis-specific translation.

**Conclusions**

Here we have characterised a sub-class of a clinically-important family of genes and identified a large number of previously unclassified/uncharacterised genes.
as potential clinically-relevant cancer biomarkers. Their identification also exposes a new cohort of genes which might have oncogenic characteristics, whose protein products might not only serve as targets for immune therapeutics, but also as new drug targets and oncogenic drivers.

**MATERIALS AND METHODS**

**Cell lines and cell culture**

The NTERA-2 (clone D1) cell line was gifted by Prof. P.W. Andrews (University of Sheffield) and are regularly authenticated within the group using standard antibody tests using anti-OCT4 antibodies and retinoic acid-induced differentiation. The A2780 cell line was provided by Prof. P. Workman (Cancer Research UK Centre for Cancer Therapeutics, Surrey, UK) and was authenticated at source. The following cell lines were purchased from the European Collection of Cell Cultures (ECACC): 1321N1, COLO800, COLO857, G-361, HCT116, HT29, LoVo, MM127, SW480 and T84. H460 was purchased from the American Type Culture Collection (ATCC), and the two ovarian adenocarcinoma cell lines, PEO14 and TO14, were obtained from Cancer Research Technology Ltd. Primary cultures of proliferating human prostate smooth muscle cells were obtained from PromoCell™ (C-12574). All cultures were used within a six month period of obtaining validated lines from external sources.

1321N1, A2780, NTERA-2 (clone D1) and SW480 cell lines were cultured in Invitrogen Dubecoo’s modified Eagle’s medium (DMEM + GLATAMAX™) supplemented with 10% foetal bovine serum (FBS). COLO800, COLO857 and H460 cell lines were cultured in Invitrogen Roswell Park Memorial Institute 1640 medium (RPMI 1640) + GLUTAMAX™ with 10% FBS. PEO14 and TO14 cell lines were cultured in RPMI 1640 + GLUTAMAX™ supplemented with 10% FBS and 2 mM sodium pyruvate, and MM127 was cultured in RPMI 1640 + GLUTAMAX™ supplemented with 10% FBS and 25 mM HEPES. Invitrogens McCoy’s 5A medium + GLUTAMAX™ supplemented with 10% FBS was used to culture the G-361, HCT116 and HT29 cell lines. Ham’s F12 + DMEM (1:1) + GLUTAMAX™ (Invitrogen™) with 10% FBS was used to culture T84 cells.

All cell lines were grown in a 37°C incubator with 5% CO2, with the exception of the NTERA-2 (clone D1) cell line which was grown at 37°C with 10% CO2.

**cDNA construction**

Total RNA preparations from the human and mouse normal tissue panels (Clontech™; 636643 and 636745 respectively). RNA from tumour tissues and cell lines were purchased from Clontech™ and Ambion™. Total RNA was also isolated from cells using TRizol (Invitrogen). Confluent cells were collected in TRizol reagent and incubated at room temperature for 5 minutes. Chloroform was added with vigorous shaking and incubated for 5 minutes at room temperature. The aqueous phase was transferred to a clean tube following centrifugation at 12,000 g for 15 minutes at 4°C. The RNA was precipitated out of solution using isopropanol (10 minutes at room temperature and centrifuged at 12,000 g for 20 minutes at 4°C). RNA preparations were re-suspended in RNase-free water containing DNase. The concentration and quality of RNA was measured using a NanoDrop (ND_1000). 1.0 µg of total RNA was reverse-transcribed into cDNA using SuperScript III First Strand synthesis kit (Invitrogen™) as per the manufacturer’s instructions.

**RT-PCR**

The sequences for each of the genes analysed were obtained from the National Center for Biotechnology (NCBI; http://www.ncbi.nlm.nih.gov/). Primers to each of the genes were designed to span exons where possible using Primer3 software (available from: www.genome.wi.mit.edu/cgi-bin/primer/primer3www.cgi; primer sequences are available upon request).

A volume of 2 µL diluted cDNA (containing ~150 ng/µl cDNA) was used for PCR in a 50 µL final volume. BioMix™ Red (Bioline™) was used for PCR amplification. Samples were amplified with a pre-cycling hold at 96°C for 5 minutes, followed by 40 cycles of denaturing at 96°C for 30 seconds, annealing at a temperature between 58-62°C for 30 seconds and extension at 72°C for 40 seconds followed by a final extension step at 72°C for 5 minutes. The products were separated on 1% agarose gels containing ethidium bromide.

**Western blot analysis**

Whole cell protein lysates were prepared from cells using lysis buffer (50 mM Tris-HCl pH 7.4, 200 mM sodium chloride, 0.5% Triton X-100, 1 mM AEBSF [4-(2-aminoethyl)-benzenesulfonyl fluoride] with complete, EDTA-free protease inhibitor cocktail (Roche)) and Laemmli buffer. The samples were boiled and an aliquot containing 60,000 cells was subjected to denaturing gel electrophoresis using a NuPAGE™ 4-12% Bis-Tris gel (Invitrogen™) and transferred to a PVDF membrane (Millipore™). Membranes were blocked for one hour using 1xPBST (0.3% Tween-20) containing 5% non-fat dry milk, followed by an overnight incubation at 4°C with rabbit polyclonal anti-PRDM9 antibody (Abcam; ab85654) at a dilution of 1:1,000, or mouse monoclonal anti-tubulin antibody (Sigma cat. no. T6074) at a...
dilution of 1:5,000, or goat polyclonal anti-lamin antibody (Santa Cruz cat. no. sc-6217) at a dilution of 1:1,000. Membranes were washed using 1xPBST and incubated with either goat, mouse or rabbit HRP-conjugated IgG antibody dependent upon the primary antibody. ECL detection reagents were then used for visualization (SuperSignal West Pico Chemiluminescent Substrate; Thermo Scientific).

ACKNOWLEDGEMENTS

JF was supported by the National Institute of Social Care and Health Research (grant HS/09/008). IA, MA, AA, NA were supported by the Saudi Arabia Government. RA was supported by a Knowledge Economy Skills Scholarship (Welsh Government’s European Science Foundation convergence programme)/UK NHS. RJM/JAW were supported by Cancer Research Wales. RJM is supported by the North West Cancer Research Fund (project grant CR888).

Editorial note: This paper was accepted based in part on peer-review in another journal (in addition to peer-review in Oncotarget).

CONFLICT OF INTERESTS

There are no conflicts of interests to declare.

REFERENCES


Meta-analysis of clinical data using human meiotic genes identifies a novel cohort of highly restricted cancer-specific marker genes - Feichtinger et al

Figure S1: RT-PCR analysis of selected human and mouse meiosis-associated genes. A. Expression profiles for five human genes originally predicted to be testis-specific. The image shows agarose gels of RT-PCR products for the five genes (HORMAD1, SYCE1, SYCE2, SYCP2, TEX12) obtained from cDNA derived from normal human tissue RNA (obtained post mortem). βACT gene expression is used as a control (bottom row). A selection of bands was subjected to DNA sequencing for validation. B. Gene expression profiles for mouse cohesion genes. The images show agarose gels of RT-PCR products for five mouse cohesion genes (RAD21, RAD21L, REC8, SMC1β, STAG3). SMC1β and RAD21L show testis-selective and testis-restricted expression profiles respectively. The other three, RAD21, REC8, STAG3, exhibit expression in an extensive range of non-meiotic tissue types. G3PDH was used as a positive control for cDNA quality. A selection of bands was subjected to DNA sequencing for validation.
Figure S2: Circus plots for single microarray analyse. A. The Circos plot showing single microarray analysis in relation to corresponding cancer types for the 25 meiCT genes and the 3 known X-CT genes (MAGE-A1, GAGE1, SSX2) covered by array sets (Supporting Information Table 4). Each connection between a gene and an individual cancer type indicates a statistically significant up regulation for that cancer type derived from a single array study for cancer tissue vs. normal tissue. B. The Circos plot showing single microarray analysis for the 21 genes which gave a testis (meiosis) only gene expression profile following RT-PCR analysis and are represented on microarrays (see Supporting Information Table 2). Each connection between a gene and an individual cancer type indicates a statistically significant up regulation for that cancer type derived from a single array study for cancer tissue vs. normal tissue.
Table S1: List of selected meiosis-associated genes used in initial study.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Functional role</th>
<th>Reference</th>
<th>Classification following validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMC1β</td>
<td>Meiotic cohesin subunit</td>
<td>Bardhan (2010) <em>Chromosome Res</em> 18: 909-924</td>
<td>Restricted CT gene</td>
</tr>
<tr>
<td>STRA8</td>
<td>Retinoic acid induced meiotic regulator</td>
<td>Anderson et al. (2008) <em>Proc Nail Acad Sci USA</em> 105: 14976-14980</td>
<td>Restricted CT gene</td>
</tr>
<tr>
<td>SYCE1</td>
<td>Synaptonemal complex component</td>
<td>Bolcun-Filas et al. (2009) <em>PLoS Genetics</em> 5: e1000393</td>
<td>Dismissed</td>
</tr>
<tr>
<td>SYCP1</td>
<td>Synaptonemal complex component</td>
<td>Poussette et al. (1997) <em>Hum Reprod</em> 12: 2414-2417</td>
<td>Restricted CT / CNS gene</td>
</tr>
<tr>
<td>SYCP2</td>
<td>Synaptonemal complex component</td>
<td>Schalk et al. (1998) <em>Chromosoma</em> 107: 540-548</td>
<td>Dismissed</td>
</tr>
<tr>
<td>TEX19</td>
<td>Meiotically up regulated</td>
<td>Kuntz et al. (2008) <em>Stem Cells</em> 26: 734-744</td>
<td>Selective CT gene</td>
</tr>
<tr>
<td>TSSK1</td>
<td>Meiotic serine/threonine kinase</td>
<td>Li et al. (2011) <em>Mol Hum Reprod</em> 17: 42-56</td>
<td>Restricted CT / CNS gene</td>
</tr>
</tbody>
</table>

Table S2: 29 genes designated as testis only expression as measured by RT-PCR validation including their coverage on arrays.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Ensembl ID</th>
<th>Uni clusters ID</th>
<th>Array coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAD1</td>
<td>ENSG00000164113</td>
<td>Hs.518957</td>
<td>231448_at, 240299_at</td>
</tr>
<tr>
<td>ARL13A</td>
<td>ENSG00000174225</td>
<td>Hs.147237</td>
<td>Not on array</td>
</tr>
<tr>
<td>ARRDC5</td>
<td>ENSG00000205784</td>
<td>Hs.574574</td>
<td>Not on array</td>
</tr>
<tr>
<td>C4orf17</td>
<td>ENSG00000138813</td>
<td>Hs.97301</td>
<td>223990_at</td>
</tr>
<tr>
<td>C4orf51</td>
<td>ENSG00000237136</td>
<td>Hs.452865</td>
<td>Not on array</td>
</tr>
<tr>
<td>C5orf47</td>
<td>ENSG00000185056</td>
<td>Hs.131469</td>
<td>1557056_at, 1557057_a_at</td>
</tr>
<tr>
<td>C5orf48</td>
<td>ENSG00000196900</td>
<td>Hs.177983</td>
<td>237428_at</td>
</tr>
<tr>
<td>C5orf50</td>
<td>ENSG00000185662</td>
<td>Hs.591740</td>
<td>Not on array</td>
</tr>
<tr>
<td>C7orf72</td>
<td>ENSG00000164500</td>
<td>Hs.99248</td>
<td>Not on array</td>
</tr>
<tr>
<td>CCDC39</td>
<td>ENSG00000175294</td>
<td>Hs.189105</td>
<td>1552335_at</td>
</tr>
<tr>
<td>CCDC79</td>
<td>ENSG00000169461</td>
<td>Hs.376505</td>
<td>1557620_a_at</td>
</tr>
<tr>
<td>CCDC105</td>
<td>ENSG00000169994</td>
<td>Hs.375985</td>
<td>1553451_at</td>
</tr>
<tr>
<td>CSF1</td>
<td>ENSG00000125815</td>
<td>Hs.121602</td>
<td>220627_at</td>
</tr>
<tr>
<td>CYLC1</td>
<td>ENSG00000183035</td>
<td>Hs.444230</td>
<td>216778_s_at, 216779_at, 216809_at</td>
</tr>
<tr>
<td>CYLC2</td>
<td>ENSG00000155833</td>
<td>Hs.332</td>
<td>207780_at</td>
</tr>
<tr>
<td>DDX4</td>
<td>ENSG00000152670</td>
<td>Hs.223581</td>
<td>221630_s_at</td>
</tr>
<tr>
<td>EFCAB9</td>
<td>ENSG00000214360</td>
<td>Hs.716824</td>
<td>Not on array</td>
</tr>
<tr>
<td>GLT6D1</td>
<td>ENSG00000204007</td>
<td>Hs.522491</td>
<td>Not on array</td>
</tr>
<tr>
<td>IQCF3</td>
<td>ENSG00000229972</td>
<td>Hs.729443</td>
<td>1555235_s_at, 236871_s_at</td>
</tr>
<tr>
<td>KCN4L1</td>
<td>ENSG00000215262</td>
<td>Hs.13861</td>
<td>237273_at</td>
</tr>
<tr>
<td>NT5C1B</td>
<td>ENSG00000185013</td>
<td>Hs.120319</td>
<td>1554368_at, 222203_s_at, 243100_at</td>
</tr>
<tr>
<td>ODF3</td>
<td>ENSG00000177947</td>
<td>Hs.350949</td>
<td>1553051_s_at, 237395_at</td>
</tr>
<tr>
<td>SYCP3</td>
<td>ENSG00000139351</td>
<td>Hs.506504</td>
<td>231618_s_at</td>
</tr>
<tr>
<td>SUNC1 (SUN3)</td>
<td>ENSG00000164744</td>
<td>Hs.406711</td>
<td>1553999_s_at, 241861_s_at</td>
</tr>
<tr>
<td>TCCT3</td>
<td>ENSG00000184796</td>
<td>Hs.733746</td>
<td>1554400_at, 1554401_s_at, 1557945_at, 232258_at</td>
</tr>
<tr>
<td>TEMEM202</td>
<td>ENSG00000187806</td>
<td>Hs.446069</td>
<td>Not on array</td>
</tr>
<tr>
<td>TEMEM225</td>
<td>ENSG00000204300</td>
<td>Hs.98377</td>
<td>244460_at</td>
</tr>
<tr>
<td>TRIML1</td>
<td>ENSG00000184108</td>
<td>Hs.348618</td>
<td>1557677_a_at</td>
</tr>
<tr>
<td>Data set name</td>
<td>Cancer type</td>
<td>Cancer sub-type</td>
<td>Tissue</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>--------------</td>
</tr>
<tr>
<td>E-GEOD-10927_ACA</td>
<td>Adrenal cancer</td>
<td>Adenoma</td>
<td>Adrenal gland</td>
</tr>
<tr>
<td>E-GEOD-10927_ACC</td>
<td>Adrenal cancer</td>
<td>Carcinoma</td>
<td>Adrenal gland</td>
</tr>
<tr>
<td>GSE12368_ACA</td>
<td>Adrenal cancer</td>
<td>Adenoma</td>
<td>Adrenal gland</td>
</tr>
<tr>
<td>GSE12368_ACC</td>
<td>Adrenal cancer</td>
<td>Carcinoma</td>
<td>Adrenal gland</td>
</tr>
<tr>
<td>GSE8514_Adrenal_Ad</td>
<td>Adrenal cancer</td>
<td>Adenoma</td>
<td>Adrenal gland</td>
</tr>
<tr>
<td>E-GEOD-21354_Brain_AC</td>
<td>Brain cancer</td>
<td>Sarcoma</td>
<td>Brain</td>
</tr>
<tr>
<td>E-GEOD-21354_Brain_EM</td>
<td>Brain cancer</td>
<td>Sarcoma</td>
<td>Brain</td>
</tr>
<tr>
<td>E-GEOD-21354_Brain_OG</td>
<td>Brain cancer</td>
<td>Sarcoma</td>
<td>Brain</td>
</tr>
<tr>
<td>GSE19728_Brain_C</td>
<td>Brain cancer</td>
<td>Sarcoma</td>
<td>Brain</td>
</tr>
<tr>
<td>E-GEOD-20086_Breast_C</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-21653_Breast_BLC</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-21653_Breast_ERBB2</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-21653_Breast_LuminaL A</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-21653_Breast_LuminalB</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-22544_Breast_C</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-5764_Breast_IDC</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-5764_Breast_ILC</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-7904_Breast_BLC</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-7904_Breast_BRCA1</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-7904_Breast_NonBLC</td>
<td>Breast cancer</td>
<td>Carcinoma</td>
<td>Breast</td>
</tr>
<tr>
<td>E-GEOD-18105_Colorectal_C</td>
<td>Colorectal cancer</td>
<td>Carcinoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-18105_Colorectal_Met</td>
<td>Colorectal cancer</td>
<td>Metastasis</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-20916_Colorectal_Ad</td>
<td>Colorectal cancer</td>
<td>Adenoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-20916_Colorectal_ADC</td>
<td>Colorectal cancer</td>
<td>Carcinoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-20916_Colorectal_C</td>
<td>Colorectal cancer</td>
<td>Carcinoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-20916_Colorectal_Ep_Ad</td>
<td>Colorectal cancer</td>
<td>Adenoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-20916_Colorectal_Ep_C</td>
<td>Colorectal cancer</td>
<td>Carcinoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-20916_Colorectal_Muc_Ad</td>
<td>Colorectal cancer</td>
<td>Adenoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-20916_Colorectal_Muc_C</td>
<td>Colorectal cancer</td>
<td>Carcinoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-23878_Colorectal_C</td>
<td>Colorectal cancer</td>
<td>Carcinoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-4183_Colorectal_C</td>
<td>Colorectal cancer</td>
<td>Carcinoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-4183_Colorectal_PreAd</td>
<td>Colorectal cancer</td>
<td>Adenoma</td>
<td>Colon</td>
</tr>
<tr>
<td>E-GEOD-12452_NPC</td>
<td>Head and neck cancer</td>
<td>Carcinoma</td>
<td>Nasopharynx</td>
</tr>
<tr>
<td>E-GEOD-17351_ESCC</td>
<td>Head and neck cancer</td>
<td>Carcinoma</td>
<td>Esophagus</td>
</tr>
<tr>
<td>E-GEOD-30784_OSSC</td>
<td>Head and neck cancer</td>
<td>Carcinoma</td>
<td>Oral tissue</td>
</tr>
<tr>
<td>E-GEOD-30784_OSSC_dysplasia</td>
<td>Head and neck cancer</td>
<td>Carcinoma</td>
<td>Oral tissue</td>
</tr>
<tr>
<td>GSE6791_OSSC</td>
<td>Head and neck cancer</td>
<td>Carcinoma</td>
<td>Esophagus</td>
</tr>
<tr>
<td>GSE68856_EAC.txt</td>
<td>Head and neck cancer</td>
<td>Carcinoma</td>
<td>Esophagus</td>
</tr>
<tr>
<td>E-GEOD-24739_CML</td>
<td>Leukemia</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>E-GEOD-26713_TALL</td>
<td>Leukemia</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>GSE19424_AML_CD4</td>
<td>Leukemia</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>GSE19424_AML_CD8</td>
<td>Leukemia</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>E-GEOD-19188_Lung_ADC</td>
<td>Lung cancer</td>
<td>Carcinoma</td>
<td>Lung</td>
</tr>
<tr>
<td>E-GEOD-19188_Lung_LCC</td>
<td>Lung cancer</td>
<td>Carcinoma</td>
<td>Lung</td>
</tr>
<tr>
<td>E-GEOD-19188_Lung_SCC</td>
<td>Lung cancer</td>
<td>Carcinoma</td>
<td>Lung</td>
</tr>
<tr>
<td>E-GEOD-19804_Lung_C</td>
<td>Lung cancer</td>
<td>Carcinoma</td>
<td>Lung</td>
</tr>
<tr>
<td>E-GEOD-31210_Lung_ADC</td>
<td>Lung cancer</td>
<td>Carcinoma</td>
<td>Lung</td>
</tr>
<tr>
<td>GSE18842_Lung_C</td>
<td>Lung cancer</td>
<td>Carcinoma</td>
<td>Lung</td>
</tr>
<tr>
<td>E-GEOD-35331_Flymph</td>
<td>Lymphoma</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>E-GEOD-6338_Lymphoma</td>
<td>Lymphoma</td>
<td>Hematological malignancy</td>
<td>Lymph node</td>
</tr>
<tr>
<td>E-MEXP-2957_CLL</td>
<td>Lymphoma</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>GSE3293_CLL</td>
<td>Lymphoma</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>GSE3293_Flymph</td>
<td>Lymphoma</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>GSE3293_MALTlymph</td>
<td>Lymphoma</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>GSE35550_MALTlymph</td>
<td>Lymphoma</td>
<td>Hematological malignancy</td>
<td>Spleen</td>
</tr>
<tr>
<td>GSE26725_CLL</td>
<td>Lymphoma</td>
<td>Hematological malignancy</td>
<td>Blood/bone marrow</td>
</tr>
<tr>
<td>E-GEOD-14001_Ovarian_C_HG</td>
<td>Ovarian cancer</td>
<td>Carcinoma</td>
<td>Ovary</td>
</tr>
<tr>
<td>E-GEOD-14001_Ovarian_C_LG</td>
<td>Ovarian cancer</td>
<td>Carcinoma</td>
<td>Ovary</td>
</tr>
<tr>
<td>E-GEOD-18520_Ovarian_C</td>
<td>Ovarian cancer</td>
<td>Carcinoma</td>
<td>Ovary</td>
</tr>
<tr>
<td>E-GEOD-27651_Ovarian_C_HG</td>
<td>Ovarian cancer</td>
<td>Carcinoma</td>
<td>Ovary</td>
</tr>
<tr>
<td>E-GEOD-27651_Ovarian_C_LM</td>
<td>Ovarian cancer</td>
<td>Carcinoma</td>
<td>Ovary</td>
</tr>
<tr>
<td>E-GEOD-22780_Pancreatic_ADC</td>
<td>Pancreatic cancer</td>
<td>Carcinoma</td>
<td>Pancreas</td>
</tr>
</tbody>
</table>
Table S3. List of the 80 data sets supported by the microarray meta-analysis including the corresponding cancer type, cancer sub-type and tissue type. (Continued)

<table>
<thead>
<tr>
<th>Data set name</th>
<th>Cancer type</th>
<th>Cancer sub-type</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE15471_Pancreatic_C</td>
<td>Pancreatic cancer</td>
<td>Carcinoma</td>
<td>Pancreas</td>
</tr>
<tr>
<td>E-GEOD-17906_Prostate_C</td>
<td>Prostate cancer</td>
<td>Carcinoma</td>
<td>Prostate</td>
</tr>
<tr>
<td>E-GEOD-30522_Prostate_C</td>
<td>Prostate cancer</td>
<td>Carcinoma</td>
<td>Prostate</td>
</tr>
<tr>
<td>E-GEOD-12606_Renal_C</td>
<td>Renal cancer</td>
<td>Carcinoma</td>
<td>Kidney</td>
</tr>
<tr>
<td>E-GEOD-12606_Renal_Met</td>
<td>Renal cancer</td>
<td>Metastasis</td>
<td>Kidney</td>
</tr>
<tr>
<td>E-TABM-282_Renal_C</td>
<td>Renal cancer</td>
<td>Carcinoma</td>
<td>Kidney</td>
</tr>
<tr>
<td>GSE11151_CRCC</td>
<td>Renal cancer</td>
<td>Carcinoma</td>
<td>Kidney</td>
</tr>
<tr>
<td>GSE11151_PRCC</td>
<td>Renal cancer</td>
<td>Carcinoma</td>
<td>Kidney</td>
</tr>
<tr>
<td>GSE11151_Renal_C</td>
<td>Renal cancer</td>
<td>Carcinoma</td>
<td>Kidney</td>
</tr>
<tr>
<td>GSE11151_Renal_Onc</td>
<td>Renal cancer</td>
<td>Carcinoma</td>
<td>Kidney</td>
</tr>
<tr>
<td>E-GEOD-6004_Thyroid_C_Center</td>
<td>Thyroid cancer</td>
<td>Carcinoma</td>
<td>Thyroid gland</td>
</tr>
<tr>
<td>E-GEOD-6004_Thyroid_C_Invasive</td>
<td>Thyroid cancer</td>
<td>Carcinoma</td>
<td>Thyroid gland</td>
</tr>
<tr>
<td>E-MEXP-2442_Thyroid_ATC</td>
<td>Thyroid cancer</td>
<td>Carcinoma</td>
<td>Thyroid gland</td>
</tr>
<tr>
<td>E-MEXP-2442_Thyroid_FAd</td>
<td>Thyroid cancer</td>
<td>Adenoma</td>
<td>Thyroid gland</td>
</tr>
<tr>
<td>E-MEXP-2442_Thyroid_FCarc</td>
<td>Thyroid cancer</td>
<td>Carcinoma</td>
<td>Thyroid gland</td>
</tr>
</tbody>
</table>

Abbreviation | Meaning
-------------|-----------------
AC           | Astrocytoma
ACA          | Adrenocortical adenoma
ACC          | Adrenocortical carcinoma
Ad           | Adenoma
ADC          | Adenocarcinoma
AdvHCC       | Advanced hepatocellular carcinoma
ATC          | Thyroid anaplastic carcinoma
BLC          | Basal-like cancer
BM           | Bone marrow
BRCA1        | BRCA1-associated
C            | Cancer
Center       | Center area
CLL          | Chronic lymphocytic leukemia
CML          | Chronic myelogenous leukemia
CRCC         | Chromophobe renal cell cancer
EAC          | Esophageal adenocarcinoma
EarlyHCC     | Early hepatocellular carcinoma
ED           | Ependymoma
Ep           | Epithelium
ESCC         | Esophageal squamous cell carcinoma
FAd          | Follicular adenoma
FCarc        | Follicular carcinoma
Flymph       | Follicular lymphoma
HG           | High grade
IDC          | Invasive ductal carcinoma
ILC          | Invasive lobular carcinoma
Invasive     | Invasive area
LCC          | Large-cell carcinoma
LG           | Low grade
LM           | Low-malignant
Met          | Metastatic
MRC          | Metastatic recurrence
Muc          | Mucosa
Neo          | Neoplasm
NPC          | Nasopharyngeal carcinoma
OG           | Oligodendro-glioma
Onc          | Oncocytoma
OSCC         | Oral squamous cell carcinoma
OTSCC        | Oral tongue squamous cell carcinoma
PRCC         | Papillary renal cell cancer
PreAd        | Precancerous adenoma
PTC          | Papillary thyroid cancer
SCC          | Squamous cell carcinoma
TALL         | T-cell acute lymphoblastic leukemia
VAdvHCC      | Very advanced hepatocellular carcinoma
VEarlyHCC    | Very early hepatocellular carcinoma
Table S4: List the 33 candidates and three control CTA genes (*MAGE-A1, GAGE, SSX2*) including their coverage on the arrays and their cancer-specific up regulation according to the microarray meta-analysis. For each candidate the meta-log 2-fold change (log2FC) and the confidence intervals (CI left, CI right) are stated.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Ensembl ID</th>
<th>Array Coverage</th>
<th>Cancer type</th>
<th>log2FC</th>
<th>CI left</th>
<th>CI right</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTRT1</td>
<td>ENSG00000123165</td>
<td>Not on array</td>
<td>Lung cancer</td>
<td>0.97</td>
<td>0.22</td>
<td>1.73</td>
</tr>
<tr>
<td>BRDT</td>
<td>ENSG00000137948</td>
<td>206787_at</td>
<td>Ovarian cancer</td>
<td>1.40</td>
<td>0.33</td>
<td>2.47</td>
</tr>
<tr>
<td>C1orf12</td>
<td>ENSG00000197651</td>
<td>236968_at</td>
<td>Ovarian cancer</td>
<td>1.14</td>
<td>−0.17</td>
<td>2.44</td>
</tr>
<tr>
<td>C1orf35</td>
<td>ENSG00000185407</td>
<td>156403_at, 231338_at</td>
<td>Ovarian cancer</td>
<td>1.05</td>
<td>0.11</td>
<td>1.99</td>
</tr>
<tr>
<td>C1orf98</td>
<td>ENSG00000214556</td>
<td>244316_at</td>
<td>Ovarian cancer</td>
<td>1.80</td>
<td>0.56</td>
<td>3.05</td>
</tr>
<tr>
<td>C1orf67</td>
<td>ENSG00000188032</td>
<td>Not on array</td>
<td>Ovarian cancer</td>
<td>1.06</td>
<td>0.01</td>
<td>2.12</td>
</tr>
<tr>
<td>C1orf79</td>
<td>ENSG00000132631</td>
<td>231134_at</td>
<td>Ovarian cancer</td>
<td>1.64</td>
<td>0.01</td>
<td>2.77</td>
</tr>
<tr>
<td>C2orf11</td>
<td>ENSG00000120160</td>
<td>1554981_at, 1554982_at, 232688_at</td>
<td>Lung cancer</td>
<td>1.21</td>
<td>0.30</td>
<td>1.95</td>
</tr>
<tr>
<td>C2orf20</td>
<td>ENSG00000171695</td>
<td>1554977_at</td>
<td>Ovarian cancer</td>
<td>1.09</td>
<td>0.06</td>
<td>2.23</td>
</tr>
<tr>
<td>C2orf24</td>
<td>ENSG000001815264</td>
<td>231617_at</td>
<td>Ovarian cancer</td>
<td>1.36</td>
<td>0.32</td>
<td>2.40</td>
</tr>
<tr>
<td>CXorf27</td>
<td>ENSG00000187516</td>
<td>215142_at</td>
<td>Brain cancer</td>
<td>1.13</td>
<td>0.30</td>
<td>1.95</td>
</tr>
<tr>
<td>DUSP21</td>
<td>ENSG00000189037</td>
<td>220515_at</td>
<td>Ovarian cancer</td>
<td>1.96</td>
<td>0.33</td>
<td>2.67</td>
</tr>
<tr>
<td>FABP9</td>
<td>ENSG00000205186</td>
<td>Not on array</td>
<td>Ovarian cancer</td>
<td>1.21</td>
<td>0.15</td>
<td>2.28</td>
</tr>
<tr>
<td>FTMT</td>
<td>ENSG00000181867</td>
<td>Not on array</td>
<td>Ovarian cancer</td>
<td>1.48</td>
<td>0.67</td>
<td>2.30</td>
</tr>
<tr>
<td>GAGE1</td>
<td>ENSG00000205777</td>
<td>207086_x_at, 207739_x_at, 208155_x_at, 208283_at</td>
<td>Ovarian cancer</td>
<td>1.23</td>
<td>−0.30</td>
<td>2.76</td>
</tr>
<tr>
<td>H3F3B</td>
<td>ENSG00000170948</td>
<td>1552458_at, 1552459_at</td>
<td>Ovarian cancer</td>
<td>2.56</td>
<td>1.45</td>
<td>3.66</td>
</tr>
<tr>
<td>MAGEA1</td>
<td>ENSG00000198681</td>
<td>207325_x_at</td>
<td>Ovarian cancer</td>
<td>1.09</td>
<td>−0.06</td>
<td>2.23</td>
</tr>
<tr>
<td>MAGEB5</td>
<td>ENSG00000188408</td>
<td>Not on array</td>
<td>Ovarian cancer</td>
<td>1.70</td>
<td>0.64</td>
<td>2.76</td>
</tr>
<tr>
<td>MBD3L1</td>
<td>ENSG000001814650</td>
<td>1552408_at, 1552409_at</td>
<td>Ovarian cancer</td>
<td>1.30</td>
<td>0.33</td>
<td>2.67</td>
</tr>
<tr>
<td>MBD3L2</td>
<td>ENSG00000181823</td>
<td>208271_at, 242158_at</td>
<td>Ovarian cancer</td>
<td>1.07</td>
<td>0.08</td>
<td>2.10</td>
</tr>
<tr>
<td>MAGEA5</td>
<td>ENSG00000182338</td>
<td>211694_at</td>
<td>Ovarian cancer</td>
<td>1.41</td>
<td>0.40</td>
<td>2.42</td>
</tr>
<tr>
<td>MAGEB5</td>
<td>ENSG00000182338</td>
<td>211694_at</td>
<td>Ovarian cancer</td>
<td>1.41</td>
<td>0.40</td>
<td>2.42</td>
</tr>
</tbody>
</table>
Table S5: Full list of all 52 meiCT genes indicating the method used to designate them and their classification from the EST screening of the original 375 human orthologues of the mouse meiosis-specific genes.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Chromosome</th>
<th>Original EST class</th>
<th>Method of meiCT designation</th>
<th>CT class</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTRT1</td>
<td>X</td>
<td>1</td>
<td>R, M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>ADAM1</td>
<td>4</td>
<td>3</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>BRDT</td>
<td>1</td>
<td>3</td>
<td>R, M</td>
<td>CT restricted</td>
</tr>
<tr>
<td>Clorf65</td>
<td>1</td>
<td>3</td>
<td>R</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>Clorf17</td>
<td>4</td>
<td>2</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>Clorf47</td>
<td>5</td>
<td>2</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>Clorf48</td>
<td>5</td>
<td>2</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>Clorf11</td>
<td>9</td>
<td>2</td>
<td>R, M</td>
<td>CT/CNS selective</td>
</tr>
<tr>
<td>Ci2orf12</td>
<td>12</td>
<td>2</td>
<td>R, S</td>
<td>CT restricted</td>
</tr>
<tr>
<td>Ci7orf98</td>
<td>17</td>
<td>1</td>
<td>R, M</td>
<td>CT restricted</td>
</tr>
<tr>
<td>Ci9orf67</td>
<td>19</td>
<td>2</td>
<td>R</td>
<td>CT restricted</td>
</tr>
<tr>
<td>Ci2orf79</td>
<td>20</td>
<td>2</td>
<td>R, S</td>
<td>CT restricted</td>
</tr>
<tr>
<td>C2orf201</td>
<td>20</td>
<td>2</td>
<td>R, M</td>
<td>CT/CNS restricted</td>
</tr>
<tr>
<td>Ci2orf33</td>
<td>22</td>
<td>2</td>
<td>R, S</td>
<td>CT restricted</td>
</tr>
<tr>
<td>Ciorf27</td>
<td>X</td>
<td>2</td>
<td>R, S</td>
<td>CT/CNS selective</td>
</tr>
<tr>
<td>CATSPER1</td>
<td>11</td>
<td>3</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>CCDC38</td>
<td>12</td>
<td>2</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>CCDC79</td>
<td>16</td>
<td>2</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>CCDC105</td>
<td>19</td>
<td>2</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>CST8</td>
<td>20</td>
<td>2</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>CYCL1</td>
<td>X</td>
<td>2</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>CYCL2</td>
<td>9</td>
<td>2</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>DDX4</td>
<td>5</td>
<td>2</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>DUSP21</td>
<td>X</td>
<td>2</td>
<td>R, S</td>
<td>CT selective</td>
</tr>
<tr>
<td>FABP9</td>
<td>4</td>
<td>2</td>
<td>R</td>
<td>CT selective</td>
</tr>
<tr>
<td>FTMT</td>
<td>5</td>
<td>2</td>
<td>R</td>
<td>CT selective</td>
</tr>
<tr>
<td>IL31</td>
<td>12</td>
<td>2</td>
<td>R</td>
<td>CT selective</td>
</tr>
<tr>
<td>IQCF3</td>
<td>3</td>
<td>2</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>LUZP4</td>
<td>X</td>
<td>2</td>
<td>R, M</td>
<td>CT restricted</td>
</tr>
<tr>
<td>MAGE-B5</td>
<td>X</td>
<td>2</td>
<td>R</td>
<td>CT restricted</td>
</tr>
<tr>
<td>MBD3L1</td>
<td>19</td>
<td>2</td>
<td>R, M</td>
<td>CT selective</td>
</tr>
<tr>
<td>NT5C1B</td>
<td>2</td>
<td>2</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>NUT</td>
<td>15</td>
<td>Manual selection</td>
<td>R, M</td>
<td>CT restricted</td>
</tr>
<tr>
<td>ODF3</td>
<td>11</td>
<td>2</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>ODF4</td>
<td>17</td>
<td>1</td>
<td>R, M</td>
<td>CT restricted</td>
</tr>
<tr>
<td>PAPOLB</td>
<td>7</td>
<td>2</td>
<td>R, M</td>
<td>CT/CNS restricted</td>
</tr>
<tr>
<td>PFNS</td>
<td>5</td>
<td>2</td>
<td>R</td>
<td>CT/CNS selective</td>
</tr>
<tr>
<td>PRDM9</td>
<td>5</td>
<td>Manual selection</td>
<td>R, M</td>
<td>CT restricted</td>
</tr>
<tr>
<td>RAD21L</td>
<td>20</td>
<td>Manual selection</td>
<td>R, S</td>
<td>CT restricted</td>
</tr>
<tr>
<td>SEPT12</td>
<td>16</td>
<td>3</td>
<td>R, M</td>
<td>CT/CNS restricted</td>
</tr>
<tr>
<td>SMCI8</td>
<td>22</td>
<td>Manual selection</td>
<td>R, M</td>
<td>CT restricted</td>
</tr>
<tr>
<td>STRA4</td>
<td>7</td>
<td>Manual selection</td>
<td>R</td>
<td>CT restricted</td>
</tr>
<tr>
<td>SYCP1</td>
<td>1</td>
<td>Manual selection</td>
<td>R, M</td>
<td>CT/CNS restricted</td>
</tr>
<tr>
<td>SYCP3</td>
<td>12</td>
<td>2</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>TCE3</td>
<td>6</td>
<td>2</td>
<td>M</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>TDRD12</td>
<td>19</td>
<td>2</td>
<td>R, S</td>
<td>CT restricted</td>
</tr>
<tr>
<td>TEPP</td>
<td>16</td>
<td>2</td>
<td>R, S</td>
<td>CT selective</td>
</tr>
<tr>
<td>TEX19</td>
<td>17</td>
<td>3</td>
<td>R, M</td>
<td>CT selective</td>
</tr>
<tr>
<td>TMEM225</td>
<td>11</td>
<td>2</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>TRIML1</td>
<td>4</td>
<td>3</td>
<td>S</td>
<td>CT restricted*</td>
</tr>
<tr>
<td>TSSK1B</td>
<td>5</td>
<td>Manual selection</td>
<td>R, S</td>
<td>CT/CNS restricted</td>
</tr>
<tr>
<td>ZCCHC13</td>
<td>X</td>
<td>2</td>
<td>R, M</td>
<td>CT/CNS selective</td>
</tr>
</tbody>
</table>

* Those genes predicted to be meiCT genes based on microarray analyses (meta or individual) have been validated for tight testis specificity by RT-PCR (see Supplementary Information Table S2).
R - Determined to be expressed in cancer samples by RT-PCR.
M - Determined to be expressed in cancer samples by microarray meta-analyses of combined microarray data sets.
S - Determined to be expressed in cancer samples by analysis of at least one individual microarray data set; these designations have the limitations imposed by statistical rigour being derived from a single microarray data set.
SUPPLEMENTARY INFORMATION: COMPUTATIONAL ANALYSES

Generation of a meiosis-specific data set

The meiosis-specific gene set was generated using: Perl 5.8.8 (available from: http://www.perl.org); and the Biomart portals (1) for GermOnline (2), MGI (3), and Ensembl (4). The initial meiosis gene set was derived from a microarray study by Chalmel et al. (5,6) whereby the meiotic transcriptome of mice was studied by analysing 17 somatic non-testicular control tissues, total testis, isolated seminiferous tubules as well as enriched populations of Sertoli cells, spermatagonia, pachytene spermatocytes, and round spermatids. 744 mouse genes were selected, which were found to be differentially expressed in testis, assigned to the meiotic or post-meiotic cluster as defined by Chalmel et al. (5,6), and not expressed in any other tissue tested. 408 human orthologues of the 744 meiosis-specific mouse genes were identified. To improve the data quality, the gene set was cross-validated with a set of human genes known to be involved in mitosis (Mitocheck) (7). The resulting 375 genes were assigned to Unigene cluster IDs.

EST analysis pipeline

The pipeline was implemented using: MySQL 5.0.77 (available from: http://www.mysql.com); and Perl 5.8.8 (available from: http://www.perl.org). The EST data derived from the Unigene database (8). For each of the 375 genes the Unigene EST profile was evaluated to determine the expression in normal and cancerous tissues. ESTs originating from cell line or uncharacterized/mixed tissue libraries were excluded as well as ESTs showing less than 90% similarity to the corresponding human protein. Genes were sorted into 5 classes according to their expression profile: (i) testis-restricted expression in normal individuals as well as cancer expression (class 1); (ii) testis-restricted expression in normal individuals (class 2); (iii) testis and brain-restricted expression in normal individuals as well as cancer expression (class 3); (iv) testis and brain-restricted expression in normal individuals (class 4); and (v) somatic expression in normal individuals (class 5). Class 5 genes were discarded.

Microarray meta-analysis pipeline

The pipeline was implemented using: R 2.12.1 (available from: http://www.cran.r-project.org) (9); the Bioconductor package (10); MySQL 5.0.77 (available from: http://www.mysql.com); and Perl 5.8.8 (available from: http://www.perl.org).

Raw data was obtained from microarray experiments using patient-derived, untreated cancer samples with corresponding normal samples deposited in the ArrayExpress (11) or the GEO (12) repository. Data sets produced from the HG-U133 Plus 2 array from Affymetrix were selected, as this type is widely used and covers a large proportion of the human genome. Obtained data sets were sub-divided into cancer subtypes/stages as appropriate. Data sets with less than three control or cancer samples or data sets analysing tissues influenced by other diseases, fetal tissues or cancer-associated cells such as the cancer microenvironment were excluded. Only cancer types with at least two data sets could be included. This resulted in 92 data sets originating from 50 experiments and covering 13 different cancer types. The quality of the arrays was further assessed using the ‘simpleaffy’ package (13) and data sets with a scale factor of 3, an ActB 3:mRNA ratio of 3 and a GAPDH 3:mRNA ratio of 1.25 were selected. Based on this, 12 data sets had to be excluded completely. Individual CEL files from of 37 data sets were excluded, as they did not fulfill the quality requirements. This resulted in 80 individual cancer data sets originating from 45 experiments and covering 13 different cancer types (Supporting Information Table 3).

The raw data of all 80 data sets were re-analysed individually to assure uniformity of the analysis process. Data were pre-processed according to methods described by Hubbell et al. (14) using the ‘affy’ package (15). The data sets were filtered with the 33 candidates, the 3 control CTA genes and the 29 meiosis-specific genes respectively, in order to reduce the number of features and to enhance the statistical power. 25 of the candidates and all 3 control CTA genes were covered by the array sets (Supporting Information Table 4), and 21 of the meiosis-specific genes were covered by the array sets (Supporting Information Table 2). For computation of differentially expressed genes, the ‘Limma’ package (16) from Bioconductor was used, with p values adjusted for multiple testing with Benjamini and Hochberg’s method to control the false discovery rate (17).

Subsequently, a meta-p value and a meta-log2-fold change value were computed for each cancer type as listed in Supplementary Table S3 using Stouffer’s method (18) and weighted linear combination (19), respectively. In order to calculate a meta-p value, the individual two-sided p values were converted to one-sided p values for up and down regulation separately, as two-sided p values are oblivious to the effect direction (20). If multiple probes mapped to the same gene identify the most extreme log 2-fold change value with its corresponding p value was selected, as it is least likely to occur by chance. We selected genes with a meta-log2-fold change > 1 or a confidence interval that does not span 0, and a meta-p value < 0.05 as potentially significant. To visualize the data of the meta-analysis, Circos plots (21) and Forest plots (22) were created.
4 CancerMA: a Web-based Tool for Automatic Meta-analysis of Public Cancer Microarray Data*

This chapter describes the implementation, usability and validation of a bioinformatic analytical web tool to automate the identification of novel candidate cancer markers/targets by means of meta-analysing the expression of user-supplied gene lists across a manually curated database of 80 publicly available cancer microarray datasets and 13 cancer types. The web tool is based on the in-house *in silico* pipeline described in the previous chapter (cf. chapter 3). The work presented in this chapter contributes to project objectives 1 and 2.

Please note that this chapter is presented as paper published in the open-access journal *Database* (available at: [http://database.oxfordjournals.org/](http://database.oxfordjournals.org/)) [234]. The content structure, layout, language and reference style follow the specifications of *Database.*
Database Tool

CancerMA: a web-based tool for automatic meta-analysis of public cancer microarray data

Julia Feichtinger1,2,*, Ramsay J. McFarlane1,3,* and Lee D. Larcombe4

1North West Cancer Research Fund Institute, Bangor University, Bangor, Gwynedd LL57 2UW, UK, 2Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Petersgasse 14, 8010, Austria, 3NISCHR Cancer Genetics Biomedical Research Unit, Bangor University, Bangor, Gwynedd LL57 2UW, UK and 4Cranfield Health, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK

*Corresponding author: Tel: +44 1248 382360; Fax: +44 1248 370731; Email: r.macfarlane@bangor.ac.uk
Correspondence may also be addressed to Julia Feichtinger. Email: julia.feichtinger@gmail.com


Submitted 8 September 2012; Revised 22 November 2012; Accepted 25 November 2012

The identification of novel candidate markers is a key challenge in the development of cancer therapies. This can be facilitated by putting accessible and automated approaches analysing the current wealth of ‘omic-scale data in the hands of researchers who are directly addressing biological questions. Data integration techniques and standardized, automated, high-throughput analyses are needed to manage the data available as well as to help narrow down the excessive number of target gene possibilities presented by modern databases and system-level resources. Here we present CancerMA, an online, integrated bioinformatic pipeline for automated identification of novel candidate cancer markers/targets; it operates by means of meta-analysing expression profiles of user-defined sets of biologically significant and related genes across a manually curated database of 80 publicly available cancer microarray datasets covering 13 cancer types. A simple-to-use web interface allows bioinformaticians and non-bioinformaticians alike to initiate new analyses as well as to view and retrieve the meta-analysis results. The functionality of CancerMA is shown by means of two validation datasets.

Database URL: http://www.cancerma.org.uk

Introduction

Cancer is a multi-factorial disease that can arise from alterations in expression levels of oncogenes and tumour suppressor genes, details of which can be elucidated by means of expression data (1). In the last decade, a large amount of microarray data for gene expression profiles has become available in public repositories such as ArrayExpress (2) and Gene Expression Omnibus (GEO) (3), which provide the opportunity to retrieve, reanalyse and integrate the data (4). Retrieval and reanalysis of publicly available data allow the development of automated pipelines to ensure a broad spectrum of users can execute rapid, homogeneous and reproducible analyses across a large number of datasets, addressing novel and specific questions. Data integration techniques, so-called meta-analyses, aim to combine the data available and integrate information from multiple independent but related microarray studies to identify significant genes [reviewed by Feichtinger et al. (5)]. Combining studies can enhance reliability and generalizability of the results (6) and can be used to obtain a more precise estimate of gene expression. In particular, the benefit of enhancing the statistical power can help to overcome the most profound limitation of microarray studies: testing tens of thousands of hypotheses, relying only on a relatively low number of samples (7, 8). For example, Arasappan et al. (9) found a refined expression signature for systemic lupus erythematosus, and Vierlinger et al. (10) reported the identification of a potential biomarker for papillary thyroid carcinoma by means of meta-analysis approaches.

© The Author(s) 2012. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Here we present CancerMA, an openly accessible integrated bioinformatic analytical pipeline with a user-friendly and intuitive web interface to automate the reanalysis of public cancer microarray datasets with user-defined sets of biologically significant and related genes. The underlying analytical approach was developed for a previous study to identify a cohort of novel cancer-specific marker genes (11) and was automated forming the core of the CancerMA tool. Further analyses and visualizations were added to aid the data interpretation. This tool allows bioinformaticians and non-bioinformaticians alike, to obtain refined and integrated differential expression for their genes of interest across a manually curated database of 80 datasets and 13 cancer types as well as to investigate the relationships between cancer types and to reveal commonalities among them. Furthermore, it can help to narrow down the excessive number of target gene possibilities presented by modern databases and system-level resources to a manageable number of putative candidates, which can be followed up in the laboratory and/or fed into an interaction network analysis. Thus, it puts a meta-analysis pipeline in the hands of those asking the biological questions. To validate our approach, we have analysed two experimentally derived datasets from the literature and could reproduce the published results.

**Methods and structure of CancerMA**

CancerMA consists of a web interface, a set of pipelined analyses and two relational databases, one holding the analysis data for each user and another one holding the gene annotation data. The general workflow is visualized in Figure 1.

**Cancer dataset retrieval**

We searched for raw data of patient-derived, untreated cancer samples with corresponding normal samples deposited in the ArrayExpress (2) or the GEO (3) repository using the HG-U133 Plus 2 array from Affymetrix. After manual assessment, we divided the retrieved datasets according to the cancer type, subtype and stage. We omitted datasets with less than three control or cancer samples as well as datasets deriving from foetal tissues, tissues influenced by other diseases or cancer-associated tissues (e.g. tumour microenvironment). We could obtain 92 datasets from 50 experiments covering 13 distinct cancer types. To allow a meta-analysis, at least two datasets per cancer type were required. Subsequently, quality control using the ‘simpleaffy’ R package (12) was used to further assess the datasets. Based on the guidelines from Affymetrix/’simpleaffy’ (available at: http://www.bioconductor.org/packages/release/bioc/vignettes/simpleaffy/inst/doc/QCandSimpleaffy.pdf), datasets with scale factors with 3-fold of one another, an ActB 3’/mid ratio <3 and a GAPDH 3’/mid ratio <1.25 were selected. Scale factors assess the comparability of the arrays, whereas the signal ratios of ActB and GAPDH can be used to measure the RNA quality. Based on this assessment, we omitted 12 datasets and excluded individual CEL files of 37 datasets. Finally, 80 individual curated cancer datasets originating from 45 experiments and covering 13 different cancer types (Supplementary Table S1) remained. For more details, refer to Feichtinger et al. (11). The full list of 80 datasets, including the GEO/ArrayExpress accession numbers as well as the 13 cancer types covered, are available on the CancerMA website (http://www.cancerma.org.uk/information.html). Additional experimental datasets can be obtained from the microarray repositories and added to the pipeline by the authors as they become available.

**The CancerMA pipeline and databases**

The pipeline handles the single microarray analysis, the meta-analysis, the GO analysis as well as the annotation and the visualizations. After manual assessment and quality control, all 80 datasets described above were individually pre-processed (background correction, normalization and computation of expression values) according to methods described by Hubbell et al. (13) using the ‘affy’ R package from Bioconductor (14), which assures uniformity of the analysis process.

For gene and probe annotation purposes, the Ensembl database (15), the HUGO Gene Nomenclature Committee (HGN) database (16) and the annotation files provided by Affymetrix (available at: http://www.affymetrix.com/support/technical/annotationfiles/main.affx) were established as a local MySQL database.
When a new job is submitted, the user-supplied gene list is used to filter the 80 pre-processed datasets in order to reduce the number of features and enhance the statistical power (17). The ‘limma’ R package (18) from Bioconductor is used to compute differentially expressed genes, and the resulting P-values are adjusted for multiple testing with Benjamini and Hochberg’s method to control the false discovery rate (19). For the single array analysis, genes with a P-value <0.05 and a log2-fold change >1 are selected as potentially significant.

Subsequently, the results of the 80 individual analyses are combined. A meta-P-value and a meta-log2-fold change value are calculated for each cancer type (Supplementary Table S1) as well as for all cancers in total using Stouffer’s method (21) and weighted linear combination (22), respectively. As two-sided P-values are oblivious to the effect direction, these P-values need to be converted to the corresponding one-sided P-values for up- and down-regulation separately (22). In case of multiple probes mapping to the same gene identifier, the most extreme log2-fold change value with its corresponding P-value are further used for feature selection. Genes with a |meta-log2-fold change| >1 or a confidence interval that does not span 0, and a meta-P-value <0.05 are considered as potentially significant.

Finally, all significantly up- and downregulated genes of the meta-analysis are fed into a gene ontology (GO) enrichment analysis using the ‘GOSstats’ R package (23) from Bioconductor.

To visualize the analysis results, Circos plots (24), forest plots (25) and Krona plots (26) are created. All data belonging to a user are stored for 30 days in the CancerMA user database, which can be accessed using the web interface during this time. This analytical approach was developed for a previous study published by the authors, and automated for the basis of the CancerMA tool. For more details, refer to Feichtinger et al. (11).

The CancerMA web interface

First, the CancerMA web interface handles the mapping of a user-supplied gene list as well as the subsequent job submission. Second, it allows the user to access the analysis results.

When submitting a new job, the user supplies a list consisting either of Ensembl IDs or of gene names, for which the identifiers are then mapped to their appropriate Affymetrix IDs by the tool to tell the user which genes can be analysed. Finally, the job can be submitted by providing an email address.

When viewing a finished job, the results of the various analyses and the visualizations are presented to the user in a simple-to-use web interface. All result files are also available for download. To view an example, visit http://www.cancerma.org.uk.

Implementation

CancerMA was developed for automated computation of the differential meta-expression for genes of interest to biologists/clinicians and, in particular, as a user-friendly and intuitive tool to view and interpret the analysis results. The CancerMA web interface for viewing the analysis data consists of three sections: a general overview, the information section as well as the result section. The general overview provides basic information about the submitted job and the data available to the user. The information section includes among others the annotated genes of interest and information about the datasets used in the analysis. The result section includes the analysis results of the meta-analysis, of the single analyses, of the single analyses (only) and of the GO enrichment analysis. The meta-analysis results comprise tables with statistical values and visualizations for the meta-upregulated as well as for the meta-downregulated genes of interest. The GO analysis results contain the enriched GO terms for the meta-up- and the meta-downregulated genes, respectively. The single analysis results show all up- and downregulations of the genes of interest in all individually analysed datasets. The single analysis (only) results, however, provide genes of interest which are either consistently up- or downregulated across the datasets. Circos and Krona plots visualize the single and meta-analysis results in their entirety to highlight relationships within the data. Furthermore, forest plots visualize the meta-analysis results for each gene separately. For a detailed documentation, please refer to the CancerMA help section (http://www.cancerma.org.uk/help.html).

Validation

We used two experimentally determined datasets providing genes differentially expressed in cancer to validate our analysis results and demonstrate the utility and the functionality of the tool: (i) 10 upregulated and 9 downregulated genes in lung cancer determined by cDNA array
Figure 2. Circos plots showing the meta-change in gene expression in relation to corresponding cancer types. The plot shows the meta-up- and meta-downregulated genes of the validation dataset from Kettunen et al. (29): (A) The expression of the genes DSP, CCNB1, PLK1, MIF, HMGA1, SFN, TIMP1 and MMP11 was found to be upregulated, whereas (B) the expression of the genes AKAP12, BMPR2, COPEB/KLF6, SOCS3, BENE/MALL, TIMP3, CAV1, CAV2 and TYROBP was found to be downregulated in lung cancer consistent with the published results. Each connection between a gene and a cancer type indicates a statistically significant mean up- or downregulation for that cancer type derived from a number of combined array studies for cancer tissue versus normal tissue. The weight of the connection corresponds to the magnitude of the meta-change in gene expression.

Figure 3. An example of a forest plot showing the expression of gene CAV1 downregulated in lung cancer. The expression of the CAV1 gene is downregulated in five of six microarray studies and upregulated in one study. The forest plot shows the meta-log 2-fold change values for the individual studies as well as the total values for lung cancer and for all cancer types combined. Each study is illustrated by a square; the position on the x-axis representing the measure estimate (log2FC ratio), the size proportional to the weight of the study and the horizontal line through it reflecting the confidence interval of the estimate.

The meta-analysis results of the 17 differentially expressed genes in lung cancer (two genes reported to be upregulated were not present on the arrays used by CancerMA) were consistent with the findings described by Kettunen et al. (29) (Figure 2). Most genes determined to be up- or downregulated in this study were reported in various other publications to be up- or downregulated accordingly (31–42). For example, the expression of the gene CAV1 was found to be highly downregulated in five of six cancer microarray datasets (Figure 3). This also provides a good example for the capability of meta-analysis techniques to identify a more valid set of differentially expressed genes, as biological, experimental and technological variations, including differences in experimental conditions, tissues, cell lines, species, platforms, sample treatment and processing can lead to inconsistencies in gene expression, which reflect the differences in the experimental setting in addition to the objective studied (43). Furthermore, interesting patterns emerge from our meta-analysis results; for example, the expression of PLK2, MMP11, CCNB1 and TIMP1 is mainly upregulated in cancer (Figure 2A), whereas the expression of AKAP12, CAV1, CAV2, COPEB/KLF6 and BENE/MALL is mainly downregulated in cancer (Figure 2B). Additionally, commonalities between cancer types can be inferred; for example, the expression pattern found in lung cancer is highly similar to the one in colorectal, ovarian and breast cancer, in particular for the upregulated genes (Figure 2A).

Our meta-analysis of the ovarian cancer validation dataset resulted in eight genes significantly upregulated, consistent with the results described by Hough et al. (30), four genes not differentially expressed and one gene...
downregulated (Figure 4). Almost all upregulated genes were reported in various other publications to also be upregulated in ovarian cancers (44–51). Several clinical trials examining the efficiency of an immunotherapy targeting the products of these genes are currently running (52, 53). According to our meta-analysis, TIMP3 expression was found to be significantly downregulated in ovarian cancer. However, this is consistent with the findings that TIMP3 is a possible tumour suppressor gene. An analysis of DNA copy number and gene expression of 22q in 18 ovarian carcinomas has shown that copy number loss across the TIMP3 locus is frequent, leading to decreased detectable TIMP3 mRNA levels (54). Furthermore, TIMP3 expression was reported to be downregulated in the lung cancer validation dataset that we used (29) and Hough et al. (30) noted that TIMP3 was not highly or consistently expressed in their tumour samples. The four genes (IGFBP2, MGP, STAT1 and SLP1) not showing significant upregulation appear to lack consistency in expression across tumour samples and/or cancer subtypes, as according to the single microarray analysis they are upregulated just in some microarray datasets (Supplementary Figure S1). This is also consistent with the findings of Hough et al. (30, 50), as they report that IGFBP2 was not consistently expressed between their tumour samples. Furthermore, STAT1 was reported to be overexpressed only in certain subtypes of serous ovarian carcinomas (55).

**Example workflow**

In our previously published work (11) we have analysed human meiotic genes using the analytical approach now implemented into CancerMA and, with RT–PCR experimental validation, identified a novel, clinically relevant subgroup of the cancer/testis gene family (the meiCT genes), which have potential as novel cancer markers and therapeutic targets. This work serves as an example workflow for potential users.

### Discussion

**Purposes and benefits of CancerMA**

CancerMA allows the automated computation of the differential meta-expression for genes of interest to biologists/clinicians across 80 cancer microarray-derived datasets.
datasets covering 13 cancer types. As shown by the validation, our meta-analysis approach enhances the statistical power by increasing the sample size and can resolve conflicting conclusions between individual studies by finding a more valid set of differentially expressed genes. Furthermore, our pipeline approach focuses on the meta-analysis on a set of related genes specified by the user, which additionally serves to enhance the significance and accuracy of the analysis, and also to narrow down the excessive number of possibilities presented by whole genome arrays to a manageable number of putative leads (17). Direct experimental evidence or other inferred relationships, such as genes involved in interaction networks, can serve as a basis to compile a set of related genes. Relationships within a gene set could include co-expression, co-regulation, affiliation to the same pathway or biological process as well as common pathological involvement.

Screening for the differential expression within a given set of genes could reveal diagnostic, therapeutic and prognostic strategies and application for specific cancer types as well as uncover common dysfunction of specific genes, gene modules or pathways across various cancer types. Furthermore, genome-scale meta-analysis can reveal common drivers of change or similar expression modules across various cancer types and therefore point towards conserved disrupted pathways or mechanisms in cancer; for example, the p53 pathway is often disrupted in cancer either due to point mutations in TP53 gene or due to one of the numerous alternative gene mutations that may lead to disruption of this pathway at key points [reviewed by Vogelstein et al. (56)]. Genetic alterations in different genes can often manifest a similar or common phenotype where these genes are related as part of the same pathway. The fact that mutations in a vast number of genes have been associated with cancer, yet disruption of only a few key pathways may give rise to the characteristics of cancer, highlights the importance of focussing on sets of related or interacting genes [reviewed by Vogelstein and Kinzler (1)].

CancerMA relies on the availability of public microarray data. Currently, we can cover 13 cancer types, but we hope that further datasets will become available in due course allowing us to expand the meta-analysis. Furthermore, we have selected datasets using the Affymetrix UG-133 Plus 2 array, as this array type is widely used and covers a large proportion of the human genome. Nevertheless, a number of genes (in particular, novel gene discoveries) are not covered by this array type and thus cannot be evaluated by this tool. However, we intend to continue the development of this tool, extending CancerMA to incorporate other Affymetrix array types and arrays form other platforms such as Illumina in due course.

Comparison to databases and tools currently available
Additionally to the repositories storing microarray data such as ArrayExpress and GEO (3), more specialized databases have become available; for example, databases such as M²DB (57) and M²PI (58) collected microarray data and uniformly pre-processed it, but do not provide data analysis and integration. Web platforms such as Oncomine (59), GEO Profiles (3), Gene Expression Atlas (available at: http://www.ebi.ac.uk/gxa/) or Gemma (60) focus on gene expression profiles across multiple conditions and tissues but do not combine the results of the individual studies. Web platforms such as GeneSapiens (61) and Genevestigator (62) combine individual studies by pooling and subsequently analysing the data with traditional techniques but do not use meta-analysis approaches. However, various microarray meta-analysis approaches are available as R packages such as metaMA (63), Rankprod (64) and metaArray (65), but require skills in statistics and R. Therefore, a simple-to-use web tool such as CancerMA providing the computation of the meta-expression profile using manually curated, patient-derived cancer microarrays for a set of genes of interests to biologists/clinicians to a wide audience is not yet available to our knowledge [for a detailed review of meta-analysis databases and tools, refer to Feichtinger et al. (5)].

Conclusion
In summary, we present CancerMA, an integrated bioinformatic analytical pipeline to automate the identification of novel candidate cancer markers/targets by means of analysing the expression of user-supplied gene lists across a manually curated database of 80 publicly available cancer microarray datasets and 13 cancer types. Such a meta-analysis enhances reliability and generalizability of the analysis results and leads to a more precise estimate of gene expression. Furthermore, the pipeline facilitates automated, homogeneous and reproducible analysis across a large number of datasets, and establishing a simple-to-use online web interface to access the pipeline puts specialist meta-analyses in the hands of biologists.

Supplementary Data
Supplementary data are available at Database online.

Acknowledgements
The authors would like to thank Dr Sarah Morgan and Dr Jane Wakeman for their assistance in proofreading the manuscript.
Funding
The Welsh National Institute for Social Care and Health Research (NISCHR) (HS/09/008 to J.F.); North West Cancer Research Fund grant (NWCRF) (CR888 to R.J.M.). Funding for open access charge: NISCHR (award number: HS/09/008).

Conflict of interest. None declared.

References
Figure S1: Circos plot showing single array gene expression analysis in relation to corresponding cancer types. The plot shows the upregulated genes of the validation dataset from Hough et al. (30). The expression of the genes GPX3, CLU, EPCAM, SPINT2, FOLR1, S100A2, APOE and CP is upregulated in numerous ovarian cancer datasets and is also significantly upregulated according to the meta-analysis (Figure 4). The genes EGFBP2, MGP, STAT1 and SLPI, however, only show upregulation in some of the ovarian datasets and thus are not significantly upregulated according to the meta-analysis (Figure 4). Each connection between a gene and a cancer type indicates a statistically significant upregulation for that cancer type derived from a single array study for cancer vs. normal tissue.
Table S1: List of cancer types covered by the microarray meta-analysis. The table lists each cancer type including the number of experiments and datasets.

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Microarray experiments</th>
<th>Microarray datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adrenal cancer</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Brain cancer</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Head and neck cancer</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Leukemia</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Pancreatic cancer</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Renal cancer</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Thyroid cancer</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>45</strong></td>
<td><strong>80</strong></td>
</tr>
</tbody>
</table>
5 CancerEST: a Web-based Tool for Automatic Meta-analysis of Public EST Data

This chapter describes the implementation, usability and validation of a bioinformatic analytical web tool for the automated identification of candidate cancer markers/ targets as well as for the investigation of tissue-specificity by means of constructing and analysing EST expression profiles of user-supplied gene lists across 36 tissue types. The web tool is based on the in-house in silico pipeline described in a previous chapter (cf. chapter 3). The work presented in this chapter contributes to project objectives 1 and 2.

Please note that this chapter is presented as manuscript to be submitted to the open-access journal Database (available at: http://database.oxfordjournals.org/). The content structure, layout, language and reference style follow the specifications of Database.
CancerEST: a Web-based Tool for Automatic Meta-analysis of Public EST Data

Julia Feichtinger*1,2, Ramsay J. McFarlane†‡1,3, and Lee D. Larcombe§4

1North West Cancer Research Fund Institute, Bangor University, Bangor, Gwynedd LL57 2UW, UK
2Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
3NISCHR Cancer Genetics Biomedical Research Unit, Bangor University, Bangor, Gwynedd LL57 2UW, UK
4Cranfield Health, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK

The identification of cancer-restricted biomarkers is fundamental to the development of novel cancer therapies and diagnostic tools. The construction of comprehensive profiles to define tissue- and cancer-specific gene expression has been central to this. To this end, the exploitation of the current wealth of “omic”-scale databases can be facilitated by automated approaches, allowing researchers to directly address specific biological questions. Here we present CancerEST, a user-friendly and intuitive web-based tool for the automated identification of candidate cancer markers/targets, for comprehensively examining tissue-specificity as well as for integrated expression profiling. CancerEST operates by means of constructing and meta-analyzing expressed sequence tag (EST) profiles of user-supplied gene sets across an EST database supporting 36 tissue types. Using a validation dataset from the literature, we show the functionality and utility of CancerEST. Database URL: http://www.cancerest.org.uk

*bspa33@bangor.ac.uk
†Author to whom correspondence should be addressed
‡r.mcfarlane@bangor.ac.uk
§leelarcombe@gmail.com
Introduction

Identifying novel candidate markers/targets is a key challenge in the development of cancer therapies (1). Tissue- and cancer-specific gene expression profiles provide information about the potential of genes to serve as clinical markers (2). Thus, accessible and automated approaches analyzing the current wealth of “omic”-scale data are required to facilitate the full exploitation of expression data. Expressed sequence tags (ESTs) are short DNA sequences (200-500 nucleotides) generated by sequencing the 5’ and/or 3’ ends of cDNAs that are subsequently clustered and counted (3). In the last decade, a large amount of EST data has been deposited in public repositories such as dbEST (4), which currently holds records of 8,692,773 human ESTs. Unigene has grouped these expression data into clusters and assigned them to genes, facilitating the indexing of the EST data (5). Pipelining the retrieval, the integration and the high-throughput investigation of such data in a fashion specifically tailored to the interests of the user, should facilitate wider application by putting EST data in the hands of researchers directly addressing focused biological questions, without requiring the involvement of bioinformaticians. Integration and subsequent investigation of EST data can not only enhance reliability and generalizability of results, but can also reveal a comprehensive expression profile across numerous tissues, which can be used to uncover information about tissue-specific expression, cancer expression and above all, about cancer marker/target potential (6). For example, Kim et al. (7) and Campagne and Skrabanek (8) identified potential cancer markers by means of EST data analyses, whereas Hofmann et al. (9) used EST data, reverse transcription polymerase chain reaction (RT-PCR) and other high-throughput gene expression data to evaluate the tissue-specificity and the cancer gene expression profiles of previously published cancer testis (CT) genes. CT genes are highly restricted, cancer-specific genes and encode a family of proteins that are widely used in clinical applications (10).

Here we present CancerEST, a freely accessible pipeline with a user-friendly and intuitive web interface to provide automated high-throughput investigation of public EST data with user-defined sets of biologically significant and related genes to determine (i) their cancer marker/target potential; (ii) their tissue-specificity; and (iii) their comprehensive expression profiles across 36 tissues (Table S1). The underlying method was developed for a previously published study, where we identified a cohort of novel cancer-specific marker genes (11), and has been improved and automated to provide the basis of CancerEST, a web-based tool with visualizations to aid data interpretation. This tool allows biologists/clinicians without skills in bioinformatics, to exploit the wealth of publicly available data presented by modern databases towards the challenge of fo-
cussing the overwhelming number of putative target genes on a manageable number of candidates, which can be followed up in the laboratory. To validate our approach, we have analyzed a list of testis-restricted genes from literature (9), and could reproduce the published results.

Methods and Structure of CancerEST

CancerEST consists of a web interface, pipelined analyses and three relational databases; one holding the analysis data, one holding the Unigene data and another one holding the gene annotation data. The principle workflow is shown in Figure 1.

**Figure 1: CancerEST workflow.** The complete Unigene database was established as a local MySQL database and subsequently used to construct meta-libraries for 36 tissue types allowing the computation of integrated expression profiles for all genes with assigned Unigene clusters. The web interface box indicates the areas, where the user provides input and/or can view the mapping or analysis results. The analysis is carried out automatically without any user input and computes integrated expression profiles tailored to the interests of the user with visualizations to aid the data interpretation.

The CancerEST Web Interface

First, the CancerEST web interface handles the user specifications and mapping of the user-supplied gene list as well as the job submission. Second, it allows the user to view and download the analysis results and visualizations.
When submitting a new job, the user provides a text file consisting either of Unigene Cluster IDs or of curated gene names, for which the identifiers are then mapped to their appropriate Unigene Cluster IDs to show the user which genes can be fed into the analysis. Furthermore, the user has to specify a tissue focus, where submitted genes are allowed to show expression in normal individuals; for example, the testis might be of interest to the user, as it is an immunologically privileged tissue (12). The user can optionally select an interfering tissue(s), where submitted genes are tolerated to show additional expression in normal individuals; for example, brain tissue could be selected by the user, as various genes that have been originally assumed to be testis-restricted are also expressed in the brain, another tissue residing in immunological privilege (13). Finally, the job can be submitted by providing an email address.

When viewing a finished job, the results of the analysis and the visualizations are presented to the user in a simple-to-use web interface. All result files are also available for download. We provide an example dataset on our website (available at: http://www.cancerest.org.uk).

**EST Data Retrieval, Data Quality and CancerEST Databases**

We obtained the complete data available from the Unigene database (Unigene Build #230) (5) and set up a local MySQL database. We excluded ESTs from normalized and subtracted cDNA libraries (6) as well as cDNA libraries deriving from uncharacterized, mixed or embryonic/fetal tissues. The exclusion of cancer cell line libraries is optional and can be specified by the user. Furthermore, we kept only libraries originating from cancerous and healthy tissues and thus excluded libraries deriving from diseases other than cancer. All ESTs of a given tissue type $t$ were merged to a meta-library. However, meta-libraries with an EST count below 10,000 were excluded to assure significance, resulting in cancer and normal meta-libraries for 36 tissue types (Table S1). For each Unigene cluster the global expression profile in cancerous and healthy tissues is computed by EST counting, following the concept of the Unigene EST profiles (5). The expression profiles in cancerous and healthy tissues are normalized by calculating the transcripts per million $tpm_{t,c}$, where $m_{t,c}$ is the number of ESTs for a given cluster $c$ and for a given tissue type $t$, and $n_t$ is the total number of ESTs for that given tissue type $t$:

$$tpm_{t,c} = \frac{m_{t,c}}{n_t} \cdot 10^6$$

For annotation purposes, the Ensembl database (14) and the HUGO Gene Nomenclature Committee (HGNC) database (15) were established as a local MySQL database.
The CancerEST Pipeline

The pipeline handles the EST meta-analysis, the annotation and the visualizations. For each of the submitted genes the expression profile is examined to determine the expression in the user-specified tissue focus, in possible interfering tissues, in all other healthy tissues as well as in all cancer-derived tissues. The weighted average $tpm_{av}$ for these four tissue groups is computed, where $w_t$ is the weight of the given tissue $t$ belonging to the set of tissues $g$, represented by the size of the meta-library:

$$tpm_{av,g,c} = \frac{\sum tpm_{t,c} \cdot w_t}{\sum w_t}$$

Genes are sorted into four classes according to their expression profile to provide information about their potential as cancer antigen encoding genes: (i) tissue focus-restricted expression in normal individuals as well as cancer expression (class 1); (ii) tissue focus-and interfering tissue-restricted expression in normal individuals as well as cancer expression (class 2); (iii) tissue focus- and/or interfering tissue-restricted expression in normal individuals but no cancer expression (class 3); and (iv) somatic expression in normal individuals (class 4). The classes are additionally designated with an ‘a’ if no focus expression was found.

The genes are also sorted into four states to provide information about their tissue-specificity: (i) tissue-specific (classes 1-3); (ii) highly tissue-selective ($tpm_{t,c} \leq 2$ for all other healthy tissues); (iii) tissue-selective ($tpm_{t,c} \leq 5$ for all other healthy tissues); and (iv) enriched (the $tpm_{av,c}$ of the tissue focus is twice the $tpm_{t,c}$ of each of the other healthy tissues).

In order to evaluate the upregulation of genes of interest in cancer, the significance of upregulation is accessed using the Fisher’s exact test (16). Genes with a p value $< 0.05$ are considered to be upregulated in these cancer types.

To visualize the analysis results, Circos plots (17) and bar charts are created. All data belonging to a user is stored for 30 days in the CancerEST user database, which can be accessed using the web interface during this time. This analytical approach was developed for a previous study published by the authors (11), and improved and automated for the basis of the CancerEST tool.
Implementation

CancerEST is running on an Intel core i7 2.66 Ghz workstation with 12 Gb RAM and installed with CentOS 5.4 GNU Linux OS (x86_64). MySQL 5.0.77 (available at: http://www.mysql.com) was used for the relational databases. The CancerEST web interface was implemented using: HTML/CSS, Twitter Bootstrap (available at: http://twitter.github.com/bootstrap/), Javascript/jQuery (available at: http://jquery.com/) and Perl 5.8.8 (available at: http://www.perl.org). The CancerEST pipeline was implemented using Perl 5.8.8 (available at: http://www.perl.org). CancerEST is freely available online at http://www.cancerest.org.uk.

Use of CancerEST

CancerEST was developed as a user-friendly and intuitive tool to compute cancer marker/target potential as well as to obtain comprehensive expression profiles and information about the tissue-specificity for genes of interest to biologists/clinicians. The CancerEST web interface for viewing the analysis results consists of three sections: the overview, the information, and the result section. The overview section provides basic information about the submitted job and a brief explanation how to interpret the results. The information section includes among others the annotated genes of interest and the 36 tissue types supported by CancerEST. The result section includes the EST meta-analysis results comprising of a ranked list of genes according to (i) their cancer marker potential; or to (ii) their tissue-specificity. Furthermore, a comprehensive expression profile across 36 healthy and cancerous tissues is available for each gene. Circos plots visualize the analysis results in their entirety to highlight relationships between the genes and the cancer types. In contrast, bar charts show the complete expression profile across 36 healthy and cancerous tissues for each gene separately. For more information, the CancerEST help section provides a detailed documentation, available at http://www.cancerest.org.uk/help.html.

Validation

We used the 39 tight testis-restricted genes determined by Hofmann et al. as a validation dataset (4 genes could not be mapped to a Unigene cluster ID or to a HGNC gene name, resulting in 35 genes that could be evaluated). Hofmann et al. have evaluated the tissue- and cancer-specific expression of 153 CT genes previously published in the CTdatabase (18) using high-throughput expression data in combination with RT-PCR data (9). We selected ‘testis’ as tissue focus and chose ‘brain’ as interfering tissue, as it
has been shown that various CT genes also exhibit expression in brain tissue (13). To be in accordance with Hofmann et al., we additionally allowed placental gene expression and included cancer cell line libraries. CancerEST determined 25 of these genes as not expressed in any healthy tissue or as tight testis-restricted (Table S2). Additionally, seven genes were found to show limited evidence for brain expression, which could have been below the threshold of Hofmann et al.; however, these seven genes are not expressed in any other healthy tissue, consistent with Hofmann et al. The remaining three genes exhibit expression in other healthy tissues, although two of the genes show expression in only one, and one of the genes in only two other healthy tissues. Hofmann et al. detected cancer expression for 21 of the 35 testis-restricted genes. CancerEST reported cancer expression for 20 of these 21 genes and additionally predicted cancer expression for GAGE6 (Table S2), which has also been reported in the literature (19). In total, CancerEST predicted that 19 genes have high cancer marker/target potential by exhibiting a testis- or testis-brain-restricted expression profile as well as cancer expression (Figure 2, Table S2). For example, the gene MAGEA1, which encodes the first CT genes to be discovered (20) is, according to CancerEST, expressed in various cancers including melanoma, lung cancer, breast cancer, bone and connective tissue sarcomas (Figure 3); an observation which is supported extensively through literature (21–25).

The results are consistent with Hofmann et al.; however, CancerEST uses a very stringent cutoff, which could explain the weak evidence for expression in the brain that was found for seven genes as well as the limited evidence for expression in healthy tissues that was found for three genes. Furthermore, with more EST data becoming available, the predictions become increasingly accurate, and CT genes originally believed to have testis-restricted expression profiles have to be adapted to testis-selective (9, 13). An alternative explanation for the limited evidence for expression in healthy tissues could be undiagnosed neoplastic change in the tissues analyzed, as many normal tissues are extracted from tissue obtained post mortem and are often pooled from tissues from a number of individuals, many of whom were aged at time of death. In support of this, Chen et al. found discrepancies concerning the expression of some genes in normal tissues, as they detected expression in tissues from one panel of normal tissues, but could not detect expression in similar tissue types from a distinct second source (26). Thus, genes with testis-selective profiles could indeed be suitable candidates and be of clinical use.

In conclusion, tissue-specificity was predicted accurately in 71% of the cases, including the genes showing expression in the brain even in 91% of the cases. Cancer expression was predicted correctly in 95% of the cases. Furthermore, Hofmann et al. reported that
Figure 2: Circos plot showing the gene expression in relation to the corresponding cancer types for the 39 testis-restricted genes determined by Hofmann et al. (9). 21 of the 39 testis-restricted genes exhibit expression in various cancer types, in particular in melanoma. Each connection between a gene and a cancer type indicates expression in a cancer. The magnitude of the connection corresponds to the transcripts per million (tpm) for the given gene in a given tissue.

The widest range of CT gene expression was found in melanoma (9), which is consistent with our results (Figure 2) and the literature (27).

In our previous work (11) we have analyzed human meiotic genes using the approach now implemented into CancerEST and, with RT-PCR experimental validation and microarray meta-analysis, identified a novel, clinically relevant subgroup of the CT gene family (the meiCT genes), whose associated proteins have potential as novel cancer markers and therapeutic targets. This work can serve as an example workflow for potential users as well as a further validation dataset.
Figure 3: An example of a bar chart showing the integrated expression profile of the MAGEA1 gene. MAGEA1 exhibits a testis-restricted gene expression profile, but is aberrantly expressed in a number of cancer types. The expression is given in transcripts per million (tpm).

Discussion

Purposes and Benefits of CancerEST

As tissue-specific gene expression plays a fundamental role in human biology and disease, the identification of genes with restricted/specific expression patterns helps to understand development, function and homeostasis of the distinct cell/tissue types as well as etiology, gene-tissue relationships and gene functions, thus aiding the discovery of novel marker/target genes (28–30). However, establishing a comprehensive map of tissue-specific expression for the complete human body poses an immense challenge due to the difficulty of obtaining such data empirically, but can be facilitated by combining publicly available high-throughput expression data. CancerEST allows the automated
construction of integrated expression profiles based on EST data across 36 tissues and thus can examine the tissue-specificity as well as identify suitable cancer marker/therapeutic targets for a set of genes of interest, as shown by our validation. CancerEST permits users to focus on a manageable number of candidate genes, which can be followed up in the laboratory and thus decreases the risk to pursue unsuitable targets. The putative candidate genes could be used for diagnostic, therapeutic and prognostic strategies for specific cancer types, or to uncover common dysfunction of gene modules across various cancer types. Analyzing a set of co-expressed, co-regulated, interacting or otherwise related genes, however, can point to conserved disrupted pathways or mechanisms in cancer, as mutations in a vast number of genes have been associated with cancer, yet disruption of only a few key pathways may give rise to the characteristics of cancer (31).

**Comparison to Tools Currently Available**

Several tools exist that exploit EST data to construct integrated expression profiles; for example TissueInfo (32) and TiGER (33) allow determining the tissue-specificity for a given gene or tissue-specific genes for a given tissue, but do not evaluate cancer expression or cancer marker/target potential, and importantly, neither allow the analysis for sets of genes. In contrast, the Unigene tool, Digital Differential Display (DDD) (5) compares EST profiles of user-defined EST libraries to identify genes with significantly different expression levels, and another Unigene tool, the EST Profile Viewer (5) shows the approximate expression profile for a given gene. However, neither of the two focuses on the cancer marker/target potential for a set of related genes. Several other tools were published but appear to be currently unavailable (DigiNorthern (34), ZooDDD (35), GBA server (36)). Therefore, a simple-to-use web tool such as CancerEST computing the cancer marker/target potential, the tissue-specificity as well as comprehensive expression profiles for a set of genes of interests to biologists/clinicians is not available to our knowledge.

**Conclusion**

In summary, we present CancerEST, an integrated bioinformatic analytical pipeline to automate the identification of novel candidate cancer markers/targets and/or to determine the tissue-specificity by means of constructing and analyzing the EST expression profiles of user-supplied gene lists across 36 tissue types. Furthermore, such an automated pipeline with a simple-to-use web interface puts an integrated EST analysis in the hands of researchers who are directly addressing biological questions.
Funding

This work was supported by the Welsh National Institute for Social Care and Health Research (NISCHR) [HS/09/008 to J.F.], and partly by North West Cancer Research Fund grant (NWCRF) [CR888 to R.J.M.]. The authors declare no conflict of interest.

Acknowledgments

The authors would like to thank Dr. Gerhard Thallinger for his support.
References


For submission to the journal Database


Table S1: List of tissue types supported by CancerEST. Normal and cancer meta-libraries were constructed for the 36 listed tissue types, allowing the construction of integrated expression profiles based on these tissues.

<table>
<thead>
<tr>
<th>#</th>
<th>Tissue type</th>
<th>#</th>
<th>Tissue type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adipose tissue</td>
<td>19</td>
<td>Nerve</td>
</tr>
<tr>
<td>2</td>
<td>Adrenal gland</td>
<td>20</td>
<td>Ovary</td>
</tr>
<tr>
<td>3</td>
<td>Blood</td>
<td>21</td>
<td>Pancreas</td>
</tr>
<tr>
<td>4</td>
<td>Bone</td>
<td>22</td>
<td>Pharynx</td>
</tr>
<tr>
<td>5</td>
<td>Bone marrow</td>
<td>23</td>
<td>Pituitary gland</td>
</tr>
<tr>
<td>6</td>
<td>Brain</td>
<td>24</td>
<td>Placenta</td>
</tr>
<tr>
<td>7</td>
<td>Connective tissue</td>
<td>25</td>
<td>Prostate</td>
</tr>
<tr>
<td>8</td>
<td>Ear</td>
<td>26</td>
<td>Skin</td>
</tr>
<tr>
<td>9</td>
<td>Eye</td>
<td>27</td>
<td>Spleen</td>
</tr>
<tr>
<td>10</td>
<td>Heart</td>
<td>28</td>
<td>Stomach</td>
</tr>
<tr>
<td>11</td>
<td>Intestine</td>
<td>29</td>
<td>Testis</td>
</tr>
<tr>
<td>12</td>
<td>Kidney</td>
<td>30</td>
<td>Thymus</td>
</tr>
<tr>
<td>13</td>
<td>Liver</td>
<td>31</td>
<td>Thyroid</td>
</tr>
<tr>
<td>14</td>
<td>Lung</td>
<td>32</td>
<td>Tonsil</td>
</tr>
<tr>
<td>15</td>
<td>Lymph node</td>
<td>33</td>
<td>Trachea</td>
</tr>
<tr>
<td>16</td>
<td>Mammary gland</td>
<td>34</td>
<td>Umbilical cord</td>
</tr>
<tr>
<td>17</td>
<td>Mouth</td>
<td>35</td>
<td>Uterus</td>
</tr>
<tr>
<td>18</td>
<td>Muscle</td>
<td>36</td>
<td>Vascular</td>
</tr>
</tbody>
</table>
Table S2: Expression profiles of the 39 testis-restricted genes determined by Hofmann et al. The table compares the tissue- and cancer-specific expression detected by Hofmann et al. to the one detected by CancerEST for the 39 genes. Results for genes highlighted in dark gray are consistent with the results of Hofmann et al. Genes highlighted in light gray show either weak evidence for additional brain expression, exhibit cancer expression, which was not initially reported by Hofmann et al., or do not show cancer expression as described by Hofmann et al. Genes not highlighted show expression in other healthy tissues, which was not initially detected by Hofmann et al. Genes designated with ‘na’ could not be assigned to a Unigene cluster or to a HGNC gene name.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Hofmann et al. classification</th>
<th>CancerEST class</th>
<th>CancerEST state</th>
<th>Expression in healthy tissues</th>
<th>Expression in cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPXCR1</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Skin</td>
</tr>
<tr>
<td>CSAG2</td>
<td>Testis-restricted, cancer expression</td>
<td>1a</td>
<td>Specific</td>
<td>Placenta</td>
<td>Prostate, mammary gland, bone, lung, connective tissue, intestine, lymph node, adrenal gland</td>
</tr>
<tr>
<td>CTAG1A</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Bone</td>
</tr>
<tr>
<td>MAGEA1</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Skin, bone, mammary gland, lung, connective tissue</td>
</tr>
<tr>
<td>MAGEA2</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Stomach, pancreas, mammary gland</td>
</tr>
<tr>
<td>MAGEC1</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Brain, placenta, bone, skin</td>
</tr>
<tr>
<td>PAGE2</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Connective tissue</td>
</tr>
<tr>
<td>PAGE2B</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Stomach</td>
</tr>
<tr>
<td>SPANXA1</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Liver, bone marrow</td>
</tr>
<tr>
<td>SPANXA2</td>
<td>Testis-restricted, cancer expression</td>
<td>1a</td>
<td>Specific</td>
<td>Testis</td>
<td>Liver</td>
</tr>
<tr>
<td>SPANXD</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Connective tissue, skin, liver</td>
</tr>
<tr>
<td>SSX3</td>
<td>Testis-restricted, cancer expression</td>
<td>1</td>
<td>Specific</td>
<td>Testis</td>
<td>Skin, bone marrow</td>
</tr>
<tr>
<td>DDX53</td>
<td>Testis-restricted</td>
<td>3</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>Gene</td>
<td>Expression Pattern</td>
<td>Specificity</td>
<td>Tissue(s)</td>
<td>Other Tissues</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>--------------------------------------------------------</td>
<td>------------</td>
<td>----------------</td>
<td>-----------------------------------</td>
<td></td>
</tr>
<tr>
<td>FTHL17</td>
<td>Testis-restricted</td>
<td>3a</td>
<td>Specific</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>MAGEB4</td>
<td>Testis-restricted</td>
<td>3</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>MAGEB5</td>
<td>Testis-restricted</td>
<td>3a</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>MAGEB6</td>
<td>Testis-restricted</td>
<td>3</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>SPANXB1</td>
<td>Testis-restricted</td>
<td>3a</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>SPANXC</td>
<td>Testis-restricted</td>
<td>3</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>SPANXN3</td>
<td>Testis-restricted</td>
<td>3</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>SPANXN4</td>
<td>Testis-restricted</td>
<td>3</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>SPANXN5</td>
<td>Testis-restricted</td>
<td>3a</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>XAGE5</td>
<td>Testis-restricted</td>
<td>3</td>
<td>Specific</td>
<td>Testis</td>
<td>None</td>
</tr>
<tr>
<td>GAGE6</td>
<td>Testis-restricted</td>
<td>1a</td>
<td>Specific</td>
<td>Stomach</td>
<td></td>
</tr>
<tr>
<td>DKKL1</td>
<td>Testis-restricted, cancer expression</td>
<td>2</td>
<td>Specific</td>
<td>Testis, brain</td>
<td>Brain, connective tissue</td>
</tr>
<tr>
<td>MAGEB1</td>
<td>Testis-restricted, cancer expression</td>
<td>2</td>
<td>Specific</td>
<td>Testis, brain</td>
<td>Skin</td>
</tr>
<tr>
<td>MAGEB2</td>
<td>Testis-restricted, cancer expression</td>
<td>2</td>
<td>Specific</td>
<td>Testis, brain</td>
<td>Skin, stomach, bone, lung</td>
</tr>
<tr>
<td>SAGE1</td>
<td>Testis-restricted, cancer expression</td>
<td>2</td>
<td>Specific</td>
<td>Testis, brain</td>
<td>Bone marrow</td>
</tr>
<tr>
<td>SSX1</td>
<td>Testis-restricted, cancer expression</td>
<td>2</td>
<td>Specific</td>
<td>Testis, brain</td>
<td>Connective tissue, skin, liver,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>bone, bone marrow, placenta</td>
</tr>
<tr>
<td>SSX2</td>
<td>Testis-restricted, cancer expression</td>
<td>2</td>
<td>Specific</td>
<td>Testis, brain</td>
<td>Liver, uterus, skin, placenta</td>
</tr>
<tr>
<td>CTAG1B</td>
<td>Testis-restricted, cancer expression</td>
<td>3a</td>
<td>Specific</td>
<td>Testis, brain</td>
<td>None</td>
</tr>
<tr>
<td>TSPY1</td>
<td>Testis-restricted</td>
<td>3</td>
<td>Specific</td>
<td>Testis, brain</td>
<td>None</td>
</tr>
<tr>
<td>CXorf48</td>
<td>Testis-restricted, cancer expression</td>
<td>4</td>
<td>Testis, brain,</td>
<td>Connective tissue, bone, skin,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>eye</td>
<td>stomach</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Expression Pattern</td>
<td>Level</td>
<td>Expression Pattern</td>
<td>Tissues</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------</td>
<td>-------</td>
<td>-----------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>MAGEA2B</strong></td>
<td>Testis-restricted, cancer expression</td>
<td>4</td>
<td>Testis, muscle, skin</td>
<td>Mammary gland, connective tissue, prostate, bone marrow, skin, pancreas, bone, adrenal gland, placenta, stomach</td>
<td></td>
</tr>
<tr>
<td><strong>MAGEB3</strong></td>
<td>Testis-restricted</td>
<td>4</td>
<td>Enriched</td>
<td>Testis, prostate</td>
<td>None</td>
</tr>
<tr>
<td><strong>CT69</strong></td>
<td>Testis-restricted</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td><strong>CT70</strong></td>
<td>Testis-restricted</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td><strong>LOC203415</strong></td>
<td>Testis-restricted</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td><strong>SPANXLE</strong></td>
<td>Testis-restricted, cancer expression</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Meta-analysis of Germline Gene Expression Points to a Soma-to-germline Transformation in Human Cancer Cells

This chapter describes the employment of the bioinformatic analytical web tools developed in the course of this thesis (cf. chapters 3-5) to investigate the expression profiles of the human homologues of *Drosophila* germline genes ectopically expressed in brain tumours caused by a mutation in the retinoblastoma pathway, inducing a soma-to-germline transformation. The results revealed that these human germline genes were overexpressed or aberrantly expressed in a wide range of human cancer types, which could indicate that human cells undergo a similar soma-to-germline transformation in the course of the development of cancer. The work presented in this chapter contributes to project objective 4.

Please note that this chapter is presented as manuscript to be submitted to the journal *Translational Oncology* (available at: [http://www.translationaloncology.org/](http://www.translationaloncology.org/)). The content structure, layout, language and reference style follow the specifications of *Translational Oncology*. 
Meta-analysis of Germline Gene Expression Points to a Soma-to-germline Transformation in Human Cancer Cells

Julia Feichtinger*, Lee D. Larcombe†, and Ramsay J. McFarlane‡§

1North West Cancer Research Fund Institute, Bangor University, Bangor, Gwynedd LL57 2UW, UK
2Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
3Cranfield Health, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK
4NISCHR Cancer Genetics Biomedical Research Unit, Bangor University, Bangor, Gwynedd LL57 2UW, UK

Cancer cells might undergo a soma-to-germline transformation, which may support the acquisition of malignant characteristics such as rapid proliferation. Such transformations have been reported in Drosophila melanogaster and Caenorhabditis elegans with mutations in chromatin regulators associated with germline-soma distinction. Here we have meta-analyzed the expression profiles of the human homologues of Drosophila germline genes ectopically expressed in brain tumors caused by a mutation in the retinoblastoma pathway and could find these human germline genes overexpressed or aberrantly expressed in a wide range of human cancer types. This points to the occurrence of a similar soma-to-germline transformation in humans and can shed light into the contribution of their gene products to a soma-to-germline transformation as well as to oncogenic events in the course of the development of cancer.
Introduction

Cancer and germ cells exhibit profound commonalities such as rapid proliferation, undifferentiated phenotype and immortality/lack of senescence [1]. Cancer cells could acquire these characteristics through a soma-to-germline transformation, which in turn could be induced through a dysfunctional control of germline-specific genes [2]. In support of this hypothesis, Janic et al. [3] recently showed that \(l(3)mbt\) tumors in \textit{Drosophila melanogaster} ectopically express germline genes; even a quarter of the up-regulated genes in \(l(3)mbt\) tumors encode proteins associated with germline functions and subsequent inactivation of these genes resulted in suppression of tumor growth. \(L(3)MBT\) is a transcriptional repressor [4] and a component of the dREAM-MMB complex [5]. Inactivation of other components of the dREAM-MMB complex also led to ectopic expression of germline genes [6]. Similar soma-to-germline transformations were found in \textit{Caenorhabditis elegans} strains with mutations in the homologues of dREAM-MMB complex components or in other functionally related repressors [7,8]. In humans, a group of genes with expression restricted to testicular cells, the so-called cancer testis (CT) genes, are aberrantly expressed in various cancer types, leading to the suggestion of a soma-to-germline transformation occurring also in human cancer cells [2]. The immunological privilege of the testis makes the CT antigens to promising candidates for immunotherapy [9] and a number of CT antigens are currently under investigation for their potential as cancer therapeutics [10,11]. Some CT genes are also expressed in placental and brain tissue, which also represent immunologically privileged areas [12,13].

As ectopic expression of germline genes could lead to gene products contributing to the acquisition of tumor characteristics, it is important to investigate the expression patterns of such genes in cancerous and healthy human tissues. In our previous work, we could already identify novel CT candidate genes by meta-analyzing the expression profiles of human homologues of mouse meiotic genes using expressed sequence tag (EST) and microarray data, validated by reverse transcription polymerase chain reaction (RT-PCR) [14]. These genes represent a novel subgroup of the CT gene family, the meiCT genes, whose associated proteins are likely to be involved in meiotic spermatogenesis, also supporting the hypothesis that human cancer cells undergo a soma-to-germline transformation. Investigating the expression of \textit{Drosophila} germline genes ectopically expressed in \(l(3)mbt\) tumors, in humans could determine if their human homologues are also ectopically expressed in various human cancers and thus may provide further support for a soma-to-germline transformation. Further evaluation of their expression profiles and tissue-specificity could also reveal new potential drug targets. A few homologues of these \textit{Drosophila} germline genes are already known to be aberrantly expressed.
in human cancer cells, as they are previously characterized CT genes; for example, SYCP1 is the human homologue of the *Drosophila* germline gene *c(3)G* [15]. Hence we evaluated the meta-expression profiles of the human homologues using our previously developed meta-analyses approaches [14] and provide evidence that 40 of 46 human homologues indeed exhibit ectopic cancer expression or upregulation in cancer, showing that these germline genes are also dysregulated/derepressed in human cancers. Furthermore, we show that 19 genes have testis- or testis/brain-restricted expression patterns and thus could potentially be used as therapeutic markers/targets.

**Materials and Methods**

**Human Homologues of the *Drosophila* Germline Genes**

We assigned the 49 *Drosophila* germline genes ectopically expressed in *l(3)mbt* tumors (3) to their human orthologues using the databases Flybase [16], Homologene [17] and Ensembl [18] as well as literature search (Table S1). We could identify human orthologues for 28 genes, resulting in 46 human genes due to numerous human paralogues. Enriched gene ontology (GO) terms for the human homologues were determined using the functional annotation tool DAVID [19].

**EST Meta-analysis**

43 of the 46 human homologues could be mapped to Unigene IDs. A comprehensive EST expression profile across 36 tissues was constructed for these genes based on a methodology developed for a previous study [14]. Briefly, all ESTs of a given tissue type *t* available from the Unigene database (Unigene Build #230) [20] were merged to a meta-library, excluding ESTs from normalized and subtracted cDNA libraries or deriving from uncharacterized, mixed or embryonic/fetal tissues. Meta-libraries with an EST count below 10,000 were excluded to assure significance, resulting in cancer and normal meta-libraries for 36 tissue types. For each Unigene cluster the global expression profile in cancerous and healthy tissues is computed by EST counting, following the concept of the Unigene EST profiles [20]. The expression profiles in cancerous and healthy tissues were normalized by calculating the transcripts per million ($tpm_{t,c} = \frac{m_{t,c}}{n_t} \cdot 10^6$), where $m_{t,c}$ is the number of ESTs for a given cluster *c* and for a given tissue type *t*, and $n_t$ is the total number of ESTs for that given tissue type *t*. Genes with expression restricted to the testis, brain and placenta as well as limited expression in one or two tissues were selected to be testis- or testis/brain restricted. The significance of upregulation in cancer was calculated using the Fisher’s exact test [21]. Genes with a p value < 0.05
or with expression in cancerous meta-libraries but not in the corresponding healthy meta-libraries were considered to be upregulated or ectopically expressed, respectively, in these cancer types. To visualize the analysis results, Circos plots [22] and bar charts were created.

**Single and Meta-analysis of Microarray Studies**

41 of the 46 human homologues could be mapped to Affymetrix array indices for the HG-U133 Plus 2 array and thus could be evaluated for their differential expression in 13 cancer types by means of a meta-analysis approach developed for a previous study [14]. Briefly, we searched for raw data of patient-derived, untreated cancer samples with corresponding normal samples deposited in ArrayExpress [23] or Gene Expression Omnibus (GEO) [24]. After manual curation and quality control [25], we obtained 80 individual cancer datasets originating from 45 experiments and covering 13 different cancer types. All datasets were preprocessed individually according to the methods described by Hubbell et al. [26] to assure uniformity of the analysis process. Subsequently, 80 datasets were filtered with the genes investigated in order to reduce the number of features and to enhance the statistical power [27]. We used the ‘Limma’ R package [28] from Bioconductor to compute differentially expressed genes and adjusted the resulting p values for multiple testing with Benjamini and Hochberg’s method to control the false discovery rate [29]. For the single array analysis, genes with a p value < 0.05 and a |log2-fold change| > 1 were selected as potentially significant. For the meta-analysis, a meta-p value and a meta-log2-fold change value were calculated for each cancer type using Stouffer’s method [30] and weighted linear combination [31], respectively. Genes with a |meta-log2-fold change| > 1 or a confidence interval that does not span 0, and a meta-p value < 0.05 were considered as potentially significant. To visualize the analysis results, Circos plots [22] and Forest plots [32] were created.

**Implementation**

The meta-analysis pipelines described above were implemented using: R 2.12.1 (available at: http://www.cran.r-project.org) [33]; the Bioconductor package (available at: http://www.bioconductor.org) [34]; MySQL 5.0.77 (available at: http://www.mysql.com); and Perl 5.8.8 (available at: http://www.perl.org).
Results

Human Homologues of the *Drosophila* Germline Genes

Janic *et al.* reported 49 *Drosophila* germline genes to be overexpressed in *l(3)mbt* tumors [3]. We could map 28 of these genes to their human orthologues, resulting in 46 human genes due to human paralogues (Table S1). In support of this, the top enriched gene ontology (GO) terms show that their gene products are mainly involved in meiosis, spermatogenesis and reproduction (Table S2).

EST Meta-analysis

We have investigated 43 human orthologues for their cancer expression, cancer marker potential and tissue-specificity based on the construction of a comprehensive expression profile (Three genes could not be mapped to Unigene IDs). Briefly, if genes show expression only in immunologically privileged tissues and in not more than two other healthy tissues, the genes are considered as testis- or testis/brain restricted. 19 genes exhibit such an expression profile (Table S3) including the previously characterized CT genes, *SYCP1* [15], *TDRD1* [35] and *PIWIL2* [36]. *MAEL*, also a previously characterized CT gene [37], however, shows expression in three normal tissues. Furthermore, 13 of these 19 genes exhibit ectopic cancer expression in at least one cancer type; for example, the gene *C16orf73* is expressed in the testis, brain and placenta as well as ectopically expressed in melanoma and sarcomas of the bone and of the connective tissue (Figure S1). In total, however, 35 of 43 human homologues exhibit ectopic expression or are upregulated in cancer (Figure 1).

Single and Meta-analysis of Microarray Studies

We evaluated the differential expression for 41 human orthologues based on a microarray meta-analysis approach across 13 cancer types (Five genes are not present on the arrays). 31 of the 41 human orthologues are significantly upregulated in eleven distinct cancer types (Figure 2, Table S4). Nine of the 19 testis- or testis/brain-restricted genes were found to be significantly upregulated, in particular in ovarian and brain cancer. For example, the gene *RN17* shows upregulation in ovarian, prostate and brain cancer (Figure S2, Table S4). Several genes were also found to be downregulated in some cancer types, including the genes *CPEB1* and *ESRP1*. Furthermore, analysis of differential expression in 80 individual microarray studies provides evidence that even 39 of the total 41 genes and 14 of the 19 testis- or testis/brain-restricted may be upregulated in cancer (Figure S3).
Cancer Expression of Human Germline Genes

Combining the results of the EST meta-analysis, of the single microarray analysis as well as of the microarray meta-analysis can provide a comprehensive picture of the cancer expression of the human germline genes investigated. In addition to the four previously characterized CT genes (*SYCP1, TDRD1, MAEL* and *PIWIL2*), 36 other germline genes show evidence for ectopic expression or upregulation in various cancer types (Table S5).
Figure 2: Circos plot showing the meta-change in gene expression in relation to corresponding cancer types for the 41 human homologues based on the microarray meta-analysis. 31 of these 41 human genes covered by the arrays exhibit an upregulation according to the microarray meta-analysis. Each connection between a gene and a cancer type indicates a statistically significant mean upregulation for that cancer type derived from a number of combined array studies for cancer vs. normal tissue. The weight of the connection corresponds to the magnitude of the meta-change in gene expression.

The investigated genes mainly belong to large germline gene families. These include among others the DAZ family genes DAZ1, DAZ2, DAZ3, DAZ4, DAZL and BOLL, the NANOS family genes NANOS1 and NANOS3, the PIWIL family genes PIWIL1 and PIWIL2, the TDRD family genes TDRD1, RNF14/TDRD4, TDRD7 and TDRD9, the TALE/TGIF family genes TGIF2 and TGIF2LY, and the OVOL family genes OVOL1 and OVOL2. Differential expression in cancer of the family member genes PIWIL3, TGIF2LY and OVOL3 could not be evaluated, as these genes are not present on the arrays investigated (Table S5). Thus, these genes need to be further investigated to determine their cancer expression.
New CT Candidate Genes

15 genes exhibit a testis- or testis/brain-restricted expression pattern according to the EST meta-analysis as well as show evidence for aberrant cancer expression based on the results of the different expression analyses (Table S5). These include the three known CT genes SYCP1, TDRD1 and PIWIL2 and mainly members of the gene families mentioned above. Their gene products have potential as cancer markers and/or therapeutic targets. Further five genes have a testis- or testis/brain-restricted expression profile, but no cancer expression could be detected so far.

Discussion

High-throughput expression data is essential for the construction of comprehensive expression profiles and/or for the investigation of differential expression across the numerous tissues of human body and also across the numerous existing cancer types, as empirical determination is costly and time-consuming. For example, several CT genes could successfully be identified to date using high-throughput expression data [12-14,38].

CT Gene Expression and Tissue-specificity

CT genes exhibit expression restricted to the testis, but are aberrantly expressed in various cancer types [2]; however, recently various CT genes were shown to be expressed also in normal brain and placenta [12,13]. As these tissues are also immunologically privileged, we included these in our screens. Moreover, we allowed expression in two healthy tissues, as limited expression could be due to undiagnosed neoplastic change in the tissues analyzed. Many normal tissues are extracted from tissues that were obtained post mortem and are often pooled from a number of individuals, many of whom were aged at time of death. In support of this, Chen et al. found discrepancies concerning the expression of some genes in normal tissues; they detected expression in tissues from one panel of normal tissues, but could not detect expression in similar tissue types from a distinct second source [38]. Thus, genes with limited expression in one or two tissues could indeed be testis- or testis/brain-restricted.

Cancer Expression of Human Germline Genes

Here we show that the Drosophila germline genes ectopically expressed in l(3)mbt tumors are also aberrantly expressed or overexpressed in a wide range of human cancers. 15 of these genes also exhibit a testis- or testis/brain restricted expression pattern, which
makes them to potential CT candidate genes. We have used the results of the EST meta-
alysis, of the single microarray analyses as well as of the microarray meta-analysis
to construct a comprehensive picture of their gene expression. Combining studies can
enhance reliability and generalizability of the results, as meta-analyses are generally
accepted to compute a more precise and reliable estimate of gene expression [39]. The
extent to which the single microarray analyses, however, reflect expression/upregulation
in cancer will rely on further analyses.

Although most known CT genes are encoded on the X chromosome [2], most of our
15 CT candidates are autosomally encoded. In general, almost all human homologues
we have investigated are autosomally encoded (Table S5). We found mainly genes be-
longing to large germline gene families to be ectopically expressed or upregulated in
cancer. Most of these family members produce proteins that are thought to be involved
in meiosis or spermatogenesis such as the NANOS or DAZ family genes [40,41]. At least
eleven genes encode proteins that are associated with meiosis, and a total of 14 gene
products may generally function in spermatogenesis (Table S2). Consistent with this,
we have recently identified a cohort of CT candidate genes, whose gene products may be
involved in meiotic spermatogenesis, by analyzing the expression of human homologues
of meiotic mouse genes [14].

We also found several genes to be downregulated in a range of cancer types such as
the genes CPEB1 and ESRP1. This is not surprising as several genes are not germline-
specific in humans. Here, the loss of function of the associated proteins could drive the
malignant state of cancer cells, as CPEB1 and ESRP1, for example, are potential tumor
suppressor genes [42,43].

Cancer Cells and Soma-to-germline Transformation

Cancer cells might undergo a soma-to-germline transformation, which in turn may sup-
port the acquisition of malignant attributes such as rapid proliferation, undifferenti-
ated phenotype and immortality. Such transformations have not only been reported
in Drosophila animals with mutations in the dREAM-MMB pathway [3,6], but also in
C. elegans strains with mutations in the homologues of the dREAM-MMB complex [8]
and in members of the nucleosome remodeling and histone deacetylase (NuRD) com-
plex [7], which are chromatin regulators of the SynMuv pathway [44,45]. Many SynMuv
proteins and their antagonistic SynMuv suppressor proteins have been associated with
histone modifications, nucleosome remodeling as well as transcriptional repression and
play a role in germline-soma distinction [44,46-48]. These data suggest that proteins
functioning in particular in the retinoblastoma pathway are responsible for repression of germline gene expression in somatic cells and thus alterations in this pathway may initialize a soma-to-germline transformation.

Many of these *Drosophila* genes are conserved in mammals [44] and the human retinoblastoma pathway is disrupted in virtually all cancer types, which is known to promote cell proliferation [49,50]. This evokes the intriguing question whether such soma-to-germline transformations occur also in humans. In human cancer, the expression of numerous CT genes may reflect the occurrence of such a soma-to-germline transformation. A few of the *Drosophila* germline genes investigated here are already known orthologues of human CT genes; for example SYCP1 is the human homologue of the *Drosophila* germline gene c(3)G [15]. Also, in humans, it has been suggested that cells become altered in genes that control germline gene expression, which could lead to an induction of a silenced gametogenic program in cancer [2]. Here we provide evidence that 40 of 46 human homologues of the *Drosophila* germline genes ectopically expressed in l(3)mbt tumors are also ectopically expressed or upregulated in a wide range of human cancers, which supports the hypothesis that human cells undergo a similar soma-to-germline transformation in the course of the development of cancer.

**Meiotic Genes as Driver of a Soma-to-Germline Transformation**

The ectopic activation of a few testis-specific factors, which act as epigenetic and transcriptional regulators, could further drive the soma-to-germline transformation [51]. The expression of germline genes, in particular of meiotic genes, is tightly regulated and mostly restricted to germline cells. The expression of meiotic genes, in particular of those encoding proteins with chromosome modulating potential or with involvement in meiosis-specific processes such as synapsis, in somatic cells could lead to perturbation of the mitotic process. This could result in inappropriate recombination events, provoking oncogenic changes such as translocations, aberrant chromosome segregation and aneuploidy [2,51,52], which in turn are hallmarks of cancer. Kalejs et al., for example, reported the upregulation of meiosis-specific genes in tumor cells, which appears to be associated with arrested mitosis and polyploidy [53]. Moreover, we have previously identified a cohort of meiotic genes with expression restricted to germ cells to be aberrantly expressed in a wide range of cancers [14]. A number of the human germline genes we have investigated here encode proteins that are also associated with meiotic functions. Their ectopic activation in mitotic dividing cells might further contribute to oncogenic events and promote a soma-to-germline transformation.
Acknowledgments

JF was supported by the Welsh National Institute for Social Care and Health Research (NISCHR) [grant number HS/09/008]. The work was partly funded by North West Cancer Research Fund grant CR888 awarded to RJM. The authors would like to thank Dr. Gerhard Thallinger for his support. The authors declare no conflict of interest.
References


For submission to the journal Translational Oncology


For submission to the journal *Translational Oncology*


For submission to the journal Translational Oncology


Figure S1: An example of a bar chart showing the integrated expression profile of the *C16orf73* gene. *C16orf73* exhibits expression restricted to the brain, placenta and testis, but is aberrantly expressed in melanoma and sarcomas of the bone and of the connective tissue. The expression is given in transcripts per million (tpm).
Figure S2: Forest plots for the gene RN17. RN17 is upregulated in (A) ovarian, (B) prostate and (C) brain cancer, according to the microarray meta-analysis. A Forest plot shows the log 2-fold change (lg2FC) values for the individual studies as well as the total values for the given cancer type and for all cancer types combined. Each study is illustrated by a square; the position on the x-axis representing the measure estimate (lg2FC ratio), the size proportional to the weight of the study, and the horizontal line through it reflecting the confidence interval of the estimate.
Figure S3: Circos plot for the 41 human homologues based on the single microarray analysis. The Circos plot shows that 39 of the total 41 human homologues covered by the arrays are upregulated in a wide range of cancer types. Each connection between a gene and an individual cancer type indicates a statistically significant upregulation for that cancer type derived from a single array study for cancer vs. normal tissue.
Table S1: List of *Drosophila* germline genes overexpressed in *l(3)mbt* tumors and their human homologues. 49 *Drosophila* germline genes were reported by Janic et al. to be overexpressed in *l(3)mbt* tumors [1]. The human homologues were determined using the databases Flybase [2], Homologene [3] and Ensembl [4], and/or literature search.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Flybase ID</th>
<th>Human orthologues</th>
<th>Human Ensembl IDs</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGO3</td>
<td>FBgn0250816</td>
<td><em>PIWIL4, PIWIL2, PIWIL3, PIWIL1</em></td>
<td>ENSG00000134627, ENSG00000197181, ENSG00000184571, ENSG00000125207</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>aub</td>
<td>FBgn0000146</td>
<td><em>PIWIL4, PIWIL2, PIWIL3, PIWIL1</em></td>
<td>ENSG00000134627, ENSG00000197181, ENSG00000184571, ENSG00000125207</td>
<td>Flybase</td>
</tr>
<tr>
<td>bam</td>
<td>FBgn0000158</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bgcn</td>
<td>FBgn0004581</td>
<td><em>YTHDC2</em></td>
<td>ENSG00000047188</td>
<td>Ensembl Biomart</td>
</tr>
<tr>
<td>BicC</td>
<td>FBgn0000182</td>
<td><em>BICC1</em></td>
<td>ENSG00000122870</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>bol</td>
<td>FBgn0011206</td>
<td><em>DAZ2, DAZ3, DAZ4, BOLL, DAZL, DAZ1</em></td>
<td>ENSG00000205944, ENSG00000187191, ENSG00000205916, ENSG00000152430, ENSG00000092345, ENSG00000188120</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>c(3)G</td>
<td>FBgn0000246</td>
<td><em>SYCP1</em></td>
<td>ENSG00000198765</td>
<td>[5]</td>
</tr>
<tr>
<td>CG15930</td>
<td>FBgn0029754</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG31755</td>
<td>FBgn0051755</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG32313</td>
<td>FBgn0052313</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG40115</td>
<td>FBgn0058115</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genes</td>
<td>Flybase IDs</td>
<td>Flybase/Ensembl Biomart</td>
<td>Comments</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>-------------------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>CG7194</td>
<td>FBgn0035868</td>
<td>ENSG00000136810,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG7795</td>
<td>FBgn00262598</td>
<td>ENSG00000091164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG9925</td>
<td>FBgn0038191</td>
<td>Flybase, Ensembl Biomart</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cona</td>
<td>FBgn0038612</td>
<td>ENSG000000104413,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>del</td>
<td>FBgn0086251</td>
<td>ENSG000000103067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dhd</td>
<td>FBgn0011761</td>
<td>TXN, TXNL1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fs(1)Yb</td>
<td>FBgn0000928</td>
<td>ENSG00000162039</td>
<td>[6] (potential functional homologues)</td>
<td></td>
</tr>
<tr>
<td>fus</td>
<td>FBgn0023441</td>
<td>ENSG00000095627,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gnu</td>
<td>FBgn0001120</td>
<td>ENSG00000132972</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hdm</td>
<td>FBgn0029977</td>
<td>C16orf73</td>
<td>Flybase, Ensembl Biomart</td>
<td></td>
</tr>
<tr>
<td>krimp</td>
<td>FBgn0034098</td>
<td>TDRD1, RNF17</td>
<td>ENSG0000016227</td>
<td></td>
</tr>
<tr>
<td>loki</td>
<td>FBgn0019686</td>
<td>CHEK2</td>
<td>Flybase, Ensembl Biomart</td>
<td></td>
</tr>
<tr>
<td>mael</td>
<td>FBgn0016034</td>
<td>MAEL</td>
<td>[7,8]</td>
<td></td>
</tr>
<tr>
<td>mia</td>
<td>FBgn0014342</td>
<td>TAF6, TAF6L</td>
<td>Flybase, Ensembl Biomart</td>
<td></td>
</tr>
<tr>
<td>mre11</td>
<td>FBgn0020270</td>
<td>MRE11A</td>
<td>Flybase, Ensembl Biomart</td>
<td></td>
</tr>
<tr>
<td>Mst57Db</td>
<td>FBgn0011669</td>
<td>ENSG00000020922</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mst77F</td>
<td>FBgn0086915</td>
<td>ENSG00000187556,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nos</td>
<td>FBgn0002962</td>
<td>NANOS3, NANOS2,</td>
<td>Flybase, Ensembl Biomart</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NANOS1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>FBgn</td>
<td>Description</td>
<td>Ensembl Accessory Numbers</td>
<td>Source</td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>------------------------------</td>
<td>---------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>orb</td>
<td>FBgn0004882</td>
<td>CPEB1, RP11-152F18.10</td>
<td>ENSG00000214575, ENSG00000260836</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>ovo</td>
<td>FBgn0003028</td>
<td>OVOL1, OVOL2, OVOL3</td>
<td>ENSG00000172818, ENSG00000125850, ENSG00000105261</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>piwi</td>
<td>FBgn0004872</td>
<td>PIWIL4, PIWIL2, PIWIL3, PIWIL1</td>
<td>ENSG00000134627, ENSG00000197181, ENSG00000184571, ENSG00000125207</td>
<td>Flybase</td>
</tr>
<tr>
<td>png</td>
<td>FBgn0000826</td>
<td>NEK1, NEK5, NEK3</td>
<td>ENSG00000137601, ENSG00000197168, ENSG00000136098</td>
<td>Ensembl Biomart</td>
</tr>
<tr>
<td>Pxt</td>
<td>FBgn0261987</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RpS19b</td>
<td>FBgn0039129</td>
<td>RPS19</td>
<td>ENSG000000105372</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>RpS5b</td>
<td>FBgn0038277</td>
<td>RPS5</td>
<td>ENSG00000083845</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>shu</td>
<td>FBgn0003401</td>
<td>FKBP6</td>
<td>ENSG00000077800</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>spn-E</td>
<td>FBgn0003483</td>
<td>TDRD9</td>
<td>ENSG00000156414</td>
<td>Flybase, Ensembl Biomart</td>
</tr>
<tr>
<td>squ</td>
<td>FBgn0002652</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stil</td>
<td>FBgn0003527</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>swa</td>
<td>FBgn0003655</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tej</td>
<td>FBgn0033921</td>
<td>TDRD7</td>
<td>ENSG000000196116</td>
<td>Ensembl Biomart</td>
</tr>
<tr>
<td>topi</td>
<td>FBgn0037751</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tor</td>
<td>FBgn0003733</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TrxT</td>
<td>FBgn0029752</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Flybase</td>
<td>Ensembl</td>
<td>Flybase</td>
<td>Flybase</td>
</tr>
<tr>
<td>------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>vasa</td>
<td>FBgn0262526</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vis</td>
<td>FBgn0033748</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDX4</td>
<td>TXN, TXNL1, TGIF2LX, TGIF2LY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000152670</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000136810</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000091164</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000153779</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000176679</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ Tub37C</td>
<td>FBgn0010097</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TUBG1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000118707</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000176679</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flybase Homologue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For submission to the journal *Translational Oncology*.
Table S2: Enriched gene ontology (GO) terms of the human homologues of the *Drosophila* genes ectopically expressed in *l(3)mbt* tumors. The enriched GO terms were determined by the means of the functional annotation tool DAVID [10]. The top enriched GO terms show that these genes encode proteins that are mainly involved in meiosis, spermatogenesis and reproduction. Please note that only the top 20 enriched GO term are shown in the following table.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>GO term</th>
<th>Gene count</th>
<th>Percentage</th>
<th>p value</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006417</td>
<td>Regulation of translation</td>
<td>12</td>
<td>28.6</td>
<td>4.61E-14</td>
<td>DAZ3, DAZ4, NANOS3, DAZ1, DAZ2, NANOS2, NANOS1, CPEB1, RPS5, BOLL, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4</td>
</tr>
<tr>
<td>GO:0051321</td>
<td>Meiotic cell cycle</td>
<td>11</td>
<td>26.2</td>
<td>8.06E-14</td>
<td>MRE11A, OVOL1, MAEL, PIWIL1, PIWIL2, PIWIL3, PIWIL4, TUBG1, SYCP1, BOLL, TDRD1</td>
</tr>
<tr>
<td>GO:0048232</td>
<td>Male gamete generation</td>
<td>14</td>
<td>33.3</td>
<td>6.55E-13</td>
<td>DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>GO:0007283</td>
<td>Spermatogenesis</td>
<td>14</td>
<td>33.3</td>
<td>6.55E-13</td>
<td>DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>GO:0007276</td>
<td>Gamete generation</td>
<td>15</td>
<td>35.7</td>
<td>7.54E-13</td>
<td>DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>GO:0007126</td>
<td>Meiosis</td>
<td>10</td>
<td>23.8</td>
<td>3.56E-12</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------</td>
<td>------</td>
<td>------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>GO:0051327</td>
<td>M phase of meiotic cell cycle</td>
<td>10</td>
<td>23.8</td>
<td>3.56E-12</td>
<td></td>
</tr>
<tr>
<td>GO:0010608</td>
<td>Post-transcriptional regulation of gene expression</td>
<td>12</td>
<td>28.6</td>
<td>5.42E-12</td>
<td></td>
</tr>
<tr>
<td>GO:0019953</td>
<td>Sexual reproduction</td>
<td>15</td>
<td>35.7</td>
<td>5.58E-12</td>
<td></td>
</tr>
<tr>
<td>GO:0048609</td>
<td>Reproductive process in a multicellular organism</td>
<td>15</td>
<td>35.7</td>
<td>1.27E-11</td>
<td></td>
</tr>
<tr>
<td>GO:0032504</td>
<td>Multicellular organism reproduction</td>
<td>15</td>
<td>35.7</td>
<td>1.27E-11</td>
<td></td>
</tr>
</tbody>
</table>

*For submission to the journal Translational Oncology*
<table>
<thead>
<tr>
<th>GO:0000279</th>
<th>M phase</th>
<th>13</th>
<th>31.0</th>
<th>3.31E-11</th>
<th>NEK3, NEK1, MRE11A, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, TUBG1, PIWIL4, TDRD1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0034587</td>
<td>piRNA metabolic process</td>
<td>5</td>
<td>11.9</td>
<td>2.23E-10</td>
<td>MAEL, PIWIL1, PIWIL2, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>GO:0022403</td>
<td>Cell cycle phase</td>
<td>13</td>
<td>31.0</td>
<td>4.69E-10</td>
<td>NEK3, NEK1, MRE11A, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, TUBG1, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>GO:0022414</td>
<td>Reproductive process</td>
<td>15</td>
<td>35.7</td>
<td>4.56E-09</td>
<td>DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>GO:0000003</td>
<td>Reproduction</td>
<td>15</td>
<td>35.7</td>
<td>4.96E-09</td>
<td>DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>GO:0022402</td>
<td>Cell cycle process</td>
<td>13</td>
<td>31.0</td>
<td>1.57E-08</td>
<td>NEK3, NEK1, MRE11A, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, TUBG1, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>GO:0031047</td>
<td>Gene silencing by RNA</td>
<td>6</td>
<td>14.3</td>
<td>2.42E-08</td>
<td>MAEL, PIWIL1, PIWIL2, PIWIL3, PIWIL4, TDRD1</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------------------------</td>
<td>-----</td>
<td>--------</td>
<td>----------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>GO:0032268</td>
<td>Regulation of cellular protein</td>
<td>12</td>
<td>28.6</td>
<td>2.89E-08</td>
<td>DAZ3, DAZ4, NANOS3, DAZ1, DAZ2, NANOS2, NANOS1, CPEB1, RPS5, BOLL, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4</td>
</tr>
<tr>
<td></td>
<td>metabolic process</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0007049</td>
<td>Cell cycle</td>
<td>14</td>
<td>33.3</td>
<td>5.70E-08</td>
<td>NEK3, NEK1, MRE11A, MAEL, CHEK2, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, TUBG1, PIWIL4, TDRD1</td>
</tr>
</tbody>
</table>
Table S3: Expression profiles based on expressed sequence tag (EST) data for the human homologues of the *Drosophila* genes ectopically expressed in *l(3)mbt* tumors across 36 tissues. 43 of the 46 human genes could be mapped to Unigene cluster IDs. Genes with expression restricted to immunologically privileged areas (testis, brain and placenta) and with expression in a maximum of two other normal tissues are considered as testis- or testis/brain-restricted (highlighted in gray). Genes exhibiting only expression in a given healthy tissue but no expression in the corresponding cancerous tissue are selected as ectopically expressed in that cancer type. Significance of upregulation in cancer was determined using Fisher’s exact test and genes with a p value < 0.05 were selected as significantly upregulated in that cancer type.

<table>
<thead>
<tr>
<th>UniGeneID</th>
<th>Gene</th>
<th>Expression in immunologically privileged areas [tpm]</th>
<th>Upregulation/ectopic expression in cancer</th>
<th>Number of other healthy tissues were expression was found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.729604</td>
<td>C16orf73</td>
<td>Brain (3), placenta (4), testis (100)</td>
<td>Bone, skin, connective tissue</td>
<td></td>
</tr>
<tr>
<td>Hs.661266</td>
<td>FKBP6</td>
<td>Brain (9), testis (174)</td>
<td>Blood, ovary, lung</td>
<td></td>
</tr>
<tr>
<td>Hs.592257</td>
<td>DAZ2</td>
<td>Testis (13)</td>
<td>Lung, stomach</td>
<td></td>
</tr>
<tr>
<td>Hs.112743</td>
<td>SYCP1</td>
<td>Brain (1), testis (182)</td>
<td>Connective tissue</td>
<td></td>
</tr>
<tr>
<td>Hs.169797</td>
<td>BOLL</td>
<td>Brain (6), testis (330)</td>
<td>Connective tissue</td>
<td></td>
</tr>
<tr>
<td>Hs.223581</td>
<td>DDX4</td>
<td>Brain (3), testis (578)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.97464</td>
<td>RNF17</td>
<td>Placenta (12), testis (165)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.112148</td>
<td>TGIF2LY</td>
<td>Testis (35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.448343</td>
<td>PIWIL3</td>
<td>Testis (13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.592220</td>
<td>TGIF2LX</td>
<td>Brain (1), testis (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.434218</td>
<td>NANOS2</td>
<td>Brain (1), testis (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.661013</td>
<td>OVOL2</td>
<td>-</td>
<td>Placenta, prostate, uterus, ovary, intestine, thyroid, stomach</td>
<td>1 (Lung)</td>
</tr>
<tr>
<td>Hs.333132</td>
<td>TDRD1</td>
<td>Brain (2), testis (113)</td>
<td>Prostate, liver, pancreas</td>
<td>1 (Trachea)</td>
</tr>
<tr>
<td>Hs.127982</td>
<td>NANOS3</td>
<td>Brain (1)</td>
<td>Mouth, connective tissue, lung</td>
<td>1 (Eye)</td>
</tr>
<tr>
<td>Hs.131179</td>
<td>DAZL</td>
<td>Brain (6), testis (426)</td>
<td>Lung, connective tissue</td>
<td>1 (Muscle)</td>
</tr>
<tr>
<td>Hs.143920</td>
<td>OVOL3</td>
<td>Testis (9)</td>
<td>Connective tissue</td>
<td>1 (Ovary)</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Tissues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.405659</td>
<td>PIWIL1</td>
<td>Brain (2), testis (512), Intestine, uterus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.614809</td>
<td>PIWIL2</td>
<td>Brain (3), placenta (4), testis (100), Testis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.672144</td>
<td>NEK5</td>
<td>Placenta (4), testis (4), 2 (Muscle, eye)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.70936</td>
<td>DAZ1</td>
<td>Testis (39), Stomach</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.591918</td>
<td>NANOS1</td>
<td>Brain (1), placenta (29), Brain, intestine, lung, muscle, ovary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.651245</td>
<td>MAEL</td>
<td>Brain (31), placenta (8), testis (1107), Liver, mammary gland</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.158745</td>
<td>BICC1</td>
<td>Brain (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.660188</td>
<td>PIWIL4</td>
<td>Brain (2), placenta (8), testis (178), Pancreas, connective tissue,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.21454</td>
<td>TDRD9</td>
<td>Brain (9), testis (122), Connective tissue, bone marrow, ovary, thyroid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.134434</td>
<td>OVOL1</td>
<td>Placenta (17), testis (96), Mammary gland, uterus, prostate, pancreas,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.547988</td>
<td>CPEB1</td>
<td>Brain (50), testis (35), Skin, liver, adrenal gland, bone marrow,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hs.714400</td>
<td>TAF6L</td>
<td>Brain (16), placenta (21), testis (26), Pancreas, thyroid, stomach,</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For submission to the journal Translational Oncology

29
<table>
<thead>
<tr>
<th>Accession</th>
<th>Gene</th>
<th>Tissues</th>
<th>Tissues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.632264</td>
<td>TGIF2</td>
<td>Brain (4), testis (9)</td>
<td>Muscle, mammary gland, mouth, thyroid, lung, ovary, eye, placenta, intestine, testis, skin, liver, brain</td>
</tr>
<tr>
<td>Hs.409989</td>
<td>NEK3</td>
<td>Brain (8), placenta (12), testis (39)</td>
<td>Stomach, mouth, liver, bone, ovary, connective tissue</td>
</tr>
<tr>
<td>Hs.291363</td>
<td>CHEK2</td>
<td>Brain (2), placenta (8), testis (22)</td>
<td>Bone marrow, thyroid, connective tissue, skin, mammary gland, uterus</td>
</tr>
<tr>
<td>Hs.487471</td>
<td>ESRP1</td>
<td>Brain (1), placenta (25), testis (9)</td>
<td>Eye, stomach, mammary gland, mouth, uterus, connective tissue, ovary, thyroid, liver, pharynx</td>
</tr>
<tr>
<td>Hs.193842</td>
<td>TDRD7</td>
<td>Brain (19), placenta (8), testis (87)</td>
<td>Stomach, lymph node, mammary gland, ovary, pharynx, connective tissue, intestine</td>
</tr>
<tr>
<td>Hs.481181</td>
<td>NEK1</td>
<td>Brain (17), testis (9)</td>
<td>Eye, ovary, adrenal gland, intestine, bone, liver</td>
</tr>
<tr>
<td>Hs.592053</td>
<td>ESRP2</td>
<td>Placenta (70), testis (17)</td>
<td>Mouth, connective tissue, pancreas, ovary, adrenal gland</td>
</tr>
<tr>
<td>Hs.231942</td>
<td>YTHDC2</td>
<td>Brain (17), placenta (17), testis (48)</td>
<td>Ovary, mouth, eye, bone marrow, thyroid, adrenal gland, lung, connective tissue</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Symbol</td>
<td>Tissues</td>
<td>Expression Levels</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>---------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Hs.192649</td>
<td>MRE11A</td>
<td>Brain (9), placenta (41), testis (4)</td>
<td>Mouth, stomach, thyroid, testis, bone, prostate, mammary gland</td>
</tr>
<tr>
<td>Hs.489309</td>
<td>TAF6</td>
<td>Brain (40), placenta (54), testis (30)</td>
<td>Eye, stomach, testis, liver, pancreas, thyroid, brain, connective tissue</td>
</tr>
<tr>
<td>Hs.435136</td>
<td>TXN</td>
<td>Brain (21), placenta (83), testis (35)</td>
<td>Brain, adrenal gland, mouth, ovary, stomach, connective tissue, kidney</td>
</tr>
<tr>
<td>Hs.279669</td>
<td>TUBG1</td>
<td>Brain (45), placenta (66), testis (287)</td>
<td>Brain, mouth, thyroid, skin, pharynx, liver, eye, lung, connective tissue, ovary</td>
</tr>
<tr>
<td>Hs.114412</td>
<td>TXNL1</td>
<td>Brain (40), placenta (25), testis (30)</td>
<td>Thyroid, stomach, uterus, liver, connective tissue, brain, mammary gland</td>
</tr>
<tr>
<td>Hs.438429</td>
<td>RPS19</td>
<td>Brain (48), placenta (75), testis (43)</td>
<td>Placenta, prostate, bone, kidney, skin, connective tissue, testis, thyroid, lymph node, pharynx, mammary gland, mouth, lung, brain, pancreas</td>
</tr>
<tr>
<td>Hs.378103</td>
<td>RPS5</td>
<td>Brain (111), placenta (199), testis (69)</td>
<td>Bone, pancreas, brain, uterus, muscle, kidney, liver, eye, placenta, prostate, skin, mammary gland, testis, lung, intestine, lymph node</td>
</tr>
</tbody>
</table>
Table S4: List the human homologues of the *Drosophila* genes ectopically expressed in *l(3)mbl* tumors and their cancer-specific upregulation according to the microarray meta-analysis. 31 of the 41 human homologues covered by the array sets were found to be significantly upregulated in eleven distinct cancer types. For each upregulated gene the weighted meta-log 2-fold change (log2FC) and the confidence intervals (CI left, CI right) are stated.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Ensembl ID</th>
<th>Cancer type</th>
<th>Weighted log2FC</th>
<th>Weighted CI left</th>
<th>Weighted CI right</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEK2</td>
<td>ENSG00000183765</td>
<td>Colorectal cancer</td>
<td>1.27</td>
<td>0.93</td>
<td>1.6</td>
</tr>
<tr>
<td>CHEK2</td>
<td>ENSG00000183765</td>
<td>Head and neck cancer</td>
<td>0.48</td>
<td>0.1</td>
<td>0.86</td>
</tr>
<tr>
<td>CHEK2</td>
<td>ENSG00000183765</td>
<td>Renal cancer</td>
<td>1.1</td>
<td>0.29</td>
<td>1.9</td>
</tr>
<tr>
<td>CHEK2</td>
<td>ENSG00000183765</td>
<td>Lung cancer</td>
<td>0.68</td>
<td>0.39</td>
<td>0.97</td>
</tr>
<tr>
<td>DAZ1</td>
<td>ENSG00000188120</td>
<td>Ovarian cancer</td>
<td>1.87</td>
<td>0.69</td>
<td>3.05</td>
</tr>
<tr>
<td>DAZ1</td>
<td>ENSG00000188120</td>
<td>Brain cancer</td>
<td>2.42</td>
<td>0.94</td>
<td>3.9</td>
</tr>
<tr>
<td>DAZ1</td>
<td>ENSG00000188120</td>
<td>Prostate cancer</td>
<td>3.3</td>
<td>2</td>
<td>4.59</td>
</tr>
<tr>
<td>DAZ3</td>
<td>ENSG00000187191</td>
<td>Ovarian cancer</td>
<td>1.87</td>
<td>0.69</td>
<td>3.05</td>
</tr>
<tr>
<td>DAZ3</td>
<td>ENSG00000187191</td>
<td>Brain cancer</td>
<td>2.42</td>
<td>0.94</td>
<td>3.9</td>
</tr>
<tr>
<td>DAZ3</td>
<td>ENSG00000187191</td>
<td>Prostate cancer</td>
<td>3.3</td>
<td>2</td>
<td>4.59</td>
</tr>
<tr>
<td>DAZ4</td>
<td>ENSG00000205916</td>
<td>Ovarian cancer</td>
<td>1.87</td>
<td>0.69</td>
<td>3.05</td>
</tr>
<tr>
<td>DAZ4</td>
<td>ENSG00000205916</td>
<td>Brain cancer</td>
<td>2.42</td>
<td>0.94</td>
<td>3.9</td>
</tr>
<tr>
<td>DAZ4</td>
<td>ENSG00000205916</td>
<td>Prostate cancer</td>
<td>3.3</td>
<td>2</td>
<td>4.59</td>
</tr>
<tr>
<td>DAZL</td>
<td>ENSG00000092345</td>
<td>Brain cancer</td>
<td>1.78</td>
<td>-0.27</td>
<td>3.83</td>
</tr>
<tr>
<td>ESRP1</td>
<td>ENSG00000104413</td>
<td>Breast cancer</td>
<td>1.24</td>
<td>0.62</td>
<td>1.86</td>
</tr>
<tr>
<td>ESRP1</td>
<td>ENSG00000104413</td>
<td>Ovarian cancer</td>
<td>4.06</td>
<td>3.23</td>
<td>4.89</td>
</tr>
<tr>
<td>ESRP1</td>
<td>ENSG00000104413</td>
<td>Lung cancer</td>
<td>1.13</td>
<td>0.74</td>
<td>1.52</td>
</tr>
<tr>
<td>Gene</td>
<td>ENSG</td>
<td>Cancer Type</td>
<td>z-score</td>
<td>Fold Change</td>
<td>p-value</td>
</tr>
<tr>
<td>---------</td>
<td>---------------</td>
<td>------------------------------------</td>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>ESRP1</td>
<td>ENSG00000104413</td>
<td>Prostate cancer</td>
<td>0.84</td>
<td>0.34</td>
<td>1.34</td>
</tr>
<tr>
<td>ESRP2</td>
<td>ENSG00000103067</td>
<td>Ovarian cancer</td>
<td>1.85</td>
<td>1.03</td>
<td>2.67</td>
</tr>
<tr>
<td>ESRP2</td>
<td>ENSG00000103067</td>
<td>Prostate cancer</td>
<td>0.95</td>
<td>0.55</td>
<td>1.35</td>
</tr>
<tr>
<td>MAEL</td>
<td>ENSG00000143194</td>
<td>Lung cancer</td>
<td>0.74</td>
<td>0.15</td>
<td>1.34</td>
</tr>
<tr>
<td>MRE11A</td>
<td>ENSG0000020922</td>
<td>Colorectal cancer</td>
<td>1.82</td>
<td>1.45</td>
<td>2.18</td>
</tr>
<tr>
<td>MRE11A</td>
<td>ENSG0000020922</td>
<td>Head and neck cancer</td>
<td>0.59</td>
<td>0.09</td>
<td>1.1</td>
</tr>
<tr>
<td>MRE11A</td>
<td>ENSG0000020922</td>
<td>Lung cancer</td>
<td>0.35</td>
<td>0.1</td>
<td>0.61</td>
</tr>
<tr>
<td>MRE11A</td>
<td>ENSG0000020922</td>
<td>Brain cancer</td>
<td>1.94</td>
<td>0.74</td>
<td>3.15</td>
</tr>
<tr>
<td>NANOS1</td>
<td>ENSG00000188613</td>
<td>Head and neck cancer</td>
<td>1.05</td>
<td>0.27</td>
<td>1.83</td>
</tr>
<tr>
<td>NANOS1</td>
<td>ENSG00000188613</td>
<td>Lung cancer</td>
<td>0.86</td>
<td>0.45</td>
<td>1.26</td>
</tr>
<tr>
<td>NANOS3</td>
<td>ENSG00000187556</td>
<td>Ovarian cancer</td>
<td>1.13</td>
<td>0.27</td>
<td>1.99</td>
</tr>
<tr>
<td>NEK1</td>
<td>ENSG00000137601</td>
<td>Ovarian cancer</td>
<td>1.71</td>
<td>0.67</td>
<td>2.75</td>
</tr>
<tr>
<td>NEK1</td>
<td>ENSG00000137601</td>
<td>Brain cancer</td>
<td>1.97</td>
<td>0.25</td>
<td>3.69</td>
</tr>
<tr>
<td>NEK1</td>
<td>ENSG00000137601</td>
<td>Leukemia</td>
<td>0.78</td>
<td>0.15</td>
<td>1.41</td>
</tr>
<tr>
<td>NEK3</td>
<td>ENSG00000136098</td>
<td>Colorectal cancer</td>
<td>1.12</td>
<td>0.79</td>
<td>1.44</td>
</tr>
<tr>
<td>OVOL1</td>
<td>ENSG00000172818</td>
<td>Breast cancer</td>
<td>1.01</td>
<td>0</td>
<td>2.02</td>
</tr>
<tr>
<td>OVOL1</td>
<td>ENSG00000172818</td>
<td>Ovarian cancer</td>
<td>2.16</td>
<td>1.32</td>
<td>3.01</td>
</tr>
<tr>
<td>OVOL1</td>
<td>ENSG00000172818</td>
<td>Lung cancer</td>
<td>0.84</td>
<td>0.45</td>
<td>1.22</td>
</tr>
<tr>
<td>OVOL2</td>
<td>ENSG00000125850</td>
<td>Ovarian cancer</td>
<td>3.06</td>
<td>2.26</td>
<td>3.85</td>
</tr>
<tr>
<td>OVOL2</td>
<td>ENSG00000125850</td>
<td>Lung cancer</td>
<td>0.46</td>
<td>0.15</td>
<td>0.76</td>
</tr>
<tr>
<td>PIWIL1</td>
<td>ENSG00000125207</td>
<td>Ovarian cancer</td>
<td>1.44</td>
<td>0.31</td>
<td>2.56</td>
</tr>
<tr>
<td>PIWIL2</td>
<td>ENSG00000197181</td>
<td>Ovarian cancer</td>
<td>2.3</td>
<td>1.45</td>
<td>3.16</td>
</tr>
<tr>
<td>PIWIL4</td>
<td>ENSG00000134627</td>
<td>Colorectal cancer</td>
<td>0.54</td>
<td>0.13</td>
<td>0.95</td>
</tr>
<tr>
<td>RNFI7</td>
<td>ENSG00000132972</td>
<td>Ovarian cancer</td>
<td>2.15</td>
<td>1.01</td>
<td>3.29</td>
</tr>
<tr>
<td>RNFI7</td>
<td>ENSG00000132972</td>
<td>Brain cancer</td>
<td>1.12</td>
<td>-0.93</td>
<td>3.17</td>
</tr>
<tr>
<td>Gene</td>
<td>ENSG</td>
<td>Cancer Type</td>
<td>Log2 Fold Change</td>
<td>Max FDR</td>
<td>Max Log Odds</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>----------------------------</td>
<td>------------------</td>
<td>-----------</td>
<td>--------------</td>
</tr>
<tr>
<td>RNF17</td>
<td>ENSG00000132972</td>
<td>Prostate cancer</td>
<td>2.51</td>
<td>0.95</td>
<td>4.07</td>
</tr>
<tr>
<td>RPS19</td>
<td>ENSG00000105372</td>
<td>Colorectal cancer</td>
<td>0.37</td>
<td>0.05</td>
<td>0.68</td>
</tr>
<tr>
<td>RPS19</td>
<td>ENSG00000105372</td>
<td>Renal cancer</td>
<td>1.05</td>
<td>0.52</td>
<td>1.58</td>
</tr>
<tr>
<td>RPS19</td>
<td>ENSG00000105372</td>
<td>Ovarian cancer</td>
<td>0.78</td>
<td>0.16</td>
<td>1.41</td>
</tr>
<tr>
<td>RPS19</td>
<td>ENSG00000105372</td>
<td>Lung cancer</td>
<td>0.66</td>
<td>0.34</td>
<td>0.98</td>
</tr>
<tr>
<td>RPS3</td>
<td>ENSG00000083845</td>
<td>Colorectal cancer</td>
<td>0.39</td>
<td>0.05</td>
<td>0.72</td>
</tr>
<tr>
<td>RPS3</td>
<td>ENSG00000083845</td>
<td>Renal cancer</td>
<td>0.79</td>
<td>0.21</td>
<td>1.37</td>
</tr>
<tr>
<td>RPS3</td>
<td>ENSG00000083845</td>
<td>Ovarian cancer</td>
<td>0.25</td>
<td>0.08</td>
<td>0.42</td>
</tr>
<tr>
<td>SYCP1</td>
<td>ENSG00000198765</td>
<td>Brain cancer</td>
<td>1.64</td>
<td>0.55</td>
<td>2.72</td>
</tr>
<tr>
<td>SYCP1</td>
<td>ENSG00000198765</td>
<td>Ovarian cancer</td>
<td>1.77</td>
<td>0.18</td>
<td>3.36</td>
</tr>
<tr>
<td>TAF6</td>
<td>ENSG00000106290</td>
<td>Lung cancer</td>
<td>0.51</td>
<td>0.29</td>
<td>0.73</td>
</tr>
<tr>
<td>TAF6</td>
<td>ENSG00000106290</td>
<td>Brain cancer</td>
<td>0.86</td>
<td>0.01</td>
<td>1.71</td>
</tr>
<tr>
<td>TAF6</td>
<td>ENSG00000106290</td>
<td>Leukemia</td>
<td>0.39</td>
<td>0.06</td>
<td>0.72</td>
</tr>
<tr>
<td>TAF6L</td>
<td>ENSG00000162227</td>
<td>Ovarian cancer</td>
<td>2.2</td>
<td>1.12</td>
<td>3.27</td>
</tr>
<tr>
<td>TAF6L</td>
<td>ENSG00000162227</td>
<td>Lung cancer</td>
<td>0.51</td>
<td>0.02</td>
<td>0.99</td>
</tr>
<tr>
<td>TGIF2</td>
<td>ENSG00000118707</td>
<td>Colorectal cancer</td>
<td>1.67</td>
<td>1.27</td>
<td>2.06</td>
</tr>
<tr>
<td>TGIF2</td>
<td>ENSG00000118707</td>
<td>Head and neck cancer</td>
<td>1.12</td>
<td>0.6</td>
<td>1.64</td>
</tr>
<tr>
<td>TGIF2</td>
<td>ENSG00000118707</td>
<td>Ovarian cancer</td>
<td>1.49</td>
<td>0.84</td>
<td>2.13</td>
</tr>
<tr>
<td>TGIF2</td>
<td>ENSG00000118707</td>
<td>Lung cancer</td>
<td>0.31</td>
<td>0.08</td>
<td>0.53</td>
</tr>
<tr>
<td>TGIF2</td>
<td>ENSG00000118707</td>
<td>Brain cancer</td>
<td>2.09</td>
<td>1.08</td>
<td>3.1</td>
</tr>
<tr>
<td>TGIF2LY</td>
<td>ENSG00000176679</td>
<td>Brain cancer</td>
<td>1.04</td>
<td>-0.52</td>
<td>2.59</td>
</tr>
<tr>
<td>TUBG1</td>
<td>ENSG00000131462</td>
<td>Colorectal cancer</td>
<td>0.86</td>
<td>0.52</td>
<td>1.2</td>
</tr>
<tr>
<td>TUBG1</td>
<td>ENSG00000131462</td>
<td>Breast cancer</td>
<td>0.61</td>
<td>0.15</td>
<td>1.06</td>
</tr>
<tr>
<td>TUBG1</td>
<td>ENSG00000131462</td>
<td>Head and neck cancer</td>
<td>0.6</td>
<td>0.23</td>
<td>0.97</td>
</tr>
<tr>
<td>TUBG1</td>
<td>ENSG00000131462</td>
<td>Lung cancer</td>
<td>0.88</td>
<td>0.64</td>
<td>1.11</td>
</tr>
<tr>
<td>Gene</td>
<td>ENSG Identifier</td>
<td>Tumor Type</td>
<td>Ratio 1</td>
<td>Ratio 2</td>
<td>Ratio 3</td>
</tr>
<tr>
<td>------</td>
<td>----------------</td>
<td>-----------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>TUBG1</td>
<td>ENSG00000131462</td>
<td>Leukemia</td>
<td>0.63</td>
<td>0.21</td>
<td>1.05</td>
</tr>
<tr>
<td>TXN</td>
<td>ENSG00000136810</td>
<td>Breast cancer</td>
<td>0.61</td>
<td>0.14</td>
<td>1.09</td>
</tr>
<tr>
<td>TXN</td>
<td>ENSG00000136810</td>
<td>Pancreatic cancer</td>
<td>0.7</td>
<td>0.17</td>
<td>1.24</td>
</tr>
<tr>
<td>TXN</td>
<td>ENSG00000136810</td>
<td>Leukemia</td>
<td>0.64</td>
<td>0.24</td>
<td>1.04</td>
</tr>
<tr>
<td>TXN</td>
<td>ENSG00000136810</td>
<td>Prostate cancer</td>
<td>0.49</td>
<td>0.19</td>
<td>0.79</td>
</tr>
<tr>
<td>TXNL1</td>
<td>ENSG00000091164</td>
<td>Adrenal cancer</td>
<td>0.39</td>
<td>0.03</td>
<td>0.75</td>
</tr>
<tr>
<td>TXNL1</td>
<td>ENSG00000091164</td>
<td>Lung cancer</td>
<td>0.2</td>
<td>0.04</td>
<td>0.35</td>
</tr>
<tr>
<td>TXNL1</td>
<td>ENSG00000091164</td>
<td>Brain cancer</td>
<td>1.11</td>
<td>-0.03</td>
<td>2.25</td>
</tr>
<tr>
<td>YTHDC2</td>
<td>ENSG00000047188</td>
<td>Leukemia</td>
<td>0.62</td>
<td>0.1</td>
<td>1.14</td>
</tr>
</tbody>
</table>
Table S5: Summary of the expression profiles for all 46 human homologues of *Drosophila* genes ectopically expressed in *l(3)mbt* tumors. The results of the EST screen, of the single microarray analyses as well as of the microarray meta-analysis were combined to construct a comprehensive expression profile. The 19 genes highlighted in gray exhibit a testis- or testis/brain-restricted expression profile. In total, 40 genes are upregulated or show ectopic expression in a wide range of cancers.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chrom.</th>
<th>Ensembl ID</th>
<th>UniGene ID</th>
<th>Array</th>
<th>Expression in immunolog. privileged tissues</th>
<th>Expression in normal tissues</th>
<th>Upregulation/ectopic expression in cancer (method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOLL</td>
<td>2</td>
<td>ENSG00000152430</td>
<td>Hs.169797</td>
<td>On array</td>
<td>Testis, brain</td>
<td>0 tissues</td>
<td>E, S</td>
</tr>
<tr>
<td>C16orf73</td>
<td>16</td>
<td>ENSG00000162039</td>
<td>Hs.729604</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>0 tissues</td>
<td>E, S</td>
</tr>
<tr>
<td>DAZ2</td>
<td>Y</td>
<td>ENSG00000205944</td>
<td>Hs.592257</td>
<td>On array</td>
<td>Testis</td>
<td>0 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>DDX4</td>
<td>5</td>
<td>ENSG00000152670</td>
<td>Hs.223581</td>
<td>On array</td>
<td>Testis, brain</td>
<td>0 tissues</td>
<td>S</td>
</tr>
<tr>
<td>FKBP6</td>
<td>7</td>
<td>ENSG00000077800</td>
<td>Hs.661266</td>
<td>On array</td>
<td>Testis, brain</td>
<td>0 tissues</td>
<td>E, S</td>
</tr>
<tr>
<td>NANOS2</td>
<td>19</td>
<td>ENSG00000188425</td>
<td>Hs.434218</td>
<td>On array</td>
<td>Testis, brain</td>
<td>0 tissues</td>
<td>-</td>
</tr>
<tr>
<td>PIWIL3</td>
<td>22</td>
<td>ENSG00000018457</td>
<td>Hs.448343</td>
<td>NA</td>
<td>Testis</td>
<td>0 tissues</td>
<td>-</td>
</tr>
<tr>
<td>RNF17</td>
<td>13</td>
<td>ENSG00000132972</td>
<td>Hs.97464</td>
<td>On array</td>
<td>Testis, placenta</td>
<td>0 tissues</td>
<td>M, S</td>
</tr>
<tr>
<td>SYCP1*</td>
<td>1</td>
<td>ENSG00000198765</td>
<td>Hs.112743</td>
<td>On array</td>
<td>Testis, brain</td>
<td>0 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>TGF2LX</td>
<td>X</td>
<td>ENSG00000153779</td>
<td>Hs.592220</td>
<td>NA</td>
<td>Testis, brain</td>
<td>0 tissues</td>
<td>-</td>
</tr>
<tr>
<td>TGF2LY</td>
<td>Y</td>
<td>ENSG00000176679</td>
<td>Hs.112148</td>
<td>On array</td>
<td>Testis</td>
<td>0 tissues</td>
<td>M, S</td>
</tr>
<tr>
<td>DAZL</td>
<td>3</td>
<td>ENSG00000092345</td>
<td>Hs.131179</td>
<td>On array</td>
<td>Testis, brain</td>
<td>1 tissue (Muscle)</td>
<td>E, M, S</td>
</tr>
<tr>
<td>NANOS3</td>
<td>19</td>
<td>ENSG00000187556</td>
<td>Hs.127982</td>
<td>On array</td>
<td>Brain</td>
<td>1 tissue (Eye)</td>
<td>E, M, S</td>
</tr>
<tr>
<td>OVOL2</td>
<td>20</td>
<td>ENSG00000125850</td>
<td>Hs.661013</td>
<td>On array</td>
<td>-</td>
<td>1 tissue (Lung)</td>
<td>E, M, S</td>
</tr>
<tr>
<td>OVOL3</td>
<td>19</td>
<td>ENSG00000105261</td>
<td>Hs.143920</td>
<td>NA</td>
<td>Testis</td>
<td>1 tissue (Ovary)</td>
<td>E</td>
</tr>
<tr>
<td>TDRD1*</td>
<td>10</td>
<td>ENSG00000095627</td>
<td>Hs.333132</td>
<td>On array</td>
<td>Testis, brain</td>
<td>1 tissue</td>
<td>E, S</td>
</tr>
</tbody>
</table>

For submission to the journal Translational Oncology
<table>
<thead>
<tr>
<th>Gene</th>
<th>ID</th>
<th>ENSG Accession</th>
<th>Gene Symbol</th>
<th>Tissues</th>
<th>Experiments</th>
<th>Tissues</th>
<th>Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEK5</td>
<td>13</td>
<td>ENSG00000197168</td>
<td>Hs.672144</td>
<td>NA</td>
<td>Testis, placenta</td>
<td>2 tissues</td>
<td>-</td>
</tr>
<tr>
<td>PIWIL1</td>
<td>12</td>
<td>ENSG00000125207</td>
<td>Hs.405659</td>
<td>On array</td>
<td>Testis, brain</td>
<td>2 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>PIWIL2*</td>
<td>8</td>
<td>ENSG00000197181</td>
<td>Hs.614809</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>2 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>DAZ1</td>
<td>Y</td>
<td>ENSG00000188120</td>
<td>Hs.522868</td>
<td>On array</td>
<td>Testis</td>
<td>3 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>MAEL*</td>
<td>1</td>
<td>ENSG00000143194</td>
<td>Hs.651245</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>3 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>NANOS1</td>
<td>10</td>
<td>ENSG0000018613</td>
<td>Hs.591918</td>
<td>On array</td>
<td>Brain, placenta</td>
<td>3 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>BICC1</td>
<td>10</td>
<td>ENSG00000122870</td>
<td>Hs.158745</td>
<td>On array</td>
<td>Brain</td>
<td>4 tissues</td>
<td>-</td>
</tr>
<tr>
<td>PIWIL4</td>
<td>11</td>
<td>ENSG00000134627</td>
<td>Hs.660188</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>5 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>TDRD9</td>
<td>14</td>
<td>ENSG00000156414</td>
<td>Hs.21454</td>
<td>On array</td>
<td>Testis, brain</td>
<td>6 tissues</td>
<td>E, S</td>
</tr>
<tr>
<td>CPEB1</td>
<td>15</td>
<td>ENSG00000214575</td>
<td>Hs.547988</td>
<td>On array</td>
<td>Testis, brain</td>
<td>7 tissues</td>
<td>E, S</td>
</tr>
<tr>
<td>OVOL1</td>
<td>11</td>
<td>ENSG00000172818</td>
<td>Hs.134434</td>
<td>On array</td>
<td>Testis, placenta</td>
<td>7 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>TAF6L</td>
<td>11</td>
<td>ENSG00000162227</td>
<td>Hs.714400</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>10 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>CHEK2</td>
<td>22</td>
<td>ENSG00000183765</td>
<td>Hs.291363</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>11 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>NEK3</td>
<td>13</td>
<td>ENSG00000136098</td>
<td>Hs.409989</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>11 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>TGIF2</td>
<td>20</td>
<td>ENSG00000118707</td>
<td>Hs.632264</td>
<td>On array</td>
<td>Testis, brain</td>
<td>11 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>ESRP1</td>
<td>8</td>
<td>ENSG00000104413</td>
<td>Hs.487471</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>12 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>TDRD7</td>
<td>9</td>
<td>ENSG00000196116</td>
<td>Hs.193842</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>13 tissues</td>
<td>E, S</td>
</tr>
<tr>
<td>ESRP2</td>
<td>16</td>
<td>ENSG00000103067</td>
<td>Hs.592053</td>
<td>On array</td>
<td>Testis, placenta</td>
<td>14 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>NEK1</td>
<td>4</td>
<td>ENSG00000137601</td>
<td>Hs.481181</td>
<td>On array</td>
<td>Testis, brain</td>
<td>14 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>MRE11A</td>
<td>11</td>
<td>ENSG00000020922</td>
<td>Hs.192649</td>
<td>On array</td>
<td>Testis, brain, placenta</td>
<td>18 tissues</td>
<td>E, M, S</td>
</tr>
<tr>
<td>Gene</td>
<td>Score</td>
<td>Entrez Gene ID</td>
<td>Hs.</td>
<td>On array</td>
<td>Tissues</td>
<td>E</td>
<td>M</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>---------------</td>
<td>-----</td>
<td>----------</td>
<td>---------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>YTHDC2</td>
<td>5</td>
<td>ENSG00000047188</td>
<td>Hs.231942</td>
<td>Testis, brain, placenta</td>
<td>18 tissues</td>
<td>E, M, S</td>
<td></td>
</tr>
<tr>
<td>TAF6</td>
<td>7</td>
<td>ENSG00000106290</td>
<td>Hs.489309</td>
<td>Testis, brain, placenta</td>
<td>19 tissues</td>
<td>E, M, S</td>
<td></td>
</tr>
<tr>
<td>TXN</td>
<td>9</td>
<td>ENSG00000136810</td>
<td>Hs.435136</td>
<td>Testis, brain, placenta</td>
<td>21 tissues</td>
<td>E, M, S</td>
<td></td>
</tr>
<tr>
<td>TUBG1</td>
<td>17</td>
<td>ENSG00000131462</td>
<td>Hs.279669</td>
<td>Testis, brain, placenta</td>
<td>22 tissues</td>
<td>E, M, S</td>
<td></td>
</tr>
<tr>
<td>TXNL1</td>
<td>18</td>
<td>ENSG00000091164</td>
<td>Hs.114412</td>
<td>Testis, brain, placenta</td>
<td>25 tissues</td>
<td>E, M, S</td>
<td></td>
</tr>
<tr>
<td>RPS19</td>
<td>19</td>
<td>ENSG00000105372</td>
<td>Hs.438429</td>
<td>Testis, brain, placenta</td>
<td>28 tissues</td>
<td>E, M, S</td>
<td></td>
</tr>
<tr>
<td>RPS5</td>
<td>19</td>
<td>ENSG00000083845</td>
<td>Hs.378103</td>
<td>Testis, brain, placenta</td>
<td>31 tissues</td>
<td>E, M, S</td>
<td></td>
</tr>
<tr>
<td>DAZ3</td>
<td>Y</td>
<td>ENSG00000187191</td>
<td>NA</td>
<td>On array</td>
<td>NA</td>
<td>NA</td>
<td>M, S</td>
</tr>
<tr>
<td>DAZ4</td>
<td>Y</td>
<td>ENSG00000205916</td>
<td>NA</td>
<td>On array</td>
<td>NA</td>
<td>NA</td>
<td>M, S</td>
</tr>
<tr>
<td>RP11-152F13.10</td>
<td>15</td>
<td>ENSG00000260836</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>

E – Determined to be (over)expressed in cancer samples by EST meta-analysis.
M – Determined to be (over)expressed in cancer samples by microarray meta-analysis of combined microarray datasets.
S – Determined to be (over)expressed in cancer samples by analysis of at least one individual microarray dataset; these designations have the limitations imposed by statistical rigor being derived from a single microarray dataset.
* – Previously characterized CT genes.
NA – No information available.
References for Supplemental Material


7 Overall Discussion

The scientific impact and importance of the separate work sections is discussed in detail in the associated chapter/paper. The aim of this chapter is to address each work objective as well as to provide a holistic picture of this work, drawing it together in the context of its potential in predicting diagnostic, prognostic and therapeutic target genes.

The search for tumour antigens (TAs) has identified a number of potential targets, some of which are currently investigated in clinical trials [30, 34, 39, 51, 235, 236]. However, effective cancer immunotherapy still has a long way to go, as researchers are continuously searching for optimal TAs and need to unravel the mechanisms of immune tolerance and suppression of anti-tumour immune responses [41]. Nevertheless, first advances have been made in this field. To date twelve therapeutic antibodies [29, 30, 34] and one cancer vaccine [40] have been approved by the US Food and Drug Administration (FDA) to treat various solid tumours or haematological malignancies.

For TAs to be potential immunotherapeutic targets, their encoding genes are required to exhibit a highly restricted expression profile in normal tissues in addition to aberrant expression in cancer so that autoimmunity can be prevented [30, 34]. A group of testis-specific genes, the cancer testis (CT) genes exhibit such restricted expression patterns, which has led to the emergence of their corresponding antigens as important oncological markers and therapeutic targets [46].

Recent studies have shown that many CT genes are additionally expressed in tissues of the central nervous system (CNS), which are also immunologically privileged [57–59]. However, many of the CT genes were subsequently determined to show even a broader expression in somatic tissues than first assumed, leading to the classification of CT genes in testis- or testis/CNS-restricted, and testis- or testis/CNS-selective [58]. The classification of CT genes remains fraught with difficulties and also highlights the need to identify optimal target genes.

In order to identify optimal target genes, comprehensive expression maps and/or differential expression profiles for all tissues of the human body and for the numerous
existing cancer types are needed to ensure that target genes indeed exhibit restricted expression patterns. Due to the difficulty in obtaining such data empirically, high-throughput expression data and data integration techniques have become essential to this field.

mRNA profiling techniques such as serial analysis of gene expression (SAGE) [237], expressed sequence tags (ESTs) [175], microarrays [238] and recent developments including RNA-seq [239] and dynamic/digital arrays [240] enable evaluation of the expression of tens of thousands of genes in parallel (Figure 7.1). In general, microarrays are the most commonly used technique for gene expression profiling, as they provide an established standard in methodology and data analysis, and are relatively inexpensive compared to sequencing-based methods [174]. Microarrays are particularly the method of choice when analysing large numbers of samples. In contrast, sequencing-based methods such as EST, SAGE and RNA-seq have the advantage that they do not require existing knowledge of the sequences and thus are additionally capable of identifying and quantifying new transcripts. In particular, RNA-seq is a powerful technology and is able to identify all types of transcripts including smallRNAs as well as to determine single nucleotide polymorphisms (SNPs) and splicing efficiency [239]. However, there is still a lack of an accepted standard for RNA-seq technology and data analysis, as it is still an evolving technique. Both microarray and RNA-seq technologies allow high throughput, whereas EST technology is limited in throughput. An important drawback of sequencing-based techniques in general is that low abundance transcripts might be suppressed during sequencing unless a high number of tags/reads has been sequenced. Thus, sequencing depth is an important determinant of reliability of differential gene expression. The more tags/reads have been sequenced, the more statistically accurate the results will be and the more low abundant transcripts will be covered [211].

As RNA-seq is based on next generation sequencing methods in contrast to EST, which is based on Sanger sequencing, higher number of reads can be produced. However, there is always a gap between technique development, data availability and widespread use. As journals now mostly require the deposition of microarray data in public repositories such as ArrayExpress [241] and Gene Expression Omnibus (GEO) [242], large amounts of raw and processed microarray data have become publicly available. Also an overwhelming amount of EST data has been deposited in public databases, with dbEST [218] being the largest one. As RNA-seq is still an evolving technique, data are not yet available to the same extent as EST and microarray data, which is essential for data integration techniques. Furthermore, lack of an accepted standard for RNA-seq technology, data analysis and reporting/annotation also impairs data mining and integration. With decreasing sequencing costs and increasing use of RNA-seq technology, RNA-seq
will become an established method with a large compendia of data publicly available in databases such as Sequence Read Archive (SRA) \[243\], and thus has great potential for re-analysis and data integration in the near future.

With this public expression data available, pipelining retrieval, integration and investigation of expression data has become an important task in the context of characterising genes according to their expression profiles as well as in the context of identifying diagnostic, prognostic and therapeutic target genes, and enables screening for such target genes without the need of tedious and costly empirical work. Most importantly, screening allows focusing on a manageable number of candidate genes, which in turn decreases the risk of pursuing unsuitable targets. Selected candidates may be followed up in the laboratory and validated by high-sensitivity methods such as (quantitative) reverse transcription polymerase chain reaction (RT-PCR/qRT-PCR) (Figure 7.1) \[244\].

![High-throughput expression profiling techniques in comparison with high-sensitivity methods.](image)

**Figure 7.1: High-throughput expression profiling techniques in comparison with high-sensitivity methods.** High-throughput methods are used for screening to discover new target genes and yield a snapshot of a certain tissue- or condition-specific transcriptome. In contrast, high-sensitivity methods are used for validating the potential target genes. Microarrays are the standard technique for discovery experiments, whereas (quantitative) reverse transcription polymerase chain reaction (RT-PCR/qRT-PCR) is commonly used for validation \[243, 245\]. (The image was adapted from an illustration by VanGuilder et al. \[244\])

The automated and integrated data analysis approaches developed in the course of this thesis include a microarray (CancerMA, available at: \[http://www.cancerma.org.uk/\])
and an EST meta-analysis pipeline (CancerEST, available at: [http://www.cancerest.org.uk/](http://www.cancerest.org.uk/), with both focusing on a set of related genes. CancerEST allows the automated construction of global integrated expression profiles based on EST data across 36 tissues and thus can examine tissue-specificity as well as predict cancer marker/target potential, whilst CancerMA enables automated computation of the differential meta-expression across 80 cancer microarray datasets based on clinically derived cancer samples and can currently cover 13 cancer types. The implementation of two in silico screening pipelines analysing high-throughput expression data successfully completes project objective 1.

Each tool can be used as a stand-alone pipeline for purposes discussed in detail in the separate work chapters/papers as well as briefly described above. Combining both tools, however, provides a powerful approach to characterise the global expression profiles of a set of genes of interest in the context of clinical relevance, to facilitate the discovery of TA candidates or to address other focused biological questions (Figure 7.2). Altogether, the user can perform automated analyses specifically tailored to his/her interests, which will hopefully also facilitate wider application by directly putting meta-analyses of high-throughput expression data in the hands of researchers.

For TA discovery (Figure 7.2), it is favourable to use the EST analysis to screen the gene set prior to subjecting it to the microarray meta-analysis pipeline, since microarray technology is limited by the number of cancer arrays available in public repositories for a given platform and a given array type as well as by the number of genes which are covered by the arrays and thus can be evaluated. Moreover, the EST screen can confirm the restricted expression patterns required for target genes and therefore functions as an additional quality filter. However, the microarray meta-analysis provides valuable information about the expression of candidate genes in clinically relevant cancer data.

The benefits of data integration and automation of the complete analysis process, as implemented here, are serveralfold. Data integration generally results in a more precise estimation of gene expression, increases the statistical power, enhances the reliability and generalizability of the results, resolves conflicting results between analogous studies and compensates for artefacts of individual studies [174, 211, 246], whereas pipelining of the analysis process allows cost-effective and fast high-throughput analysis.

As mentioned before, both pipelines analyse a set of related genes (e.g., a set of co-expressed, co-regulated or interacting genes), which has several advantages. First, it allows the user to focus on his/her genes of interest. Second, it may reveal candidates of
Figure 7.2: Examples of applications and workflows for both pipelines. CancerMA and CancerEST might be used separately or in combination. CancerEST allows inferring tissue-specificity and gene-tissue relationships. CancerMA, in contrast, allows analysing the differential expression across 13 cancer types deriving from clinically relevant microarray data. Combing both tools facilitates the discovery of tumour antigens (TAs) or may be used to characterise a gene set of interest.

Clinical relevance for specific cancer types. Third, it may serve to get insights into common dysfunction of specific genes or pathways across various cancer types. Furthermore, filtering microarray data represents a profound additional advantage for CancerMA, as it enhances the significance and accuracy of the analysis [247].

The underlying EST data for CancerEST were obtained from Unigene [222], as it is the most comprehensive and most frequently updated database holding clustered and indexed EST data [222, 227]. Appropriate EST libraries were selected for the computation of meta-libraries as shown in Figure 7.3. For the computation of the global expression profile in cancerous and healthy tissues, the concept of the Unigene EST profiles [222] was adopted, because it is a well-established approach and allows stringent determination of tissue-specificity in comparison with other methods, which calculate mainly highly enriched or predominately expressed genes in a given tissue [187, 230].
This is of great importance, as highly restricted expression patterns are a requirement for TA candidates to be of clinical relevance. CancerEST sorts the genes according to their expression profiles into four classes to provide information about their cancer marker/target potential as well as into four states to classify their predicted tissue-specificity. This should facilitate the interpretation, sorting and filtering of the results and allows the user to focus on his/her research interest. Initially, protein similarities were used to ensure that ESTs were properly assigned to genes. However, testing both approaches showed that using all ESTs mapped to a given gene produced better results. This led to the assumption that erroneously assigned ESTs appear to be less frequent than ESTs incorporating untranslated regions (UTRs), which obviously will exhibit a low protein similarity.

**Figure 7.3: Data selection process for CancerEST.** The complete data available from the Unigene database (Unigene Build #230) was retrieved, resulting in 8724 cDNA libraries as a starting point. First, only libraries originating from cancerous and healthy tissues were kept and thus libraries deriving from diseases other than cancer were excluded. ESTs from normalised and subtracted libraries were also omitted as well as cell line libraries for the computation of healthy meta-libraries. Second, cDNA libraries deriving from uncharacterised, mixed or embryonic/foetal tissues were disregarded. All ESTs of a given tissue type were merged to a metatlibrary. However, meta-libraries with an EST count below 10,000 were excluded to assure significance, resulting in cancer and normal meta-libraries for 36 tissue types based on 4301 and 2531 libraries, respectively.
Figure 7.4: Data selection process for CancerMA. The microarray data for CancerMA derives from the two largest microarray repositories available, ArrayExpress [241] and Gene Expression Omnibus (GEO) [242]. The HG-U133 Plus 2 array from Affymetrix was chosen as array type due to its widespread use and large coverage of the human genome. As a starting point all 1404 experiments in ArrayExpress based on the HG-U133 Plus 2 array and relating to cancer were selected (accessed August 2012). To obtain experiments solely based on patient-derived cancer samples, all experiments using cell lines were excluded, resulting in 634 experiments. Subsequently, all experiments where no raw data was available for download were omitted, as the meta-analysis approach developed requires raw data. The first manual curation step was based on manual inspection of the publications and descriptions of the experiments and only experiments assessing patient-derived, untreated cancer samples with corresponding normal samples were selected. This resulted in 73 experiments including 15 experiments that were manually selected from GEO and also fulfilled our criteria as described. In the second curation step, the data was manually inspected and datasets with less than three control or cancer samples as well as datasets deriving from foetal tissues, tissues influenced by other diseases or cancer-associated tissues (e.g. tumour microenvironment) were excluded. To allow a meta-analysis, at least two datasets per cancer type are required, which also led to the exclusion of 4 experiments. After manual assessment, the retrieved datasets were divided according to the cancer type, subtype and stage, resulting in 92 datasets from 50 experiments covering 13 distinct cancer types. Subsequently, quality control using the ‘simpleaffy’ R package [256] was used to further assess the datasets. Finally, 80 individual curated cancer datasets originating from 45 experiments and covering 13 different cancer types remained.
The microarray data for CancerMA derives from the two largest microarray repositories available, ArrayExpress [241] and GEO [242]. The HG-U133 Plus 2 array from Affymetrix was chosen as array type due to its widespread use and large coverage of the human genome. Affymetrix platforms have also shown the most consistent results across multiple laboratories [248]. The data selection process for CancerMA was based on manual curation and quality assessment of the cancer microarray experiments available in ArrayExpress [241] and GEO [242] to ensure the underlying quality of the data (Figure 7.4). Nevertheless, one limitation of this tool is that a number of genes (in particular novel gene discoveries) are not covered by this array type and thus cannot be evaluated by this tool. However, CancerMA was designed for the incorporation of data from multiple array and platform types. The individual microarray analyses and the meta-analysis are independent processes in the CancerMA pipeline and the meta-analysis itself is based on a common identifier (Ensembl IDs). Thus, by further extending this tool it would be possible to combine data from various sources.

In general, microarray analyses highly depend on the preprocessing steps, annotation approaches and analysis methods used [174, 246]. Therefore, raw data was gathered, although processed data is more frequently available [249]. But this ensures the required standardisation of the analysis process producing comparable results, which in turn is essential for a meta-analysis. The obtained raw data was preprocessed, analysed and annotated using well-established techniques to ensure qualitative results [174, 250, 251]. The meta-analysis approaches implemented are based on Stouffer’s method [252] and weighted linear combination [253], both also well-established meta-analysis techniques [253]. Stouffer’s method is used to combine the individual p values, computing a meta-p value. The closely related method, Fisher’s sum of logs performed well compared to other methods in a meta-analysis comparison study [254], in which Stouffer’s method itself was not evaluated. Stouffer’s method additionally allows weighing each study and thus has increased power when combining studies of different size, which is usually the case when performing a meta-analysis. Weighted linear combination is used to obtain an overall estimate of the effect size (fold change in this case). Advantages of this method are that it also allows weighting of each study [246] and that fold change values have been reported to be more robust and reproducible across studies [255]. As for most analyses, the results are characterised by a significance estimate as well as an effect size estimate. Most meta-analyses, however, compute either meta-p or meta-fold change values. But there is a clear distinction between the two; although large effect sizes are usually correlated with highly significant values, this is not always the case and thus may not be sufficient as selection criteria [253]. Using both, a combined significance estimate and a combined fold change estimate therefore indeed ensures significant results.
Yet a number of points need to be considered. Both CancerEST and CancerMA rely on the availability of public expression data. Currently, CancerEST covers 36 tissue types, whereas CancerMA supports 13 cancer types. Further EST libraries and microarray datasets will become available in due course allowing expansion of the meta-analyses. Moreover, certain cancer subtypes or grades might be over- or underrepresented in CancerMA, respectively. Genes associated with subtypes that are underrepresented will not score well in the meta-analysis. Similarly, genes involved in processes uncovered or covered only by few experiments will either not score well in the meta-analysis. However, the results of the individual analyses allow the user to explore the results of a specific dataset of interest. It should further be considered that CancerEST shows only approximate gene expression patterns. Similar to CancerMA, certain subtypes and grades might not be present in the underlying EST repository and thus the expression profile may not be complete. Finally, the underlying data of both pipelines might be biased in many ways, among others due to lack of standards, inadequate experimental procedures and reporting of the studies, which in turn could result in poor data quality. According to a study by Larsson and Sandberg [249], only 23% of the raw microarray data in GEO and ArrayExpress meet the quality requirements for RNA integrity and hybridisation sensitivity to be considered as reliable datasets. Thus, the raw microarray data used for CancerMA has been curated and assessed [256] and low quality data has been removed. The EST data for CancerEST has also been assessed to ensure its quality. First, Uni-gene itself assesses and filters low quality data before clustering to ensure the quality of the clusters, which are used as underlying data for CancerEST. Second, normalised and subtracted EST libraries have been removed. Nevertheless, for a variety of reasons some biases may have been concealed from removal and might still influence the data.

To develop CancerMA and CancerEST, agile software development methodology was used, as the requirements and solutions evolved during the course of the thesis and rapid response to change was required. Altogether, the development of the pipelines was a continuous process, which included multiple major improvements and advances of the implementation as well as of the underlying analysis approaches used. Initially, the pipelines were developed as command line tools for in-house use and the tools had to rapidly evolve according to the demands of the research group. CancerMA was initially based on a simple vote counting strategy and was immensely improved in the course of the thesis by implementation of the current meta-analysis calculations as well as by addition of the visualisations. CancerEST’s original purpose was to provide a simple screening tool to limit the overwhelming number of candidate genes. The initial tool did not estimate the level of expression, evaluate the differential expression, compute
states according to the tissue-specificity or generate visualisations. It simply grouped the genes into classes corresponding to their expression in cancerous and healthy meta-libraries and thus was extensively improved in the course of the thesis. The decision to implement web interfaces was made after the main functionalities of the tools had been implemented and consequently the tools had to be adapted to be suitable for a multi-user environment. As the implementation of the web interfaces occurred very late in the development process, it should be noted that such adaptations have their limitations, as the tools were originally not designed for this purpose. The performance of both tools was improved by distributing separate jobs to run on separate cores, avoiding that jobs are processed in a linear fashion. As the server possesses four processor cores, four jobs can run in parallel at a time, immensely speeding up the completion of accumulated jobs. The computationally most intensive analysis step of CancerMA represents the preprocessing of microarray raw data, which can simply be avoided by circumventing this step and loading previously preprocessed data. This is possible, as not until after the preprocessing step the individual analysis become distinct due to filtering the data with the gene set of interest to the user. CancerEST was improved in a similar fashion. The computationally most intensive analysis step of CancerEST is the EST counting. This step can also be bypassed by beforehand establishing and storing the EST counts of each Unigene cluster for each healthy and cancerous meta-library. To support further developments of the tools, a comprehensive documentation has been written, as it is essential for a successful project.

The website makes use of cookies to ensure that a user only has access to his/her own data and thus can access secure areas of the website. Hence cookies are used to identify the users and to maintain data uniquely linked to the user while navigating through the pages. The data belonging to a user are stored on the server and linked to a unique cookie, which ensures data security. The user is informed about the use of cookies by an implied cookie consent [257]. However, hackers might be able to attack the website, for example, by interception of emails, cookie hijacking or MySQL injection. To improve the security of the website, one possibility would be the use of Hypertext Transfer Protocol Secure (HTTPS) as communications protocol for secure communication.

The Freedom of Information Act gives anyone the right to ask public organisations for the recorded information they have on any subject [258]. Sensitive information, including personal information such as email addresses are protected and do not need to be released. The Data Protection Act, however, applies to personal data held on electronic systems, which must be kept confidential. If a user asks about personal information stored about herself/himself, the organisation is required to provide any information.
stored about this user. CancerMA and CancerEST do not store any information about
the user herself/himself apart from the email address and thus are not able to link the
name of the user to the email address. An email address is not sufficient for identifi-
cation of a user, as email addresses can be generated without providing correct names.
Furthermore, the user can delete the job submitted at any time, which will lead to the
elimination of all data belonging to user including the provided email address. All jobs
older than 30 days are also erased completely from our system including the email ad-
dress. As the long-term location of CancerMA and CancerEST have not been set yet,
this issue has not been handled in more detail, but will be appropriately handled in
due course. One option would be the omission of notifications. A session ID may be
generated when a job is submitted and the user needs to return to CancerMA without
any notification. The drawback of this option is that the user will not be notified when
the analysis results are available. Another possibility would be the deletion of the email
address as soon as the user has first logged into the system. Furthermore, it would be
possible to create a more sophisticated login system, so that each user has to register,
which in turn will enable us to link the email address to the name of the user and thus
would allow us to answer to any information requests.

To visualise the results, CancerMA creates Circos [259], Krona [260] and Forest [261]
plots. Circos [259] and Krona [260] plots present the single and meta-analysis results in
their entirety to highlight relationships within the data (Figure 7.5). Forest plots [261],
in contrast, visualise the meta-analysis results for each gene separately to provide more
details about a gene of interest (Figure 7.5). This user-friendly output is one the key
features of this tool, allowing non-bioinformatics to get a quick overview of the results
and to grasp the available information. For advanced users with knowledge in this field,
a Cytoscape input file is available; Cytoscape [262] is an optimal programme to view
and explore the relationships within such data and also offers more control and manip-
ulation of the data (Figure 7.5). Also the development of CancerEST was emphasised
on intuitive visualisation. CancerEST creates Circos [259] plots and bar charts. Circos
[259] plots show the analysis results in their entirety and highlight relationships between
genes and cancer types, whereas bar charts show the global expression profile across 36
healthy and cancerous tissues for each gene separately. Furthermore, it is intended to
add the generation of Krona plots as well as Cytoscape input files in the near future.

In order to make these tools available for a broad audience, the pipelines needed to
be adapted for a multi-user environment with an underlying relational database, which
allows storing and fast access to user-specific data. As multiple users may use the tools
at the same time, a multi-user environment is essential for web tools to avoid conflicts.
Overall Discussion

Figure 7.5: Visualisations of the results generated by CancerMA. Visualisations are a key feature of the CancerMA tool and help the user to grasp the available information. (A & C) CancerMA creates Circos [259] and Krona [260] plots to present the single and meta-analysis results in their entirety and to highlight relationships within the data. (B) Forest plots [261], in contrast, are further generated to visualise the meta-analysis results for each gene separately to provide more details about a gene of interest. (D) Additionally a Cytoscape input file is produced, which can be fed into Cytoscape [262] to draw a gene expression network.

User-friendly web interfaces for both tools were designed and developed for intuitive handling, viewing and interpretation of the analysis results. Comprehensive help sections complement and support the user-friendly interface. Twitter Bootstrap [263] was partially used as a frontend framework for web development. The advantage of using
Twitter Bootstrap is not only the attractive design, but also the enabling of intuitive and fast web development by providing a comprehensive set of tools. An important aspect was to ensure that the rendering of the web pages is independent of the data processing and thus can be completed before the server-side processing of the data is finished. Therefore, JQuery Datatables [264] was used for most tables, which also enables advanced interaction controls to any HTML table. Sorting, filtering and searching of the data can be handed off to the server, which is important for large data amounts. Furthermore, only the user interface has to be adapted to different native clients (e.g., devices), because the required functions of the backend have already been implemented. Testing and debugging is an essential step in web development. Hence several users have tested the web interfaces at a number of sites and using a number of distinct browsers, ensuring the functionality and the clear and intuitive use of the tools.

Both tools have been validated using experimentally derived test datasets from literature [58, 265, 266]. To validate CancerEST, a tissue-specific dataset was required, however, most tissue-specific datasets in literature actually consist of enriched or predominately expressed genes in a given tissue due to the difficulty in obtaining such data experimentally. Thus, we used the tight testis-restricted genes determined by Hofmann et al. [58] using high-throughput expression data in combination with RT-PCR data as a validation dataset. For CancerMA, in contrast, a set of genes differentially expressed in cancer is needed. Such datasets are more frequently available in literature and we chose a set of differentially expressed genes in lung cancer determined by cDNA array analysis and partially validated by RT-PCR [265] as well as another set of upregulated genes in ovarian cancer validated by RT-PCR [266]. Both tools performed well and could reproduce the experimental results. Thus, the validation demonstrates the functionality of the tools and shows that the output of both pipelines is significant and meaningful, accomplishing successfully project objective 2.

The first aim of this thesis, the development of an integrative bioinformatic analytical approach to automate and optimise the identification of novel TA candidates has been successfully completed. Thus, the second aim of this thesis, testing the hypothesis that a group of meiosis-specific genes is aberrantly expressed in cancer, could be accomplished by the employment of the developed pipelines and by the analysis of germline-associated datasets. The first step in this process was to generate a human meiosis-specific gene set. The dataset was generated based on a mouse microarray study by Chalmel et al. [128, 129] that was subsequently mapped to human orthologues. This dataset is the first comprehensive human meiosis-specific gene set to our knowledge and will hopefully also serve to gain insight into meiosis and gametogenesis. The dataset was further
Overall Discussion

Cross-validated with a set of human genes known to be associated with mitosis [267]. As many gene products may have functions in both meiosis and mitosis, this filtering step is important to improve the data quality and to weed out non-meiosis-specific genes. The developed EST pipeline was used to screen and further refine this meiosis-specific dataset by evaluating the tissue-specificity as well as the cancer expression. This generated a tight meiosis-specific gene set, which was validated by fellow Ph.D. students using RT-PCR on RNA isolated from a range of normal human tissues as well as from some tumour tissues and many cell lines. This verified 62 testis- or testis/CNS-restricted genes, with 33 of them also exhibiting aberrant cancer expression. The successful generation and refinement of a meiosis-specific dataset accomplishes project objective 3. To explore the clinical relevance of these 33 validated genes, they were subjected to the developed microarray meta-analysis, which could show that many of these genes (15 of 25 genes covered by the array sets) are frequently expressed in cancer, above all in ovarian cancer (e.g., PRDM9). As immunotherapy based on CT antigens has shown positive results [236], these genes could serve as further targets for immunotherapy, in particular for ovarian cancer immunotherapy. Furthermore, the meta-analysis could show that many of 29 genes ascribed as testis-specific without any apparent expression in any of the cancer cells tested by RT-PCR analysis (nine of 21 genes covered by the arrays sets), are upregulated in clinically derived microarray data (e.g., SYCP3). Altogether, the use of the EST screen, of the microarray meta-analysis as well as of RT-PCR validation has constructed a comprehensive picture of gene expression and thus led to the postulation of a novel group of human meiosis-specific CT genes (the meiCT genes). The employment of both screening pipelines to identify novel CT genes successfully meets project objective 4.

meiCT genes are unique in two respects. First, they are mainly autosomally encoded and thus are not subjected to meiotic X inactivation. Most CT genes identified so far, however, are encoded on the X chromosome and their gene products should have largely non-meiotic roles in the testis, as they are silenced in meiotic spermatocytes. Second, the meiCT genes are not only restricted to the testis, but are likely to be further restricted to the meiotic spermatocytes, which are protected by the blood-testis-barrier (BTB), forming a highly immunologically privileged area. Therefore, their corresponding proteins represent optimal targets for diagnostic, prognostic and therapeutic strategies [57].

As mentioned before, the approach taken to select meiCT genes for validation was relatively stringent, as only genes that according to the EST screen did not exhibit expression in healthy tissues apart from the testis and tissues of the CNS, were used, which ensured tightly restricted expression patterns. Running the EST screen with less
stringent parameters (e.g., allowing expression in two healthy tissues) could identify further candidates. Many of the candidates will indeed be expressed in these tissues, but a fraction could exhibit expression in one or two healthy tissues due to undiagnosed neoplastic changes in these tissues. RNA from normal tissues are usually pooled and extracted from tissues obtained post mortem, which are often retrieved from aged individuals. In support of this, varying expression of several genes could be detected in distinct RNA panels of the same normal tissues [56]. Thus, further, less stringent analyses and subsequent validation of these candidates could identify additional meiCT genes. Furthermore, a post-transcriptional mechanism could exist, capable of degrading meiotic mRNAs in mitotic cells. Such a mechanism is known to exist in *Schizosaccharomyces pombe*, whereby the protein Mmi1 binds to a cis-acting region in meiotic mRNAs to confer their removal [268, 269], but whether this mechanism is conserved or whether a similar one exists in humans remains yet to be proven. Since a number of genes encoding proteins with meiosis-specific functions (e.g., *SPO11*, *REC8* and *STAG3*) show expression not only restricted to the testis, such a mechanism could explain why their expression is not testis-restricted. Such genes could be aberrantly activated in cancer due to a dysfunction of such a post-transcriptional mechanism. However, this would also entail that transcriptional profiling alone could miss some meiosis-specific genes. Nevertheless, the most comprehensive human meiosis-specific dataset to date has been generated in the course of this thesis. Most importantly, however, the second aim of this thesis was accomplished by showing that a number of meiosis-specific genes are indeed aberrantly expressed in cancer, forming a novel group of CT genes.

The aberrant expression of meiotic genes in cancer evokes intriguing questions about the underlying causes and the subsequent consequences. In general, expression of meiotic genes is tightly regulated to ensure tissue-specific expression. This is of great importance as the expression of meiotic genes in somatic cells could have severe impacts and could lead to perturbation of the mitotic process. Genes such as *RAD21L*, *SYCP1*, *SYCP3*, *SMC1β* and potentially several other genes here detected with gene products of yet unknown functionality, encode proteins involved in synaptonemal complex (SC) formation [78, 152, 270, 271]. They could trigger oncogenic events such as inappropriate recombination events and aberrant chromosome segregation [31, 129, 172]. Moreover, the aberrant activation of genes encoding factors acting as epigenetic and transcriptional regulators such as *PRDM9* [131, 149] or *BRDT* [132], could result in altered transcriptional activity and/or epigenetic programming, which in turn could lead to further expression of genes with oncogenic characteristics and thus drive tumorigenesis [172]. The expression of *PRDM9* is particularly intriguing, as it encodes a protein capable of binding to degenerate 13-mer motifs in the DNA sequence as well as of trimethylating
lysine 4 of histone H3 (H3K4me3), which marks DSB initiation sites (meiotic hotspots) [130, 131, 149]. Thus, PRDM9 plays an important role in regulating meiotic hotspot chromatin activation. Moreover, it could have a function in transcriptional activation of meiosis-specific genes (e.g., RIK), as it is known of its orthologue in mice [272]. Its aberrant activation in cancer, however, could lead to unstable chromatin lesions or to aberrant expression of oncogenic genes such as the CT genes.

CT gene expression is frequent in a wide range of cancer types and many of these genes are co-expressed [31, 46]. A potential explanation for the underlying cause of their expression could be the intriguing commonalities between cancer and germs cells, as both exhibit attributes such as rapid proliferation, undifferentiated phenotype and immortality, leading to the suggestion that these characteristics usually unique to germline cells could be hijacked through aberrant expression of genes originally specific to germline cells. Altogether, this could reflect a soma-to-germline transformation in cancer cells, which has already been reported in Drosophila animals with a mutation in the dREAM-MMB complex causing brain tumours [168]. The dREAM-MMB core components and several related proteins function as chromatin regulators in the retinoblastoma (Rb) pathway and belong to a class of proteins called synMuv B [273–275]. Many of synMuv B mutants have been reported to lead to aberrant expression of germline-specific genes [168–170, 275, 276] and could cause such soma-to-germline transformations.

To further assess this view, the Drosophila germline genes ectopically expressed in l(3)mbt brain tumours [168] were analysed to evaluate the expression of their human orthologues in cancer. Here, the EST and the microarray meta-analysis pipelines were employed to construct a holistic picture of gene expression, which further contributes to successful accomplishment of project objective 4. Although this analysis was solely based on computational techniques, it serves to provide first insights; 40 of 46 human orthologues show indeed expression in cancer and 15 genes also exhibit a testis-restricted expression pattern. Among these are also eleven genes that are associated with meiosis such as SYCP1 [78] and BOLL [277]. In conclusion, this analysis provides additional evidence supporting the postulation that meiotic genes are frequently expressed in cancer, forming a novel subset of the CT genes, as well as supports the view that cancer cells might undergo a soma-to-germline transformation, which in turn could drive tumorigenesis and may also contribute to the acquisition of tumour characteristics.

Julia Feichtinger
8 Conclusions

As discovery of targets for cancer diagnostic, prognostic and therapeutic strategies represents an ongoing challenge in cancer research, the purpose of this thesis is to provide a novel approach to facilitate their identification, characterising this project by two specific aims: (i) development of an integrative bioinformatic analytical approach to automate the discovery of novel tumour antigen (TA) candidates; and (ii) further employment of the developed approach to test the hypothesis that a number of meiosis-specific genes are aberrantly expressed in cancer and thus represent a novel cohort of the cancer testis (CT) genes. In conclusion, the work presented here has met the initial aims as defined for this project (cf. chapter 2).

8.1 Project Aim I

The first aim of this project was accomplished by the development of two integrative pipelines, CancerEST and CancerMA, exploiting two different high-throughput expression data resources. Both tools use different approaches to aid the identification of TA candidates and thus complement each other. CancerEST automatically constructs global integrated expression profiles based on expressed sequence tag (EST) data across 36 tissues. This allows characterising a gene set of interest in two respects, as tissue-specificity and cancer marker/target potential can be predicted. CancerMA, in contrast, automatically estimates the differential meta-expression across 13 cancer types and is based on 80 cancer microarray datasets of patient-derived cancer samples with corresponding normal samples. Therefore, CancerMA enables the evaluation of gene expression in cancer and, most importantly, can interfere clinical relevance of the TA candidates. As shown by validations, the meta-analysis approaches produce significant results and can enhance their reliability and generalizability. To facilitate wider application of both tools, intuitive web interfaces have been developed, which make the analyses directly accessible to the end-user. Altogether, combining both tools provides a powerful approach to discover TA candidates, and can govern experimental research.
8.2 Project Aim II

The substantiation of the hypothesis that a group of meiosis-specific genes are aberrantly expressed in cancer has been achieved in two steps. First, a meiosis-specific gene set has been successfully defined, and second, the cancer expression of this gene set has been analysed using the pipelines developed in the course of this thesis. This verified the existence of a novel cohort of human meiosis-specific CT genes, which are indeed frequently expressed in a wide range of cancer types and whose associated proteins represent potential clinically relevant cancer marker/therapeutic targets. In summary, a comprehensive picture of the gene expression profiles has here been presented based on the developed integrative pipelines in combination with reverse transcript polymerase chain reaction (RT-PCR) validation performed by fellow PhD students.

The discovery of CT genes specific to the meiotic programme also reveals a new group of genes, which might have oncogenic characteristics and whose gene products can potentially drive tumorigenesis. CT antigens in general are thought to cause oncogenic events, leading to the suggestion of soma-to-germline transformations occurring in human cancer cells, which in turn may contribute to the acquisition of tumour characteristics. The analysis of the human orthologues of *Drosophila* germline genes ectopically expressed in brain tumours provided further evidence of such soma-germline transformations in cancer cells. The human orthologues indeed exhibit expression in cancer, also exposing further meiosis-specific CT genes, as many generate proteins associated with meiotic functions.
9 Further Works

A large amount of further work arises from this project. Developments in the short term would focus on basic additions to elaborate the tools. In particular, enhancement of three major points could contribute to the value of the tools.

1. The functionality and usability of the tools could be enhanced by linking and connecting the tools, which would allow the user to feed the results of CancerEST directly into CancerMA as well as to easily compare the results of both tools.

2. CancerEST was developed at a later date of progression, whereas CancerMA was developed early on. Accordingly, CancerEST could profit from features already implemented in CancerMA. First, the usability of CancerEST would be enhanced through further visualisations such as Krona plots [260]. Cytoscape input files should also be generated to offer advanced users the possibility to export the results to Cytoscape [262]. Second, the tool would benefit from a gene ontology (GO) analysis, as provided by CancerMA.

3. CancerEST could further be improved in two respects. First, automatic updates of the local Unigene database would ensure that the tool is always based on the latest version of the data. Second, a user-specified threshold allowing a stringency definition of tissue-specificity according to the interests of the user could be incorporated into the tool.

In the long term, however, the scope of the project exhibits a wide range of possible improvements and extensions.

1. Additional data resources could be incorporated into CancerMA and CancerEST, which would intensify the effects of data integration and resolve some of the limitations of the tools as discussed before (e.g., limited coverage of cancer types). The implementation of CancerMA is suitable for the incorporation of data from multiple array and platform types and would also allow incorporation of RNA-seq data, as the individual microarray analyses and the meta-analysis are independent
processes and the meta-analysis itself is based on a common identifier. Similarly, CancerEST could be extended, for example, by incorporating serial analysis of gene expression (SAGE) data.

2. The generation and analysis of further sets of genes with specific expression in immunologically privileged tissues such as the brain or with developmental-specific expression could also reveal more genes encoding novel tumour antigen (TA) candidates. Additionally, creation and analysis of a non-meiotic testis-specific gene set may identify new cancer testis (CT) candidate genes that are not associated with meiosis and would further complete the picture of the current analyses.

3. Further bioinformatic analyses can characterise the identified meiotic CT candidate genes and their associated proteins to uncover information about regulation, structure, functionality and protein localisation, which would lead to a better understanding of the associated proteins in biology and disease. Preliminary data have been generated to characterise the candidates in two respects. First, existing tools and sequence analysis has been employed to examine the presence of nuclear transport signals (NLS) (Appendix A) [278, 279]. Second, sequence analysis of flanking regions was used to discover putative binding sites for two epigenetic regulators potentially associated with CT gene expression (Appendix B) [68, 130]. A further interesting analysis represents the search for sequence elements responsible for selective removal of mRNAs in humans. Such a post-transcriptional mechanism is known in Schizosaccharomyces pombe, capable of degrading meiotic mRNAs in mitotic cells [205, 269]. This mechanism could be conserved in humans, representing an additional level of regulating meiosis-specific gene expression.

4. Ultimately, the gene expression profile of the discovered meiotic CT candidate genes must be further investigated by protein-expression analyses of their associated proteins. Based on considerations of physicochemical properties, folding potential and secondary structure, synthetic antigens for these candidates could be predicted, which could be used for the production of antibodies against these potential CT antigens.
Although translation takes place on cytoplasmic ribosomes, many proteins function in the nucleus and are involved in processes such as transcription, DNA replication/repair and RNA processing. These proteins contain nuclear localisation signals (NLS) and are directed into the nucleus by means of several nuclear import pathways [278, 279].

Preliminary data predicting the nuclear localisation have been generated for the validated 33 cancer testis (CT) antigen candidates using existing tools [280, 281], the protein database Uniprot [282] and sequence analysis searching for published NLS motifs [278, 279] in the protein sequence. The preliminary data are summarised in Table A.1 and show that many of the candidates (45%) are potentially directed to the nucleus. Nuclear transported proteins may directly or indirectly interact with the DNA throughout the cell cycle, potentially causing oncogenic changes, whereas cytoplasmic proteins could only do so during cell division.
Table A.1: Preliminary data showing potential nuclear localisation for the validated 33 cancer testis (CT) antigen candidates. Nuclear localisation has been predicted using the tools NLStradamus [278] and Nucleo [280], the protein database Uniprot [282] as well as sequence analysis based on published NLS motifs [278, 279]. Proteins for which two sources predict nuclear localisation or experimental evidence is available are considered as nuclear proteins.

<table>
<thead>
<tr>
<th>Ensembl Gene ID</th>
<th>Protein</th>
<th>NLStradamus</th>
<th>Nucleo</th>
<th>Sequence Analysis</th>
<th>Uniprot</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000123165</td>
<td>ACTRT1</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000137948</td>
<td>BRDT</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N*</td>
<td>N</td>
</tr>
<tr>
<td>ENSG00000197651</td>
<td>C12orf12</td>
<td>N</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000184507</td>
<td>C15orf55</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N/C*</td>
<td>N/C</td>
</tr>
<tr>
<td>ENSG00000214556</td>
<td>C17orf98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000188032</td>
<td>C19orf67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000178395</td>
<td>C1orf65</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N/C</td>
<td></td>
</tr>
<tr>
<td>ENSG00000171695</td>
<td>C20orf201</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000132631</td>
<td>C20orf79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000120160</td>
<td>C9orf11</td>
<td></td>
<td></td>
<td></td>
<td>C/ME</td>
<td></td>
</tr>
<tr>
<td>ENSG00000187516</td>
<td>CXorf27</td>
<td>N</td>
<td>N</td>
<td></td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ENSG00000189037</td>
<td>DUSP21</td>
<td>N</td>
<td>N</td>
<td>N/C*</td>
<td>N/C</td>
<td></td>
</tr>
<tr>
<td>ENSG00000205186</td>
<td>FABP9</td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>ENSG00000181867</td>
<td>FTMT</td>
<td>N</td>
<td></td>
<td>M*</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>ENSG00000204671</td>
<td>IL31</td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>ENSG00000102021</td>
<td>LUZP4</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N*</td>
<td>N</td>
</tr>
<tr>
<td>ENSG00000188408</td>
<td>MAGEB5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000170948</td>
<td>MBD3L1</td>
<td>N</td>
<td></td>
<td>N</td>
<td>N*</td>
<td>N</td>
</tr>
<tr>
<td>ENSG00000184650</td>
<td>ODF4</td>
<td></td>
<td></td>
<td></td>
<td>ME</td>
<td></td>
</tr>
<tr>
<td>ENSG00000218823</td>
<td>PAPOLB</td>
<td></td>
<td></td>
<td></td>
<td>N*</td>
<td>N</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Symbol</td>
<td>Location</td>
<td>Tissue</td>
<td>Side 1</td>
<td>Side 2</td>
<td>Side 3</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------</td>
<td>----------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>ENSG00000196570</td>
<td>PFN3</td>
<td>N/C*</td>
<td>N/C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000164256</td>
<td>PRDM9</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000244588</td>
<td>RAD21L1</td>
<td>N</td>
<td>N</td>
<td>N*</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ENSG00000140623</td>
<td>SEPT12</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000077935</td>
<td>SMC1B</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000146857</td>
<td>STRA8</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000198765</td>
<td>SYCP1</td>
<td>N</td>
<td>N</td>
<td>N*</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>ENSG00000173809</td>
<td>TDRD12</td>
<td>N</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000159648</td>
<td>TEPP</td>
<td>S</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000182459</td>
<td>TEX19</td>
<td>N</td>
<td>N*</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000185264</td>
<td>TEX33</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000212122</td>
<td>TSSK1B</td>
<td>C*</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSG00000187969</td>
<td>ZCCHC13</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The 33 validated cancer testis (CT) antigen candidates were identified using meta-analysis approaches in combination with reverse transcription polymerase chain reaction (RT-PCR) validation as described before (cf. chapters 3-5) [57]. All protein sequences were obtained from Ensembl Biomart [283]. The sequence analysis was implemented in Perl 5.8.8 (available at: [http://www.perl.org](http://www.perl.org)).

**Abbreviations:**
- N – Nucleus
- C – Cytoplasm
- M – Mitochondrion
- S – Secreted
- ME – Membrane
- * – Experimental evidence
Appendix B

Epigenetic regulation has a crucial share in regulating gene expression and functions either by DNA methylation or histone modification, which influence the transcriptional efficiency usually by modulating accessibility to transcription factors (TFs) [284]. CTCFL/BORIS and its parologue, CTCF are both known epigenetic regulators, and BORIS has recently been postulated as cancer testis (CT) antigen [68]. In contrast to the ubiquitously expressed CTCF gene, gene expression of BORIS is restricted to the testes. BORIS competes with CTCF over binding sites, as both can bind to the same DNA sequences. The two paralogues exhibit antagonistic features and BORIS is thought to be responsible for epigenetic reprogramming during germ cell development [68, 88, 285, 286]. Furthermore, BORIS is associated with promoting the expression of CT genes (e.g., MAGEA1) by binding directly to their promoters leading to demethylation as well as by recruiting chromatin remodelling complexes resulting in histone modifications activating transcription [68, 88, 285, 287–289].

Meiotic recombination is initiated by double stranded breaks (DSBs), which are concentrated at specific sites in the genome [104, 135, 147]. These DSB initiation sites (meiotic hotspots) appear to be both genetically and epigenetically marked. The meiosis-specific gene PRDM9 encodes a protein, which is capable of binding to degenerate 13-mer motifs in the DNA sequence as well as of trimethylating lysine 4 of histone H3 (H3K4me3), marking such DSB initiation sites [130, 131, 149]. Thus, PRDM9 plays an important role in regulating meiotic recombination. Moreover, studies in mice showed that PRDM9 leads to transcriptional activation of meiosis-specific genes such as the RIK gene [269]. Its aberrant activation in cancer, however, could lead to oncogenic changes such as unstable chromatin lesions or aberrant expression of oncogenic genes.

As both BORIS and PRDM9 are capable of altering epigenetic programming, their aberrant activation in cancer cells could lead to oncogenic events, provoking the expression of oncogenic genes such as the CT genes [172]. Thus, preliminary data predicting potential binding motifs of these two proteins have been generated for the validated 33 meiotic CT candidate genes as well as for the 29 additional meiotic candidate genes [57] by means of sequence analysis searching for their published binding motifs [130, 290].
in the upstream flanking regions of the genes. The preliminary data are summarised in Table 2 and interestingly, the frequency of the binding motifs of both proteins tends to be lower in the flanking regions of the meiotic CT candidate genes than in the flanking regions of all genes in the genome or of the CT genes encoded on the X chromosome (CT-X). These findings could point to a BORIS/PRDM9-independent transcriptional regulation pathway. However, these finding are not significant and the employed sequence analysis is very basic and preliminary. Therefore, the results do not necessarily reflect the true presence of binding motifs in the flanking regions.
Table B.1: Preliminary data showing the percentage of the validated 33 meiotic cancer testis (CT) candidate genes and of the 29 additional meiotic candidate genes with potential binding motifs for PRDM9 and CTCF/BORIS in their upstream flanking regions. The frequency of binding motifs for PRDM9 and BORIS tends to be lower in the flanking regions of meiotic CT candidate genes than in the flanking regions of all genes in the genome or of the CT genes encoded on the X chromosome (CT-X).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Target genes</th>
<th>Number of genes</th>
<th>Number of genes with motif</th>
<th>% genes with motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRDM9</td>
<td>All genes in genome</td>
<td>21946</td>
<td>4464</td>
<td>20.34</td>
</tr>
<tr>
<td>PRDM9</td>
<td>Meiotic CT candidate genes</td>
<td>52</td>
<td>8</td>
<td>15.38</td>
</tr>
<tr>
<td>PRDM9</td>
<td>Strict CT-X</td>
<td>31</td>
<td>7</td>
<td>22.58</td>
</tr>
<tr>
<td>CTCF/BORIS</td>
<td>All genes in genome</td>
<td>21946</td>
<td>11226</td>
<td>51.15</td>
</tr>
<tr>
<td>CTCF/BORIS</td>
<td>Meiotic CT candidate genes</td>
<td>52</td>
<td>22</td>
<td>42.31</td>
</tr>
<tr>
<td>CTCF/BORIS</td>
<td>Strict CT-X</td>
<td>31</td>
<td>16</td>
<td>51.61</td>
</tr>
</tbody>
</table>

The 33 validated cancer testis (CT) candidate genes as well as the 29 additional meiotic candidate genes were identified using meta-analysis approaches in combination with reverse transcription polymerase chain reaction (RT-PCR) validation as described before (cf. chapters 3-5) [57]. The 29 additional meiotic candidate genes were ascribed as testis-specific according to the RT-PCR validation, but expression in cancer for some of the genes could be shown by the microarray meta-analysis. The CT-X genes derive from a study by Hofmann et al. All nucleotide sequences for the flanking regions (1000bp upstream) were obtained from Ensembl Biomart [283]. The sequence analysis was implemented in Perl 5.8.8 (available at: [http://www.perl.org](http://www.perl.org)).
Appendix C

Figure C.1: Database schema of CancerMA. Visualisation of the database schema of the underlying database used by CancerMA.
Figure C.2: Database schema of Unigene. Visualisation of the database schema of the local Unigene database.
Figure C.3: Database schema of CancerEST. Visualisation of the database schema of the underlying database used by CancerEST.
Appendix D

The CancerMA Pipeline

The external resources and script languages used for the CancerMA project are listed below.

Perl

Description

The pipeline CancerMA was implemented in Perl 5.8.8 (available at http://www.perl.org/). The following Perl modules were used in the project. The modules are available at CPAN (available at http://www.cpan.org/).

Packages

- strict
- warnings
- DBI
- Switch
- File::Remove
- Cwd
- List::MoreUtils
- TryCatch
- vars
- LWP::Simple
- Mail::Sender
Appendix D

• File::Copy
• POSIX
• Pod::2::html
• Spreadsheet::WriteExcel
• XML::DOM::XPath
• XML::DOM

MySQL

Description

MySQL 5.0.77 was used as relational database management system for this project (available at http://www.mysql.com/).

R/Bioconductor

Description

The microarray and statistical analyses were implemented in R 2.12.1 (available at http://www.r-project.org/) and Bioconductor (available at http://www.bioconductor.org/). The following R libraries were used for this project and are available at the links stated above.

Libraries

• GOstats
• hgu133plus2.db
• DBI
• RMySQL
• WriteXLS
• RColorBrewer
• rmeta
• gplots
Appendix D

- Biobase
- affy
- genefilter
- limma
- gtools
- gdata
- bitops
- caTools
- grid
- AnnotationDbi
- gcrma
- affyPLM
- simpleaffy

Citation

R Development Core Team [291]
Gentleman et al. [246]

Circos Plots

Description

Circos is a software package written in Perl and is employed in visualising data and information (available at [http://circos.ca/](http://circos.ca/)).

Input and Output

Input: Text file with data to be visualised.

Citation

The Circos plot software package was implemented by Krzywinski et al. [256].
Krona Plots

Description

Krona allows hierarchical data to be explored with zoomable pie charts (available at http://sourceforge.net/p/krona/home/krona/). The pie charts are created using HTML5, JavaScript and Perl.

Input and Output

Input: Text file with data to be visualised.

Citation

The Krona plot software package was implemented by Ondov et al. [257].

Ensembl database

Description

Ensembl Human is a genome database (available http://www.ensembl.org/index.html). The Ensembl database was downloaded (accessed January 2012, available at http://www.biomart.org/) and read into a local MySQL database.

Citation

Flicek et al. [292].

HGNC database

Description

The HUGO Gene Nomenclature Committee (HGNC) database is a curated online repository of HGNC-approved gene nomenclature and gene information (available at http://www.genenames.org/). The HGNC database was downloaded (accessed June 2012) and read into a local MySQL database.

Citation

Seal et al. [293].
Appendix D

The CancerMA Web Interface

The external resources and script languages used for the CancerMA web interface are listed below.

Perl

Description

The pipeline CancerMA was implemented in Perl 5.8.8 (available at http://www.perl.org/). The following Perl modules were used in the project. The modules are available at CPAN (available at http://www.cpan.org/).

Packages

- strict
- warnings
- DBI
- CGI
- CGI::Carp
- Digest::MD5
- File::Basename
- LWP::UserAgent
- Scalar::Util
- List::MoreUtils
- Cwd
- Data::Dumper
- JSON::XS
- vars
- Spreadsheet::WriteExcel
- Pod::2::html
Appendix D

- POSIX
- Mail::Sender
- Switch
- Archive::Zip

MySQL

Description

MySQL 5.0.77 was used as relational database management system for this project (available at http://www.mysql.com/).

HTML/CSS

Description

The web pages were created using HyperText Markup Language (HTML) and Cascading Style Sheets (CSS). Twitter Bootstrap was partially used for the layout (available at http://twitter.github.com/bootstrap/index.html).

Javascript/jQuery

Description

Javascript was employed to enhance the web pages. Javascript/jQuery libraries were used among others for displaying dynamic tables and for validation. This project made use of the Javascript libraries listed below.

Libraries

- jQuery library (available at http://jquery.com/)
- Validation jQuery Plugin (available at http://bassistance.de/jquery-plugins/jquery-plugin-validation/)
- Bootstrap JS library (available at http://twitter.github.com/bootstrap/index.html)
- Chained Selects jQuery Plugin (available at http://www.appelsiini.net/projects/chained)
Appendix D

The CancerEST Pipeline

The external resources and script languages used in the CancerEST project are listed below.

Perl

Description

The pipeline CancerEST was implemented in Perl 5.8.8 (available at [http://www.perl.org/](http://www.perl.org/)). The following Perl modules were used in the project. The modules are available at CPAN (available at [http://www.cpan.org/](http://www.cpan.org/)).

Packages

- strict
- warnings
- DBI
- Spreadsheet::WriteExcel
- List::MoreUtils
- List::Util
- File::Remove
- XML::DOM
- XML::DOM::XPath
- Switch
- TryCatch
- Cwd
- Mail::Sender
- File::Copy
- POSIX
- Math::Round
Appendix D

- Math::Pari
- Pod::2::html
- Chart::Bars
- Chart::HorizontalBars

MySQL

Description

MySQL 5.0.77 was used as relational database management system for this project (available at [http://www.mysql.com/](http://www.mysql.com/)).

Circos Plots

Description

Circos is a software package written in Perl and is employed in visualising data and information (available at [http://circos.ca/](http://circos.ca/)).

Input and Output

Input: Text file with data to be visualised.

Citation

The Circos plot software package was implemented by Krzywinski et al. [256].

Unigene database

Description

Unigene computationally identifies transcripts (mainly ESTs) from the same locus (available at [http://www.ncbi.nlm.nih.gov/unigene/](http://www.ncbi.nlm.nih.gov/unigene/)). The Unigene database was downloaded (UniGene Build #230) and read into a local MySQL database.

Citation

Pontius et al. [222].
Appendix D

Ensembl database

Description

Ensembl Human is a genome database (available http://www.ensembl.org/index.html). The Ensembl database was downloaded (accessed January 2012, available http://www.biomart.org/) and read into a local MySQL database.

Citation

Flicek et al. [292].

HGNC database

Description

The HUGO Gene Nomenclature Committee (HGNC) database is a curated online repository of HGNC-approved gene nomenclature and gene information (available at http://www.genenames.org/). The HGNC database was downloaded (accessed June 2012) and read into a local MySQL database.

Citation

Seal et al. [293].
The CancerEST Web Interface

The external resources and script languages used for the CancerEST web interface are listed below.

Perl

Description

The web interface of CancerEST was implemented in Perl 5.8.8 (available at [http://www.perl.org/](http://www.perl.org/)). The following Perl modules were used in the project. The modules are available at CPAN (available at [http://www.cpan.org/](http://www.cpan.org/)).

Packages

- strict
- warnings
- DBI
- CGI
- CGI::Carp
- Digest::MD5
- File::Basename
- List::MoreUtils
- LWP::UserAgent
- File::Remove
- Mail::Sender
- Switch
- Math::Round
- Data::Dumper
- JSON::XS
- Scalar::Util
Appendix D

- Spreadsheet::WriteExcel
- Cwd
- vars
- Pod::2::html

MySQL

Description

MySQL 5.0.77 was used as relational database management system for this project (available at http://www.mysql.com/).

HTML/CSS

Description

The web pages were created using HyperText Markup Language (HTML) and Cascading Style Sheets (CSS). Twitter Bootstrap was partially used for the layout (available at http://twitter.github.com/bootstrap/index.html).

Javascript/jQuery

Description

Javascript was employed to enhance the web pages. Javascript/jQuery libraries were used among others for displaying dynamic tables and for validation. This project made use of the Javascript libraries listed below.

Libraries

- jQuery library (available at http://jquery.com/)
- Validation jQuery Plugin (available at http://bassistance.de/jquery-plugins/jquery-plugin-validation/)
- Bootstrap JS library (available at http://twitter.github.com/bootstrap/index.html)
- Chained Selects jQuery Plugin (available at http://www.appelsiini.net/projects/chained)
Appendix E

Screenshots of the CancerMA Web Interface

Figure E.1: Homepage of CancerMA. The screenshot shows the homepage of CancerMA (available at http://www.cancerma.org.uk/).
Figure E.2: Submission of a new CancerMA job. The screenshot shows the job submission form of the CancerMA pipeline. When submitting a new job, the user supplies a list consisting either of Ensembl IDs or of gene names.
Appendix E

Figure E.3: Mapping results of CancerMA. The screenshot shows the mapping results presented to the user. The identifiers provided by the user are mapped to their appropriate Affymetrix IDs by the tool to tell the user which genes can be analysed. Finally, the job can be submitted by providing an email address.
Appendix E

Figure E.4: Overview section of CancerMA. The screenshot shows the overview section of CancerMA. The general overview provides basic information about the submitted job and the data available to the user. The navigation bar on the left hand side allows the user to go through the results and the information section. The information section includes among others the annotated genes of interest and information about the datasets used in the analysis. The result section includes the analysis results of the meta-analysis, of the single analyses, of the single analyses (only) and of the gene ontology (GO) enrichment analysis.
Appendix E

Figure E.5: Meta-analysis results of CancerMA. The screenshot shows the meta-analysis results of CancerMA. The meta-analysis results comprise tables with statistical values and visualisations for the meta-upregulated as well as for the meta-downregulated genes of interest. All visualisation may be viewed in this section. Circos and Krona plots visualise the single and meta-analysis results in their entirety to highlight relationships within the data. Furthermore, forest plots visualise the meta-analysis results for each gene separately. The Cytoscape input file may also be downloaded in this section.
Figure E.6: Meta-analysis results of CancerMA (continued). The screenshot shows the meta-analysis results of CancerMA (continued). The meta-analysis results comprise tables with statistical values and visualisations for the meta-upregulated as well as for the meta-downregulated genes of interest.
**Appendix E**

**Figure E.7: Gene ontology results of CancerMA.** The screenshot shows the gene ontology (GO) results. The GO analysis results contain the enriched GO terms for the meta-up- and the meta-downregulated genes, respectively.
Figure E.8: Available downloads. The screenshot shows the downloads available to the user. All result files may be downloaded separately or as zip archive.
Appendix E

Figure E.9: Help section of CancerMA. The screenshot shows the detailed documentation available in the CancerMA help section (available at [http://www.cancerma.org.uk/help.html](http://www.cancerma.org.uk/help.html)).
Appendix E

Screenshots of the CancerEST Web Interface

Figure E.10: Homepage of CancerEST. The screenshot shows the homepage of CancerEST (available at [http://www.cancerest.org.uk/] ).
Figure E.11: Submission of a new CancerEST job. The screenshot shows the job submission form of the CancerEST pipeline. When submitting a new job, the user provides a text file consisting either of Unigene Cluster IDs or of curated gene names.
Figure E.12: Mapping results of CancerEST. The screenshot shows the mapping results presented to the user. The identifiers provided by the user are mapped to their appropriate Unigene Cluster IDs by the tool to tell the user which genes can be analysed. Finally, the job can be submitted by providing an email address.
Figure E.13: Overview section of CancerEST. The screenshot shows the overview section of CancerEST. The overview section provides basic information about the submitted job and a brief explanation how to interpret the results. The navigation bar on the left hand side allows the user to go through the results and the information section. The information section includes among others the annotated genes of interest and the 36 tissue types supported by CancerEST. The result section includes the EST meta-analysis results comprising of a ranked list of genes according to (i) their cancer marker potential; or to (ii) their tissue-specificity.
Figure E.17: Results of CancerEST. The screenshot shows the meta-analysis results of CancerEST. The result section includes the EST meta-analysis results comprising of a ranked list of genes according to (i) their cancer marker potential; or to (ii) their tissue-specificity. Furthermore, a comprehensive expression profile across 36 healthy and cancerous tissues is available for each gene. All visualisation may be viewed in this section. Circos plots visualise the analysis results in their entirety to highlight relationships between the genes and the cancer types. In contrast, bar charts show the complete expression profile across 36 healthy and cancerous tissues for each gene separately.
Figure E.18: Meta-analysis results of CancerEST (continued). The screenshot shows the meta-analysis results of CancerEST (continued). The result section includes the EST meta-analysis results comprising of a ranked list of genes according to (i) their cancer marker potential; or to (ii) their tissue-specificity.
**Figure E.19: Available downloads.** The screenshot shows the downloads available to the user. All result files may be downloaded separately or as zip archive.
Appendix E

Figure E.20: Help section of CancerEST. The screenshot shows the detailed documentation available in the CancerEST help section (available at http://www.cancerest.org.uk/help.html).
Appendix F

Code Snippets in R

Individual Microarray Analysis Using Limma

# Limma package: moderated t-test analysis
# Goal: identification of genes that are differentially expressed between
# two groups
# Limma makes use of linear models
# It requires two matrices:
# 1) design matrix: defines the different RNA samples used
# 2) contrast matrix: defines contrasts, which allow comparisons of interest
# limmaUsersGuide() - get the user guide
# Empirical bayes moderation: for moderated t-statistics
# The empirical Bayes approach is equivalent to shrinkage of the estimated
# sample variances towards a pooled estimate, resulting in far more stable
# inference when the number of arrays is small

# define the levels (different RNA samples - normal/cancer)
limma_group <- factor(names, levels = c("normal", "cancer"))

# create a design matrix
limma_design <- model.matrix(~0 + limma_group)

# define the column names
colnames(limma_design) <- c("normal", "cancer")

# fit the linear model to the data (to each gene)
# abCorrect_subset - preprocessed and filtered expression data
limma_fit <- lmFit(abCorrect_subset, limma_design)

# make the contrast matrix
#define which difference should be calculated (cancer-normal)
lmma_contrasts <- makeContrasts(cancer-normal, levels=limma_design)

#apply the contrasts of interest to the MArrayLM object
#calculates the difference between cancer and normal
limma_fit2 <- contrasts.fit(limma_fit, limma_contrasts)

#moderated t-statistics
#applies empirical Bayes smoothing to the standard errors
limma_fit2 <- eBayes(limma_fit2)

**Meta-analysis Approach Using Stouffer’s Method**

#Stouffer’s method
#This so-called inverse normal method was introduced by Stouffer (1949)
#Here the individual p values are combined to a meta-p value
#cf. Burns 2004 for R implementation

#pValues - data frame holding all p values of the individual studies
#(one-tailed for up and down)
#valuesLgth - number of individual studies
metapValUp <- pnorm(sum(qnorm(pValues[1:valuesLgth,2]))/sqrt(valuesLgth))
metapValDown <- pnorm(sum(qnorm(pValues[1:valuesLgth,3]))/sqrt(valuesLgth))

**Meta-analysis Approach Using Linear Combination**

#Linear combination (Morgan 2010)
#Log 2 fold change values are aggregated by means of linear combination
#and weighted by the variance in the effect size within each study.
#The confidence intervals are combined with the same weights.

#values - data frame holding the log 2 fold change values as well as
#the confidence intervals of the individual studies
#valuesLgth - number of individual studies
weights <- 1/values[1:valuesLgth,3]
weightedFC <- sum(weights*(values[1:valuesLgth,2]))/(sum(weights))
weightedCIleft <- sum(weights*(values[1:valuesLgth,4]))/(sum(weights))
weightedCIright <- sum(weights*(values[1:valuesLgth,5]))/(sum(weights))
Appendix F

Forest Plot

#Creation of a forest plot (Lewis and Clarke 2001)

#up_metaData - description data
#up_FC - fold change values of the individual studies
#up_CIleft, up_CIright - the confidence intervals of the individual studies
forestplot(up_metaData, up_FC, up_CIleft, up_CIright, xlab = "lg2(FC)",
zero=0, col=meta.colors(box="royalblue",line="darkblue", summary="royalblue"),
align= c("1","1","1"), is.summary=c(rep(FALSE,(dim(values_up)[1])), TRUE, TRUE))

Code Snippets in Perl

Krona Plots

#Creation of a Krona plot (Ondov 2011)

#$header - header
#$kronaFile, $kronaFile2 - file names
my $command = $dir . "krona/bin/ktImportText";
system("$command $kronaFile -o $kronaFile2 -n $header >/dev/null");

Circos Plots

#Creation of a Circos plot (Krzywinski 2009)

#$userID - user ID
#$fn - file name
system("cat input_$userID/$fn.txt | bin/parse-table -no-field_delimCollapse -blank_means_missing > data_$userID/tmp[$userID].txt");
system("cat data_$userID/tmp[$userID].txt | bin/make-conf -dir data_$userID >/dev/null");
system("../../../../bin/circos -conf etc/circos$userID.conf >/dev/null");
Electronic Appendices

Please find the following data on the enclosed CD.

Appendix G
Appendix G contains all Perl and R files generated for the two developed screening pipelines, CancerEST and CancerMA. Each file is commented and includes a description. For more information, refer to Appendix J, which provides a complete documentation for each pipeline, describing the purpose, synopsis, input/output and functions of each file.

Appendix H
Appendix H is a collection of all HTML/CSS, Perl CGI, Javascript and image files for the two developed web tools. Each file is commented and includes a description. For more information, refer to Appendix J, which provides a complete documentation for each web tool. Furthermore, help sections as well as example datasets are available for each tool. Please visit CancerMA (available at: http://www.cancerma.org.uk/) and CancerEST (available at: http://www.cancerest.org.uk/)

Appendix I
Appendix I includes all Perl files and MySQL setup files to establish the underlying databases for the two developed pipelines. Please note that most raw data (e.g., microarray raw data) could not be included due to the large file sizes. However, the data may be downloaded directly from ArrayExpress, Affymetrix, Unigene, HGNC and Ensembl. For more information, refer to Appendix J, which provides a complete documentation for each script file, including the data source.

Appendix J
Appendix J is a complete documentation for files contained in Appendices G-I, describing the purpose, synopsis, input/output and functions of each file. It further includes a description of all external resources and programming languages used in each project.
Appendix K

Appendix K includes the Perl files for conducting the sequence analyses, as described in Appendix A and B. Please note that these are preliminary scripts.
References


Julia Feichtinger
References


References


References


References


References


107. Székvölgyi L, Nicolas A: From meiosis to postmeiotic events: homologous recombination is obligatory but flexible. *FEBS J.* 2010, **277**:571–89.


References


155. Liu JG, Yuan L, Brundell E, Björkroth B, Danenholt B, Höög C: Localization of the N-terminus of SCP1 to the central element of the synaptonemal complex and evidence for


References


References


Julia Feichtinger
References


References


References


References


258. GOV.UK [https://www.gov.uk/].


263. Bootstrap [http://twitter.github.com/bootstrap/].

264. DataTables (table plug-in for jQuery) [http://datatables.net/].


References


